·系列讲座·

倾向指数
第二讲　倾向指数常用研究方法

王永吉　蔡宏伟　夏结来　蒋志伟

【关键词】 倾向指数；匹配；分层；协变量调整

Propensity score (Ⅱ) Three commonly used methods on propensity score

WANG Yong-ji[1], *CAI Hong-wei*[2], *XIA Jie-lai*[1], *JIANG Zhi-wei*[1]. 1 *Department of Health Statistics, Faculty of Preventative Medicine*, 2 *Information Center, School of Stomatology, Fourth Military Medical University, Xi'an* 710032, China

Corresponding author: *CAI Hong-wei*, Email: hwcai@fmmu.edu.cn; *XIA Jie-lai*, Email: xiajielai@fmmu.edu.cn

This work was supported by a grant from the National Natural Science Foundation of China (No. 30800952)

【Key words】 Propensity score; Matching; Stratification; Covariate adjustment

倾向指数(编者注：国内有学者译为倾向评分)方法最主要作用是通过均衡暴露组和对照组间的协变量分布来控制组间偏倚[1-3]。通过倾向指数调整，倾向指数相同或相近的两个研究对象，虽然可能拥有不同的特征变量，但进入倾向指数模型调整的多个特征变量之间总体上是均衡的。在第一部分已经介绍了倾向指数方法的概念和研究步骤，这部分将主要介绍倾向指数的常用方法，并比较几种常用的倾向指数方法。倾向指数主要研究方法包括匹配、分层和协变量调整等。

1. 匹配法：倾向指数匹配法在医学研究中应用最为广泛。传统匹配是从对照中选出与暴露组特征相近的个体进行配对，常用的方法是马氏距离(Mahalanobis distance)匹配，即用马氏距离来评价个体特征相近的程度，马氏距离排除了变量间相关性的干扰，消除了量纲，但马氏距离夸大了变异微小的变量作用，而且不适用于协变量较多的情况[4,5]。而倾向指数可以综合表示多个协变量共同作用的结果，克服了马氏距离匹配的缺点。

倾向指数匹配是在通过模型估计倾向指数后，从对照组中选出与暴露组倾向指数相同或相近的个体进行配对，以达到均衡组间协变量的目的。从匹配范围上，倾向指数匹配法可分为局部匹配(local algorithms)和全局匹配(global algorithms)。所谓局部匹配，也称最近可用匹配(nearest available neighbor matching)，是指暴露组从第一个个体开始，在对照组中寻找倾向指数与其最接近的个体，直到暴露组所有个体都有匹配的个体，其优点在于匹配集的最大化，最大程度保留了研究样本的信息。如果在最近匹配的基础上加一个限制条件，即暴露组与对照组个体倾向指数差值在事先设定的某范围内才能进行匹配，就是卡钳匹配(caliper matching)。卡钳的设定会影响匹配集的样本量，卡钳值越大，能够完成匹配的个体越多，匹配集就越大，但是可能会产生一些不良匹配(bad matches)，也就是倾向指数差值较大的暴露组和对照组个体形成匹配，会增大估计处理效应的偏倚；反之，卡钳设置过小，会减少匹配样本，降低估计处理效应的准确性。Cochran和Rubin研究，卡钳值取两组倾向指数标准差的60%可以减少86%～91%的偏倚，取两组倾向指数标准差的20%可以减少98%～99%的偏倚。Austin[6]通过蒙特卡罗模拟比较了研究者实际应用中经常选用的卡钳值：模拟试验分别将卡钳值设为两组倾向指数标准差的20%、60%，或者直接将两组个体倾向指数匹配的最大绝对差值分别设定为0.005、0.01、0.02、0.03和0.1。研究结果表明最合适的卡钳值是取两组倾向指数标准差的20%或者取两组间倾向指数绝对差值(卡钳值)为0.02或0.03。

