Python 实现倾向性评分匹配

孔 岩，尹 椽，张逸清

摘 要：倾向性评分匹配是一种常用于医学和公共卫生领域的统计方法，旨在控制混杂偏差，提高处理因素和观察结局之间关系分析的有效性。目前，该方法主要在 R 语言中得到应用，而在 Python 中可用的第三方库 psmatching 的匹配实现模式较为单一，应用场景受到较大的限制。因此，本课题的主要目标是复现 psmatching 库并参考 R 语言，增加更多的功能与方法。我们最终实现了倾向性评分匹配，并增加了多种加权方法。我们的代码在功能评估中表现良好，基本实现了研究课题的目标。

1 背景介绍

倾向性评分匹配是一种统计方法，用于进行干预效应分析。在医学和公共卫生领域的研究中，为了比较处理因素和结局之间的关系，通常需要设置对照组。然而，如果对照组和处理组之间存在混杂因素，这将影响观察结果。为了控制混杂偏差，倾向性评分匹配通过使处理组和对照组中的混杂因素分布趋于平衡来实现，从而更准确地分析处理因素与观察结局之间的关系。

倾向性评分匹配的大致流程包括计算得分、匹配和加权[1]。

2 测试数据集与方法

2.1 测试数据集

我们目前使用的测试数据集有十余个，覆盖临床医学、社会学、经济学领域。

临床医学的代表是来源于范德比尔特生物统计学数据集，其中包含右心导管插入术的数据。该数据集可在 https://hbiostat.org/data/ 上获取。该数据集一共有 63 列数据，其中第 9 列（是否死亡）和第 45 列（是否插入右心导管）是比较关键的数据。此外，数据集中还包含一些其他重要的协变量，如性别、年龄等。总体样本数量为 5735，样本容量足够大，适合用于进行分析。

其余的数据集如 COMPAS（Correctional Offender Management Profiling for Alternative Sanctions）数据集、adults 数据集（也称为 Census Income 数据集）、

美国国家卫生和营养调查（NHANES）数据集等，也都是公开可用的，并且包含了许多变量和指标，可以用于各种类型的倾向性评分匹配研究。

2.2 原理方法

1.得分计算方法：二分类 logistic 回归分析法

二分类 logistic 回归分析法是一种研究因变量为二项分类结果与某些影响因素之间关系的回归分析方法。公式如下：

$$\mathrm{logit}(P(Y=1|X)) = \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_k X_k$$

其中，因变量 y 为二分类变量，取值为 0 或 1。影响 y 取值的 k 个混杂因素分别为 $X_1, X_2, \cdots\cdots X_k$。在 k 个混杂因素作用下进入实验组的概率为 $P(Y=1|X)$[2]。

2.匹配

a.最邻近匹配

最邻近匹配是倾向性评分匹配最直接的匹配方法。其主要思路是首先将处理组的研究对象随机排序，后从处理组的第一个研究对象开始，为其在对照组中寻找一个倾向性评分值最邻近的个体作为匹配对象，直到所有处理组的对象均在对照组找到匹配对象为止。这种匹配方法所得到的两组结果在协变量的影响下差异仍较大。

b.半径匹配

半径匹配实现了对最邻近匹配的进一步精确化。通过设置卡尺，实现了对于处理组和对照组之间评分差异的控制，评分差异小于卡尺的对象才可以进行匹配，减少了对象之间受协变量影响的差异性。

c.一对多匹配和重复匹配

一对多的匹配是指处理组中的一个对象可以与多个对照组中的对象进行匹配，而重复匹配是指每个对象可以被重复匹配。这两种匹配方式提高了匹配的灵活性。

3.倾向性评分加权

倾向性评分加权根据目标人群的不同，通过不同的加权方法利用 PS 值对研究人群赋予权重，生成协变量均衡分布的虚拟人群。相较于匹配实现的个体间的 1:1 的均衡，倾向性评分加权通过加权实现群体之间全局性的均衡[3]。我们将采取以下几种加权方法。

(1)逆概率条件加权(IPTW)

