《统计软件杂志》 MMMMMM YYYY年，第VV卷，第II期。

0.i00

arXiv:2305.02009v1 [stat.ML] 3 May 2023

fairml：统计学家对公平机器学习建模的看法

Marco ScutariIstituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)

摘要

在必须确保公平性和问责制的应用中，机器学习的采用促使文献中涌现了大量模型提案。这些提案大多被构造为带约束的优化问题，旨在减少或消除敏感属性对响应变量的影响。尽管从理论角度看，这种方法非常灵活，但由此产生的模型在某种程度上具有黑箱性质：人们对其统计性质、应用中的最佳实践，以及如何将其扩展到最初设计目标之外的问题知之甚少。此外，从软件工程的角度看，每个模型的估计都需要专门实现并采用适当的求解器，这并不理想。

本文介绍**fairml** R 包，它实现了我们先前的研究（Scutari、Panero和Proissl（2022））以及文献中的相关模型。**fairml** 以经典统计模型（广义线性模型）和惩罚回归结果（岭回归）为基础，构建可解释且性质明确的公平模型。用于实现公平性的约束与模型估计彼此正交，因此可以针对每种应用灵活组合所需的模型族和公平性定义。此外，**fairml** 还提供模型估计、模型选择和验证功能，包括诊断图。

关键词：公平性、岭回归、广义线性模型。

1. 引言

在许多应用中，确保基于机器学习模型所做决策的可解释性、可说明性和公平性至关重要：例如刑事司法（Berk、Heidari、Jabbari 等 2021）、信用风险建模（Fuster、Goldsmith-Pinkham、Ramadorai和Walther（2022））以及求职申请筛选（Raghavan、Barocas、Kleinberg和Levy（2020））。未能做到这一点，已经导致了基于种族的歧视（Angwin,

Larson、Mattu 和 Kirchner（2016年））、性别（Lambrecht 和 Tucker（2019年））、国籍（Tolan、Miron、Gómez 和 Castillo（2019年））以及年龄（Díaz、Johnson、Lazar 等人 2019年；Hort 和 Sarro（2022年））等。因此，美国、欧盟和英国最近都出台了相关立法，以规范机器学习模型的使用（Cath、Wachter、Mittelstadt 等人2018年）。欧盟委员会还发布了首个人工智能使用法律框架（欧盟委员会 2021），目前该框架正在接受成员国的修订。从更广泛的层面来看（包括地理和时间维度），提高公平性和减少不平等是联合国可持续发展目标的重要组成部分（联合国 2015年）。

法律要求以及 AI for Good（联合国 2023）等倡议的出台，推动了算法公平性作为一个独立研究领域的发展。除了对公平性进行不同的数学刻画之外，研究人员还提出了许多模型和算法，通过利用（受约束的）优化和信息论，在确保不基于敏感属性（或法律限制的属性）歧视个人的同时，实现尽可能高的预测准确度。这些模型和算法通常仅由目标函数和约束集合构成，分别用于表示拟合优度和公平性，而非完整的概率模型：它们的统计特性、模型选择和验证的最佳实践、显著性检验等，都必须针对每个模型从头重新推导。在我们此前的研究中（Scutari 等 2022），我们采取了相反的观点，认为可以调整经典统计模型以实现公平性，同时保留其众所周知的特性及其在实际应用中相关的最佳实践。我们证明，将广义线性模型（GLM）与惩罚回归相结合，非常适合实现这一目标。如今，我们为 R （R 核心团队 2022）提供了这些模型及相关模型的生产级实现，并将其整合到 **fairml** 软件包中（Scutari 2022）。R 为统计建模提供了丰富的集成环境，而结构良好且经过充分测试的 R 软件包，比算法公平性文献中通常提供的Python 概念验证脚本更加可靠，也更适合一般用途。

本文旨在概述 **fairml**。在第 2 节中，我们将简要回顾公平机器学习模型的方法学文献（第 2.1 节）以及软件包的可用情况（第 2.2 节），尤其关注线性模型。在第 3 节中，我们将介绍先前工作中的公平岭回归（FRRM）和公平广义岭回归（FGRRM）模型（Scutari 等人，2022年），以及 **fairml** 中实现的其他模型。在第4 节中，我们将介绍该软件包的软件架构和功能。最后，我们将在第5节说明相关函数的用法，并在第6节总结我们的结论。

2. 公平机器学习模型与软件

算法公平性研究包括两个互补的主题：探索公平性的数学表征，以及高效估计能够产生公平预测的模型。

文献中各种公平性标准及其特征，近期已由Mehrabi、Morstatter、Saxena 等人 (2021)、Del Barrio、Gordaliza 和 Loubes(2020年) 以及 Pessach 和 Shmueli (2022年) 进行了综述。总体而言，这些标准分为两种路径： 群体公平和 个体公平。群体公平标准要求由敏感属性划分的各群体之间的预测相似。对于给定的模型，这些标准通常表示为

即某个响应变量 $\mathbf{y}$ 的拟合值 $\hat{}$与敏感属性 $\mathbf{S}$之间的条件独立关系。两个常见例子是统计 或人口统计均等($\hat{}\mathbf{y} \perp\!\!\!\perp \mathbf{S}$) 以及机会均等 ($\hat{}\mathbf{y} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{y}$)。另一方面，个体公平标准要求相似的个体获得相似的预测，并表示为不同群体中的成对个体（由下文的 $d_2(\cdot)$ 标识）之间预测差异的累积值（下文的 $d_1(\cdot)$）： i, j

$$f(\boldsymbol{\alpha}\mathbf{y}, \mathbf{S}) = \sum_{i,j} d_1(y_i, y_j) d_2(\mathbf{s}_i\boldsymbol{\alpha}, \mathbf{s}_j\boldsymbol{\alpha}). \tag{1}$$

群体公平和个体公平最初都是在一个简单场景中定义的：其中$\mathbf{S}$ 包含单一的二元敏感属性，但针对特定模型，这些定义已扩展至多个敏感属性和连续敏感属性（例如参见Komiyama、Takeda、Honda 和 Shimao 2018年；Zafar、Valera、Gomez-Rodriguez 和 Gummadi 2019年）。

在公平建模方面，人们已尝试在模型选择、估计和验证过程的不同阶段，以及针对不同类别的模型，采取多种方法来解决歧视问题。就前者而言，我们可以区分以下几类（D'Alessandro、O'Neil 和 LaGatta 2017年）：

- 预处理方法旨在转换数据以消除其潜在的歧视，从而确保基于转换后数据拟合的任何模型都是公平的。采用这一方法的奠基性研究是Calmon、Wei、Vinzamuri等人(2017年，该研究阐述了如何学习一种概率变换，在推动数据实现群体公平和个体公平的同时，惩罚较大的特征变化，以保持数据保真度。

- 处理中 方法通过修改模型估计过程来消除歧视，可以通过改变其目标函数（通常为对数似然），或对其参数施加约束来实现。

- 后处理 方法利用留出集评估先前估计的模型（将其视为黑箱），并调整其预测结果，使其更加公平。 Hardt、Priceric 和 Srebro（2016年）例如展示了如何从（用于二分类的）非公平分类器中导出公平的贝叶斯最优预测器。

处理中方法的性质很大程度上取决于我们正在估计的模型类型。学习公平的黑箱机器学习模型（如深度神经网络）最具挑战性（例如，参见 Choraś、Pawlicki、Puchalski 和 Kozik 2020年），并且正在可解释人工智能领域进行研究（Barredo Arrieta、Díaz-Rodríguez、Del Ser 等 2020年）。因此，文献中的很大一部分都聚焦于更简单的模型。在许多场景中，这类模型更可取，因为可用数据有限、存在计算限制，或者它们更易于解释。fairml fair ml 软件包和我们之前的工作（Scutari 等 2022年）便属于这种情况：下一节将更详细地概述这些类型的公平模型。

