基于 Group Bridge 组变量选择方法的血压影响因素实证分析

李中秋, 张汝飞, 鲁亚军

(中国人民大学 统计学院, 北京 100872)

摘 要: 高血压是常见的心血管疾病, 针对引起血压显著变化的影响因素开展深入研究对预防高血压及其并发症均具有重要意义. 为此, 根据 11624 个样本数据, 选用 Group Bridge 方法对血压及其年龄、文化程度等 8 组共 35 个影响因素进行拟合分析. 结果显示: Group Bridge 方法能够提供科学有效的稀疏拟合结果; 拟合结果共选定 6 组中的 18 个影响因素, 其中存在正向影响关系的因素 15 个, 负向影响关系的因素 3 个; 综合考虑影响因素数量和强度, 发现体格方面的影响因素对血压的影响最为重要, 其次是体脂指标及生活方式和行为方面, 再次为疾病家族史、年龄及文化程度, 最后是婚姻状况及收入水平.

关键词: 血压; Group Bridge; 影响因素

1 引言

据 2002 年全国居民营养与健康状况调查资料显示, 我国有高血压患者约 1.6 亿人, 其中成人高血压患病率更是高达 18.8%, 高血压已经成为严重危害我国人民群众身体健康最主要的疾病之一, 如何控制高血压、治疗高血压、防御高血压显得尤为迫切, 为此, 我们应当首先探明引起血压变动的影响因素, 在此方面很多专家学者已经取得了丰硕的研究成果. 如胡世红等人 (2002) 在分析不同腰围个体血压值后, 发现随着腰围的增加, 收缩压 (SBP) 和舒张压 (DBP) 也会显著增加 (P<0.001) [1]. 韩春姬等人 (2005) 通过研究延吉市朝鲜族和汉族老年人内脏脂肪指数与血压的关系后, 指出无论是朝鲜族还是汉族的老年人, 其血压值和内脏脂肪均呈正相关关系 [2]. Puddey 等人 (2006) 对饮酒与血压的关系进行研究后发现, 经常饮酒会导致血压上升, 并建议男性高血压患者每日酒精摄入量不超过 20 克, 女性不超过 10 克 [3]. Snodgrass 等人 (2008) 采用最小二乘估计方法分析得出西伯利亚人群的基础代谢率与血压之间存在正向关系 [4]. 李新立等人 (2003) 通过研究不同年龄人群血压和动脉顺应性关系, 证实了随着年龄增加, 舒张压 (DBP) 也会增加 [5]. 郝唯蔚等人 (2009) 通过研究中老年人群不同血压水平与血肌酐的相关性, 认为与正常血压人群相比, 高血压患者的血肌酐水平有明显升高 ($P < 0.05$). [6]Munger 等人 (1988) 对比有高血压家族史和无家族史孩子的血压后, 指出母亲与子女血压之间的相关性通常比父亲与子女血压之间的相关性大 [7]. 李东光等人 (1995) 采用病例对照分析方法发现教育水平越高患有高血压的风险越低 [8].

这些研究成果对于推动血压影响因素的研究均具有重要价值, 但细心观察不难发现普遍

收稿日期: 2015-05-30

存在两点不足：第一, 少有系统全面的分析. 上述研究往往均是就某一种影响因素与血压的关系开展分析研究, 鲜有组织多个影响因素与血压开展系统分析. 第二, 研究方法单一. 以往对血压影响因素的研究方法通常局仅限于传统最小二乘估计方法 (OLS).

针对上述两方面的不足, 本文尝试采用 Group Bridge 方法对血压及年龄、文化程度、婚姻状况、收入水平、生活方式和行为、疾病家族史、体格、体脂指标 8 组共 35 个影响因素进行拟合, 以期从系统角度探寻血压的真实影响因素及其作用效果.

后续结构为：第二部分, Group Bridge 方法介绍；第三部分, 影响因素实证分析；第四部分, 全文结论.

2 Group Bridge

为了解决高维数据变量选择的问题, 1996 年 Tibshirani 提出了一种以系数压缩为特征的变量选择方法, 称为 Least Absolute Shrinkage and Selection Operator, 即 LASSO [9]. 其核心思想是用模型的绝对系数函数作为调谐项来压缩模型系数, 使得绝对值较小的系数直接压缩为 0, 从而实现变量降维、变量选择、参数估计和提供稀疏解的目的.