逆概率加权（IPW）是一种用于解释由于非随机选择观测值或人群信息的非随机缺失而造成的缺失和选择偏差的方法。IPW 基于假设整：个研究人群都有可以预测纳入概率（非遗漏）的个体信息，因此，在考虑到这些信息后，我们可以仅从非遗漏的观察值开始对整个目标人群进行推断。IPTW 公式如下：

$$W_i = \frac{Z_i}{ps_i} + \frac{1 - Z_i}{1 - ps_i}$$

Wi 表示第 i 个受试者的均衡权重,本例中其形式为逆概率权重;Zi 表示第 i 个受试者的分组(Z=1 处理组;Z=0 对照组);psi 表示第 i 个受试者的倾向性分值 PS[4]。

(2)标准死亡比加权（SMRW）

标准死亡加权同样通过加权的方式，使一组中样本的属性与标准人群趋于一致，以下是两种以不同人群为标准人群的加权方式的公式：

a.以实验组作为“标准人群”

$$W_i = Z_i + \frac{(1 - Z_i)ps_i}{1 - ps_i}$$

b..以对照组作为“标准人群”

$$W_i = (1 - Z_i) + \frac{Z_i(1 - ps_i)}{ps_i}$$

两种加权方法分别在处理组和对照组中，对与对照组和处理组更为相似的样本加以更大的权重，以实现组间属性差异的减小[5]。

(3)重叠加权(OW)

OW 为每个样本分配与该其属于相反组的概率成比例的权重。公式如下：

$$W_i = Z_i(1 - ps_i) + (1 - Z_i)ps_i$$

IPTW 和 SMRW 的主要缺陷是若样本的倾向性评分接近 0 或 1，可能出现权重趋于无穷大的情况。OW 则能够克服这一缺陷，但其具有的局限性为：对于倾向性评分值接近 0 或 1 的样本，OW 加以的权重很小，导致数据的损失[6]。

4.评估 p 值

在进行倾向性评分匹配后，我们需要评估匹配后实验组和对照组之间的差异是否具有统计学意义。这时，我们可以使用 K^2 检验来比较两组之间的差异。首先建立实验组和对照组在各协变量上的值相同的零假设。接着，采用 K^2 检验对实验组和对照组数据进行分类并计算得到 p 值，设定显著性水平为 0.01，若 p 值小于等于 0.01，则拒绝零假设，认为实验组和对照组之间存在显著差异。

5. 组间差异衡量标准

采用标准均值误差（SMD）来衡量同一特征的组间差异，对于连续型协变量，SMD 的计算公式为：

$$d = \frac{\bar{x}_{treatment} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

其中，$\bar{x}_{treatment}$和$\bar{x}_{control}$代表协变量在实验组和对照组受试者中的样本均值，$s^2_{treatment}$和$s^2_{control}$代表协变量在实验组和对照组受试者中的样本方差。

对于二分类协变量，SMD 的计算公式为：

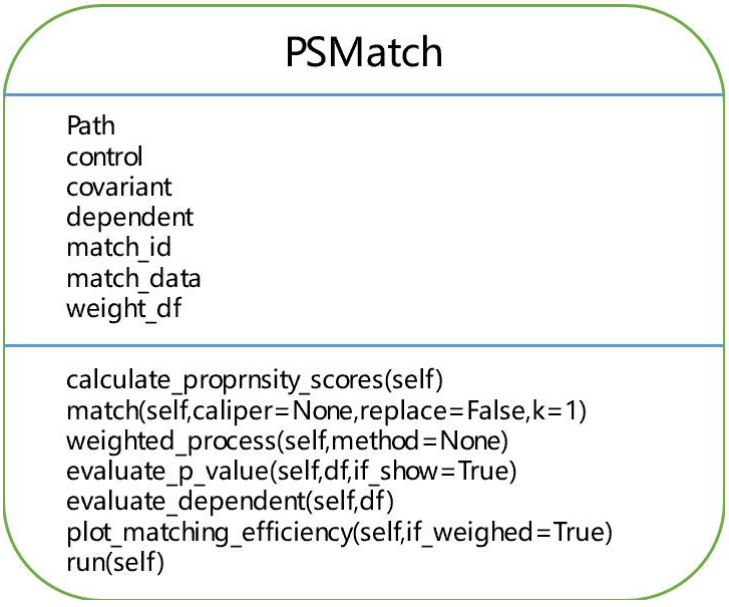
$$d = \frac{\hat{p}_{treatment} - \hat{p}_{control}}{\frac{\hat{p}_{treatment}(1-\hat{p}_{treatment}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}$$

