基于广义增强模型的倾向性评分重叠权重加权方法研究*

涂博祥[1]　秦婴逸[1△]　徐　宵[2]　赵艳芳[1]

【提　要】 **目的**　本研究构建了基于广义增强模型的倾向性评分重叠权重加权模型(GBM-OW)。**方法**　通过模拟数据探讨在混杂因素与处理因素间关系复杂的情况下,不同样本量,不同倾向性评分值重叠程度下 GBM-OW 模型在均衡混杂因素、效应估计等方面的表现,并与多因素调整模型及其它三种倾向性评分加权模型进行比较。**结果及结论**　从模拟结果来看,当变量之间关系复杂、样本量大、倾向性评分值重叠程度小的情况下,GBM-OW 模型在各方面均有较好的表现,可应用于观察性研究中。

【关键词】 倾向性评分　重叠权重　广义增强模型

【中图分类号】 R191.5　　【文献标识码】 A　　**DOI**　10.11783/j.issn.1002-3674.2024.05.009

Research of Generalized Boosting Model Combined with Propensity Score Overlap Weighting

Tu Boxiang, Qin Yingyi, Xu Xiao, et al (*Department of Military Health Statistics, Naval Medical University, Shanghai* 200433)

【**Abstract**】 **Objective**　This study constructed the generalized boosting model combined with propensity score overlap weighting(GBM-OW). **Methods**　Within the situations that there are complex relationships between confounders and treatment factors, and different sample size and different propensity score overlap, we explored the performance of GBM-OW model in balance confounders and estimate effect. And compared with multivariate adjusted model and other three propensity score weighting models. **Results and Conclusion**　From the simulation results we concluded that when the relationship between variables is complex, the sample size is large, and the propensity score value overlap is small, the GBM-OW model has a good performance in all aspects and can be used in observational studies.

【**Key words**】 Propensity score; Overlap weighting; Generalized boosted model

在观察性研究中,倾向性评分(propensity score, PS)加权作为一种均衡组间混杂因素的方法已得到越来越多的应用,其基本思想是利用 PS 值对研究对象进行加权,生成混杂因素特征分布均衡的虚拟人群。倾向性评分加权的一般分析步骤可以概括为:①估计 PS 值;②利用 PS 值对研究对象进行加权;③检验加权后的均衡性;④处理效应估计[1-2]。

第一步中估计 PS 值是以混杂因素为自变量,处理因素为因变量构建回归模型,最常用的为 logistic 回归模型[3]。在利用 logistic 回归模型估计 PS 值时需要在模型中正确设定混杂因素与处理因素的关系(包括线性关系、非线性关系、交互关系等),但实际上混杂因素与处理因素的关系往往是复杂且未知的,因此通常在模型中只放入混杂因素的主效应,但这可能会导致 PS 值估计不准确。广义增强模型(generalized boosted models, GBM)是一种自迭代的回归模型,可通过不断迭代找到自变量与因变量之间的作用关系,相较于 logistic 回归方法其简便之处在于无需人为设置自变量与因变量之间的关系,GBM 利用自适应算法自动估计自变量与因变量之间的关系,包括线性关系、非线性关系、交互关系等。McCaffrey 等人于提出用 GBM 估计 PS 值[4],当混杂因素与处理因素之间的线性、非线性或交互作用等函数形式无法确定时,只需通过一定的参数设置以及足够多的迭代次数即可得到更加准确的 PS 值估计。

第二步中常用的加权方法为逆概率加权法(inverse probability weight, IPW),目标人群为全部研究对象,估计的效应为平均处理效应(average treatment effect, ATE),权重的取值范围为(1,+∞)。当组间 PS 值重叠较少时,IPW 会出现极端权重的问题,导致个别研究对象对处理效应估计的影响较大,从而导致处理效应的偏倚。重叠权重加权(overlap weighting, OW)的目标人群为 PS 值重叠区域的研究对象,这部分研究对象被分到处理组和对照组的概率相当,估计的效应为重叠人群平均处理效应(ATE in the overlap, ATO),重叠权重的取值范围为(0,1),减弱了极端权重人群对效应估计的影响[5-6]。