全局匹配法是把匹配问题转化为运筹学中网络流(network flows)问题[7]，把暴露组和对照组个体看作节点(node)，把匹配转化为求最小化节点间的总距离，不保证每个处理都能找到最优的匹配，也就是说，与暴露组个体匹配的对照组个体倾向指数的差值并不是最小的，但是能保证匹配集倾向指数总体差值的最小化，这个优势是其他匹配方法无法比拟的。全局匹配法不需要设定卡钳或者半径，但是当数据为海量时，需要建立巨大的距离矩阵，影响执行效率，所以在实际应用中并不常见。也有研究者把全局匹配转化为指派问题(assignment problem)来处理[8]。

此外，局部匹配还存在是否允许放回(replacement)的问题。所谓允许放回，指在匹配过程中允许重复利用对照，允许放回使匹配样本倾向指数总体差异最小化。在卡钳匹配中，允许放回能匹配更多个体，增大匹配数据集，同时也能在一定程度上减少不良匹配，特别是在对照组个体倾向指数只有少部分与暴露组相近时。例如，暴露组个体倾向指数都很高，而对照组只有一部分与其接近，如果强行匹配，结果中会有很多倾向指数差值较大的不良匹配[9]。但如果允许放回，在选择分析方法时需要考虑匹配后对照组内包含重复个体，是否应该在分析中考虑对照组个体缺乏独立性的特点，以及选用何种统计方法来进行分析等问题都存在争议，有待进一步验证和讨论[1]，所以在实际应用中，一般不允许放回，即一旦一对个体匹配完成，对照组个体就不会被考虑做重复

DOI: 10.3760/cma.j.issn.0254-6450.2010.05.026
基金项目：国家自然科学基金(30800952)
作者单位：710032 西安，第四军医大学预防医学系卫生统计学教研室(王永吉、夏结来、蒋志伟)，口腔医学院信息中心(蔡宏伟)
通信作者：蔡宏伟，Email: hwcai@fmmu.edu.cn；夏结来，Email: xiajielai@fmmu.edu.cn

配对。

2. 分层法：传统分层法根据协变量把样本分层，通过分层来抵消因为组间某个或某些协变量不均衡对研究结果的影响，在层内进行组间比较。在实际应用中，传统分层法存在一些问题：首要问题是当变量增多时，层数成指数趋势显著增加，例如每个变量都是二分类，分层后，n个变量会产生出2^n层，可能有部分层不含暴露组或者对照组个体，导致该层内无法估计处理效应，所以当协变量很多时，传统分层法并不适用；还有一个问题是分层的变量只能是分类变量，而不适用于连续性变量，如果把连续性变量转化为分类变量进行分层，会损失研究样本的信息，从而影响对处理效应的估计。

倾向指数分层法是把倾向指数作为分层的惟一标准，通过模型估计倾向指数后，确定倾向指数界值的范围，然后按倾向指数分为若干区间，视区间为层进行分析，层内组间协变量分布应该是均衡的，将各层处理效应赋予权重后相加来估计处理效应，并检验各层内暴露组和对照组间每个协变量的均衡性[10]。分层法也可以做层内比较，层内比较的结果可以很直观、简洁地用直方图来表示。

倾向指数分层法在应用中要注意的问题是分层数和权重的设定。可以通过比较层内组间倾向指数的均衡性来检验所选定的层数是否合理，如果层内组间倾向指数是不均衡的，说明分层数可能不够，需要增加层数。权重衡量各层处理效应估计对总体处理效应估计的作用大小，一般由各层样本占总样本量的比例来确定，也有研究者认为通过倾向指数分层后，最高层和最低层组间处理分配是不均衡的，如果用样本比例来确定权重，在估计总体处理效应时会增大偏倚，建议用各层内处理效应方差的倒数来确定权重[11]。倾向指数分层法与传统分层法相比，优点在于协变量的增多不会影响层数，因此可以应用于协变量很多的情况，也不受协变量类型的影响。根据文献研究[12]，如果协变量为连续性变量，五层均等分层法，即按倾向指数把样本平均分为五层，能减少90%以上的偏倚，这也是分层法中最常用的方法。