2.1. 公平回归和分类模型的相关研究

文献中已对所有常用的经典线性模型进行了改造，以确保公平性：公平分类模型比公平回归模型更为常见，而公平回归模型又比其他广义线性模型（GLM）类别更为常见。

分类模型通常基于（二元）逻辑回归：例如，Wood- worth、Gunasekar、Ohannessian 和 Srebro (2017年) 直接对二元敏感属性施加机会均等约束；而 Zafar 等人(2019年) 则研究了统计均等条件下逻辑回归和支持向量机中决策边界的不公平性。Agarwal、Beygelzimer、Dudik 等人 (2018年) 使用逻辑回归和梯度提升树的集成方法，将公平分类转化为一系列代价敏感分类问题。多项逻辑回归尚未得到同等程度的研究，因为其公式构造更加复杂（Narasimhan 2018；Cotter、Gupta、Jiang 等人 2019年）。

对于公平线性回归模型，研究者通过辅助生成模型（Fukuchi、Sakuma 和 Kamishima 2013年）、聚类（Calders、Karim、Kamiran 等人2013年）和核正则化（Pérez-Suay、Laparra、Mateo-García 等人 2017年）来实现公平性。这些方法及其他方法（Agarwal、Dudik 和 Wu 2019年；Chzhen、Denis、Hebiri等人 2020年）借鉴公平分类文献中的方法，将响应变量、敏感属性或二者离散化，或将模型限制为仅包含一个离散（通常为二元）敏感属性。同样，许多此类方法将公平性定义为机会均等；Berk、Heidari、Jabbari 等人(2017年) 是一个显著的例外，它同时实现个体公平和群体公平。

除我们的研究外，另外两项公平回归研究已在 fairml：Komiyama 等人(2018年) 和 Zafar 等人 (2019年) 中实现。两者均将公平性定义为统计均等。前者提出了一种针对公平线性回归模型的二次优化方法，通过限制由敏感属性解释的方差相对比例，对最小二乘估计施加约束。后者则对各个敏感属性与拟合值之间的相关性设定界限。与上述方法相比，这两种方法有若干优势：

- 预测变量和敏感属性既可以是连续的，也可以是离散的（使用对比编码）；
- 模型中可以纳入任意数量的预测变量和敏感属性；
- 用户可以直接控制公平性水平，无需通过模型校准来进行经验估计；
- 公平性以与模型损失函数相同的方式进行数学编码，这意味着敏感属性的影响不会在模型估计过程中被意外捕获。

Do、Putzel、Martin 等人（2022年）和Scutari 等人（2022年）采用了类似的方法，并将其扩展至广义线性模型，以涵盖线性回归和分类：前者使用广义线性模型对数似然的一阶近似，后者则使用岭惩罚（我们将在第3节进一步详细讨论）。

2.2. 相关软件

写作本文时，可用于处理公平模型的软件包并不多。除了R 生态系统之外，IBM AI Fairness 360（Bellamy、Dey、Hind 等 2019年）是唯一提供大量预处理和后处理方法，以及一些处理中方法的选择。

至于R，**mlr3fairness** (Pfisterer, Siyi, andLang 2022) 软件包为**mlr3** R 包（Lang, Binder, Richter 等 2019年）扩展了预处理和后处理方法，并且

为**fairml** 提供了用于处理中方法的前端。公平性(Kozodoi and Varga2021) 和 **fairmodels** (Wiśniewski and Biecek 2022) 软件包实现了公平性指标的计算与可视化。 **predfairness** (de B.G. de Oliveira, Vieira, Silva 等 2021）对模型预测进行后处理。

3. 公平广义岭回归模型

fairml 的核心是Scutari等人（2022）提出的公平岭回归模型（FRRM）和公平广义岭回归模型（FGRRM）。FRRM 定义为

$$\mathbf{y} = \mu + \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{with} \qquad \widehat{\mathbf{U}} = \mathbf{X} - \mathbf{B}^{\mathrm{T}}\mathbf{S}, \tag{2}$$

其中，$\mathbf{y}$ 是响应变量，$\mathbf{X}$ 是预测变量的设计矩阵，$\mathbf{S}$ 是敏感属性的设计矩阵$\widehat{}$，而 $\boldsymbol{\varepsilon}$ 是模型的残差。矩阵$\mathbf{U}$ 表示预测变量$\mathbf{X}$ 中与敏感属性$\mathbf{S}$正交的部分，通过用普通最小二乘法将前者对后者进行回归并取残差计算得出。

与$\mathbf{S}$ 相关的回归系数 $\boldsymbol{\alpha}$ 受到岭惩罚，以减弱敏感属性对$\mathbf{y}$的$\widehat{}$影响；而 $\boldsymbol{\beta}$ 不受惩罚，因为 $\mathbf{U}$ 与$\mathbf{S}$ 正交，因此不受歧视效应影响。[1] 换言之， $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 的估计方式为

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mu - \mathbf{S}\boldsymbol{\alpha} - \widehat{\mathbf{U}}\boldsymbol{\beta}\|_2^2 + \lambda(r)\|\boldsymbol{\alpha}\|_2^2 \tag{3}$$

其中， $\lambda(r) \in \mathbb{R}^+$ 是惩罚系数，令

$$R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \frac{\mathsf{VAR}(\mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}})}{\mathsf{VAR}(\widehat{\mathbf{y}})} = \frac{\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_{\mathrm{FRRM}}\mathsf{VAR}(\mathbf{S})\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}}{\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_{\mathrm{FRRM}}\mathsf{VAR}(\mathbf{S})\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} + \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{FRRM}}\mathsf{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}} \tag{4}$$

从用户指定的不公平程度 $r \in [0, 1]$ 开始。也就是说，用户选择模型应当具备多大程度的公平性，而 FRRM 会在内部选择 $\lambda(r)$ 的值，在这一约束条件下获得最佳拟合优度$\widehat{}$。对于给定的 $\lambda(r)$， $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ 和 $\boldsymbol{\beta}_{\mathrm{FRRM}}$ 有通常的闭式表达式

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda(r)\mathbf{I}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y}, \qquad \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} = (\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}})^{-1}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \tag{5}$$

因此，从 r 数值求解 $\lambda(r)$ 在计算上十分高效。
不公平程度定义为拟合值 $\widehat{}\mathbf{y}$ 的方差中由敏感属性 $\mathbf{S}$ 解释的比例： $r = 0$ ($\lambda(r) \to +\infty$) 对应于一个完全公平的模型，在该模型中， $\widehat{}\mathbf{y}$ 与 $\mathbf{S}$ 相互独立；而根据构造， $r = 1$ ($\lambda(r) = 0$) 是一个非活跃约束，因为 $R^2_{\mathbf{S}} \leqslant 1$。在实际应用中，施加 $r = 0$ 通常不可行，因此更倾向于较低的 r 值，因为这能在公平性和预测准确度之间实现更好的权衡。

(3) 中的 FRRM 模型可以重写为如下优化问题

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathsf{E}\left[(\mathbf{y} - \widehat{\mathbf{y}})^2\right] \quad \text{such that} \quad \|\boldsymbol{\alpha}\|_2^2 \leqslant t(r),$$

[1] μ 也不会受到惩罚，这符合惩罚回归模型的标准做法。

这与Komiyama 等人（2018年）提出的非凸线性模型（NCLM）类似：

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathsf{E}\left[(\mathbf{y}-\widehat{\mathbf{y}})^2\right] \text{ such that } R^2_{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant r. \tag{6}$$

后者直接约束 $R^2_{\mathbf{S}}$，并使用二次约束二次规划求解器估计 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$，而不是仅约束 $\boldsymbol{\alpha}$。因此，它不容易扩展为广义线性模型；而且，估计得到的 $\boldsymbol{\beta}_{\text{NCLM}}$ 的行为不符合直觉，因为尽管 $\mathbf{U}$ 不存在歧视效应，它们仍会随 $\widehat{}$ 和 r 变化。

另一方面，借助广义岭回归模型的相关文献（Friedman、Hastie 和 Tibshirani 2010；Simon、Friedman、Hastie 和 Tibshirani 2011），并将 $\mathbf{S}$ 所解释的方差比例替换为相应的偏差比例，将 FRRM 扩展为 FGRRM 很容易。也就是说，我们估计 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 为

$$(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}}\, D(\boldsymbol{\alpha},\boldsymbol{\beta}) + \lambda(r)\|\boldsymbol{\alpha}\|_2^2. \tag{7}$$