目前 R 软件[7]中有多款程序包可实现倾向性评分加权分析,如 PSW 包[8]和 PSweight 包[9]可实现基于 logistic 回归的加权,PSweight 包和 twang 包[10]可实现基于 GBM 的加权。其中 PSweight 包可实现基于

* 基金项目:国家自然科学基金项目(82003558);海军军医大学"深蓝"人才工程"启航计划";上海市 2022 年度"科技创新行动计划"启明星项目(22QA1411400)

1. 海军军医大学军队卫生统计学教研室(200433)
2. 同济大学医学院
△通信作者:秦婴逸,E-mail:yingyi_qin@163.com

GBM 的逆概率加权（GBM-IPW）和重叠权重加权（GBM-OW），但其直接采用最后一次迭代下的拟合，并未对所有迭代下的拟合进行比较和选择。twang 包中 GBM-IPW 模型的应用是根据加权后混杂因素的均衡性对各迭代进行比较，选择了均衡性最好的迭代为最佳拟合，但 twang 包中尚不能进行 GBM-OW 模型的应用。

关于各种倾向性评分加权方法及 R 软件的实现已有详细介绍[11]，本研究将通过模拟生成处理因素为二分类，结局为连续型资料的模拟数据，构建 GBM-OW 模型，并根据均衡性检验的结果选择最佳拟合，与多因素调整模型（adjusted）、基于 logistic 回归的逆概率加权（logistic-IPW）、基于 logistic 回归的重叠权重加权（logistic-OW）、GBM-IPW 等模型比较在均衡混杂因素、效应估计等方面的表现。

模拟数据集

1.混杂因素

本研究模拟数据集共设置 8 个混杂因素 $W_1 \sim W_8$，其中 $W_1 \sim W_4$ 分别为服从 $Bern(0.2)$、$Bern(0.4)$、$Bern(0.5)$、$Bern(0.6)$ 分布的二分类变量，$W_5 \sim W_8$ 分布为服从 $N(1,1)$、$N(2,1)$、$N(3,1)$、$N(4,1)$ 标准正态分布的连续型变量。

2.处理因素

处理因素 A 为服从 $Bern(P_A)$ 分布的二分类变量，$A=1$ 表示处理组，$A=0$ 表示对照组，P_A 为 $A=1$ 的概率。设定 P_A 由 8 个混杂因素构建回归方程得到，回归方程如下：

$$\ln\left(\frac{P_A}{1-P_A}\right)=\beta_0+\gamma W_A$$

上式中 β_0 为常数，可通过调整 β_0 来调整 P_A 的大小，本研究设定 P_A 的平均值为 0.4，使处理组所占比例大致为 0.4，W_A 为 8 个混杂因素构建的方程，通过调整 γ 的大小可设置处理组和对照组间混杂因素特征的重叠程度，γ 越小则重叠程度越大，γ 越大则重叠程度越小。为了探讨当混杂因素与处理因素间关系复杂时各模型的表现，本研究设定 W_A 中有混杂因素的主效应外，还有两个二次项和两个交互作用，具体公式如下：

$$W_A=\beta_1W_1+\beta_2W_2+\beta_3W_3+\beta_4W_4+\beta_5W_5+\beta_6W_6+\beta_7W_7+\beta_8W_8+\beta_{67}W_6W_7+\beta_{58}W_5W_8+\beta_{66}W_6^2+\beta_{88}W_8^2$$