其中$\hat{p}_{treatment}$和$\hat{p}_{control}$代表二分类变量在实验组和受试者中的使用率[7]。

3 结果汇报

3.1 程序设计

1. 类图



类图

2. 程序方法

详见 psmatch reference.pdf

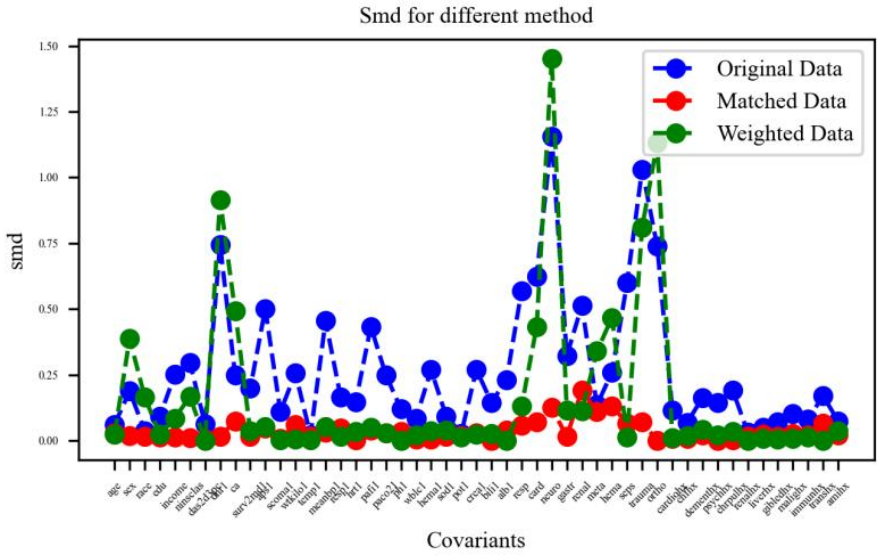
3.2 测试代码案例与运行结果

右心导管插入数据集在十四个测试案例中具有典型性，数据量始终，协变量中连续变量和非连续变量兼具，有非常适配的加权方法 IPTW-P,因此以其为例，进行展示。

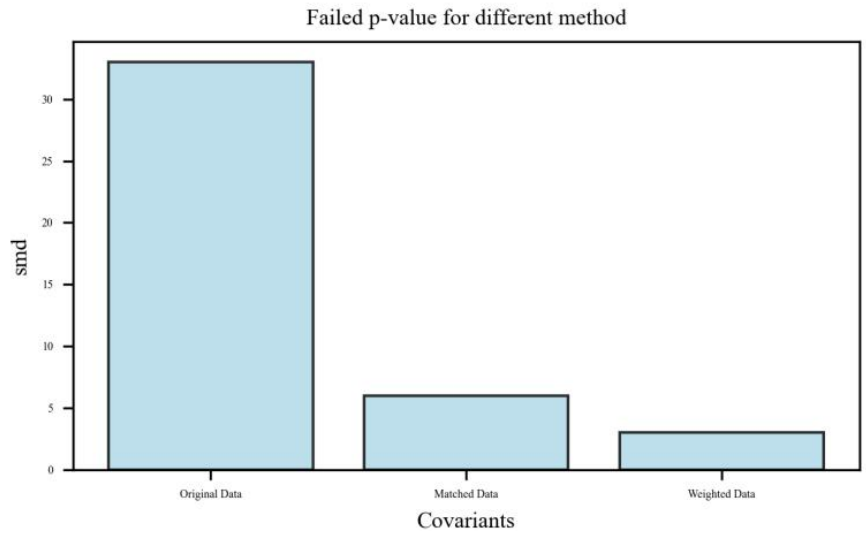
1. 测试案例代码

```python
import psmatching2 as psm

path = "data-rhc.csv"
control="swang1"
cov = ["age", "sex", "race", "edu", "income", "ninsclas", "das2d3pc", "dnr1", "ca", "surv2md1", "aps1", "scoma1",
       "wtkilo1", "temp1", "meanbp1", "resp1", "hrt1", "pafi1", "paco21", "ph1", "wblc1", "hema1", "sod1", "pot1",
       "crea1", "bili1", "alb1", "resp", "card", "neuro", "gastr", "renal", "meta", "hema", "seps", "trauma", "ortho",
       "cardiohx", "chfhx", "dementhx", "psychhx", "chrpulhx", "renalhx", "liverhx", "gibledhx", "malighx", "immunhx",
       "transhx", "amihx"]
dep=["death"]

m=psm.PSMatch(path,control,cov,dep)
m.df.replace({"RHC":1,"No RHC":0},inplace=True)
m.calculate_proprnsity_scores()
m.match(caliper=0.2)
m.weighted_process(method="IPTW-P")
print(m.evaluate_dependent(m.df))
print(m.evaluate_dependent(m.matched_data))
print(m.evaluate_dependent(m.weighted_df))
m.evaluate_p_value(m.df)
m.evaluate_p_value(m.matched_data)
m.evaluate_p_value(m.weighted_df)
m.plot_matching_efficiency()
```