其中 $D(\cdot)$ 为偏差，并且我们选择 $\lambda(r)$，使得

$$\frac{D(\boldsymbol{\alpha},\boldsymbol{\beta}) - D(\mathbf{0},\boldsymbol{\beta})}{D(\boldsymbol{\alpha},\boldsymbol{\beta}) - D(\mathbf{0},\mathbf{0})} \leqslant r. \tag{8}$$

其中$\mathbf{0}$ 是零向量。具体而言：

- 对于高斯广义线性模型，（7）与（3）完全相同，因为偏差就是残差平方和，并且（8）可简化为 $R^2\mathbf{s}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r$。FGRRM 只需恢复为 FRRM。
- 对于采用典则逻辑连接函数的二项分布或多项广义线性模型，即（二项）逻辑回归，(8) 限制了 $\mathbf{S}\boldsymbol{\alpha}$ 对分类优势比的影响。
- 对于采用典则对数连接函数的泊松广义线性模型，即对数线性回归，(8) 限制了强度的差异（也就是单位时间内的预期到达次数）。
- 对于 Cox比例风险模型，我们可以将风险函数写作

  $$h(t;\widehat{\mathbf{U}},\mathbf{S}) = h_0(t)\exp(\mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta})$$

  其中 $h_0(t)$ 是时间 t 时的基线风险。相应的偏差可像 (7) 和 (8) 中那样使用，以实现期望的公平性水平，通过敏感属性影响的差异来限制风险比。

此外，F(G)RRM 具有完全模块化的特点，将模型估计与 $\widehat{}$ $\widehat{}$公平性约束分离开来。参数估计值 $\{\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \boldsymbol{\beta}_{\text{FRRM}}\}$（$\{\widehat{\boldsymbol{\alpha}}_{\text{FGRRM}}, \boldsymbol{\beta}_{\text{FGRRM}}\}$，分别）仅取决于约束 $R^2_{\mathbf{S}} \leqslant r$ 通过 $\lambda(r)$产生的影响。因此，任何能在 r 与 $\lambda(r)$之间建立单调关系的公平性约束，都可以直接代入 F(G)RRM，而无需对（5）作任何修改。来自 Berk等人(2017年）的个体公平约束满足这一条件；同样，也可以轻易修改（4）以构造满足该条件的机会均等约束（第4节）。这不同于文献中的其他方法：它们将公平性约束深度整合到估计过程中，从而无法灵活组合不同的模型族与公平性约束。

函数	模型	参考文献
`frrm()`, `fgrrm()`	F(G)RRM	Scutari 等（2022年）
`nclm()`	NCLM	Komiyama等 （2018年）
`zlm()`, `zlm.orig()`	ZLM	Zafar 等（2019年）
`zlrm()`, `zlrm.orig()`	ZLRM	Zafar 等（2019年）

表 1：fairml 中实现的公平模型及其规范参考文献。

NCLM、FRRM 和 FGRRM 约束 **S** 的总体影响，以使模型更加公平。作为替代方案，Zafar 等人(2019年）约束各个敏感属性的影响S_i：在线性模型（ZLM）中

$$\min_{\boldsymbol{\beta}} \mathsf{E}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^2\right] \text{ such that } |\mathsf{COV}(\mathbf{X}\boldsymbol{\beta}, S_i)| < c, c \in \mathbb{R}^+ \tag{9}$$

以及在逻辑回归模型（ZLRM）中

$$\max_{\boldsymbol{\beta}} D(\boldsymbol{\beta}) \text{ such that } |\mathsf{COV}(\mathbf{X}\boldsymbol{\beta}, S_i)| < c, c \in \mathbb{R}^+. \tag{10}$$

显然，限制 **S** 的总体效应 S_i，就限制了各个 S_i的效应；但前者并未强制所有 S_i都受到相同的界限约束；而且 c 没有界限，因此更难^进行选择。此外，Zafar 等人(2019年) 使用**X** 而非**U** 作为解释变量，导致受到约束、必须达到近乎完美公平的模型预测准确度灾难性下降（$c \to 0$），正如Scutari 等人(2022年) 所展示的。更一般地说，任何采用这种方法的模型都会出现这一问题，因为约束会同时移除**X** 中的非歧视性信息和歧视性信息：两者并未被区分，而是关联到相同的回归系数上。不过，ZLM 和 ZLRM 的优点是其模型方程不涉及**S**，这意味着我们无需为想要预测的新观测值收集敏感属性。因此，当存在这一问题时，它们是 FRRM 和 FGRRM 的有效替代方案。

4. 特性与软件架构

该**fairml** 包围绕实现上一节所讨论模型的函数构建：FRRM 由 `frrm()`实现，FGRRM 由 `fgrrm()`实现，NCLM 由 `nclm()`实现，ZLM 由 `zlm()` 实现，ZLRM 则由 `zlrm()`实现。所有这些函数均汇总于表1 ，且具有相似的签名：

```
frrm(response, predictors, sensitive, unfairness, definition = "sp-komiyama",
lambda = 0, save.auxiliary = FALSE)
fgrrm(response, predictors, sensitive, unfairness, definition = "sp-komiyama",
family = "binomial", lambda = 0, save.auxiliary = FALSE)
nclm(response, predictors, sensitive, unfairness, covfun, lambda = 0,
save.auxiliary = FALSE)zlm(response, predictors, sensitive, unfairness)
zlrm(response, predictors, sensitive, unfairness)
```

公平性	Type	参考资料
"sp-komiyama"	统计均等	Komiyama等 (2018年)
"eo-komiyama"	机会均等	Scutari 等人(2022年)
"if-berk"	个体公平	Berk 等人（2017年）

表 2：在 fairml 中作为约束实现的公平性定义。

在上述内容中， response 是响应变量（在第 y 节中记为 3）， predictors 是非敏感解释变量（$\mathbf{X}$）， sensitive 是敏感属性（$\mathbf{S}$），而unfairness 是模型允许的不公平程度（$r \in [0,1]$）。 predictors和 sensitive都会在内部转换为相应的设计矩阵。对于 zlm() 和 zlrm()，公式（9）和（10）中的协方差约束会替换为相应的相关性约束，以便将其重新缩放至 $[0,1]$。至于其他参数：

- save.auxiliary 控制模型 $\hat{\mathbf{U}} = \mathbf{X} - \mathbf{B}_{\mathrm{T}}\mathbf{S}$（用于生成（2）中的去相关预测变量（并隐含地用于（6）和（7））的模型）是否与主模型一同返回。
- definition 是一个标签，用于指定所使用的公平性定义（表2）：
  - "sp-komiyama" 是（4）中的公平性约束，该约束也用于 NCLM；– "eo-komiyama" 表示拟合值的方差或偏差中可由敏感属性解释、但无法由原始响应变量解释的比例；– "if-berk" 是（1）中的约束，其实现和归一化形式如下

$$f(\boldsymbol{\alpha}, \mathbf{y}, \mathbf{S}) = \frac{\sum_{i,j} |y_i - y_j| (\mathbf{s}_i \boldsymbol{\alpha} - \mathbf{s}_j \boldsymbol{\alpha})^2}{\sum_{i,j} |y_i - y_j| (\mathbf{s}_i \hat{\boldsymbol{\alpha}}_{\mathrm{OLS}} - \mathbf{s}_j \hat{\boldsymbol{\alpha}}_{\mathrm{OLS}})^2} \in [0,1],$$

  其中， $\hat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$ 是在未施加任何公平性约束的情况下估计的模型中与 $\mathbf{S}$ 相关的系数（即 $r = 1$）。

  作为替代，用户可以提供一个函数来实现自定义公平性定义，如第 5.3 节所示。

- lambda 是一个可选的岭惩罚项，用于对 β（在 FRRM 和 FGRRM 中）或对（α， β）（在 NCLM 中）施加惩罚，以对所得模型进行正则化。
- covfun 是 NCLM 内部使用的协方差函数。其默认值为 cov()；其他选项包括 cov.shrink() 中的 James-Stein 估计量，该估计量来自 **corpcor**（Schafer、Opgen-Rhein、Zuber 等人于 2021年提出的估计量，或类似于 Komiyama等人(2018年) 使用的核估计量。