通过不同的 γ（0.1、0.3、0.5、1）设置 4 种不同的组间 PS 值重叠程度，分别为重叠程度大、较大、较小和小，四种 PS 值重叠情况如图 1 所示。

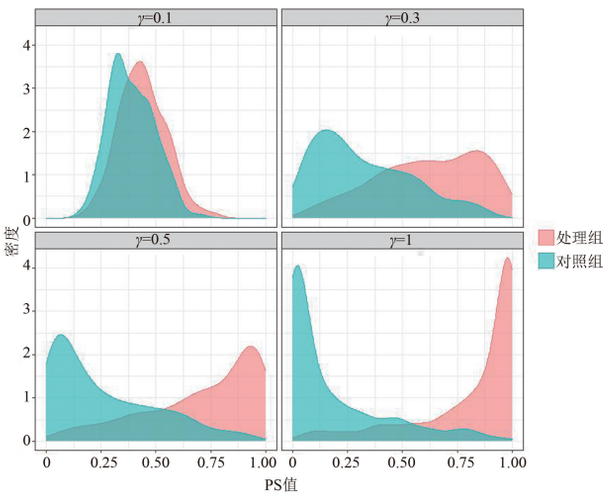

图 1 模拟数据中四种不同 PS 值重叠程度示例

3.结局变量

本研究结局变量 Y 为连续型变量，构建结局变量 Y 与处理因素 A 及混杂因素间的线性回归方程：

$$Y=\beta_0+W_Y+\beta_AA+\varepsilon$$

其中 β_0 为常数项，ε 为误差项，W_Y 的设置与 W_A 类似。β_A 为处理因素的系数，即为处理因素效应值真值，本研究中设置 $\beta_A=1$。

4.样本量

共设置 6 种样本量，分别为 500、1000、2000、3000、4000、5000，探讨不同样本量情况下各模型的表现。

因此，4 种 PS 值重叠程度，6 种样本量，共 24 种场景，每种场景下模拟 1000 个数据集。模拟数据示例见表 1。

表 1 24 种场景模拟数据示例

W1	W2	W3	W4	W5	W6	W7	W8	A	Y
0	1	1	1	−1.1585	2.4657	0.1781	2.6967	0	15.3799
0	0	0	0	2.2378	2.9790	4.4507	4.6115	1	17.2423
0	0	1	0	1.8048	1.0878	1.8582	3.9203	0	14.8737
1	1	0	1	1.3740	2.8633	3.5291	5.3630	0	16.7308
0	0	0	1	2.4693	1.5715	2.8725	3.0473	0	17.2862
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0	0	0	0	1.1900	1.1676	1.8082	2.7379	1	16.0471
0	0	1	1	0.8498	1.4903	2.7357	3.6748	1	16.3782
0	1	1	1	0.3586	1.6051	2.8382	2.6780	0	16.1875
0	0	1	1	1.2144	2.2953	2.2311	4.1473	1	17.9208
0	0	1	0	−0.2760	1.0021	3.4007	3.6057	0	16.9359

构建 GBM-OW 模型

GBM-OW 模型的构建步骤大致如下：①构建以混杂因素为自变量，处理因素为因变量的 GBM 模型估计 PS 值；②根据重叠权重的加权公式 $OW=A(1-PS)+(1-A)PS$ 对研究对象进行加权；③检验加权后混杂因素的均衡性；④迭代步骤 1~3；⑤找出均衡性最好的迭代；⑥用最佳迭代下的 PS 值进行重叠权重加权；⑦加权后的效应估计（图 2）。

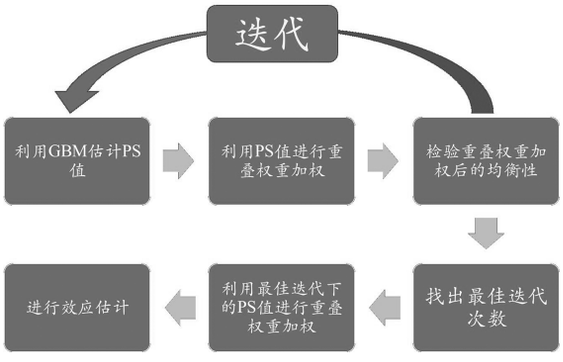

图 2　GBM-OW 模型的构建步骤

其中均衡性检验的指标有绝对标准均值差（absolute standardized mean difference，ASMD）[12-13]、Kolmogorov-Smirnov 统计量（简称 KS 值）[14]等，ASMD 值或 KS 值越小，则均衡性越好[15-17]。本研究中利用 KS 值选择最佳迭代。

