文章编号：2095-9982(2017)02-0184-04

【EHP 专栏】

使用线性回归统计方法分析暴露组与健康相关性的系统比较(待续)

摘要：

[背景] 暴露组构成了一个有前景的框架，通过明确考虑多重检验、避免选择性报告来提升关于环境暴露对健康影响的理解，但暴露组研究在同时考虑多种相关暴露方面受到了挑战。

[目的] 比较线性回归法在分析暴露组与健康相关性中的性能。

[方法] 在模拟研究中，设定237个暴露协变量，它们具有真实的相关结构并且对其中0~25个协变量呈线性相关的健康结局。主要比较统计方法的假阳性率(false discovery proportion，FDP)和灵敏度。

[结果] 在所有模拟设定中，弹性网络和稀疏偏最小二乘回归法的灵敏度为76%，FDP为44%；图形单元进化随机搜索(Graphical Unit Evolutionary Stochastic Search，GUESS)和删除/替换/添加(deletion/substitution/addition，DSA)算法的灵敏度为81%，FDP为34%。全环境关联分析(environment-wide association study，EWAS)尽管灵敏度更高，但平均FDP达到86%，但性能弱于前两者。当评估协变量间高度相关的暴露组暴露矩阵时，其性能明显下降。

[结论] 暴露之间的相关性是暴露组研究的一项挑战。在真实的暴露组情况下，本研究所考察的统计学方法有效区分真实预测变量与相关协变量的能力有限。尽管GUESS和DSA在灵敏度和FDP之间平衡方面稍好，但他们在所有考察的情景及属性上并未比其他多因素统计方法效力更好，在选择这些方法时也应考虑计算的复杂性和灵活性。

1　前言

环境因素包括广泛的物理、化学、生物和社会应激源。在双生子和移民研究中，环境可以解释大部分慢性病或连续健康效应的风险变化(Rappaport等，2014；Willett，2002)。目前为止，环境流行病学研究通常使用分别考虑每种环境暴露的方法评估环境暴露与健康之间的关联，对于环境与健康相关性只能展现零星的表象(Buck Louis 等，2013；Rappaport 2011；Vrijheid等，2014；Greenland，1994；Lenters等，2015为例外)。这些方法的结果可能由于(被忽略的)联合暴露、选择性报告或发表偏移而存在混杂(Patel和Ioannidis，2014；Slama和Vrijheid，2015)。暴露组的概念最初由Wild(2005)提出，包含自产前期起的所有环境暴露，并主张对所有暴露同时进行整体考虑(Wild，2012)。

大多数暴露组与健康关系的研究依赖于全环境关联分析(EWAS，单一暴露因素与分别估计的结局之间的相关性)(Patel等，2010)，有时增加一个包含所选自变量的多因素回归步骤(Patel等，2013)。已有多种基于多因素回归的统计学方法，能够解释多种暴露对健康的潜在联合作用(Chadeau-Hyam等，2013)。例如稀疏偏最小二乘法(sparse partial least squares sPLS；Chun和Keles，2010)最近被用于男性生育力的研究中(Lenters等，2015)，弹性网络(elastic net，ENET；Zou和Hastie，2005)被用于研究多种环境污染物与出生体重之间的关联(Lenters等，2016)。就我们所知，暴露组研究尚未应用其他多因素回归统计方法。

上述统计方法在暴露组框架下的效力仍有待系统地评估。最近的一项模拟研究(Sun等，2013)以少量暴露($n \leqslant 20$)对多项多因素回归方法进行研究，这些暴露最多只具有中等程度的相关性(Pearson相关性<0.57)。但(未来的)暴露组研究可能考虑更多的协变量，并且在大型的暴露组数据库，例如NHANES(Patel 等，2010，2013；Patel 和 Ioannidis，2014)中经常观察到更强的相关性(通常>0.6)。因此，本研究把Sun等的工作扩展到真实的暴露组情况中，旨在为今后的暴露组研究比较不同线性回归方法的性能。

