arXiv:2303.01754v1 [stat.ME] 3 Mar 2023

复杂调查数据的逻辑回归参数估计：基于真实数据的模拟研究

Amaia Iparragirre
*[1]Irantzu Barrio[1,3]
Jorge Aramendi[2]
Inmaculada Arostegui[1,3]

[1] 数学系。巴斯克大学 UPV/EHU[2] Eustat - 巴斯克统计局 - 巴斯克自治区统计研究所[3] BCAM - 巴斯克应用数学中心

摘要

在复杂调查数据中，每个抽样观测值都被赋予一个抽样权重，表示其在总体中所代表的单位数量。在模型参数的估计过程中是否应考虑抽样权重，是一个仍在不同领域研究人员之间引发广泛讨论的问题。我们旨在通过在逻辑回归模型框架下开展一项基于真实数据的模拟研究，为这一争论作出贡献。为研究其表现，我们考虑了三种用于估计逻辑回归模型系数的方法：a）未加权模型；b）加权模型；以及 c）未加权

*通讯作者：电子邮件：amaia.iparragirre@ehu.eus，地址：数学系。巴斯克大学 UPV/EHU。

混合模型。结果表明，应使用加权逻辑回归模型，这凸显了在估计模型参数时使用抽样权重的重要性。

关键词： 复杂调查数据、抽样权重、逻辑回归、模型参数估计、基于真实数据的模拟研究。

1 引言

复杂调查数据在不同领域的研究人员和分析人员中越来越受欢迎。特别是，复杂调查数据通常用于拟合预测模型，以及其他用途。这些数据通常通过某种复杂抽样设计，对调查所关注的有限总体进行抽样而获得。这类数据的特征之一是抽样权重，它表示每个抽样观测在有限总体中所代表的单位数。在处理复杂调查数据时，在实施传统统计技术之前——其中大多数技术都是为简单随机样本设计的——应当分析这些技术是否适用于此类数据（Skinner 等人，1989年）。

特别是，在拟合预测模型时是否使用抽样权重，是许多研究人员在文献中广泛讨论的问题（Brewer 和 Mellor，1973；Smith，1981年）。将预测模型拟合到调查数据时，可以采用不同的视角，这些视角通常被称为基于模型和设计的方法（Binder 和 Roberts，2009；Chambers 和 Skin-ner，2003）。一方面，采用基于设计的视角的研究人员警告说，如果在模型参数的估计过程中不考虑复杂抽样设计，尤其是不考虑抽样权重，则方差往往会被低估，并且可能得到有偏估计（Binder 和 Roberts，2009；Heeringa 等，2017年）。因此，他们主张在模型参数的估计过程中应当考虑抽样权重。

另一方面，从基于模型的观点来看，如果模型设定得当，

系数估计值必然是无偏的，即使在估计过程中没有直接考虑抽样权重；而考虑抽样权重可能会增加估计值的标准差，尤其是在样本量较小时（Chambers and Skinner，2003；Korn and Graubard，1995；Reiter 等人，2005；Scott and Wild，1986）。在此背景下，Rubin（1976年）；Scott（1977年）；Sugden and Smith（1984年）确立了在何种条件下可以出于推断目的而忽略抽样设计。正如Skinner 等人（1989年）所解释的，使设计具有可忽略性的一个条件是其为非信息性的。如果即使在考虑了将纳入模型的协变量之后，响应变量仍与抽样权重相关，则称抽样设计具有信息性（Pfeffermann and Sverchkov，2009）。从基于模型的视角出发，人们提出了不同的方法，以确保设计具有可忽略性且模型设定得当（Pfeffermann and Sverchkov，2009）。采取这一视角的研究人员提出，除其他技术外，还应将抽样过程中考虑的所有设计变量及其交互作用作为协变量纳入模型（例如，参见DeMets and Halperin（1977年）；Gelman（2007年）；Nathan 和 Holt（1980））。

尽管 Chambers and Skinner（2003年）早已指出这一点，但两种观点之间的讨论至今仍未平息。一些较新的研究，如 Reiter 等人（2005年）、Masood 等（2016年）以及Lumley andScott（2017年），表明这场争论仍令研究人员感到困惑，也使他们难以决定是否在分析中使用抽样权重。大多数研究人员都认同，如果样本具有信息性，或模型设定不当，就不宜忽略抽样权重；但与此同时，他们也主张在抽样权重并非绝对必要时忽略它们。难点通常在于判断，依据特定的调查数据估计模型参数时，抽样权重是否必要；换言之，设计是否具有信息性。正如Pfeffermann and Sverchkov（2009年）所解释的，信息性不仅取决于抽样设计，还取决于拟合的模型、该模型的响应变量以及纳入模型的协变量。因此，通常很难确定待分析调查数据的抽样设计对于拟合某一特定

模型是否具有信息性。此外，由于多种原因，例如信息缺乏、设计变量数量过多，以及将设计变量作为协变量纳入模型可能会损害其科学可解释性，因此并非总能将所有设计变量及其交互作用都纳入模型（Pfeffermann and Sverchkov, 2009年）。因此，时至今日，在实践中仍难以决定估计模型参数时是否应考虑抽样权重。基于此，我们认为该领域有必要开展更多研究；尤其是，我们认为有必要通过基于真实数据的模拟研究提供见解，以补充迄今受到广泛讨论的理论结果和案例研究。

本文始终聚焦于模型参数估计，尤其是二分类响应变量的逻辑回归框架。尽管总体而言，关于线性回归模型的研究更多（例如，DeMets and Halperin (1977年); Hausman andWise (1981年); Holt 等人(1980年); Nathan and Holt(1980年)），在逻辑回归模型领域，也有不少研究致力于解决复杂抽样所引发的这一问题。具体而言，Scott and Wild(1986年, 2002年)使用受病例对照研究启发的模拟数据开展研究。值得注意的是，病例对照研究会依据二分类响应变量对数据进行分层，因此始终基于信息性抽样设计。但是，如果我们不知道某一抽样设计对于拟合特定模型是否具有信息性，该怎么办？如上所述，实际处理真实复杂调查数据时，通常就会遇到这种情况。 Chambless and Boyle(1985年); Lumley and Scott (2017年)和Reiter 等人(2005年)在利用真实调查数据进行分析时提出了这一问题，并采用基于模型和基于设计的两种视角比较了多种估计方法，最终选出最适合其分析的模型。然而，在实际中，我们如何判断这些估计值之间的差异是大还是小？如果差异较大，又该选择哪个估计值？本文旨在更进一步，通过基于真实数据的模拟研究分析不同方法之间的差异，从而在真实场景下开展研究，为上述论文中的相关工作作出补充；在该场景中

我们可以将系数估计值与理论值进行比较。因此，数据基于实际调查生成；对于拟合不同模型而言，这些数据是否具有信息性，我们事先并不知晓。我们的目标是通过模拟研究分析一种实践中经常出现的情形，并分析和评估在每种情形下，决定是否纳入抽样权重以估计逻辑回归模型系数所带来的后果或影响。本研究比较了几种常用于估计逻辑回归模型系数的估计方法的表现（例如，Lumley and Scott(2017年)）。具体而言，我们比较以下方法获得的系数估计值：a）非加权逻辑回归模型；b）加权逻辑回归模型；以及 c）带随机截距的非加权逻辑回归混合模型。我们依据以下因素设定了不同情景：a）两组真实调查数据；以及 b）模型中的协变量/参数数量。这些实际调查由巴斯克官方统计局（Eustat）设计并收集，采用一步分层，并在每一层内进行简单随机抽样。

本文其余部分安排如下。第2 节介绍促成本研究的两个原始实际调查：ESIE 和 PRA 调查。第3 节介绍用于估计模型参数的方法。关于模拟过程、所抽取的情景以及所得结果的信息，见第4节。第5节将所述方法应用于真实调查数据，以作说明。最后，本文在第6节进行讨论并作总结。