在利用 gbm 包实现 GBM 模型时，需要设置的参数有：最大迭代次数 *n*.trees、最大交互作用阶数 interaction.depth、收缩系数 shrinkage 等。若最大迭代次数过小，则模型可能在最大迭代次数前尚未达到最佳拟合，若最大迭代次数过大则会浪费计算力。interaction.depth 为模型中允许的混杂因素之间的最大交互阶数，2 个变量间的交互作用为 2 阶交互，3 个变量间的交互作用为 3 阶交互。收缩率反应迭代过程中的精细化程度，收缩率越小，迭代越精细，不过需要更多的计算力。经过综合考虑，本研究设置参数如下：*n*.trees = 20000，interaction.depth = 3，shrinkage = 0.01。其中加权后的 KS 值和 ASMD 值通过自编程序获取，效应估计的标准误利用 sandwich 包[18-19]通过夹心方差法获得。

GBM-OW 模型与其它模型的比较见表 2。

表 2　GBM-OW 模型与其它模型的比较

模型	PS 值模型	权重	结局模型
adjusted	无	1	$Y=\beta_0+\beta_A A+\beta_1 W_1+\cdots+\beta_8 W_8$
logistic-IPW	logistic 模型（只纳入混杂因素的主效应）	$\frac{A}{PS}+\frac{1-A}{1-PS}$	$Y=\beta_0+\beta_A A$
logistic-OW	logistic 模型（只纳入混杂因素的主效应）	$A(1-PS)+(1-A)PS$	$Y=\beta_0+\beta_A A$
GBM-IPW	GBM 模型	$\frac{A}{PS}+\frac{1-A}{1-PS}$	$Y=\beta_0+\beta_A A$
GBM-OW	GBM 模型	$A(1-PS)+(1-A)PS$	$Y=\beta_0+\beta_A A$

评价指标

1.均衡性评价

对于每个模型，均计算所有混杂因素的 KS 值和 ASMD 值，其中多因素调整模型为未加权的指标，logistic-OW 与 logistic-IPW 为加权后的指标，GBM-OW 与 GBM-IPW 模型则为基于设定规则下最佳拟合加权后的指标。选择所有混杂因素中 KS 值与 ASMD 值的最大值参与比较。

2.效应估计评价

（1）相对偏倚

相对偏倚为效应估计值与所设真值差值与真值比值的绝对值，即：

$$\delta(\%)=\left|\frac{(\hat{\beta}-\beta)}{\beta}\right|100\%$$

（2）均方根误差（root mean squared error，RMSE）

均方根误差是估计值与真值偏差的平方和与模拟次数 N 比值的平方根，即：

$$RMSE=\sqrt{\frac{\sum(\hat{\beta}-\beta)^2}{N}}$$

RMSE 既考虑了估计的偏倚，又考虑了估计的变异，本研究中模拟次数 $N=1000$。

（3）95% 置信区间（confidence interval，CI）覆盖率

上述相对偏倚和 RMSE 均是对点估计的评价，除此之外，还要对区间估计进行评价，即效应估计值的 95%*CI* 是否覆盖了真值，而在所有模拟中，95%*CI* 覆盖了真值的模拟次数与总的模拟次数的百分比即为 95%*CI* 覆盖率。理论上来说，一个好的模型的 95%*CI* 覆盖率应接近于 95%[20]。

结　果

五种模型的模型结果见表 3。

表 3 五种模型的模拟结果

模型	KS	ASMD	δ（%）	RMSE	95%*CI* 覆盖率（%）
adjusted	0.36	0.90	16.28	0.18	39.49
logistic-IPW	0.14	0.23	23.18	0.64	85.28
logistic-OW	0.07	0.00	15.60	0.18	86.82
GBM-IPW	0.13	0.36	156.49	1.58	19.94
GBM-OW	0.01	0.03	8.98	0.13	98.44

不同 PS 值重叠程度及不同样本量下各模型在效应估计方面的表现详见图 3。

讨 论

从表 3 可以看出，GBM-OW 模型的 5 项指标中，除 ASMD 值比 logistic-OW 模型大，其它 4 个指标均为五种模型中最好的。logistic-OW 模型的 ASMD 值接近于 0，这符合 OW 精确均衡的特点，Li Fan 等人已在文献中介绍过[21]。ASMD 评价的是均值的差异，而 KS 值评价的是分布的差异，GBM-OW 模型在这两个指标上均有良好的表现，说明 GBM-OW 模型具有较好的均衡混杂因素的能力。总的来说，两种 OW 模型在均衡性上的表现要好于两种 IPW 模型，考虑这是由于 OW 与 IPW 的目标人群不同导致的，IPW 的目标人群为全人群，而 OW 的目标人群是 PS 值重叠人群，在分析和解释结果时要注意目标人群的问题[22]。