假设 $\mathbf{X}_k = (X_{1k}, \cdots, X_{nk})'$, $k = 1, \cdots, d$ 是设计矩阵, $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ 是响应变量, 那么两者线性关系可以用下式表示：

$$\mathbf{Y} = \mathbf{X_1}\beta_1 + \cdots + \mathbf{X_d}\beta_d + \varepsilon \tag{1}$$

其中, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)'$ 为残差项. 对于列向量 $\mathbf{v} \in \Re^d$, $d \geq 1$, 假设 2- 范数 L_2 的形式定义为 $\|\mathbf{v}\|_2 = (\mathbf{v}'\mathbf{v})^{1/2}$, 则平方损失函数 $R(\beta)$ 形式如下：

$$R(\beta) = \left\| \mathbf{Y} - \sum_{k=1}^{d} \mathbf{X}_k \beta_k \right\|_2^2 \tag{2}$$

LASSO 估计为：

$$\hat{\beta}^{Lasso} = \arg\min \left\| \mathbf{Y} - \sum_{k=1}^{d} \mathbf{X}_k \beta_k \right\|_2^2, s.t. \sum_{k=1}^{d} |\beta_k| \leq \lambda \tag{3}$$

其中, λ 为调谐参数且 $\lambda \geq 0$.

与传统变量选择方法相比, LASSO 的优点在于把小的系数往 0 方向压缩, 一旦某个系数被压缩到 0, 对应的变量就被删除. 随后, 统计学家陆续提出许多类似 LASSO 的模型. Fan 和 Li(2001) 提出了绝对偏差的光滑剪切方法 (Smooth Clipped Absolute Deviation, 简称 SCAD) 模型, 认为基于调谐函数得到的估计应该具有 3 个性质, 即：无偏性 (Unbiasedness)、稀疏性 (Sparsity) 以及连续性 (Continuity) [10]. 采用 SCAD 方法得到的估计是无偏估计, 这点优于 LASSO 方法. Zou 和 Hasie 针对 LASSO 方法变量间不能有效处理高维问题以及当变量间存在相关性时, 参数估计有偏等缺点, 提出 Elastic Net 方法 [11]. 上述方法均是基于变量间选择, 而当研究对象成组时, 即组内变量之间可能存在相关性, 组间变量之间是相互独立的, 运用上述变量选择方法便会得到错误的估计. 为此, Yuan 和 Lin(2006) 提出了 Group LASSO 方法 [12]. 而 Group LASSO 的缺点在于只能对组水平进行选择, 不能对组内重要变量进行选择. 换言之, 一旦组未被选中, 那么组内所有变量系数均为 0. 为了同时既考虑组的选择又考虑组

内重要变量的选择, Huang 等人 (2009) 提出了 bi-level 组变量选择方法 —Group Bridge 模型[13].

Group Bridge 的具体模型为:

令 $A_1,\cdots,A_J$ 是 $\{1,\cdots,d\}$ 的子集, 对于 $m\times 1$ 维向量 $\mathbf{a}$, 1- 范式 L_1 的形式定义为 $\|\mathbf{a}\|_1=|a_1|+\cdots+|a_m|$, 那么 Group Bridge 组变量选择的系数估计可以根据计算下式得到:

$$\hat{\beta}_{A_j}=\arg\min\{\left\|\mathbf{Y}-\sum_{k=1}^{d}\mathbf{X}_k\beta_k\right\|_2^2\}+\lambda\sum_{j=1}^{J}c_j\left\|\beta_{A_j}\right\|_1^{\gamma} \tag{4}$$

其中, λ 为调谐参数, c_j 是常数, 一般假定 $c_j\propto|A_j|^{1-\gamma}$, $\gamma\in(0,1)$.