2 激励性数据集

本节介绍促成本研究开展的两项复杂抽样调查。这些调查由巴斯克官方统计局（Eustat）设计、实施并收集数据。

一方面，信息社会调查[1]（企业），因其西班牙语简称而通常称为 ESIE 调查，该调查面向巴斯克地区（BC）的企业开展，旨在收集有关技术使用情况的信息。具体而言，本研究关注的响应变量是一个二分类响应变量，用于表示企业是否拥有自己的网页（1 表示有，0 表示没有）；我们旨在通过所有权、经营活动或企业单位员工人数等协变量对其进行建模。另一方面，活动相关人口调查（PRA）[2] 面向巴斯克地区年满 16 岁的居民开展，旨在估算巴斯克地区劳动力的比例。具体来说，本研究考虑的响应变量表示每个人是否属于劳动力（1 表示是，0 表示否）。我们发现，最重要的协变量包括年龄、教育程度、国籍和性别。

在这两项调查中，两个有限总体均采用单阶段分层抽样，并在每层内进行简单随机抽样。也就是说，总体被划分为不同的层，然后从每层中随机抽取一定数量的单位（数量事先确定）。不过，这两项调查的分层方式截然不同。在 ESIE调查中，分层依据三个分类变量的组合确定，分别是：1）公司所在的省（分为3类）；2）公司的经营活动（分为65类）；3）员工人数（分为3类）。因此，划分出了大量规模较小的层，共585层。不过需要指出的是，其中一些层在总体中没有单位，因此实际上总共只有515层。相比之下，在 PRA 调查中，层就是 BC 的23个地区。这使得 ESIE调查中响应变量在不同层内的分布各不相同，而在 PRA 中，响应变量的分布在各层之间没有差异。在 ESIE 和 PRA 调查中，按照上述抽样过程从有限总体中抽取样本后，都会为每个抽样单位赋予一个抽样权重。

在 ESIE调查中，从由 195222 家公司构成的有限总体中抽取了 7 725 家公司

[1]https://en.eustat.eus/estadisticas/tema_150/opt_1/tipo_7/temas.html[2]
https://en.eustat.eus/estadisticas/tema_37/opt_0/temas.html

作为样本（这些数据收集于 2010年）。具体而言，有限总体中的层大小范围为 1 至 14 535，中位数为 38，四分位距为$7 - 185.5$。各层的抽样概率范围为 0.0061 至 1，中位数为 0.2830，四分位距为 $0.0970 - 0.8417$。相比之下，在 PRA 调查中，从总计 1 851 316 名个体中抽取了 10 609 名作为样本（相关信息截至 2016年最后一个季度）。具体而言，层大小范围为 2 768 至 438 595，中位数为 44 335， $22\,024 - 72\,834$ 四分位距。抽样概率范围为 0.0041 至 0.0488，中位数为 0.0063（四分位距为 $0.0055 - 0.0102$）。

3 方法

本节将介绍我们为估计复杂调查数据的逻辑回归系数而考虑的方法。

令 Y 表示二分类响应变量，该变量取值为 1 表示所关注的事件发生（否则为 0），令 $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ 表示 p 协变量向量。令 $U = \{1, \ldots, N\}$ 表示一个有限总体，其中 N组随机变量（$Y, \boldsymbol{X}$）的实现值彼此相关，即 $\{(y_i, \ldots, \boldsymbol{x}_i)\}_{i=1}^{N}$。令 S 表示通过一步分层从有限总体 U 中抽取的包含 n 个观测值的样本。令$h = 1, \ldots, H$ 表示不同的层。与每个抽样单位 $i \in S$ 相关联的抽样权重记为 w_i。

我们将真实的总体逻辑回归模型定义如下：

$$logit(p_i) = \ln\left[p_i/(1 - p_i)\right] = \boldsymbol{x}_i^T \boldsymbol{\beta}^{True} \tag{1}$$

其中， $p_i = P(Y = 1|\boldsymbol{x}_i)$表示在给定协变量 $\boldsymbol{x}_i$($\forall i \in U$)和模型系数 $\boldsymbol{\beta}^{True} = (\beta_0^{True}, \beta_1^{True}, \ldots\ldots, \beta_p^{True})^T$的取值时，单位 i 发生该事件的概率；这些系数通过最大化总体似然计算得出：

$$L_{pop}(\boldsymbol{\beta}) = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{2}$$

然而，需要注意的是，通常只有抽样单位 y_i 的响应 $i \in S$才是已知的。因此，应基于样本 S对模型进行估计。本文比较了几种常用于估计二分类响应变量逻辑回归模型系数的估计方法（Lumleyand Scott，2017年）。我们的目标是将这些估计值与 $\boldsymbol{\beta}^{True}$ 进行比较，以分析各方法的性能。

在此情境下，可以对复杂抽样调查样本 S拟合一个简单的逻辑回归模型，其定义如下：

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}. \tag{3}$$

可以采用不同的方法，根据 S估计回归系数向量$\boldsymbol{\beta} = (\beta_0, \ldots\ldots, \beta_p)^T$：

M1. 非加权逻辑回归模型

此方法通过最大化方程（4）中的似然函数来估计模型系数，可借助迭代算法实现，例如迭代加权最小二乘（IRLS）算法（McCullagh andNelder，1989年）：

$$L(\boldsymbol{\beta}) = \prod_{i \in S} p_i^{y_i} \left(1-p_i\right)^{1-y_i}. \tag{4}$$

$\hat{}$ 以下将通过极大似然法估计的系数记为 $\boldsymbol{\beta}_{\text{M1}}$ 。

M2. 加权逻辑回归模型

该方法通过估计使方程（5）中的伪似然函数达到最大值的系数（Binder，1981年，1983），并将抽样权重纳入考量 w_i：

$$PL(\boldsymbol{\beta}) = \prod_{i \in S} p_i^{y_i w_i} \left(1-p_i\right)^{(1-y_i)w_i}. \tag{5}$$

伪似然函数也通过迭代算ˆ法求得最大值（Heeringa 等，2017年；Wolter，2007年）。将 $\boldsymbol{\beta}_{\mathrm{M2}}$ 记为基于该方法获得的系数估计值。

除了上述方法外，另一种选择是将复杂抽样设计视为第二层单位来拟合混合模型（例如，参见Lum- ley 和 Scott(2017年）；Masood 等(2016年）。尤其是在本研究中，我们以与Lumley and Scott（2017年）相同的方式考虑随机截距模型。令$i = 1$，.....，n_h表示属于层 $h\,(\forall h \in \{1, \ldots, H\})$ 的抽样单位，而 $\boldsymbol{x}_{hi}$ 和 y_{hi} 分别表示层 h中 i 的协变量向量和响应变量的取值。于是，我们的目标是对样本 S 拟合以下模型：

$$\mathrm{logit}(p_{hi}) = \ln\left(\frac{p_{hi}}{1-p_{hi}}\right) = \boldsymbol{x}_{hi}^T\boldsymbol{\gamma} + u_h, \quad u_h \sim N(0, \sigma_u^2). \tag{6}$$

其中 $p_{hi} = P(Y = 1|\boldsymbol{x}_{hi}, \ u_h) = \dfrac{e^{\boldsymbol{x}_{hi}^T\boldsymbol{\gamma}+u_h}}{1+e^{\boldsymbol{x}_{hi}^T\boldsymbol{\gamma}+u_h}}$。

M3. 带随机截距的非加权逻辑回归模型

在这种情况下，似然函数定义如下：

$$L_{\mathrm{mix}}(\boldsymbol{\gamma}, \sigma_u^2) = \prod_{h=1}^{H}\int_{-\infty}^{+\infty} f(y_{hi}|\boldsymbol{x}_{hi}, u_h)f(u_h)du_h, \tag{7}$$

其中 $f(y_{hi}|\boldsymbol{x}_{hi}, u_h) = \prod_{i=1}^{n_h} p_{hi}^{y_{hi}}(1-p_{hi})^{1-y_{hi}}$和 $f(u_h) = \dfrac{1}{\sigma_u\sqrt{2\pi}}e^{-u_h^2/2\sigma_u^2}$。参数 $\boldsymbol{\gamma}$ 和 σ_u^2 通常通过数值最大化式（7）中的似然函数来估计，通常采用拉普拉斯近似（Lee 和 Nelder，2001年）。下文分别将这些估计值记为 ˆ$\boldsymbol{\gamma}$ 和 ˆ σ_u^2 。

然而，条件随机效应模型与相应边际模型所得到的系数并不容易直接比较（Lee 和 Nelder，2004）。对于逻辑随机截距模型，

边际系数 $\boldsymbol{\beta}$ 可根据条件参数 $\boldsymbol{\gamma}$ 按如下方式获得：

$$\boldsymbol{\beta} = \frac{\boldsymbol{\gamma}}{\sqrt{1 + c^2\sigma_u^2}}, \tag{8}$$