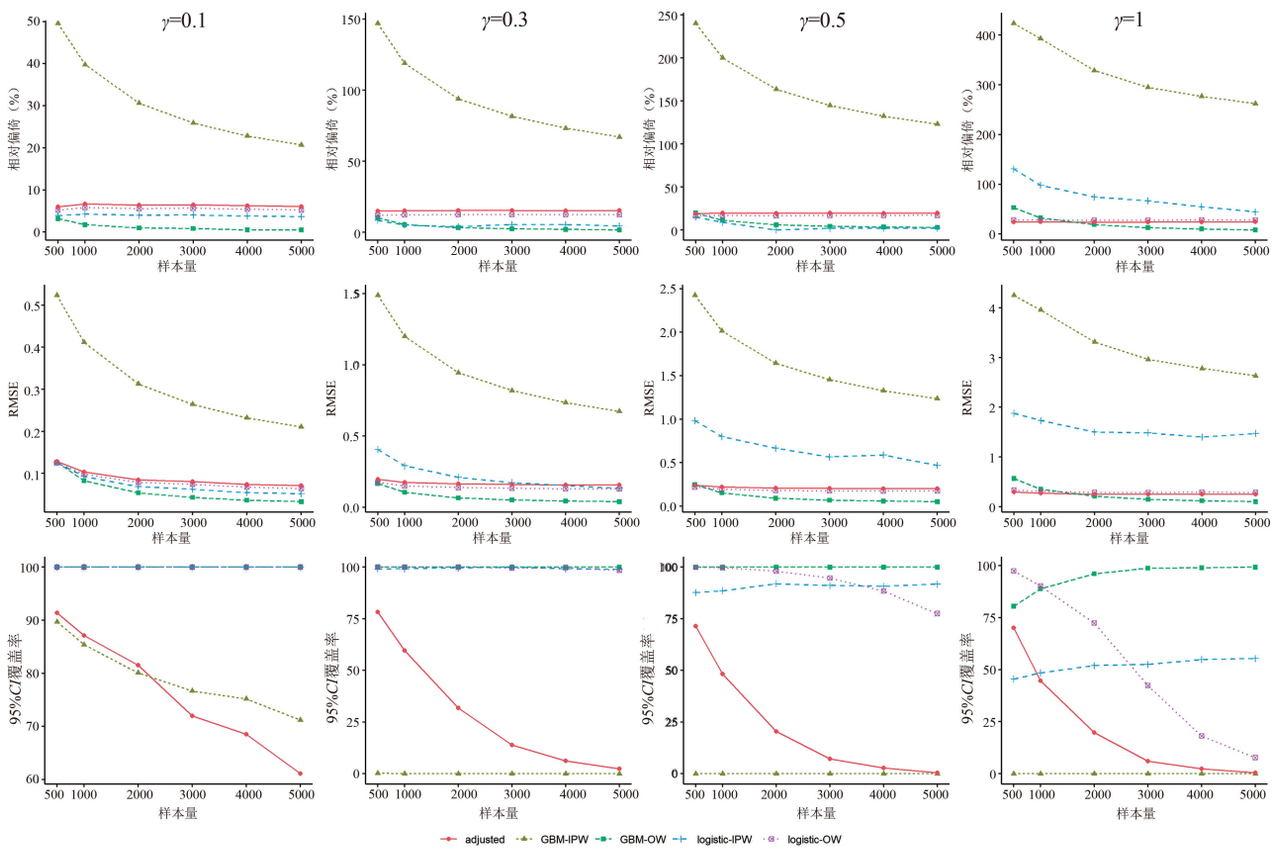

图 3 五种模型在效应估计方面的表现比较

从图 3 中可以看出，GBM-OW 模型在相对偏倚及 RMSE 上的表现随着样本量的增大而变佳，考虑是因为 GBM 模型迭代拟合需要一定的样本量支持，提示 GBM-OW 模型尤其适用于大样本的数据，如观察性研究。当 PS 值重叠程度小时，其它模型的表现不佳，此时 GBM-OW 模型仍能有较好的表现。从本研究模拟结果来看，GBM-IPW 模型的表现并不好，甚至大多数指标均为几种模型中最差的，考虑可能是相对 logistic 回归模型，GBM 模型会进一步放大 IPW 中极端数据的影响，具体原因还需今后进一步研究。虽然本文只展示了混杂因素与处理因素间关系复杂情况下的模拟结果，不过本研究还模拟了其它情况下的数据，当变量间关系较简单的情况下，几种模型的表现相对于数据复杂情况下的表现要好，GBM-OW 模型的表现不差于其它模型，只是当数据情况复杂时更加体现出 GBM-OW 模型优越的表现。

GBM-OW 模型的 95%*CI* 覆盖率为98.44%，略高于 95%，考虑是因为利用 GBM 估计 PS 值，没有关于 PS 值模型的回归系数，因此在利用夹心方差法计算的是保守方差，标准误偏大，置信区间较宽。但总的来说，GBM-OW 模型在 95%*CI* 覆盖率上的表现优于其他模型，这与其点估计的准确性比其它模型高也有一定的关系。

目前 GBM-IPW 模型可以通过 twang 包实现并选

择最佳拟合,GBM-OW 模型实现最佳拟合的选择尚需通过自编程序实现,不过从本研究的模拟结果来看,GBM-OW 选择的最佳拟合的迭代次数均接近最大次数,因此也可考虑通过 PSweight 包实现 GBM-OW 模型,直接选择最大迭代次数下的拟合。

综上所述,GBM-OW 模型在各种数据情况下均有较好的表现,尤其是当变量间关系复杂、PS 值重叠程度小以及样本量大的情况下,因此适用于观察性研究。目前可利用 gbm 包,结合自编程序选择 GBM-OW 的最佳拟合,也可通过 PSweight 包直接选择最大迭代次数下的拟合。在利用 GBM-OW 模型进行分析和结果解释时,要注意目标人群为 PS 值重叠人群而不是全人群。本研究的结局为连续型资料,后续还将对其它类型的结局,如二分类结局、生存结局等进行模拟研究,探讨 GBM-OW 模型在各种类型结局数据下的表现。在此之前,已将 GBM-OW 模型用于实例分析中的敏感性分析[23],与 adjusted、倾向性评分匹配、logistic-OW 等模型相比,其结果具有稳健性,今后将进一步优化程序,探索如何更好地将该模型应用于实例中。

参 考 文 献

[1] 王永吉, 蔡宏伟, 夏结来, 等. 倾向指数 第一讲 倾向指数的基本概念和研究步骤[J]. 中华流行病学杂志, 2010,31(3): 347-348.

[2] 王永吉, 蔡宏伟, 夏结来, 等. 倾向指数 第二讲 倾向指数常用研究方法[J]. 中华流行病学杂志, 2010,31(5): 584-585.

[3] 吴美京, 吴骋, 王睿, 等. 倾向性评分法中评分值的估计方法及比较[J]. 中国卫生统计, 2013, 30(3): 440-444.

[4] Mccaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies[J]. Psychological Methods, 2004, 9(4): 403-425.

[5] Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights[J]. Am J Epidemiol, 2019, 188(1): 250-257.

[6] Mlcoch T, Hrnciarova T, Tuzil J, et al. Propensity Score Weighting Using Overlap Weights: A New Method Applied to Regorafenib Clinical Data and a Cost-Effectiveness Analysis[J]. Value Health, 2019, 22(12): 1370-1377.

[7] R Core Team. R: A language and environment for statistical computing[EB/OL]. R Foundation for Statistical Computing, Vienna, Austria. 2020,https://www.R-project.org/.