此外，fairml 提供了两个函数 zlm.orig()和 zlrm.orig()，使用协方差约束（9）和（10）的原始形式实现 ZLM 和 ZLRM。 frrm()和 fgrrm() 基于 glmnet 包构建（Friedman 等，2010；Simon 等，2011；Tay、Narasimhan 和 Hastie，2023）；zlm() zlrm()、 zlm.orig()和 zlrm.orig() 基于 CVXR 包构建（Fu、Narasimhan 和 Boyd，2020）； nclm() 基于 cccp 包提供的求解器构建（Pfaff，2022）。

所有这些函数都会返回一个 "fair.model" 类对象，并附加一个用于标识模型估计器的类（"frrm"、"fgrrm"、 "nclm"、 "zlm"或 "zlrm"）。这些类用于将模型分派给 stats 包为内置模型提供的方法：

- print() 和 summary()，用于打印模型的关键信息；
- coef()、 fitted()、 residuals()和 sigma()，用于从模型中提取相关参数；
- deviance()、 logLik() 和 nobs()，用于评估模型的拟合优度，并使AIC()和 BIC() 能够正常运行；
- plot() 用于显示模型验证的诊断图；
- all.equal() 用于比较两个模型；
- predict() 用于预测新观测值的响应变量值。

此外，fairml 提供了两种用于模型选择的工具。第一种是实现交叉验证、用于评估模型预测准确度的函数，

fairml.cv(response, predictors, sensitive, method = "k-fold", ..., unfairness, model, model.args = list(), cluster)

其中：

- response、 predictors、 sensitive 和 unfairness 与模型估计函数中的相同；
- method 是将用于重采样的交叉验证方案（"k-fold"、 "hold-out" 或 "custom-folds"），而 ... 则包含其可选参数；
- model 和 model.args 接收模型名称（字符串）以及一个可选参数列表，并将其传递给模型；
- cluster 是一个可选对象，通过 parallel 软件包中的 makeCluster() 函数创建，用于启用并行处理。

第二个是一个函数，用于生成剖面图，以评估估计模型如何随公平性约束变化，

fairness.profile.plot(response, predictors, sensitive, unfairness,legend = FALSE、type = "coefficients"、 model、 model.args = list()、 cluster)

其中， type 决定绘制的内容（"coefficients"、 "constraints"、 "precision-recall"或 "rmse"），legend 则控制是否显示带有变量名称的图例。其余参数的含义与 fairml.cv() 中相同。

最后，**fairml** 附带了一套全面的真实世界数据集，这些数据集曾用于相关文献中（表3）：其中许多来自 UCI机器学习知识库（Dua 和 Graff，2023）。提供这套数据集是为了更便于探索和基准测试公平

Name	分布族	响应变量	敏感属性 属性
成人 (adult)	二项分布	收入	性别、种族、年龄
Bank (bank)	二项分布	已订阅	年龄，婚姻状况
社区与犯罪 ঘটনা (communities.and.crime)	高斯	犯罪率	% 黑色 ， % 外国出生人口
COMPAS (compas)	二项分布	再犯	性别、种族、年龄
药物消费 (drug.consumption)	多项式	消费	年龄、性别、种族
游离轻链 （flchain，生存分析中）	Cox比例风险模型	距死亡天数	年龄、性别
健康与退休 (health.and.retirement)	泊松	得分	婚姻， 种族、年龄 性别，
法学院招生 (law.school.admissions)	高斯	GPA	年龄、种族
全国纵向调查 (national.longitudinal.survey)	高斯	收入	性别、年龄
肥胖程度 (obesity.levels)	多项式	肥胖程度	性别、年龄

表3：文献中用于评估公平机器学习模型的真实世界数据集，这些数据集已纳入 fairml；flchain 实际上由 survey 程序包提供，但 fairml 中记录了其在公平建模中的用途。对于每个数据集，我们报告其名称（包括 R 对象名称）、公平模型的广义线性模型族、响应变量和敏感属性。

机器学习模型，包括 fairml 中的模型，同时也为未来文献中的模型比较提供依据。所有数据集都经过了最低限度的预处理和清理，以保留其原始特征，具体记录于相应的手册页中。

5. 公平建模示例

在本节最后，我们将使用 drug.consumption 和 national.longitudinal.survey数据集（见表3），演示如何使用 **fairml** 来选择和估计公平模型。

5.1. 药物消费

该药物消费数据集最初由 Fehrman、Muhammad、Mirkes等人(2017年) 研究，来源于一项在线调查，旨在根据人格特质、冲动性（Impulsivity）、感觉寻求（SS）以及人口统计信息，评估个人的药物消费风险。人格特质采用“五因素模型”进行建模，该模型包含神经质（Nscore）、外向性（Escore）、经验开放性（Oscore）、宜人性（Ascore）和尽责性（Cscore）的得分。至于人口统计信息，该调查记录了每位受访者的年龄（Age）、性别（Gender）、种族（Race）和教育水平（Education）。该数据集包含有关消费情况的信息

18 种精神活性药物，包括安非他命、可卡因、快克、摇头丸、海洛因、氯胺酮等。在本例中，我们将重点关注 LSD 的使用。
受访者是一个自我选择的样本：他们在招募期间自行报名。因此，自然会产生这样一个问题：样本是否存在偏差，因为个体由于年龄、性别或种族而未能报名，或报名后拒绝回答某些问题？这种形式的抽样偏差也在调查抽样领域得到过研究，关于如何使用抽样权重消除这种偏差，文献中已有深入探讨（例如，Lohr 2021）。

加载数据后，

```
R> data(drug.consumption)
```

并提取响应变量（r）、敏感属性（s）和预测变量（p），

```
R> r = drug.consumption[, "LSD"]
R> s = drug.consumption[, c("Age", "Gender", "Race")]
R> p = drug.consumption[, c("Education", "Nscore", "Escore", "Oscore",
+                           "Ascore", "Cscore", "Impulsive", "SS")]
```

w我们合并在 Education 和响应变量中样本数量较少的类别。 .

```
R> levels(p$Education) =
+    c("at.most.18y", "at.most.18y", "at.most.18y", "at.most.18y",
+      "university", "diploma", "bachelor", "master", "phd")
R> levels(r) = c("never", ">=1y", ">=1y", "<1y", "<1m", "<1m", "<1m")
```

因此，这两个变量的所有水平在1885个样本中至少有89个样本被观测到，其中大多数水平的观测样本数超过200个。
然后，我们可以根据 r、 s 和 p 拟合一个多项式 FGRRM 模型，使不公平性（$r = 0.05$）达到最低，并额外施加较小的岭惩罚以进行平滑处理。

```
R> m = fgrrm(response = r, sensitive = s, predictors = p,
+        family = "multinomial", unfairness = 0.05, lambda = 0.1)
R> summary(m)


Fair Linear Regression Model

Method: Fair Generalized Ridge Regression

Call:fgrrm(response = r, predictors = p, sensitive = s, unfairness = 0.05,
family = "multinomial", lambda = 0.1)


Coefficients:
                      never       >=1y        <1y         <1m
(Intercept)       1.1079621  0.1549633 -0.4910178 -0.7719076
```

```
Age25.34                0.0044356   0.0002583  -0.0019688  -0.0027251
Age35.44                0.0031367   0.0170390  -0.0104855  -0.0096902
Age45.54                0.0052437   0.0149946  -0.0117205  -0.0085178
Age55.64                0.0026464   0.0192400  -0.0123804  -0.0095060
Age65.                  0.0326250  -0.0116477  -0.0116190  -0.0093583
 [ reached getOption("max.print") -- omitted 19 rows ]

Ridge penalty (sensitive attributes): 9.601 (predictors): 0.1
Log-likelihood: -1944
Komiyama's R^2 (statistical parity): 0.05 with bound: 0.05
```