本研究使用大量暴露协变量(237个)之间的实证相关结构生成暴露数据，并假定其中0~25个暴露线性地影响连续健康结局而没有效应修饰(相互作用)。

本文作者、作者单位及通信作者详见原文。

参与比较的统计方法是：a)EWAS方法；b)EWAS后跟一个含有已识别自变量的多因素回归步骤；c)ENET，一种惩罚回归法；d)sPLS回归，一种监督降维方法；e)图形单元进化随机搜索(Graphical Unit Evolutionary Stochastic Search，GUESS)算法，一种计算优化的贝叶斯变量选择方法(Bottolo等；2013)；f)删除/取代/添加(deletion/substitution/addition，DSA)顺序算法(Sinisi和van der Laan，2004)。基于6项既定准则和2项修正准则对所选方法的统计学效力进行系统比较，评估其变量选择和点估计能力。本研究还调查这些方法对生成暴露数据的实证相关结构修饰作用的敏感性。

2 方法

模拟模型的依据是一个有关虚拟人群暴露变量X的矩阵。在该矩阵中依据线性回归模型生成健康结局Y；根据真实预测变量的数量定义7种情景。使用一组预先选择的统计方法估计每个模拟X和Y之间的关联，在各情景下对方法的统计效力进行评估，并使用下文中的指标进行比较。在每种情景下模拟100个独立的数据集。

2.1 暴露组的生成

为了生成具有真实相关结构的暴露变量，本研究依据现有的INMA(Infancia y Medio Ambiente)母婴队列(Guxens等，2012)，通过问卷调查、地理空间建模和生物监测方式，共评估了母亲孕期237个环境因素。从所有成对相关的矩阵中，计算最接近的正定矩阵(Higham，2002)，将此估值作为基准相关矩阵Σ(见图S1)。使用Σ产生X，即1 200名真实研究对象的暴露组[该人群为正在进行的欧洲暴露组项目的研究人群，包含INMA队列(Vrijheid等，2014)]，X满足一个均值中心化的多因素常态分布：$X\sim N(0, \Sigma)$，其中N为多元高斯分布。该队列的数据包含5个二分类变量(其他是连续变量)，因此这些变量在模拟数据集也被分成两类，以重现在原始数据中观察到的阳性率。

2.2 健康结局的生成

根据以下公式生成一个关于暴露组的健康效应Y函数：

$$Y=\sum_{i=1}^{237}\beta_i X_i+\in,\ \in\sim N(0,\sigma) \qquad (1)$$

其中X_i是包含所有预测变量值i的向量，e代表回归模型的残差。回归系数β_i均设为0，除了k个随机选择且假定与健康效应有因果相关的变量(以下称为“真实预测变量”)，其$\beta_i=1$。研究考察了7个情景，每个情景都设定了不同数量的真实预测变量：k=0，1，2，3，5，10和25。残差方差σ^2被定义为：变量被真实预测变量(R^2)解释的比例等于$3\%\times k$。在此条件下，一个既定情景中所有模拟的信噪比是相同的，并且从不相关的真实预测变量的未校正分析中选择一个真实预测变量的能力在所有情景中是恒定的(参见补充材料S1)。

共生成七个情景组。情景组1满足上述定义的条件。情景组2和3旨在评估相关性水平在真实预测变量之间的影响，这可能引起识别性问题。与情景组1不同，情景组2确保所有真实预测变量之间的相关性绝对值<0.2，情景组3为>0.5。情景组4和5旨在评估整个暴露组的相关结构对统计方法效力的影响；不同于情景组1根据Σ生成X，情景组4使用除对角线外Σ系数除以2而得到的相关矩阵Σ^-，情景组5通过将Σ系数乘以2、系数上限设定为1且计算最接近的半定矩阵而得到Σ^+。情景组6生成与情景组1类似的条件，不同之处在于其暴露数据是从INMA队列中通过自举法获得的真实环境数据，以考察自正态分布暴露的假设偏离(即包含可能的偏态分布和异常值)。情景组7考察方法在不均等效应下的稳健性，生成与情景组1类似的条件，不同之处在于根据[0.5，1.5]的均匀分布绘制真实预测变量的效应大小(即β_i)。

所有情景下的健康结局均如上所述生成。当真实预测变量的数量一定时，可被真实预测变量解释的方差比例在全部七组情景中均相同。

2.3 评估暴露组与健康相关性的统计学方法

使用6种基于线性回归的统计学方法评估Y和X间的相关性。

2.3.1 全环境关联分析

EWAS(Patel等，2010)是依赖于对每个协变量独立拟合的线性回归模型，对多重比较进行校正后，使用双侧p值评估预测变量和效应之间的统计学关联。以广泛使用的Benjamini和Yekutieli(2001)校正作为基准，控制假阳性率(FDR)在所需水平(本文为0.05)。此外，EWAS中有意义的协变量被纳入一个多因素线性回归模型，如果其双侧p值小于0.05则被保留(Tzoulaki等，2012)。这个两步法被称为EWAS-多元线性回归(EWAS-MLR)。