式 (4) 不能直接求导, 因为 $\gamma\in(0,1)$ 时调谐函数不再是凸函数, 所以模型的参数估计只能寻找特殊算法进行迭代求解. Huang 等人 (2009) 对这一问题进行了详细探讨, 认为如果 $\lambda=\tau^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, 则式 (4) 可等价转化为

$$\hat{\beta}_{A_j},\hat{\theta}=\arg\min\{\left\|\mathbf{Y}-\sum_{k=1}^{d}\mathbf{X}_k\beta_k\right\|_2^2\}+\sum_{j=1}^{J}\theta_j^{1-1/\gamma}c_j^{1/\gamma}\left\|\beta_{A_j}\right\|_1+\tau\sum_{j=1}^{J}\theta_j \tag{5}$$

其中, τ 是调谐函数, $\theta\geq 0$. 由 (5) 式可知, 调谐函数可以看作是自适应 L_1 调谐, 它使得系数 β 的估计值变得稀疏的, 而如果 θ_j 很小, 那么会导致 $\beta_{A_j}=0$, 这样也完成了对组的选择.

Huang 等人 (2009) 进一步给出了迭代算法:

第 1 步: 获得 β 的初始估计 $\beta^{(0)}$. 一般选用最小二乘估计值作为 β 的初始值;

第 2 步: 计算 $\theta_j^{(s)}=c_j(\frac{1-\gamma}{\tau\gamma})^{\gamma}\left\|\beta_{A_j}^{(s-1)}\right\|_1^{\gamma}$, 其中 $s=1,2,\cdots$;

第 2 步: 计算 $\hat{\beta}_{A_j},\hat{\theta}=\arg\min\{\left\|\mathbf{Y}-\sum_{k=1}^{d}\mathbf{X}_k\beta_k\right\|_2^2\}+\sum_{j=1}^{J}\theta_j^{1-1/\gamma}c_j^{1/\gamma}\left\|\beta_{A_j}\right\|_1$;

第 4 步: 重复迭代第二、三步直至收敛.

在变量选择中, 调谐参数 λ 的选择是一个重要问题. λ 值越小, 模型调谐力度就越小, 模型中变量数目就越多; λ 值越大, 模型调谐力度就越大, 导致系数收缩量就越大, 选出的变量数目就越少. 在 Group Bridge 方法中, λ 的选择方法一般有 3 种: AIC 准则、BIC 准则以及 GCV 准则, 3 种统计量具体形式如下所示:

$$AIC(\lambda)=\log(\left\|\mathbf{Y}-x\hat{\beta}_n(\lambda)\right\|_2^2/n)+2d(\lambda)/n$$

$$BIC(\lambda)=\log(\left\|\mathbf{Y}-x\hat{\beta}_n(\lambda)\right\|_2^2/n)+\log(n)d(\lambda)/n$$

$$GCV(\lambda)=\frac{\left\|\mathbf{Y}-x\hat{\beta}_n(\lambda)\right\|_2^2}{n(1-d(\lambda)/n)^2}$$

其中, $x=(\mathbf{x}_1,\cdots,\mathbf{x}_d)$ 是 $n\times d$ 维协变量矩阵, $\hat{\beta}_n(\lambda)$ 是 β 的估计量, $d(\lambda)=trace(x_\lambda[x_\lambda'x_\lambda+0.5W_\lambda]^{-1}x_\lambda')$, x_λ 是给定 λ 条件下系数 $\hat{\beta}_n(\lambda)$ 为非 0 协变量, W_λ 为一对角矩阵.

选取 λ 的一般做法是找一些格子点, 分别计算这些点上的 AIC、BIC、GCV 值, 求出达到最小值时, 所对应的 λ 值即为模型所需调谐参数值.