[8] Mao HZ, Li L. PSW: Propensity Score Weighting Methods for Dichotomous Treatments[EB/OL]. R package version 1.1-3. 2018,https://CRAN.R-project.org/package=PSW.

[9] Zhou TH, Tong GY, Li F, et al. PSweight: Propensity Score Weighting for Causal Inference with Observational Studies and Randomized Trials. R package version 1.1.5. 2021,https://CRAN.R-project.org/package=PSweight.

[10] Cefalu M, Ridgeway G, McCaffrey D, et al. twang: toolkit for weighting and analysis of nonequivalent groups. R package version 2.5. 2021, https://CRAN.R-project.org/package=twang.

[11] 涂博祥, 秦婴逸, 吴骋, 等. 倾向性评分加权方法介绍及 R 软件实现. 中国循证医学杂志, 2022, 22(3): 365-372.

[12] Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies[J]. Med Decis Making, 2009, 29(6): 661-677.

[13] Schober P, Mascha EJ, VetteR TR. Statistics From A(Agreement) to Z(z Score): A Guide to Interpreting Common Measures of Association, Agreement, Diagnostic Accuracy, Effect Size, Heterogeneity, and Reliability in Medical Research[J]. Anesthesia & Analgesia, 2021,133(6):1633-1641.

[14] 胡克震. Kolmogorov—Smirnov 检验法的应用[J]. 中国卫生统计, 1985,2(3): 12-17.

[15] Mccaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models[J]. Statistics in Medicine, 2013, 32(19): 3388-3414.

[16] 王永吉, 蔡宏伟, 夏结来, 等. 倾向指数 第三讲 应用中的关键问题[J]. 中华流行病学杂志, 2010,31(7): 823-825.

[17] Zhang Z, Kim HJ, Lonjon G, et al. Balance diagnostics after propensity score matching[J]. Annals of Translational Medicine, 2019, 7(1):16.

[18] Zeileis A, Köll S, Graham N. Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R [J]. Journal of Statistical Software, 2020, 95(1): 1-36.

[19] Zeileis A. Econometric Computing with HC and HAC Covariance Matrix Estimators [J]. Journal of Statistical Software, 2004, 11(10): 1-17.

[20] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods[J]. Statistics in Medicine, 2019, 38(11): 2074-2102.

[21] Li F, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting[J]. Journal of the American Statistical Association, 2018, 113(521): 390-400.

[22] 秦宇辰, 郭威, 阮一鸣, 等. 重叠加权法在医学研究混杂因素控制中的应用[J]. 中国卫生统计, 2020, 37(3): 363-366.

[23] Tu B, Tang Y, Cheng Y, et al. Association of Prior to Intensive Care Unit Statin Use With Outcomes on Patients With Acute Kidney Injury[J]. Frontiers in Medicine, 2021, 8:810651.