3. 协变量调整：匹配和分层法主要用于均衡组间协变量，使组间具有可比性。而协变量调整法是把倾向指数引入模型，直接作为回归分析的一个协变量，或者把代表多个协变量的倾向指数作为回归分析的惟一协变量，以结局变量为应变量来构建模型，估计处理效应。

4. 评价：倾向指数匹配法、分层法和协变量调整法在医学研究中都有不同程度应用，其中倾向指数匹配法应用最为广泛。倾向指数匹配法应用广泛有以下一些原因：首先，以往的实际应用和理论研究表明，在不遗漏混杂因素的情况下，倾向指数匹配法能得到处理效应的无偏估计，但是无论选择何种倾向指数模型，分层法都会产生处理效应的有偏估计，匹配法比分层法能更大程度地减少偏倚。第二，倾向指数匹配法对协变量的均衡能力优于分层法[10]，组间协变量能得到更好的平衡，即使运用五层均等分层法，分层后组间协变量在最高层和最低层也可能是不平衡的。第三，倾向指数匹配法能直接比较匹配数据集暴露组和对照组间协变量的均衡性，从而确定暴露组和对照组间的可比性，与其他方法相比更为直观，而分层法只能在层内比较，不能直接比较研究样本的均衡性，协变量调整法则无法比较。第四，针对倾向指数匹配法的灵敏度检验方法已经提出并应用，灵敏度检验用来分析潜在混杂因素引起的偏倚对估计处理效应产生的影响，而针对倾向指数协变量调整和加权法的灵敏度检验还没有实质性的进展[6]。第五，协变量调整法是基于模型的分析，它要求建模正确，增加了结果的不确定性，丧失了倾向指数方法易于理解、结果便于解释的特点，这是倾向指数方法与传统多元方法的最大区别，而匹配法和分层法不需要建模。第六，Rubin研究表明在暴露组和对照组间协变量方差不齐的情况下，协变量调整法可能会增加偏倚，而组间协变量方差不齐在观察性研究中很常见，所以协变量调整法要谨慎运用[4]。

倾向指数分层法和协变量调整法的优势是没有损失样本，最大限度地保留了原有信息，这也正是匹配法的缺点，匹配后因为排除了无法匹配的样本而减少了样本量，如果暴露组和对照组间样本量差别较大，可能会造成匹配样本占原始样本的比例过小，从而改变样本特征，会降低估计处理效应的准确性。

参 考 文 献

[1] Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med, 2008, 27:2037-2049.

[2] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med, 2007, 26:734-753.

[3] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika, 1983, 70:41-55.

[4] D' Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med, 1998, 17:2265-2281.

[5] Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Eco Stat, 2002, 84: 151-161.

[6] Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biomet J, 2009, 51: 171-184.

[7] Rosenbaum PR. Optimal matching for observational studies. J Am Stat Asso, 1989, 84: 1024-1032.

[8] Ming K, Rosenbaum PR. A note on optimal matching with variable controls using the assignment algorithm. J Computat Graphic Stat, 2001, 3:455-463.

[9] Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? J Economet, 2005, 125: 305-353.

[10] Austin PC, Mamdani1 MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med, 2006, 25:2084-2106.

[11] Katherine HH, Thomas AL. Propensity score modeling strategies for the causal analysis of observational data. Biostatistics, 2002, 2:179-193.

[12] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Asso, 1984, 79:516-524.

[13] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med, 2004, 23:2937-2960.

[14] Perkins SM, Tu W, Underhill MG, et al. The use of propensity scores in pharmacoepidemiologic research. Pharmacoepidemiol Drug Safety, 2000, 9:93-101.

（收稿日期：2009-12-07）
（本文编辑：张林东）