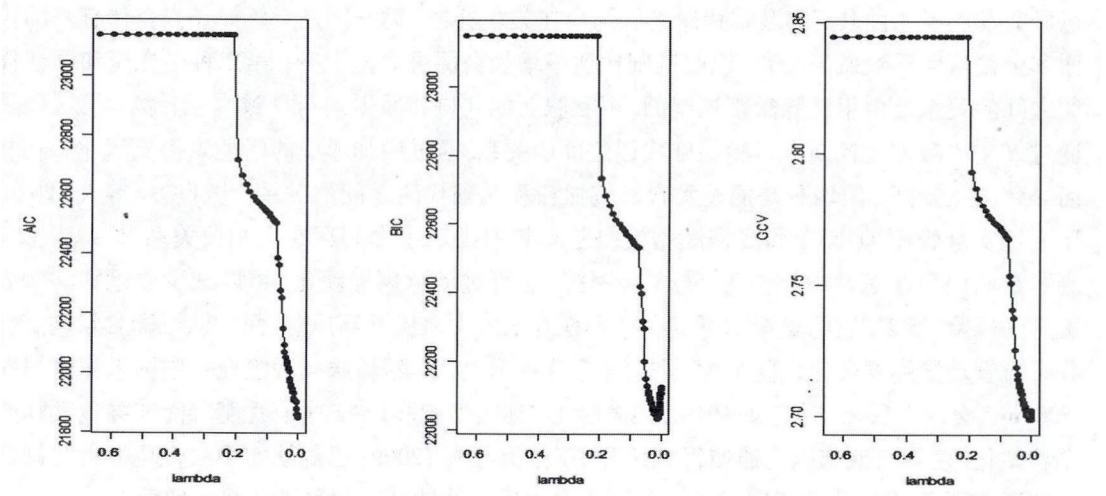
3　实证分析

1) 数据来源

本文选取年龄、文化程度、婚姻状况、收入水平、生活方式和行为、疾病家族史、体格以及体脂指标 8 组共 35 个变量作为影响血压的影响因素. 数据来源于中国医学科学院、阜外心血管病医院以及卫生部心血管病防治研究中心的心血管病及危险因素调查数据. 母体数据中存在多项缺失值, 经剔除后剩余样本量为 11624, 对样本进行标准化处理, 按 70/30 的比例把样本划分为训练集和测试集, 训练集用于模型参数估计, 测试集用于模型预测效果评价. 其中, 舒张压 (DBP) 为血压值, x1~x2 为年龄组变量, x3 为文化程度组变量, x4 为婚姻状况组变量, x5 为收入水平组变量, x6~x10 为生活方式和行为组变量, x11~ x20 为疾病家族史组变量, x21~ x25 为体格组变量, x26~ x35 为体脂指标组变量. 具体情况如表 2 所示.

2)λ 的确定

图 1 为调谐参数 λ 在 AIC、BIC 以及 GCV 三种统计量下的选择轨迹图, 可以看到基于 BIC 准则选出的 λ 值较 AIC 准则和 GCV 准则选出的 λ 值更远离 0, 为 0.012, 这说明使用 BIC 统计量选定的 λ 进行 Group Bridge 估计时, 所得模型结果最为稀疏, 这最符合我们进行变量降维、变量选择的初衷. AIC、BIC 以及 GCV 三种统计量下确定的 λ 值及 Group Bridge 估计结果详情如表 1 所示.



(a)AIC 统计量　　(b)BIC 统计量　　(c)GCV 统计量

图 1　Group Bridge 模型调谐参数 λ 的选择轨迹图

表 1 为所构造的 200 套验证集的预测效果对比表. 可以看到, 采用基于 BIC 准则选出的调谐参数 λ 来拟合 Group Bridge 模型所得结果最为稀疏, 且根均方误差 (RMSE) 值最小, 为 0.99, 这与 Huang(2009) 等人的模拟结果不谋而合.

综上所述, 本文最终选定 $\lambda = 0.012$.

3) 结果分析

Group Bridge 模型拟合结果如表 2 所示, 可以看到, 最终选定 6 组中的 18 个影响因素,

占全部 8 组 35 个影响因素的 51.43%. 其中, 年龄组 1 个影响因素, 文化程度组 1 个影响因素, 生活方式和行为组 3 个影响因素, 疾病家族史组 2 个影响因素, 体格组 5 个影响因素, 体脂指标组 6 个影响因素. 从影响因素数量来看, 体格和体质指标两组包含的数量最多.

表 1 验证集预测效果比较

	AIC	BIC	GCV
组系数为 0 平均数目	3.175	5.945	4.055
变量系数为 0 平均数目	4.395	14.565	6.575
RMSE 均值	1.029	0.99	1.027

从与血压影响强度来看, 影响强度较大 (系数大于 0.1) 的因素有: 基础代谢 (0.25)、甘油三酯 (0.2)、饮食 (0.16)、腰围 (0.14)、饮酒 (0.13) 以及身体脂肪率 (0.10)6 项, 其中, 2 个影响因素属于生活方式和行为组, 3 个影响因素属于体格组, 1 个影响因素属于体脂指标组.