附录:关键程序的 R 代码

```
###第一部分 模拟数据生成###
DataGenerate<- function(n,gammaA){
#n 为样本量,gammaA 为 γ
  p1 <- 0.2; p2 <- 0.4; p3 <- 0.5; p4 <- 0.6
  m5 <- 1; m6 <- 2; m7 <- 3; m8 <- 4
  w1 <- rbinom(n, size=1, prob=p1)
  w2 <- rbinom(n, size=1, prob=p2)
  w3 <- rbinom(n, size=1, prob=p3)
  w4 <- rbinom(n, size=1, prob=p4)
  w5 <- rnorm(n, m5,1)
  w6 <- rnorm(n, m6,1)
  w7 <- rnorm(n, m7,1)
  w8 <- rnorm(n, m8,1)
  W_A <- 0.3*w1+0.2*w2+0.4*w3+0.1*w4-0.4*w5
+0.1*w6+0.1*w7+0.1*w8+0.5*w6*w6+0.2*w8*w8+
0.4*w6*w7+0.4*w5*w8
```

```
W_a <- 0.3 * p1+0.2 * p2+0.4 * p3+0.1 * p4-0.4 * m5+0.1 * m6+0.1 * m7+0.1 * m8+0.5 * m6 * m6+0.2 * m8 * m8+0.4 * m6 * m7+0.4 * m5 * m8
W_Y <- 0.4 * w1+0.3 * w2+0.2 * w3+0.1 * w4+0.4 * w5-0.3 * w6+0.2 * w7+0.1 * w8+0.2 * w6 * w6+0.3 * w8 * w8+0.5 * w6 * w7+0.2 * w5 * w8
W_y <- 0.4 * p1+0.3 * p2+0.2 * p3+0.1 * p4+0.4 * m5-0.3 * m6+0.2 * m7+0.1 * m8+0.2 * m6 * m6+0.3 * m8 * m8+0.5 * m6 * m7+0.2 * m5 * m8
#设置常数项
probTreat=0.4
beta0A=log(probTreat/(1-probTreat))- gammaA * W_a
probA <- plogis(beta0A+gammaA * W_A)
A <- rbinom(n, size=1, prob=probA)
#结局
Y.1 <- rnorm(n, mean=15+1+W_Y, sd=1)
Y.0 <- rnorm(n, mean=15+0+W_Y, sd=1)
Y <- Y.1 * A+Y.0 * (1-A)
return(data.frame(w1,w2,w3,w4,w5,w6,w7,w8,A,Y))
}
#生成模拟数据,以样本量为 1000,γ 为 0.1 为例
data <- DataGenerate(n=1000, gammaA=0.1)

###第二部分 模型构建###
#加载需要用的程序包
library(gbm)
library(sandwich)
library(survey)
#进行 GBM 拟合
gbm <- gbm(A~w1+w2+w3+w4+w5+w6+w7+w8,
           data=data,
           distribution="bernoulli",
           n.trees=20000,
           interaction.depth=3,
           n.minobsinnode=10,
           shrinkage=0.01,
           bag.fraction=1,
           train.fraction=1,
           verbose=FALSE,
           keep.data=FALSE)
#导出 PS 值(以第 20000 次迭代为例)
ps <- predict(gbm, newdata=data, n.trees=20000,type="response")
#利用 PS 值进行重叠权重加权
data$OW[data$A==0] <- ps[data$A==0]
data$OW[data$A==1] <- 1-ps[data$A==1]
#计算加权后的 KS 值(以变量 W5 为例)(参考 twang 包中计算加权后 KS 值的程序)
data$w[data$A==1] <- with(subset(data, A==1), OW/sum(OW))
data$w[data$A==0] <- with(subset(data, A==0), -OW/sum(OW))
ind <- order(data[,"w5"])
cumv <- abs(cumsum(data$w[ind]))
cumv <- cumv[diff(data[,"w5"][ind])!=0]
ks <- ifelse(length(cumv)>0, max(cumv), 0)
#计算加权后的 ASMD 值(以二分类变量 w1 与连续型变量 w5 分别举例)
design <- svydesign(ids=~1, weight=~w, data=data.frame(x=data[,c("w1","w5")], t=data$A, w=data$OW, sampw=rep(1,nrow(data))))
design.t <- subset(design,t==1)
design.c <- subset(design,t==0)
attach(data)
##二分类变量的合并标准差
sd.denorm.w1 <- with(data, sqrt((mean(w1[A==1])*(1-mean(w1[A==1]))+mean(w1[A==0])*(1-mean(w1[A==0])))/2))
##连续型变量的合并标准差
sd.denorm.w5 <- with(data,sqrt((var(w5[A==1])+var(w5[A==0]))/2))
detach(data)
ASMD_w1 <- abs(svymean(~x.w1, design.t, na.rm=TRUE)[[1]]-svymean(~x.w1, design.c, na.rm=TRUE)[[1]])/sd.denorm.w1
ASMD_w5 <- abs(svymean(~x.w5, design.t, na.rm=TRUE)[[1]]-svymean(~x.w5, design.c, na.rm=TRUE)[[1]])/sd.denorm.w5
#效应估计
lm <- lm(Y~A, data=data, weights=data$OW)#拟合以 OW 加权后的线性回归方程
effect <- coefficients(lm)[[2]]#提出效应估计值
SE <- sqrt(diag(sandwich(lm)))[[2]]#以夹心方差法计算标准误
```

(责任编辑:郭海强)