其中 $c = (16\sqrt{3})/(15\pi)$ (Diggle 等人，2002年)。我们将基于 $\hat{\boldsymbol{\beta}}_{\mathrm{M3}}$ 和 $\hat{\boldsymbol{\gamma}}$ 获得的系数估计值记为 $\hat{\sigma}_u^2$。

目标是通过将估计值 $\boldsymbol{\beta}_{\mathrm{M1}}$、$\boldsymbol{\beta}_{\mathrm{M2}}$ 和 $\boldsymbol{\beta}_{\mathrm{M3}}$ 与真实的有限总体系数 $\boldsymbol{\beta}^{True}$ 进行比较，分析上述方法的表现。

4 模拟研究

本节介绍我们开展的模拟研究，旨在分析第 3 节所述估计方法在不同情景下，基于复杂调查数据估计逻辑回归模型系数时的表现。如前所述，本研究旨在真实数据情景下，将系数估计值与真实的有限总体系数进行比较。

第 4.1 节详细描述模拟过程，第 4.2 节展示模拟研究所得结果。

4.1 情景与设置

本节介绍开展模拟研究的不同情景以及我们遵循的步骤。模拟过程分步描述如下：

步骤 1. 生成伪总体 U 中 N 来自随机变量集合的Y，$\boldsymbol{X}$)（见附录 A）：$\{(y_i, \boldsymbol{x}_i)\}_{i=1}^{N}$。

步骤 2. 通过最大化式（$\boldsymbol{\beta}^{True}$ 2）中的总体似然来计算。

对于 $r = 1, \ldots, R$，重复以下步骤：

步骤 3. 通过一步分层抽样获得样本 $S^r \subset U$，并分配相应的抽样权重 $w_i,\ \forall i \in S^r$（见附录 B）。

步骤 4. 使用 (4)、(5) 和 (7) 中的似然函数将模型拟合到 S^r，并分别得到 $\hat{\boldsymbol{\beta}}^r_{\mathrm{M1}}$、$\hat{\boldsymbol{\beta}}^r_{\mathrm{M2}}$ 和 $\hat{\boldsymbol{\beta}}^r_{\mathrm{M3}}$。

最后，对于基于样本 $r = 1, \ldots, R$ 得到的结果，以及每种方法 $\forall m \in \{$M1,M2,M3$\}$，我们将系数向量估计值的偏差定义如下：

$$\mathrm{bias}^r_d = \hat{\beta}^r_{d,m} - \beta^{True}_d, \quad \forall d = 0, 1, \ldots, p. \tag{9}$$

那么，$\forall r = 1, \ldots, R$ 的平均偏差（AvBias）和均方误差（MSE）分别由方程（10）和（11）定义：

$$\mathrm{AvBias}_d = \frac{1}{R}\sum_{r=1}^{R}(\mathrm{bias}^r_d) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}^r_{d,m} - \beta^{True}_d\right), \quad \forall d = 0, 1, \ldots, p, \tag{10}$$

$$\mathrm{MSE}_d = \frac{1}{R}\sum_{r=1}^{R}(\mathrm{bias}^r_d)^2 = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}^r_{d,m} - \beta^{True}_d\right)^2, \quad \forall d = 0, 1, \ldots, p. \tag{11}$$

基于第 2 节所述的两项实际调查，定义了两种情景：ESIE（以下简称情景1）和 PRA（以下简称情景2）。根据每项调查（见步骤 1。，参见附录 A）生成了一个有限伪总体。根据 Eustat 在相应实际调查中采用的复杂抽样设计（定义见步骤 3。，参见附录 B），从这些总体中抽取样本。每个伪总体共获得 个样本。

此外，对于每项具有不同协变量数量的 survey，均针对有限总体和样本拟合了两个不同的模型（**步骤2**）。具体而言，在情景1中，分别拟合了包含 $p = 1$（X_1）个协变量和 $p = 3$（X_1、X_2 和 X_3）个协变量的模型。同样，在情景2中，分别拟合了包含$p = 1$（X_1）个协变量和 $p = 4$（X_1、X_2、X_3 和 X_4）个协变量的模型。

需要注意的是，所有协变量均为类别变量，且每个类别均估计一个系数，但被视为参考类别的类别除外。具体而言，在情景1中，包含 $p = 1$ 个协变量的模型共估计 $l = 7$ 个参数（包括截距、 β_0），而 $p = 3$则估计 $l = 14$ 个参数。同样，在情景2中， $p = 1$ 估计 $l = 7$ 个参数，而 $p = 4$估计 $l = 14$个参数。

所有计算均使用 (64 位) R 4.0.5（R 核心团队，2021）完成，所用工作站配备 32GB 内存、Intel i7-8700 处理器（3.20 Ghz）和 Windows 10 操作系统。具体而言，非加权逻辑回归模型 (M1) 使用 `glm` 函数，来自 `stats` 程序包；加权逻辑回归模型 (M2) 使用 `svyglm` 函数，来自 `survey`程序包（Lumley，2019）；带随机截距的未加权混合模型 (M3) 则使用 `glmer` 函数，来自 `lme4` 程序包（Bates 等人，2015）。

4.2 结果

本节介绍我们在两个情景中获得的结果：情景1（基于 ESIE调查）和情景2（基于 PRA 调查）。如第 4.1 节所述，在每个情景中，我们都拟合了协变量数量不同的两个模型。我们的目标是比较第 3 节所述三种系数估计方法（即未加权逻辑回归 (M1)、加权逻辑回归 (M2) 和带随机截距的未加权逻辑回归 (M3)）所得的估计值与真实有限总体系数（β^{True}），并从偏差和均方误差（MSE）方面进行比较。

由于获得的结果数量众多，我们先重述主要发现。比较三种方法在各情景中的表现时，可以发现结果因情景而异。在情景1中，M2在偏差和均方误差方面优于M1和M3，但M2得到的估计值方差大于M1或M3得到的估计值。另一方面，在情景2中，三种方法所得结果之间没有差异。结果还表明，M2方法在

两种情景中都表现良好，且就偏差（所有情景中的偏差均可忽略不计）和均方误差而言，结果相当相似。然而，M1和M3方法在偏差（进而也在均方误差）方面的表现因情景而异，在情景2中远低于情景1。下面我们将分析与各情景相关的图形和数值结果。

图1 展示了情景1中包含 $p = 1$ 个（图1 （a））和 $p = 3$ 个（图1 （b））协变量的模型采用方法 M1、M2 和 M3 得到的估计值偏差箱线图。从图中可以看出，无论模型包含 $p = 1$ 个还是 $p = 3$ 个协变量，M2 在偏差方面的表现都是最佳的。这一点也可从表1中看出。该表列出了情景1中分别包含 $p = 1$ 个和 $p = 3$个协变量的模型所得到的估计值的均值、标准差、平均偏差和均方误差，以及真实有限总体系数。可以看出，M2方法得到的估计值与有限总体中得到的真实系数（β^{True}）相当接近（因此该方法的平均偏差较低），而M1和M3方法得到的估计值则存在显著差异。例如，对于包含$p = 1$ 个协变量的模型，以系数$\beta_{1为例，6}$ 采用M2方法得到的平均偏差为-0.095，显著低于M1方法（0.378）和M3方法（-1.379）的结果。还可以看到，模型中纳入 $p = 3$ 个协变量后，所有方法的平均偏差均有所降低（M1和M3尤为明显）。这一结果与Nathan andHolt (1980年)一致。具体而言，在包含 $p = 3$ 个协变量的模型中，与类别$\beta_{1相关的系数估计值的平均偏差降至M1方法的0.050、M2方法的0.007和M3方法的-0.700。6}$