模型摘要包含多项逻辑回归为响应变量的每个水平拟合的系数组；[2] 还包含对敏感属性和预测变量施加的整体岭惩罚；并确认已达到所需的（不）公平性水平。

这个模型对数据的拟合效果好吗？由以下工具生成的诊断图

```
R> plot(m, support = TRUE)
```

并在图中展示1 表明，模型受到从未使用过 LSD 的人数与使用过 LSD 的人数不均衡的影响（左上面板）。针对响应变量的每个水平分别计算的一对其余 ROC 曲线都相似，表明模型对数据的拟合效果尚可（右下面板）。按照Hastie、Tibshirani 和 Friedman (2009年)的建议，进行标准的 10 次 10 折交叉验证后，我们确认精确率和召回率都不算高。

```
R> fairml.cv(response = r, sensitive = s, predictors = p, model = "fgrrm",
+    unfairness = 0.05, method = "k-fold", k = 10, runs = 10,
+    model.args = list(family = "multinomial", lambda = 0.1))


  k-fold cross-validation for fair models

  model:
                                      Fair Generalized Ridge Regression
  number of folds:                       10
  number of runs:                        10
  average loss over the runs:
    precision:                           0.6443056
    recall:                              0.2144495
  standard deviation of the loss:
    precision:                           0.009535637
    recall:                              0.006509704
```

[2] nnet 中的 `multinom()` 函数会返回响应变量除第一个水平外其他所有水平的系数，并减去第一个水平对应的系数。glmnet 返回所有水平的原始系数，fairml 也遵循相同的约定。

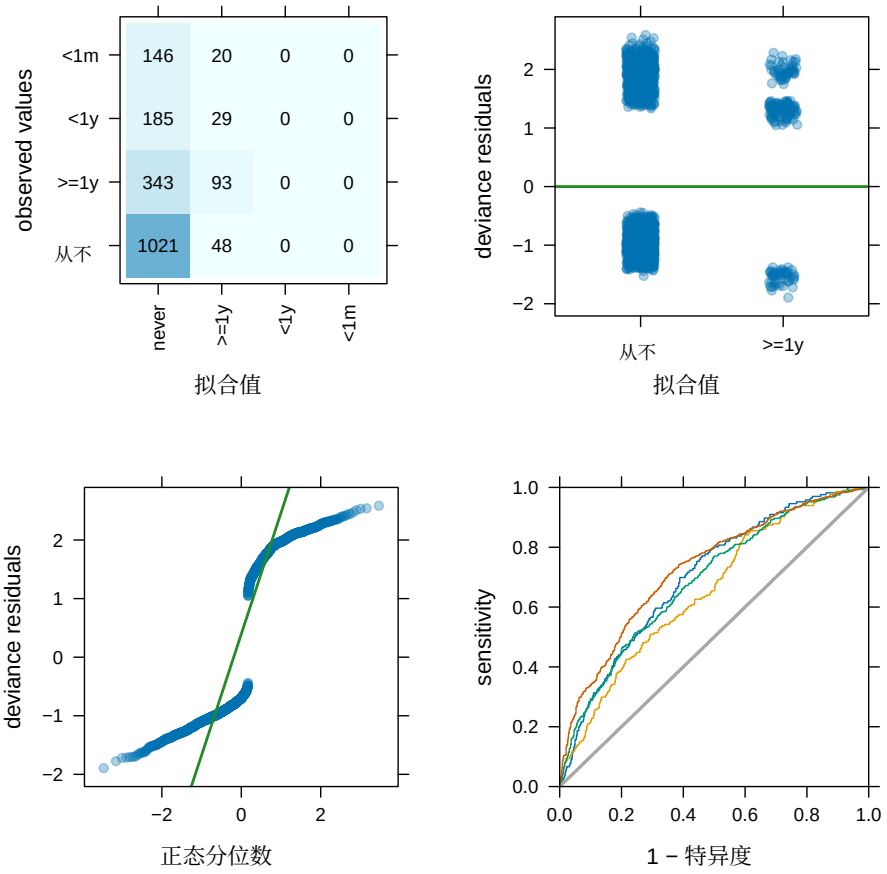


图1： plot(m)生成的诊断图：混淆矩阵（左上）、偏差残差与拟合值的关系图（右上）、偏差残差的QQ图（左下），以及响应变量四个水平的ROC曲线。

这让我们面临两个问题：

1. 公平性约束对模型的拟合优度和预测性能有显著影响吗？2. 将响应变量的各个水平合并为二元变量“曾使用”/“从未使用”，使数据更加均衡，是否会改善模型?

我们通过在图中绘制估计回归系数随公平性约束变化的曲线，来探讨第一个问题。2。

```
R> fairness.profile.plot(response = r, sensitive = s, predictors = p,
+       model = "fgrrm", type = "coefficients",
+       model.args = list(family = "multinomial", lambda = 0.1))
```

当 unfairness较小时，所有回归系数都明显向零收缩：当模型在 unfairness处不受限制时，它们的取值（绝对值）会大得多。

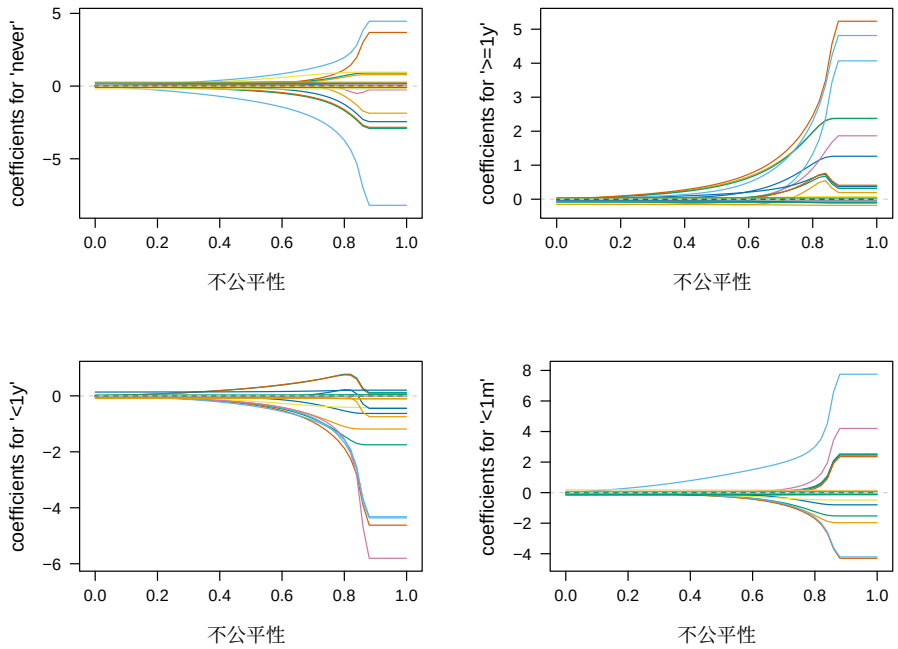


图2：对于响应变量的所有水平，回归系数随公平性约束变化的剖面图。

等于 1。这表明，除了对响应变量的直接影响外， Age、 Gender 和 Race 还与预测变量密切相关。当unfairness 等于 0.85 左右时，它们的曲线趋于平缓，这意味着当公平性约束变为非活动状态时，敏感属性可以解释模型约 85% 的偏差。至于通过将响应变量转换为二元变量来重新平衡响应变量，

```
R> levels(r) = c("never", "used", "used", "used")
R> table(r)
```

```
r
never  used
1069 816
```

通过交叉验证测试（不再是多项式的）FGRRM逻辑回归模型的预测性能，结果显示召回率显著提高，而精确率基本相同。

```
R> fairml.cv(response = r, sensitive = s, predictors = p, model = "fgrrm",
+     unfairness = 0.05, method = "k-fold", k = 10, runs = 10,
+     model.args = list(family = "binomial", lambda = 0.1))
```

```
  k-fold cross-validation for fair models
```

```
model:
                                    Fair Generalized Ridge Regression
number of folds:                      10
number of runs:                       10
average loss over the runs:
  precision:                          0.6866465
  recall:                             0.5515931
standard deviation of the loss:
  precision:                          0.002800798
  recall:                             0.007156275
```