作为灵敏度分析，本研究测试了若干方法以校正多重假设检验：基于置换的方法(Patel等，2010)、Benjamini和Hochberg(1995)法、Bonferroni(1936)校

正。本研究还在不应用多重比较校正的情况下测试了EWAS方法，以阐述对每个暴露协变量进行独立研究将得到什么结果。

2.3.2　弹性网络　ENET(Zou和Hastie, 2005)是一种依赖于广义线性结构的惩罚回归模型，使用最小绝对值收缩和选择算子(least absolute shrinkage and selection operator, LASSO)(Tibshirani, 1996)和岭(Hoerl, Kennard, 1970)惩罚的加权混合。LASSO惩罚可提高稀疏性，并通过收缩实施变量选择：由对应于最少信息预测变量的最低回归系数得到一个零值。岭惩罚调节相关变量并确保数值稳定。调整参数的标化、整体惩罚和两种惩罚的混合比例均由最小化的预测变量均方根误差(RMSE)决定，使用10倍交叉验证(即数据被分成10个子集；每个子集的数据都在其他9个子集上进行训练并在给定的排除子集中拟合，以估计RMSE)。为了防止过度拟合，最优标化参数被定义为在最小RMSE的1个标准误范围内生成RMSE并在此条件下提供最大稀疏模型(以非零回归系数的数值衡量)的参数(Meinshausen和Bühlmann, 2006)。

2.3.3　稀疏偏最小二乘回归　偏最小二乘回归是一种监督降维技术，其将汇总变量建立为一组原始变量的线性组合。为了确保生成的低维数据与研究结局相关，变量被反复定义使其尽可能多地解释预测变量和健康结局之间的剩余协方差。sPLS方法建立原始预测变量的稀疏线性组合，同时具有良好的预测性能和恰当的变量选择(Chun和Keleş, 2010)。在线性组合系数的估计时纳入惩罚(η)来诱导稀疏性，即绝对值低于最大绝对系数的某个分数η的所有系数收缩为零，该过程称为软阈值法(Lenters等, 2015)。只有第一个K作为协变量纳入线性回归模型中。K和η值经过5倍交叉验证(默认实现)的最小化RMSE标化。为完成模型比较，本研究类推参考实例使其能够包括空模型(K=0)。

2.3.4　图形单元进化随机搜索　作为贝叶斯变量选择方法的一部分，GUESS寻求预测健康结局的最优模型，每个模型由独立的协变量组合定义(Bottolo和Richardson, 2010)。要求使用进化蒙特卡洛算法，在$2p$(p指协变量的总数)个可能的协变量组合中识别最相关模型，并结合调节的多因素链与遗传算法共同运行。这种蒙特卡洛算法同时确保了采样器的混合优化及跨链的信息交换(Bottolo等, 2013)。

对每个模拟数据集运行GUESS算法20 000次迭代，丢弃前5 000个测试迭代。设置链数量为3。为了易于收敛并防止大量参数标化，注意E(先验预期模型大小)和ρ(其方差)。当$k<5$，设置E=3和ρ=3；当$k\geqslant 5$，$E=k+2$和ρ=5。在访问的模型中，为保守估计，保留后验概率>0.01的模型。

从保留模型包含的暴露集合全体中，选择无效假设(即没有协变量与结局相关)条件下边际后验纳入概率(marginal posterior probability of inclusion, MPPI；变量被纳入任意保留模型中的概率)大于MPPI分布(1–0.05/237)分位数的暴露。

GUESS的最初目标是寻找协变量的最佳组合来预测结局，其最新应用(Liquet等, 2016)允许在给定模型中进行系数估计值的后验模拟。然而，在本研究模拟条件下，真实预测变量在各数据集之间均有不同，这种间接(即在变量选择的条件下)的估计过程需要在所有模型上进行整体后验，这意味着非常大的计算量，因而不适用于直接的系数估计。作为一个保守的替代方案，本研究应用一个由GUESS选择变量的岭回归附加步骤来估计方法的系数，但这一步骤可能降低估计数据的质量。