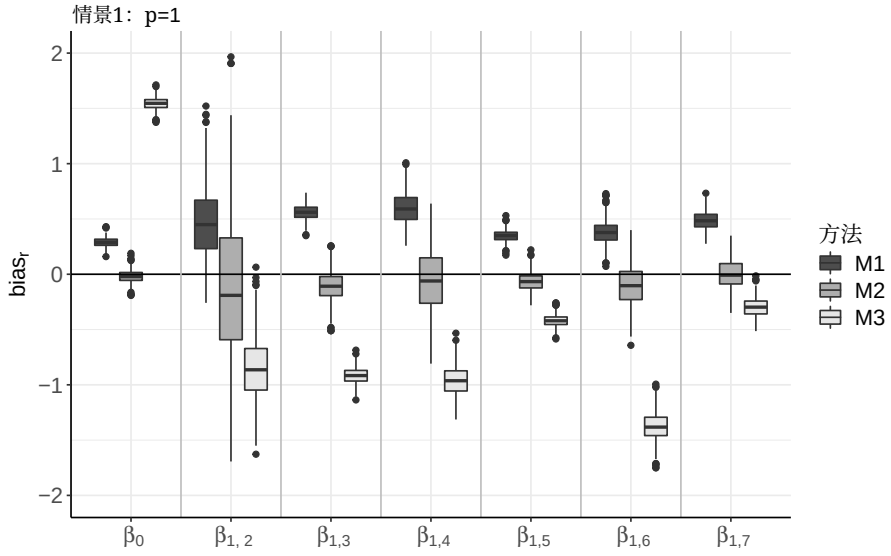
在图 1 中还可以看出，与其余方法相比，基于 M2方法获得的估计值的变异性最大。表 1 也显示了这一点，其中这些估计值的标准差最高可达 M1 和 M3 的两倍。例如，$\beta_{1,\ 3}$ 对应的估计值的标准差分别为 M1 的 0.063、M2 的 0.132 和 M3 的 0.070。先前研究中也观察到了加权估计的较大变异性（例如，参见 Scott and Wild（1986年））。这种变异性的来源可能

也与数据有关。总体而言，所有方法中系数 $\beta_{1,\ 2}$ 的估计值变异性都十分显著，尤其是 M2。需要注意的是，在情景1中，协变量X_1 的类别2中单位数量非常少。具体而言，总体中有 450 个单位（占有限总体单位总数的 0.2%）在该协变量上取此类别，而在样本中，这一数量从 27（0.3%）到 53（0.7%）不等（结果未显示）。这可能影响参数 $\beta_{1,\ 2}$的估计值，尤其是对于 M2方法。估计值 $\beta_{1,\ 4}$ 的表现也可以用同样的方式解释；同样观察到了较大的变异性，尤其是 M2（有限总体中有 2008 个单位（1.0%），样本中则从 178（2.5%）到 232（3.2%）不等）。此外，在表1中还应注意，对于所有方法而言，含有 $p = 3$ 的模型中三种方法的标准差均略高于含有 $p = 1$协变量的模型。

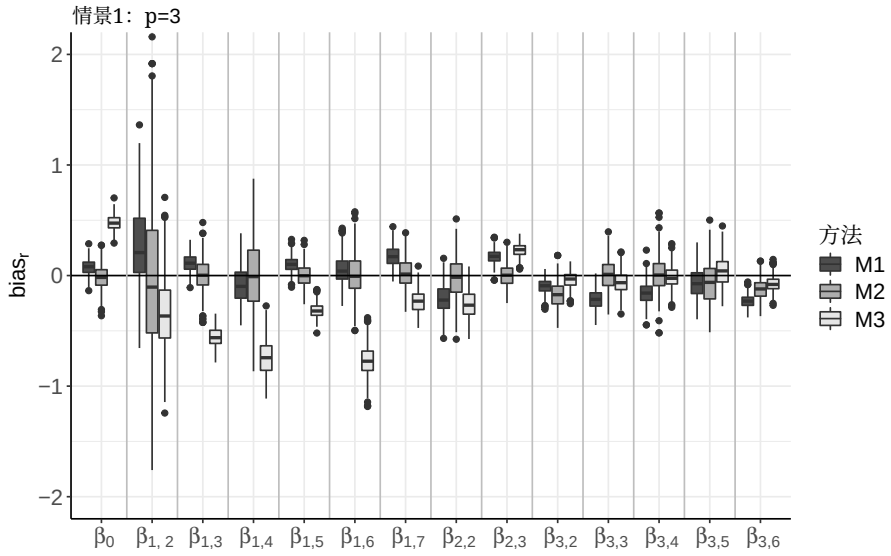
最后，如表1所示，在大多数情况下，M3方法的均方误差最大，这是由于基于该方法的估计值存在较大偏差。例如，在包含 $p = 1$ 的模型中，与类别$\beta_{1,\ 4}$ 对应的系数，其采用M3方法时的均方误差为0.722，而采用M2方法和M1方法时的均方误差分别为0.085和0.378。由于对于M1方法和M3方法，随着协变量的增加，偏差会减小，因此均方误差也会以相同方式减小。对于同一系数，当 $p = 3$时，M3方法对应的均方误差降至0.536。在包含 $p = 1$ 和$p = 3$ 协变量的两个模型中，M2方法的均方误差相当接近。比较M2方法和M1方法的均方误差可以发现，当 $p = 1$时，M1方法的均方误差更大。然而，在情景1中，当 $p = 3$时，由于M2估计值的变异性较大，尽管其偏差较小，M1方法和M2方法在均方误差方面并无差异。

图2 展示了情景2中含有 $p = 1$ 和 $p = 4$ 个协变量的模型采用方法 M1、M2 和 M3 得到的估计值偏差箱线图。在这种情况下，如图 2所示，三种方法在偏差和变异性方面的表现相当相似。不同方法之间以及不同模型之间（分别拟合了 $p = 1$ 和 $p = 4$个协变量）的差异都不显著。表 2 描述了均值、标准差、

平均偏差和这些估计值的均方误差，以及 $p = 1$ 和 $p = 4$ 在情景2中的真实有限总体系数。所有方法在两个模型中的平均偏差都很小，无论模型包含 $p = 1$ 还是 $p = 4$ 个协变量。观察到的最大平均偏差为 -0.067，对应于采用 M2方法时含有 $p = 4$ 个协变量的模型中的系数 $\beta_{4,\ 2}$ 。M2方法得到的估计值的变异性通常略高于其他方法。不过，如上所述，这些差异非常小。估计值标准差和均方误差方面的最大差异出现在含有 $p = 4$ 的模型中，对应于类别 $\beta_{2,\ 5}$的系数估计值。M2方法得到的估计值标准差为 0.185，而 M1 和 M3 方法对应的标准差为 0.174。同样，M2方法对该系数的均方误差为 0.035，而 M1 和 M3 方法的均方误差为 0.030。可以得出结论，所有受研究的方法都能在情景2中恰当地估计有限总体模型系数。

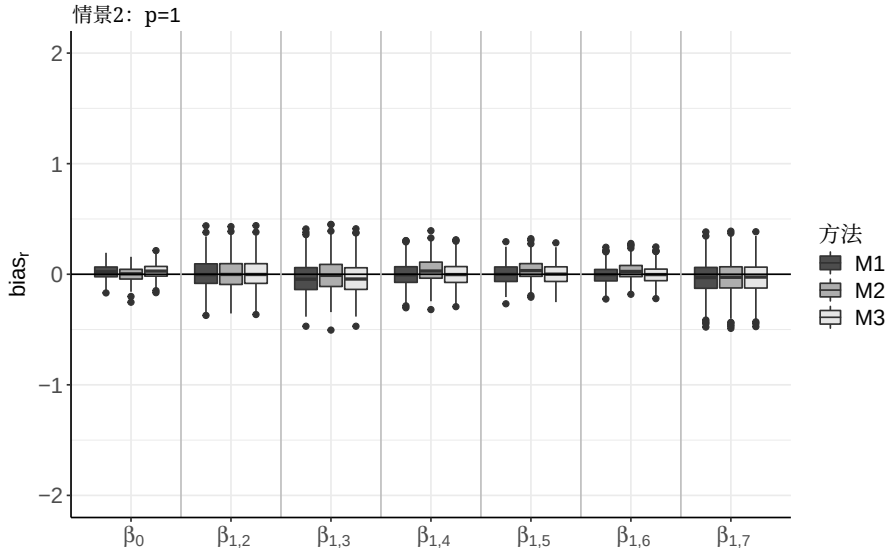


(a)