因此，我们可以得出结论：原始响应变量的不平衡对多项式 FGRRM 模型的预测准确度产生了明确的影响。

5.2. 收入与劳动力市场

全国纵向调查数据集包含美国劳工统计局为收集劳动力市场活动信息而开展的一项调查结果（美国劳工统计局 2023）。除了 1990 年、1996 年和 2006 年的收入（income90、income96和 income06）外，该调查还记录了受访者（我们的敏感属性）的性别（gender）、年龄（age）和种族（race）、他们的身体特征（身高、体重、总体健康状况）、犯罪记录（违法行为和指控的数量）以及受教育程度（grade90）。与第5.1节的情况一样，我们希望通过抽样偏差及其他机制消除敏感属性引入的偏见。

加载数据、合并观测数量较少的 race 和 grade90 的类别，并分离响应变量（本例中为 income90 ）、预测变量和敏感属性后，

```
R> data(national.longitudinal.survey)
R> nlsy = national.longitudinal.survey
R> levels(nlsy$grade90)[1:5] = "7TH GRADE"
R> levels(nlsy$race)[c(3, 14, 15)] = "CHINESE"
R> r = nlsy[, "income90"]
R> s = nlsy[, c("gender", "age", "race")]
R> p = nlsy[, setdiff(names(nlsy),
+               c("income90", "income96", "income06", "gender", "race", "age"))]
```

我们可以像文献中通常的做法那样拟合一个 FRRM 模型（或等价地，拟合一个高斯 FGRRM）（例如，参见Komiyama等人，2018年）。

```
R> m = frrm(response = r, sensitive = s, predictors = p, unfairness = 0.05)
R> summary(m)


Fair Linear Regression Model

Method: Fair Ridge Regression
```

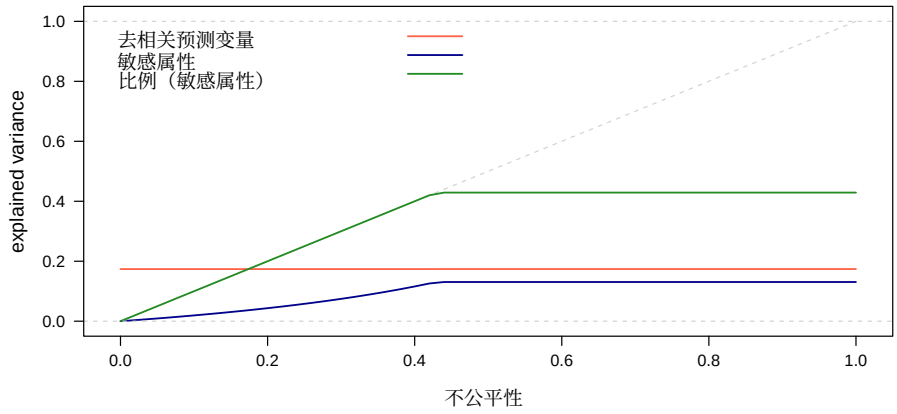


图 3：去相关预测变量和敏感属性所解释的方差比例剖面图。

```
Call:
frrm(response = r, predictors = p, sensitive = s, unfairness = 0.05)

Coefficients:
 (Intercept)  genderFemale           age     raceBLACK   raceCHINESE
   1.6102737    -0.1873255     0.0186816    -0.0980725     0.0765317
 [ reached getOption("max.print") -- omitted 62 entries ]

Ridge penalty (sensitive attributes): 4.081 (predictors): 0
Log-likelihood: -7835
Residual standard error: 1.202
Multiple R^2: 0.2337
Komiyama's R^2 (statistical parity): 0.05 with bound: 0.05
```

由 R^2 报告的系数 `summary()` 表明，该模型对数据的拟合效果不佳。这是公平性约束造成的吗？将敏感属性和其他预测变量所解释的方差比例与 `fairness.profile.plot()`进行比较后，我们可以从图3 中看出，在没有公平性约束的情况下（该约束在$r \approx 0.45$时变为非活动状态），敏感属性解释的变异几乎与预测变量一样多。

```
R> fairness.profile.plot(response = r, sensitive = s, predictors = p,
+      model = "frrm",  type = "constraints",  legend = TRUE)
```

虽然移除或放宽公平性约束几乎会使模型的 R^2 翻倍，但由 `plot()` （图 4）生成的诊断图突显了数据中的两个潜在问题：较高的收入被截断至 7.428，这会影响 QQ图右尾的残差，并在前两幅图中形成点状模式；此外，响应变量的下界为零。

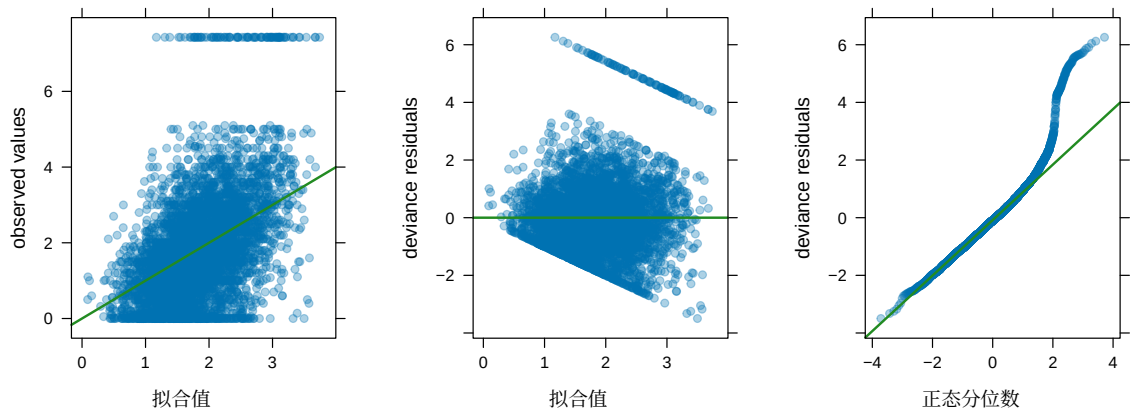


图 4： `plot()` 为具有连续响应变量的 FGRRM 模型生成的诊断图：观测值与拟合值（左）、偏差残差与拟合值（中）以及偏差残差的 QQ图（右）。

为解决这些问题，我们可以考虑移除被截断的观测值，并基于数据拟合泊松 FGRRM 模型，这对于非负响应变量而言是一个自然的选择。

```
R> large.income = (r == max(r))
R> table(large.income)

large.income
FALSE  TRUE
4816 92

R> m = fgrrm(response = r[!large.income], sensitive = s[!large.income, ],
+        predictors = p[!large.income, ], family = "poisson",
+        unfairness = 0.05)
```

然而，使用相同数据（并使用相同的折叠）通过交叉验证比较这两个模型后发现，与泊松 FGRRM 相比，原始 FRRM 模型在所选公平性水平下似乎具有更高的预测准确度。

```
R> xval.lm =
+    fairml.cv(response = r[!large.income], sensitive = s[!large.income, ],
+      predictors = p[!large.income, ], model = "frrm", unfairness = 0.05,
+      method = "k-fold", k = 10, runs = 10)R> xval.glm =
+    fairml.cv(response = r[!large.income], sensitive = s[!large.income, ],
+      predictors = p[!large.income, ], model = "fgrrm", unfairness = 0.05,
+      model.args = list(family = "poisson"), method = "custom-folds",
+      folds = cv.folds(xval.lm))R> summary(cv.loss(xval.lm))
```

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.9908 0.9929 0.9937 0.9938 0.9950 0.9967
```

```
R> summary(cv.loss(xval.glm))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.128 1.130 1.131 1.131 1.133 1.134
```

fairml 中实现的另外两个模型**fairml**，`zlm()` 来自 Zafar 等人 (2019年) 和 `nclm()` 来自 Komiyama 等人 (2018年)，似乎也不如 FRRM 拟合数据。

```
R> m2 = nclm(response = r[!large.income], sensitive = s[!large.income, ],
+         predictors = p[!large.income, ], unfairness = 0.05)
```