2.3.5　删除–替换–加法算法　DSA是一种迭代线性回归模型检索算法(Sinisi和van der Laan, 2004)。可能的模型集的建立，受到三种使用者指定的规定的约束：预测变量间交互的最大顺序、给定预测变量的最大效力和最大模型大小。每次迭代遵循以下三个步骤：a)移除项，b)替换为另一项，c)向当前模型添加项。寻找最佳模型从截距模型开始，并为每个模型大小确定最佳模型。通过使用5倍交叉验证的数据最小化RMSE值来选择最终模型。不允许多项式或交互项，并且模型考虑的协变量多达40个(然而，模拟中并未达到这个数量)。

使用统计方法的R工具，分别在stats、glmnet、spls、R2GUESS和DSA软件包中可获得(R Project for Statistical Computing 3.1.1版本；DSA 2.15.3版本)。本文作者开发的R代码和相关矩阵Σ可见于补充材料S2和Excel表S1。

2.4　统计性能评估

使用衡量变量选择相关性和点估计值质量的关键标准来评估各统计方法的效力。

计算每个情景和模拟条件下方法的灵敏度，作为给定方法实际选择的真实预测变量的比例。以同样的方法计算特异性，作为未被选择的不相关暴露

的比例。

假阳性率(false discovery proportion, FDP)定义为与结局没有真正相关性的选择变量的比例。当在给定的运行中没有变量被选择时,就认为没有错误地选择变量,FDP为0%。具有0个真实预测变量的场景不计算FDP和灵敏度。

根据237个系数的平均绝对偏差计算所估计系数的准确性:

$$\frac{1}{237}\sum_{i=1}^{237}|\beta_i-\hat{\beta}_i| \quad (2)$$

其中β_i表示模拟中使用的系数,$\hat{\beta}_i$表示相应的估计值。对真实预测变量和无关暴露(即非真实预测变量)也分别计算平均绝对偏差。

由于暴露之间可能存在强相关性,有观点认为,选择另一个高度相关的变量而非真实的预测变量不应被视为完全错误的选择,因为统计学方法并未遗漏全部信息。为证实这一观点,本研究根据真实预测变量和统计学方法选择的协变量之间的最大绝对相关性,设定了灵敏度和FDP的替代衡量指标:

$$AltSens=\frac{1}{k}\sum_{i\in A}\max_{j\in B}\{|\widehat{corr}(X_i,X_j)|\}$$
$$AltFDP=1-\frac{1}{n_B}\sum_{j\in B}\max_{i\in A}\{|\widehat{corr}(X_i,X_j)|\} \quad (3)$$

其中A是真实预测变量的集合,B是由方法选择的变量的集合(也称为自变量),k和n_B是它们各自的大小。*AltSens*衡量一个真实预测变量和方法选择的任一变量之间的平均最大绝对相关性值,*AltFDP*表示一个被选择变量和任一真实预测变量之间的平均最大绝对相关性值。*AltFDP*等于1减去一个被选择变量和任一真实预测变量之间的平均最大绝对相关性值。如果所选协变量的集合包括所有真实预测变量,则这些替代指标与经典的灵敏度和FDP一致。由于$|\widehat{corr}(X_i,X_j)|\leqslant 1$,*AltSens*总是大于灵敏度,而*AltFDP*总是小于FDP。

2.5 变量选择的扩展方案

有观点认为,为了增加灵敏度并避免丢失重要数据,不应只关注所选暴露,而是所有与真实预测变量高度相关的暴露(即水平>α,其中α为0.6~0.9)。根据这种方法计算灵敏度和FDP。

3 结果

3.1 生成暴露的相关结构

Σ矩阵被定义为最接近INMA相关结构的正定矩阵,仅与其略有不同:75%的绝对差异<0.01,而95%的绝对差异<0.05。Σ矩阵中暴露之间的绝大部分绝对相关性(83%)<0.2,但78%的暴露与至少一种其他暴露有相关性(水平>0.6)(图S1)。

(待续)

翻译:窦冠珅;审校:金泰廙

参考文献(略)

本文原文刊登于*EHP*杂志,需要者务必引用英文原文,详见 Agier L, Portengen L, Chadeau-Hyam M, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. Environ Health Perspect 124(12): 1848-1856.

本文原文及参考文献请浏览http://dx.doi. org/10.1289/EHP172

(编辑:汪源;校对:陶黎纳)