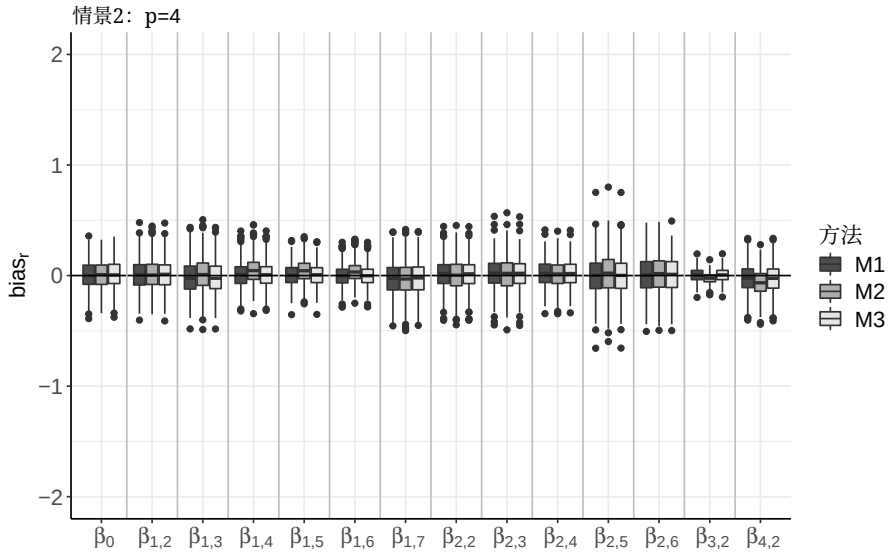


(b)

图 1： 情景1中，方法 M1、M2 和 M3 所得估计值的偏差箱线图：模型包含 (a) $p = 1$($l = 7$) 个协变量和 (b) $p = 3$($l = 14$) 个协变量， $\forall r = 1$、.. . 、 R。



(a)



(b)

图 2： 情景2中，方法 M1、M2 和 M3 对含有 (a) $p = 1(l = 7)$ 个协变量和 (b) $p = 4(l = 14)$ 个协变量的模型中各系数所得估计值的偏差箱线图 $\forall r = 1, \ldots, R$。

Table 1: True finite population model coefficients ($\boldsymbol{\beta}^{True}$) and the average (mean), standard deviation (sd), average bias (AvBias) and MSE of the estimates obtained by the M1, M2 and M3 methods for the models with $p = 1$ and $p = 3$ covariates in Scenario 1 for $R = 500$ samples.

	$\boldsymbol{\beta}^{True}$	M1			M2			M3		
		Mean(sd)	AvBias	MSE	Mean(sd)	AvBias	MSE	Mean(sd)	AvBias	MSE
p = 1 (l = 7)										
β_0	-1.015	-0.727(0.043)	0.288	0.085	-1.035(0.058)	-0.021	0.004	0.529(0.056)	1.544	2.387
$\beta_{1,2}$	1.184	1.658(0.325)	0.474	0.330	1.050(0.659)	-0.134	0.451	0.334(0.290)	-0.850	0.807
$\beta_{1,3}$	1.360	1.920(0.063)	0.560	0.318	1.252(0.132)	-0.108	0.029	0.443(0.070)	-0.916	0.845
$\beta_{1,4}$	1.342	1.942(0.138)	0.600	0.378	1.283(0.286)	-0.060	0.085	0.382(0.141)	-0.960	0.942
$\beta_{1,5}$	0.537	0.885(0.055)	0.348	0.124	0.471(0.085)	-0.067	0.012	0.118(0.053)	-0.420	0.179
$\beta_{1,6}$	1.908	2.286(0.110)	0.378	0.155	1.813(0.184)	-0.095	0.043	0.531(0.129)	-1.377	1.912
$\beta_{1,7}$	0.470	0.956(0.080)	0.487	0.243	0.471(0.121)	0.001	0.015	0.174(0.088)	-0.296	0.095
p = 3 (l = 14)										
β_0	-0.959	-0.883(0.069)	0.076	0.010	-0.979(0.111)	-0.020	0.013	-0.484(0.066)	0.475	0.230
$\beta_{1,2}$	0.602	0.850(0.356)	0.248	0.188	0.582(0.668)	-0.020	0.445	0.257(0.330)	-0.344	0.227
$\beta_{1,3}$	0.824	0.937(0.078)	0.112	0.019	0.830(0.146)	0.006	0.021	0.269(0.082)	-0.556	0.316
$\beta_{1,4}$	0.926	0.840(0.158)	-0.086	0.032	0.926(0.310)	0.000	0.096	0.184(0.161)	-0.742	0.576
$\beta_{1,5}$	0.382	0.482(0.066)	0.101	0.014	0.383(0.096)	0.002	0.009	0.065(0.064)	-0.317	0.104
$\beta_{1,6}$	1.145	1.195(0.124)	0.050	0.018	1.153(0.197)	0.007	0.039	0.374(0.138)	-0.771	0.614
$\beta_{1,7}$	0.355	0.526(0.089)	0.172	0.037	0.373(0.125)	0.018	0.016	0.119(0.097)	-0.235	0.065
$\beta_{2,2}$	-0.630	-0.843(0.136)	-0.212	0.064	-0.647(0.182)	-0.016	0.033	-0.892(0.121)	-0.262	0.083
$\beta_{2,3}$	0.036	0.207(0.058)	0.170	0.032	0.040(0.098)	0.004	0.010	0.266(0.056)	0.230	0.056
$\beta_{3,2}$	0.042	-0.055(0.065)	-0.097	0.014	-0.128(0.115)	-0.171	0.042	0.005(0.067)	-0.038	0.006
$\beta_{3,3}$	1.514	1.297(0.086)	-0.217	0.054	1.520(0.143)	0.006	0.020	1.448(0.091)	-0.066	0.013
$\beta_{3,4}$	1.418	1.260(0.096)	-0.158	0.034	1.428(0.150)	0.010	0.023	1.402(0.098)	-0.016	0.010
$\beta_{3,5}$	1.564	1.491(0.134)	-0.073	0.023	1.497(0.190)	-0.067	0.040	1.601(0.133)	0.037	0.019
$\beta_{3,6}$	2.700	2.465(0.053)	-0.234	0.058	2.577(0.093)	-0.123	0.024	2.623(0.069)	-0.077	0.011

Table 2: True finite population model coefficients ($\boldsymbol{\beta}^{True}$) and the average (mean), standard deviation (sd), average bias (AvBias) and MSE of the estimates obtained by the M1, M2 and M3 methods for the models with $p = 1$ and $p = 4$ covariates in Scenario 2 for $R = 500$ samples.