```
Loading required namespace: cccp
```

```
R> m3 = zlm(response = r[!large.income], sensitive = s[!large.income, ],
+         predictors = p[!large.income, ], unfairness = 0.05)
```

```
Loading required namespace: CVXR
```

```
R> summary(m2)
```

```
Fair Linear Regression Model

Method: Komiyama et al. (2018)

Call:nclm(response = r[!large.income], predictors = p[!large.income,],
 sensitive = s[!large.income, ], unfairness = 0.05)


Coefficients:
 (Intercept)  genderFemale           age     raceBLACK   raceCHINESE
   1.4635456    -0.1974729     0.0162676     0.0076029     0.2002531
 [ reached getOption("max.print") -- omitted 62 entries ]

Ridge penalty: 0
Custom covariance matrix: FALSE
Log-likelihood: -6757
Residual standard error: 0.9912
Multiple R^2: 0.248
Komiyama's R^2 (statistical parity): 0.04984 with bound: 0.05
```

```
R> summary(m3)
```

```
Fair Linear Regression Model

Method: Zafar's Linear Regression

Call:zlm(response = r[!large.income], predictors = p[!large.income,],
 sensitive = s[!large.income, ], unfairness = 0.05)


Coefficients:
(Intercept)    grade90.L    grade90.Q    grade90.C    grade90.4
  1.4581758    1.3300355    0.0028437   -0.2713558    0.0125724
 [ reached getOption("max.print") -- omitted 34 entries ]

Log-likelihood: -6940Residual standard error: 1.027
Multiple R^2: 0.1887Marginal correlation (disparate impact):
genderFemale          age    raceBLACK  raceCHINESE  raceENGLISH
0.032564 0.007244 0.036622 0.013714 0.041502
[ reached getOption("max.print") -- omitted 23 entries ]
with bound: 0.05
```

因此，我们可以得出结论：在我们考察的模型中，FRRM 对 `unfairness = 0.05` 数据的拟合效果最好。

5.3. 使用不同的公平性约束

我们将展示的 **fairml** 最后一项特性，是能够将自定义的公平性数学表征接入 `frrm()`和 `fgrrm()`。内置的公平性定义以表 2中列出的标签标识。自定义定义可通过具有以下签名的函数提供`function(model, y, S, U, family)`其中 `model` 是我们正在评估其公平性的模型，`family`是该模型所属的广义线性模型族，而 `y`、 `S`、 `U`分别是 **y**、**S**、**U**，取自（3）。借助 `str()` 和一个虚拟函数，我们可以看到， `model` 对象包含若干描述待评估模型的关键量。

```
R> dummy.fairness =
+    function(model, y, S, U, family) { str(model); return(c(value = 0)) }R>
R> m = fgrrm(response = r, sensitive = s, predictors = p, unfairness = 0.05,
+        definition = dummy.fairness,  family = "gaussian")


List of 5
 $ coefficients: Named num [1:67] 0.6526 -0.7423 0.0754 -0.0635 0.4791 ...
  ..- attr(*, "names")= chr [1:67] "(Intercept)" "genderFemale" "age" ...
 $ deviance    : num 6350
 $ loglik      :Class 'logLik' : -7596 (df=68)
```

```
 $ fitted      : num [1:4908] 1.86 2.92 1.63 2.32 3.5 ...
 $ residuals   : num [1:4908] 0.1419 -0.6214 0.0683 0.6846 0.5017 ...
```

此函数的唯一要求是返回一个包含名为"value"的元素的数组，该元素的取值范围为 0（完全公平）到 1（没有约束）。内置函数会返回其他元素，这些元素随后会用于`fairness.profile.plot()`中，但对于自定义公平性定义，除名为 "value" 的元素外，其他元素都会被忽略。因此，例如，我们可以像Berk等人(2017年所做的那样，将统计均等和个体公平加权求平均来定义公平性，方法是重复使用**fairml** 用于估算它们的内部函数。

```
R> combined.fairness = function(model, y, S, U, family) {
+
+    value =
+      0.4 * fairml:::fgrrm.sp.komiyama(model, y, S, U, family)[["value"]] +
+      0.6 * fairml:::fgrrm.if.berk(model, y, S, U, family)[["value"]]
+
+    return(c(value = value))
+
+  }#COMBINED.FAIRNESS
```

或者，我们也可以采用一种完全不同的公平性定义，例如，按照 Zafar等人（2019年） 的思路，对拟合值与敏感属性之间的相关性设定一个界。

```
R> custom.fairness =
+    function(model, y, S, U, family)
+      return(c(value = max(abs(cor(model$fitted, S)))))
```

或者，假设数据中只有一个二元敏感属性，我们可以对由敏感属性识别的两个群体的拟合值分布之间的 Kolmogorov-Smirnov 检验的 p 值进行界定。