2. 运行结果



各方法下各协变量的 smd



各方法下检验失败的协变量个数

详见 example result.pdf

3.分析与结论

比较原始数据和卡尺为 0.2 的匹配后数据，我们发现几乎各协变量的 smd 均下降，p 值均上升，检验失败的协变量个数下降。

比较原始数据和匹配加权后的数据，我们发现对结果影响力较强的协变量的 smd 进一步下降，p 值进一步上升，如这里的年龄，对结果影响力较弱的协变量的 smd 下降程度较弱甚至会上升，p 值上升程度较弱，设置会下降如这里的性别、种族等。但如果找到较合适的加权方法，仍能进一步降低检验失败的协变量个数。

总体来说，程序能显著减小各协变量的组间差异，实现了课题预期目标。

3.3 程序评估

1.程序功能效率

我们共测试了 14 组主流数据集，均能得到较理想结果。

与原始数据比较，匹配处理后，各协变量 smd 平均下降 88.51%，检验失败的协变量个数减少 74.09%；进一步加权处理后，各协变量 smd 平均下降 38.27%，检验失败的协变量个数减少 81.89%。

2.程序运行速度

由于本程序在进行数据处理时，调用 numpy、scipy、statsmodels 等库，程序运行速度较快。

3.程序的兼容性

本程序只调用了 numpy、pandas、scipy、matplotlib、statsmodels 库，对库的版本无特殊要求，支持 widows、linux 操作系统。程序的兼容性较好。

4 总结

在本次项目中，我们成功实现了倾向性评分匹配（Propensity Score Matching, PSM），并且添加了多种匹配和加权方法，充分实现了我们项目开始阶段所设定的目标。项目过程中，我们对各种匹配方法进行了研究，以适应不同场景下的需求。此外，为了验证我们实现的方法的有效性和适用性，我们利用多个不同的数据集检测了代码的运行效果。实验结果显示，我们的方法在不同数据集下均表现出较好的性能，得到了积极的反馈。为了让用户更好地使用和理解 PSMatch，我们还编写了一份用户使用手册。同时，我们还对程序进行了性能测试，通过优化代码结构和算法实现，提高了程序的运行速度和稳定性，为用户带来了更好的使用体验。

总之，在本项目中，我们通过实现倾向性评分匹配及其多种匹配和加权方法，有效地解决了处理观察数据中选择偏误问题。我们的方法在多个数据集上的表现以及性能测试均证明了我们的工作的有效性。

参考文献

[1] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

[2] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[3] Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical statistics, 10(2), 150-161.

[4] Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. American journal of epidemiology, 168(6), 656-664.

[5] Zhou, B., & Wand, M. P. (2017). Semi-parametric maximum likelihood estimation with data-driven smoothing for propensity score-based weights. Biometrika, 104(2), 325-338.

[6] Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11(5), 550-560.

[7] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.