	$\boldsymbol{\beta}^{True}$	M1			M2			M3		
		Mean(sd)	AvBias	MSE	Mean(sd)	AvBias	MSE	Mean(sd)	AvBias	MSE
p = 1 (l = 7)										
β_0	-0.965	-0.945(0.068)	0.020	0.005	-0.965(0.065)	0.000	0.004	-0.941(0.068)	0.025	0.005
$\beta_{1,2}$	2.676	2.681(0.127)	0.005	0.016	2.680(0.133)	0.003	0.018	2.682(0.127)	0.006	0.016
$\beta_{1,3}$	3.369	3.337(0.149)	-0.032	0.023	3.366(0.152)	-0.003	0.023	3.338(0.149)	-0.031	0.023
$\beta_{1,4}$	3.539	3.539(0.110)	0.000	0.012	3.576(0.109)	0.038	0.013	3.539(0.109)	0.001	0.012
$\beta_{1,5}$	3.034	3.035(0.093)	0.002	0.009	3.071(0.091)	0.037	0.010	3.036(0.093)	0.003	0.009
$\beta_{1,6}$	1.571	1.568(0.082)	-0.003	0.007	1.600(0.075)	0.029	0.006	1.569(0.082)	-0.002	0.007
$\beta_{1,7}$	-2.854	-2.886(0.139)	-0.033	0.020	-2.883(0.146)	-0.029	0.022	-2.884(0.139)	-0.030	0.020
p = 4 (l = 14)										
β_0	-1.529	-1.523(0.125)	0.006	0.016	-1.524(0.127)	0.005	0.016	-1.516(0.125)	0.012	0.016
$\beta_{1,2}$	2.478	2.490(0.134)	0.012	0.018	2.490(0.140)	0.012	0.020	2.490(0.134)	0.012	0.018
$\beta_{1,3}$	3.206	3.188(0.153)	-0.017	0.024	3.221(0.156)	0.015	0.024	3.189(0.153)	-0.017	0.024
$\beta_{1,4}$	3.329	3.342(0.121)	0.013	0.015	3.379(0.122)	0.050	0.017	3.341(0.121)	0.013	0.015
$\beta_{1,5}$	2.780	2.785(0.103)	0.005	0.011	2.824(0.103)	0.044	0.012	2.786(0.103)	0.006	0.011
$\beta_{1,6}$	1.379	1.376(0.097)	-0.003	0.009	1.413(0.092)	0.034	0.010	1.378(0.097)	-0.001	0.009
$\beta_{1,7}$	-2.943	-2.974(0.149)	-0.031	0.023	-2.974(0.157)	-0.031	0.026	-2.970(0.150)	-0.028	0.023
$\beta_{2,2}$	0.702	0.715(0.134)	0.013	0.018	0.704(0.144)	0.002	0.021	0.714(0.134)	0.012	0.018
$\beta_{2,3}$	1.391	1.407(0.140)	0.016	0.020	1.404(0.148)	0.013	0.022	1.406(0.140)	0.015	0.020
$\beta_{2,4}$	0.811	0.827(0.117)	0.016	0.014	0.819(0.126)	0.008	0.016	0.827(0.117)	0.016	0.014
$\beta_{2,5}$	1.620	1.618(0.174)	-0.001	0.030	1.638(0.185)	0.018	0.035	1.618(0.174)	-0.001	0.030
$\beta_{2,6}$	1.664	1.669(0.159)	0.005	0.025	1.678(0.165)	0.014	0.027	1.670(0.159)	0.006	0.025
$\beta_{3,2}$	-0.427	-0.422(0.057)	0.005	0.003	-0.451(0.049)	-0.024	0.003	-0.421(0.057)	0.005	0.003
$\beta_{4,2}$	-0.481	-0.506(0.129)	-0.025	0.017	-0.548(0.123)	-0.067	0.019	-0.508(0.129)	-0.027	0.017

5 应用于真实数据集

本节将第3节所述的方法应用于第 2节介绍的真实调查数据。目标是比较通过不同方法获得的系数估计值。请注意，在本例中，真实的有限总体系数未知。

针对每项调查分别拟合一个模型。具体而言，我们将包含三个协变量（$p = 3$）的模型应用于 ESIE调查，将包含四个协变量（$p = 4$）的模型应用于 PRA 调查。这些协变量是两项调查模拟研究中所考虑的变量，也包含在 Eustat 实际应用的模型中。为拟合这些模型，我们采用了第3 节所述的三种方法：未加权逻辑回归（M1）、加权逻辑回归（M2）以及带随机截距的非加权逻辑回归（M3）。表 3 和表4 分别展示了针对 ESIE调查和 PRA 调查拟合的模型所得的系数估计值及其标准误。

表 3：使用方法 M1、M2 和 M3 对含有 $p = 3$ 个协变量的 ESIE调查所得的系数估计值（估计值）及其标准误（SE）。

	ESIE调查					
	M1		M2		M3	
	估计值	SE	估计值	SE	估计值	SE
β_0	-2.261	0.097	-2.482	0.133	-2.217	0.140
$\beta_{1,2}$	1.892	0.338	1.293	0.444	1.697	0.368
$\beta_{1,3}$	2.490	0.107	2.718	0.161	2.337	0.119
$\beta_{1,4}$	2.248	0.196	2.577	0.299	2.151	0.215
$\beta_{1,5}$	1.550	0.084	1.721	0.111	1.458	0.094
$\beta_{1,6}$	2.260	0.146	2.544	0.206	2.092	0.181
$\beta_{1,7}$	1.341	0.103	1.130	0.133	1.197	0.119
$\beta_{2,2}$	-0.774	0.148	-0.613	0.189	-0.883	0.329
$\beta_{2,3}$	0.453	0.073	0.358	0.107	0.538	0.123
$\beta_{3,2}$	0.669	0.069	0.632	0.097	0.750	0.077
$\beta_{3,3}$	0.996	0.096	0.965	0.132	1.124	0.134
$\beta_{3,4}$	1.479	0.114	1.452	0.152	1.698	0.149
$\beta_{3,5}$	2.230	0.182	2.205	0.241	2.461	0.209
$\beta_{3,6}$	2.454	0.143	2.532	0.151	2.787	0.195

As s如表 3所示，系数估计值及其标准误 s,

表4：使用方法 M1、M2 和 M3 对含有 $p = 4$ 个协变量的 PRA 调查数据进行分析所得的系数估计值（估计值）及其标准误（SE）。

	PRA 调查					
	M1		M2		M3	
	估计值	SE	估计值	SE	估计值	SE
β_0	-2.039	0.176	-2.040	0.171	-2.037	0.179
$\beta_{1,2}$	2.508	0.164	2.523	0.172	2.515	0.164
$\beta_{1,3}$	3.106	0.179	3.105	0.191	3.113	0.179
$\beta_{1,4}$	3.191	0.121	3.292	0.126	3.194	0.122
$\beta_{1,5}$	2.836	0.114	2.934	0.118	2.835	0.114
$\beta_{1,6}$	1.455	0.103	1.543	0.108	1.454	0.103
$\beta_{1,7}$	-3.170	0.184	-3.102	0.199	-3.182	0.184
$\beta_{2,2}$	1.005	0.174	0.899	0.177	1.016	0.174
$\beta_{2,3}$	1.689	0.178	1.587	0.182	1.700	0.179
$\beta_{2,4}$	1.167	0.171	1.056	0.170	1.170	0.172
$\beta_{2,5}$	2.123	0.207	1.970	0.227	2.128	0.208
$\beta_{2,6}$	2.357	0.192	2.177	0.201	2.360	0.193
$\beta_{3,2}$	-0.596	0.063	-0.546	0.067	-0.596	0.063
$\beta_{4,2}$	0.547	0.158	0.530	0.190	0.551	0.159

采用上述三种方法得到的结果在ESIE调查中存在显著差异。需要指出的是，估计值及其标准误的这些差异可能导致沃尔德统计量（定义为这些参数之间的比值）出现显著差异。然而，在本例中，这些差异并未影响模型参数的显著性，所有参数在统计上均显著（结果未显示）。在大多数情况下，M2方法得到的标准误最大。此外，与系数$\beta_{1,\ 2}$ 相关的标准误大于其他所有系数，这与模拟研究中观察到的该系数在情景1下具有较大变异性相符。根据模拟研究的结果，我们可以得出结论，在本例中，采用M2方法拟合的模型更为合适。

相比之下，PRA 调查所得的系数估计值及其标准误彼此非常相似，如表4所示。这也与模拟研究（情景2）中观察到的结果一致。正如预期，M2 得出的估计值的标准误通常略大于其他方法，不过总体而言差异并不大。

6 讨论

本研究通过一项基于真实数据的模拟研究，比较了三种不同方法在复杂调查数据的逻辑回归框架下估计模型系数的表现。总体而言，我们得到的结果与相关研究的结果一致；这些研究采用了逻辑回归框架（Chamblessand Boyle、1985；Lumley andScott、2017；Reiter 等人、2005；Scott and Wild、1986、2002）或线性回归框架（DeMets and Halperin、1977；Holt 等人、1980；Nathan andHolt、1980；Smith、1981）。不过，本研究与上述研究之间也存在一些差异。接下来我们将讨论这些相似之处与差异。

本研究与前文所述研究最大的区别之一在于，本研究是基于真实调查数据开展的模拟研究。研究旨在构建一个真实的情景，以便将获得的结果与有限总体模型的真实系数进行比较。模拟研究的数据基于巴斯克自治区官方统计局（Eustat）开展的两项真实调查模拟生成。两项调查中的有限总体均采用一步分层进行抽样，但分层的定义方式截然不同。ESIE调查根据三个类别较多的分类变量的组合定义层，共形成585个较小的层。相比之下，PRA 调查根据每个个体所属的地区定义层，共有23个不同的层。除了抽样设计之外，本研究还分析了模型中纳入的协变量数量和参数数量所产生的影响。值得注意的是，在本模拟研究中，我们并不知道生成有限总体的理论模型。因此，我们将基于所研究方法得到的模型估计值，与通过对有限总体拟合模型而获得的真实系数值进行比较。