```
R> custom.fairness = function(model, y, S, U, family) {
+
+    group1 = model$fitted[S == 0]
+    group2 = model$fitted[S == 1]
+    value = ks.test(group1, group2)$p.value
+
+    return(c(value = value))
+
+  }#CUSTOM.FAIRNESS
```

本质上，自定义函数参数中提供的信息结合其他 R 软件包的功能，能够带来极大的灵活性。

6. 总结与讨论

算法公平性是一个具有重要实际应用价值且研究领域迅速发展的课题。**fairml** 软件包的核心是我们此前在 Scutari 等人 (2022年) 的工作，该工作将公平线性模型的估计与公平性的数学刻画分离开来。一方面，这种模块化设计选择实现了紧凑的实现，并支持所有广义线性模型族，因此 **fairml** 可用于各种应用领域。另一方面，无需重新实现模型估计，即可将任意公平性定义嵌入模型的能力，是研究中的一项宝贵优势。此外，**fairml**生成的模型具有广为人知的统计性质和最佳实践（与构成大部分文献的、基于数值优化器的黑箱估计器不同），并提供用于验证和评估这些模型的标准工具。

参考文献

Agarwal A、Beygelzimer A、Dudik M 等（2018年）。《一种实现公平分类的归约方法》。《机器学习研究论文集》，80，60–69。第35届国际机器学习会议（ICML）。Agarwal A、Dudik M、Wu ZS（2019年）。《公平回归：定量定义与基于归约的算法》。《机器学习研究论文集》，97，120–129。第36届国际机器学习会议。Angwin J、Larson J、Mattu S、Kirchner L（2016年）。《机器偏见：全国各地都在使用预测未来罪犯的软件，而它对黑人存在偏见》。ProPublica。网址 https://www.propublica.org/article/machine-bias-riskassessments-in-criminal-sentencing。Barredo Arrieta A、Díaz-Rodríguez N、Del Ser J 等（2020年）。《可解释人工智能（XAI）：概念、分类体系、机遇与迈向负责任人工智能的挑战》。《信息融合》，58，82–115。Bellamy RKE、Dey K、Hind M 等（2019年）。《AI Fairness 360：用于检测和缓解算法偏见的可扩展工具包》。《IBM研究与开发杂志》，63(4/5)，1–15。 doi:10.1147/JRD.2019.2942287。Berk R、Heidari H、Jabbari S 等（2017年）。《公平回归的凸框架》。载于《机器学习中的公平性、问责制与透明度》（FATML）。Berk R、Heidari H、Jabbari S 等（2021）。《刑事司法风险评估中的公平性：现状》。《社会学方法与研究》，50(1)，3–44。doi:10.1177/0049124118782533。Calders T、Karim A、Kamiran F 等（2013年）。《控制线性回归中的属性效应》。载于《第13届IEEE国际数据挖掘会议论文集》，第71–80页。 doi:10.1109/ICDM.2013.114。Calmon F、Wei D、Vinzamuri B 等（2017）。《用于歧视防止的优化预处理》。载于《神经信息处理系统进展》，第30卷，第3992–4001页。

Cath C、Wachter S、Mittelstadt B 等（2018）。《人工智能与“美好社会”：美国、欧盟和英国的做法》。《科学与工程伦理》，24(2)，505–528。 doi:10.1007/s11948-017-9901-7。Choraś M、Pawlicki M、Puchalski D、Kozik R（2020）。《机器学习——结果并非唯一重要的事情！安全性、可解释性和公平性又该如何考量？》载于《国际计算科学会议》（ICCS）论文集，第615–628页。doi:10.1007/978-3-030-50423-6_46。Chzhen E、Denis C、Hebiri M 等（2020）。《通过插件估计量和重新校准实现公平回归，并提供统计保证》。载于《神经信息处理系统进展》第33卷，第19137–19148页。Cotter A、Gupta M、Jiang H 等（2019）。《训练具有良好泛化能力的分类器，以满足公平性度量和其他数据依赖约束》。《机器学习研究论文集》，97，1397–1405。第36届国际机器学习会议。D’Alessandro B、O’Neil C、LaGatta T（2017）。《审慎分类：数据科学家关于歧视感知分类的指南》。《大数据》，5(2)，120–134。 doi:10.1089/big.2016.0048。de B G de Oliveira T、Vieira LP、Silva GRL 等（2021）。predfairness：机器学习模型的歧视缓解。网址 https://cran.r-project.org/package=predfairness。Del Barrio E、Gordaliza P、Loubes JM（2020）。机器学习公平性的数学框架综述。网址 https://arxiv.org/abs/2005.13755。Díaz M、Johnson I、Lazar A 等（2019）。《解决情感分析中的年龄相关偏见》。载于第二十八届国际人工智能联合会议论文集，第6146–6150页。 doi:10.24963/ijcai.2019/852。Do H、Putzel P、Martin AS 等（2022）。《带凸惩罚项的公平广义线性模型》。《机器学习研究论文集》，162，5286–5308。第39届国际机器学习会议。Dua D、Graff C（2023）。UCI机器学习知识库。加利福尼亚大学信息与计算机科学学院。网址 https://archive.ics.uci.edu/ml/。欧盟委员会（2021）。《人工智能协调规则法规提案》。网址 https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence。Fehrman E、Muhammad AK、Mirkes EM 等（2017）。《人格五因素模型与药物使用风险评估》。载于《数据科学》，第231–242页。 doi:10.1007/978-3-319-55723-6_18。Friedman J、Hastie T、Tibshirani R（2010）。《通过坐标下降求解广义线性模型的正则化路径》。《统计软件杂志》，33(1)，1–22。 doi:10.18637/jss.v033.i01。

Fu A, Narasimhan B, Boyd S（2020）。“CVXR：用于规范凸优化的 R 包。”《统计软件杂志》，94(14)，1–34。 doi:10.18637/jss.v094.i14。Fukuchi K, Sakuma J, Kamishima T（2013）。“基于模型中立性的预测。”载于《欧洲机器学习与数据库知识发现联合会议》（ECML PKDD）论文集，第499–514页。施普林格。 doi:10.1587/transinf.2014EDP7367。Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A（2022）。“可预见的不平等？机器学习对信贷市场的影响。”《金融学杂志》，77(1)，5–47。doi:10.1111/jofi.13090。Hardt M, Priceric E, Srebro N（2016）。“监督学习中的机会均等。”载于《神经信息处理系统进展》第29卷，第3315–3323页。Hastie T, Tibshirani R, Friedman J（2009）。《统计学习要素：数据挖掘、推断与预测》。第2版。施普林格。 doi:10.1007/978-0-387-84858-7。Hort M, Sarro F（2022）。“特权组与非特权组：年龄属性对公平性影响的实证研究。”载于《国际公平数据与技术研讨会》论文集，第17–24页。 doi:10.1145/3524491.3527308。Komiyama J, Takeda A, Honda J, Shimao H（2018）。“带公平性约束的回归非凸优化。”《机器学习研究论文集》，80，2737–2746。第35届国际机器学习会议。Kozodoi N, Varga TV（2021）。fairness：算法公平性指标。网址 https://cran.r-project.org/package=fairness。Lambrecht A, Tucker C（2019）。“算法偏见？对 STEM 职业广告展示中表面性别歧视的实证研究。”《管理科学》，65(7)，2966–2981。 doi:10.1287/mnsc.2018.3093。Lang M, Binder M, Richter J 等（2019）。“mlr3：R 中现代的面向对象机器学习框架。”《开源软件期刊》，4(44)，1903。 doi:10.21105/joss.01903。Lohr SL（2021）。《抽样：设计与分析》。第3版。CRC 出版社。 doi:10.1201/9780429298899。Mehrabi N, Morstatter F, Saxena N 等（2021）。“机器学习中的偏见与公平性综述。”ACM计算调查，54(6)，115。 doi:10.1145/3457607。Narasimhan H（2018）。“复杂损失函数与约束下的学习。”《机器学习研究论文集》，84，1646–1654。第21届国际人工智能与统计学会议。Pérez-Suay A, Laparra V, Mateo-García G 等（2017）。“公平核学习。”载于《欧洲机器学习与数据库知识发现联合会议》（ECML PKDD）论文集，第339–355页。施普林格。 doi:10.1007/978-3-319-71249-9_21。

Pessach D, Shmueli E（2022）。“机器学习公平性综述。”ACM计算调查，55(3)，51。doi:10.1145/3494672。Pfaff B（2022）。cccp：锥约束凸优化问题。网址 https://cran.r-project.org/package=cccp。Pfisterer F, Siyi W, Lang M（2022）。mlr3fairness：mlr3 的公平性审计与去偏。https://mlr3fairness.mlr-org.com, https://github.com/mlr-org/mlr3fairness。R 核心团队（2022）。R：一种用于统计计算的语言与环境。奥地利维也纳：R 统计计算基金会。网址 https://www.R-project.org/。Raghavan M, Barocas S, Kleinberg J, Levy K（2020）。“缓解算法招聘中的偏见：评估主张与实践。”载于第三届公平性、问责制与透明度会议论文集，第469–481页。 doi:10.1145/3351095.3372828。Schafer J, Opgen-Rhein R, Zuber V 等（2021）。corpcor：协方差与（部分）相关性的高效估计。网址 https://cran.r-project.org/package=corpcor。Scutari M（2022）。fairml：机器学习中的公平模型。网址 https://cran.r-project.org/package=fairml。Scutari M, Panero F, Proissl M（2022）。“通过简单的岭惩罚实现公平性。”《统计学与计算》，32，77。 doi:10.1007/s11222-022-10143-w。Simon N, Friedman J, Hastie T, Tibshirani R（2011）。“通过坐标下降法求解 Cox 比例风险模型的正则化路径。”《统计软件杂志》，39(5)，1–13。 doi:10.18637/jss.v039.i05。Tay JK, Narasimhan B, Hastie T（2023）。“适用于所有广义线性模型的弹性网络正则化路径。”《统计软件杂志》，106(1)，1–31。doi:10.18637/jss.v106.i01。Tolan S, Miron M, Gómez E, Castillo C（2019）。“机器学习为何可能导致不公平。”载于第十七届国际人工智能与法律会议论文集，第83–92页。doi:10.1145/3322640.3326705。联合国（2015）。可持续发展目标：可持续发展议程。网址 https://www.un.org/sustainabledevelopment/development-agenda/。联合国（2023）。AI for Good。网址 https://aiforgood.itu.int/。美国劳工统计局（2023）。全国青年纵向调查数据集。网址https://www.bls.gov/nls/。Wiśniewski J, Biecek P（2022）。“fairmodels：用于二元分类模型偏差检测、可视化与缓解的灵活工具。”《R期刊》，14，227–243。 doi:10.32614/RJ-2022-019。Woodworth B, Gunasekar S, Ohannessian MI, Srebro N（2017）。“学习非歧视性预测变量。”《机器学习研究论文集》，65，1920–1953。学习理论会议。

Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP（2019年）。“公平性约束：一种实现公平分类的灵活方法。”《机器学习研究杂志》，**20**，1–42。

所属机构：
Marco Scutari Istituto Dalle Molle di Studi sull’
Intelligenza Artificiale（IDSIA）大学中心 卢加诺东校区 圣塔
街1号 卢加诺6962，瑞士 电子邮箱： scutari@bnlearn.com网
址： https://www.bnlearn.com/