本研究的主要结论是，加权逻辑回归（M2）在两种情景下均表现良好，所得估计值也无偏。相比之下，未加权逻辑回归（M1）和未加权

带随机截距的加权逻辑回归（M3）的表现取决于具体情景以及模型中估计的协变量/参数数量。在与 ESIE调查相关的情景中，与基于 PRA 调查的情景不同，采用这两种方法得到的估计值存在偏差。这些结果与Holt 等人(1980年); Nathan and Holt (1980年); Scott and Wild (1986年) 等人的研究一致；这些研究同样指出，无论在线性回归还是逻辑回归框架中，未加权系数估计值都可能存在偏差。Scott and Wild (1986年) 声称，当拟合样本的模型与生成数据的真实理论模型完全相同时，未加权系数估计值的偏差会小于拟合模型“合理但并不完美”时的偏差。如前所述，我们并不知道生成有限总体的理论模型。不过，本研究也观察到，随着模型中纳入更多协变量，偏差会减小，这与上述研究的结果一致。然而，这一偏差仍大于通过加权逻辑回归得到的偏差。

加权逻辑回归模型所得估计值的变异性大于通过未加权逻辑回归模型（有无随机截距均包括在内）所得估计值的变异性，这与 Cham- bless 和 Boyle (1985年）；Lumley andScott (2017年）；Scott and Wild (1986年）的结论一致。在大多数情况下，这些差异并不大。然而，我们观察到，当分类变量的某个类别中个体数量较少时，该类别对应系数的加权估计的变异性可能远大于未加权估计的变异性。我们认为，当分类变量中各类别的个体分布不平衡时，应当谨慎处理。

我们还将所研究的三种方法应用于真实调查数据，所得估计值与模拟研究中观察到的结果一致。一方面，在 PRA 调查中，各方法所得估计值相当接近，这些估计值的标准差之间也没有太大差异，因此我们得出结论，所有受研究的方法在这种情况下都能正常发挥作用。另一方面，在 ESIE 调查中，不同方法所得估计值存在较大差异

不同方法所得参数的估计值存在差异。鉴于模拟研究与真实数据集应用结果之间的相似性，并考虑到这些结果也与类似实证研究（例如Chambless andBoyle （1985年）和Lumley and Scott （2017年）的结果一致，我们可以认为，在处理 ESIE 调查数据时，加权逻辑回归应是首选方法。

现在，我们将讨论这项工作的局限性。首先，在本次模拟研究中，我们无法得知数据源自哪个理论模型，因为我们旨在基于真实调查数据开展模拟研究，因此重点比较了根据样本得到的估计值与对有限总体拟合模型所得的真实系数。需要指出的是，处理调查数据的目的往往是对特定有限总体得出结论，因此，在这一背景下进行这种比较是有意义的。对于有兴趣将结果与理论上的无限总体模型进行比较的读者，我们建议查阅Scott andWild (2002年)。其次，如上所述，一些作者建议将设计变量及其交互作用作为协变量纳入模型。然而，在本研究中，尤其是对于 ESIE调查而言，由于模型需要估计的参数数量过多（总计 585 个），这一做法并不可行。因此，我们决定拟合混合模型，以此复现 LumleyandScott(2017年)在真实数据集上所做的比较。此外，模型中纳入的一些协变量与分层变量相关。还需指出的是，本次模拟研究采用了样本量相当大的调查，这在官方统计中十分常见。尽管如此，我们也认为，基于样本量较小的实际调查开展模拟并比较结果会很有意义，尤其应关注估计值的变异性。最后，本研究聚焦于逻辑回归模型参数的估计。其他值得关注的问题，例如协变量的选择，或这些差异对个体估计概率的影响，均不在本研究的讨论范围内。

总而言之，加权逻辑回归在我们设定的所有情景中都表现良好。相比之下，未加权逻辑回归（无论是否包含随机截距）的表现取决于具体情景。因此，根据模拟研究的结果，我们认为，在需要使用抽样权重时不使用权重，所导致的结果会比在不需要使用权重时使用权重更差。基于此，我们建议在复杂调查数据中使用加权逻辑回归模型。

致谢
本研究部分经费由巴斯克政府教育、语言政策与文化部通过资助项目 ´IT1456-22] 提供；另由科学与创新部通过 BCAM Severo Ochoa 认证[CEX2021-001142-S / MICIN / AEI / 10.13039/501100011033] 以及项目[PID2020-115882RB-I00 /AEI / 10.13039/501100011033] 提供支持。该项目由国家研究署资助，简称“S3M1P4R”。此外，本研究还获得了巴斯克政府 BERC 2022—2025 计划的支持。AI 的研究工作获得了资助[PIF18/213]的支持。

我们谨此感谢巴斯克官方统计局（Eustat）向我们提供 ESIE 和 PRA 调查数据。

利益冲突
作者声明不存在任何利益冲突。

参考文献

Bates, D., M. M¨achler, B. Bolker, 和 S. Walker (2015). 使用 lme4 拟合线性混合效应模型。《统计软件杂志》*67*(1), 1–48。

Binder, D. A. (1981). 关于复杂调查中渐近正态估计量的方差。《调查方法学》*7*(3), 157–170。

Binder, D . A.（1983）。论复杂调查中渐近正态估计量的方差 om

。《国际统计评论》/《国际统计评论》*51*(3), 279–292。Binder, D. A. 和 G. Roberts（2009）。模型参数的基于设计和模型的推断。载于《统计学手册》，第29卷，第33–54页。荷兰：Elsevier。Brewer, K. R. W. 和 R. W. Mellor（1973）。样本结构对分析调查的影响。《澳大利亚统计学杂志》*15*(3), 145–152。Chambers, R. L. 和 C. J. Skinner（2003）。《调查数据分析》。纽约：约翰·威立父子公司。Chambless, L. E. 和 K. E. Boyle（1985）。复杂样本数据的最大似然法：逻辑回归和离散比例风险模型。《统计学通讯：理论与方法》*14*(6), 1377–1392。DeMets, D. 和 M. Halperin（1977）。对通过子抽样程序获得的样本中的简单回归系数进行估计。《生物统计学》*33*(1), 47–56。Diggle, P.、K.-Y. Liang 和 S. L. Zeger（2002）。《纵向数据分析（第*2*版）》。牛津：牛津大学出版社。Gelman, A.（2007）。调查加权与回归建模面临的挑战。《统计科学》*22*(2), 153–164。Hausman, J. A. 和 D. A. Wise（1981）。内生变量的分层与估计：加里收入维持实验。载于《离散数据的结构分析及其计量经济学应用》，第365–391页。马萨诸塞州剑桥：麻省理工学院出版社。Heeringa, S. G.、B. T. West 和 P. A. Berglund（2017）。《应用调查数据分析》。佛罗里达州博卡拉顿：Chapman and Hall/CRC。

Holt, D.、T. M. F. Smith 和 P. D. Winter（1980年）。复杂调查数据的回归分析。英国皇家统计学会杂志：*A*辑（综合）*143*(4)，474–487。Korn, E. L. 和 B. I. Graubard（1995年）。抽样调查中不同加权与未加权估计值的示例。美国统计学家 *49*(3)，291–295。Lee, Y. 和 J. A. Nelder（2001年）。层次广义线性模型：广义线性模型、随机效应模型与结构化离散度的综合。*Biometrika 88*(4)，987–1006。Lee, Y. 和 J. A. Nelder（2004年）。条件模型与边际模型：另一种视角。《统计科学》*19*(2)，219–238。Lumley, T.（2019年）。survey：复杂调查样本分析。R 软件包版本 3.35-1。Lumley, T. 和 A. Scott（2017年）。将回归模型拟合于调查数据。《统计科学》*32*(2)，265–278。Masood, M.、T. Newton 和 D. Reidpath（2016年）。复杂调查数据的四种分析策略比较：西班牙数据案例研究。流行病学、生物统计学与公共卫生 *13*(1)，1–7。McCullagh, P. 和 J. A. Nelder（1989年）。广义线性模型（第*2*版）。伦敦：Chapman & Hall。Nathan, G. 和 D. Holt（1980年）。调查设计对回归分析的影响。英国皇家统计学会会刊：*B*辑（方法学）*42*(3)，377–386。Pfeffermann, D. 和 M. Sverchkov（2009年）。信息性抽样下的推断。载于《统计学手册》，第29卷，第455–487页。荷兰：Elsevier。Reiter, J. P.、E. L. Zanutto 和 L. W. Hunter（2005年）。复杂调查中工作实践的分析建模。*ILR*评论 *59*(1)，82–100。

Rubin, D. B.（1976年）。推断与缺失数据。*Biometrika 63*(3)，581–592。Scott, A. J.（1977年）。关于调查抽样中随机化问题的探讨。*Sankhya 39*，1–9。Scott, A. J. 和 C. J. Wild（1986年）。在病例对照抽样或基于选择的抽样下拟合逻辑斯蒂模型。英国皇家统计学会会刊：*B*辑（方法学）*48*(2)，170–182。Scott, A. J. 和 C. J. Wild（2002年）。关于加权方法在拟合病例对照数据模型时的稳健性。英国皇家统计学会会刊：*B*辑（统计方法学）*64*(2)，207–219。Skinner, C. J.、D. Holt 和 T. M. F. Smith（1989年）。复杂调查分析。纽约：约翰·威立父子公司。Smith, T. M. F.（1981年）。复杂调查的回归分析。载于调查抽样当前主题，第267–292页。纽约：学术出版社。Sugden, R. A. 和 T. M. F. Smith（1984年）。调查抽样推断中的可忽略设计与信息性设计。*Biometrika 71*(3)，495–506。Wolter, K.（2007年）。方差估计导论（第*2*版）。纽约：施普林格出版社。

伪总体生成

本节介绍用于第4节模拟研究的伪总体生成过程，涉及情景1（基于ESIE调查）和情景2（基于PRA survey）。

情景2中应用的、与PRA 调查相关的伪总体实际上是一个真实的有限总体，其中响应变量以及其余解释变量均为已知。该伪总体由Eustat构建并提供。

对于情景1，我们基于ESIE调查的真实有限总体和样本生成了一个伪总体。将 S_{ESIE} 记为原始调查样本，将 U_{ESIE} 记为规模为 N ($S_{\text{ESIE}} \subset U_{\text{ESIE}}$)的真实有限总体。如第2节所述，通过结合三个分类变量的信息，共定义了 H 个层（即 $\{1,\ldots,H\}$），这三个变量分别记为 X_1、X_2 和 X_3。因此，有限总体可以划分为由这些层定义的子集，即$U_{\text{ESIE}} = \bigcup_{h=1}^{H} U_{\text{ESIE},\,h}$。 $\forall h \in \{1,\ldots,H\}$ 将N_h 记为有限总体中第 h 层的规模 U_{ESIE} （$U_{\text{ESIE},\,h}$），将 n_h 记为该层在样本中的规模 S_{ESIE}。那么，来自第 h 层的单位$j \in S_{\text{ESIE}}$ 对应的抽样权重如下：

$$w_j = \frac{N_h}{n_h}. \tag{12}$$

我们的目标是根据已知的真实 ESIE 调查数据生成一个伪总体（U），其中包含协变量 $X_1,\ldots,X_p$和响应变量 $Y_1,\ldots,Y_q$的所有信息。这个新的伪总体 U将与真实 ESIE 总体（N）规模相同。为简化记号，伪总体中的变量名称与真实有限总体中的相同；真实 ESIE 总体中的单位记为 $j \in U_{\text{ESIE}}$，而为伪总体人工生成的单位记为 $i \in U$。

原始 survey 中有多个二分类响应变量，其中之一是响应变量 Y，即我们在模拟研究中采用的变量。我们考察了这些响应变量的所有可能组合。例如，假设 Y_1、.....、 Y_q 是 survey 中所有可用的响应变量（其中 $Y \in \{Y_1,\ldots,Y_q\}$），对于某个 $j \in S_{\text{ESIE}}$: $\boldsymbol{y}_j = (y_{1,\,j}$、.....、$y_{q,\,j}) = \alpha, \forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$，其中 $\{\alpha_1,\ldots,\alpha_A\}$ 是所有可能的响应组合构成的集合。对于每一层 $h \in \{1,\ldots,H\}$ 以及响应的每一种可能组合（即 $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$），我们在伪总体（$U$）中生成 $N_{h,\,\alpha}$ 个单位，具体如下：

$$N_{h,\alpha} = \sum_{j \in S_{\text{ESIE}}} w_j 1_{U_{\text{ESIE},\,h}}(j)\left[\boldsymbol{y}_j = \alpha\right], \tag{13}$$

其中，

$$1_{U_{\text{ESIE},\,h}}(j) = \begin{cases} 1, & \text{if } j \in U_{\text{ESIE},\,h}, \\ 0, & \text{if } j \notin U_{\text{ESIE},\,h}, \end{cases} \tag{14}$$

and

$$[\boldsymbol{y}_j = \alpha] = \begin{cases} 1 & \text{if } \ \boldsymbol{y}_j = \alpha, \\ 0 & \text{if } \ \boldsymbol{y}_j \neq \alpha. \end{cases} \tag{15}$$

这样，$N_{h,\,\alpha}$ 表示伪总体 U 中第h层内响应取值为（$y_{1,\,j}$，...，$y_{q,\,j}$）$=\alpha$的单位数。对 $\forall h \in \{1,\ldots,H\}$ 和 $\forall \alpha \in \{\alpha_1,\ldots,\alpha_A\}$ 重复这一过程后，我们便根据响应变量（Y等）和层的信息生成了一个包含 $N = \sum_{h\in\{1,\ldots,H\}} \sum_{\alpha\in\{\alpha_1,\ldots,\alpha_A\}} N_{h,\,\alpha} = \sum_{j\in S_{\text{ESIE}}} w_j$ 个单位的伪总体（因此，设计变量 X_1、 X_2 和 X_3 的信息也将随之生成）。

最后，我们按如下方式生成其余协变量： $\forall d \in \{4,\ldots,p\}$ 假设 X_d 是一个共有 K 个类别的分类变量： $\{1,\ldots,K\}$ 个类别。然后，对于在伪总体（$\forall i \in U$）中生成的每个单位 i ，我们根据分类分布生成$x_{di} \in \{1,\ldots,K\}$ （即 $x_{di} \sim Cat$（π_{d1}，...， π_{dK}））， 其中对应于每个类别 $k \in \{1,\ldots,K\}$ 的概率根据已知的 ESIE 有限总体 U_{ESIE}计算如下。假设$i \in U_{ESIE,\,h}$， $\forall h \in \{1,\ldots,H\}$。那么，

$$\pi_{dk} = \frac{\sum_{j\in U_{\text{ESIE}}} 1_{U_{\text{ESIE},\,h}}(j)\,[x_{dj} = k]}{\sum_{j\in U_{\text{ESIE}}} 1_h(j)}, \forall k \in \{1,\ldots,K\}, \tag{16}$$

其中$1_{U_{\text{ESIE},\,h}}$（j）在式（14）中定义，并且，

$$[x_{dj} = k] = \begin{cases} 1 & \text{if } \ x_{dj} = k, \\ 0 & \text{if } \ x_{dj} \neq k, \end{cases} \quad \forall j \in U_{\text{esie}} \text{ and } \forall k \in \{1,\ldots,K\}. \tag{17}$$

这样，基于 ESIE调查 的伪总体便已生成，其中包含响应变量 Y、解释变量向量 $\boldsymbol{X}$ 和层。

B 伪总体抽样过程

这两个伪总体采用一步分层抽样进行抽样，方式与第2节所述的真实调查数据相同。

为了从情景1的伪总体中抽样，首先，我们确定在真实调查样本 S_{ESIE} 中，从层 $h, \forall h \in \{1, \ldots, H\}$ 抽取了多少个单位（将这个数量记为 n_h）。然后，我们从伪总体 U的层 h 中随机抽取 n_h 个单位。对$\forall h \in \{1, \ldots, H\}$ 重复相同的过程，我们得到一个样本 S.

最后，按如下方式为每个抽样单位分配抽样权重。对于 $\forall i \in S$，假设 i 是层 $h, \forall h \in \{1, \ldots, H\}$中的一个单位，则：

$$w_i = \frac{N_h}{n_h}, \tag{18}$$

其中， N_h 表示 U中层 h 的单位数， n_h 表示 S中层 h 的单位数。