就影响因素与血压的作用方向来看, 存在正向影响关系的因素有 15 个, 分别为年龄、饮酒、饮食、母亲有高血压、母亲有高血脂、体重、腰围、基础代谢、身体脂肪率、内脏脂肪指数、甘油三酯、APOB、血肌酐、超敏 C 反应蛋白和微量白蛋白; 存在负向影响关系的因素有 3 个, 分别为文化程度、高密度脂蛋白胆固醇 (HDLC) 和吸烟.

查阅文献发现, 许多专业学者也得出了与本文类似的结果. 如 Snodgrass 等人 (2008) 通过研究发现了 4 种血压和基础代谢之间的内在关联机制: 第一种是交感神经兴奋伴有基础代谢率上调从而导致血压上升, 说明基础代谢率增加会促进血压上升; 第二种是血压和基础代谢之间的关系是由甲状腺激素控制的, 甲状腺功能亢进和减退均可导致血压升高; 第三种是通过氧化应激方式控制血压和基础代谢之间的关系, 通过增加形成的活性氧提高氧化损伤从而导致血压升高; 第四种是通过发育影响血压和基础代谢之间的关系 [4]. ODA 等人 (2010) 研究证实舒张压 (DBP) 和高密度脂蛋白胆固醇 (HDLC) 之间存在负相关关系 [14]. 高密度脂蛋白胆固醇在医学中被称为 "好" 胆固醇, 它可抗动脉粥样硬化, 可以减少患冠状动脉心脏病的风险. 所以, 它的量增加可以在一定程度上减少血压升高风险. 而甘油三酯含量增加会导致血管动脉粥样硬化风险加大, 进而导致血压升高. 国际糖尿病联盟规定中国人腰围男性 ≥90cm、女性 ≥80cm 时定义为中心性肥胖 (或称腹型肥胖)[15]. 中心性肥胖能够导致动脉变厚和硬化, 使得血液难以流通使得血压升高, 张梅等人 (2009) 已经证实中心性肥胖者患高血压的风险更高 [16]. 这些研究成果均验证了本文模型估计结果的科学性及合理性.

4 简要结论

本文根据 11624 个样本数据, 选用 Group Bridge 方法对血压及其年龄、文化程度、婚姻状况、收入水平、生活方式和行为、疾病家族史、体格、体脂指标 8 组共 35 个影响因素进行拟合. 结果显示:

1)Group Bridge 方法能够提供科学有效的稀疏拟合结果.

2) 最终选定 6 组中的 18 个影响因素, 其中存在正向影响关系的因素 15 个, 负向影响关系的因素 3 个.

表 2 Group Bridge 模型拟合结果

组别	变量	AIC	BIC	GCV
		$\lambda = 0.004$	$\lambda = 0.012$	$\lambda = 0.002$
年龄	(x1) 年龄	0.81	0.09	0.65
	(x2) 年龄 ^{2}	-0.73	0	-0.56
文化程度	(x3) 文化程度	-0.103	-0.09	-0.11
婚姻状况	(x4) 婚姻状况	0.02	0	0
收入水平	(x5) 收入水平	0	0	0
生活方式和行为	(x6) 吸烟	-0.13	-0.04	-0.12
	(x7) 饮酒	0.09	0.13	0.09
	(x8) 饮茶	0.01	0	0.003
	(x9) 饮食	0.16	0.16	0.17
	(x10) 劳动强度	0	0	0
疾病家族史	(x11) 高血压家族史 (父)	0.007	0	0
	(x12) 高血压家族史 (母)	0.009	0.01	0.03
	(x13) 高血脂家族史 (父)	0	0.	0
	(x14) 高血脂家族史 (母)	0.013	0.01	0.009
	(x15) 糖尿病家族史 (父)	0.01	0	0.01
	(x16) 糖尿病家族史 (母)	0	0	0
	(x17) 冠心病家族史 (父)	-0.0077	0	0
	(x18) 冠心病家族史 (母)	-0.002	0	0
	(x19) 脑卒中家族史 (父)	-0.009	0	0
	(x20) 脑卒中家族史 (母)	-0.002	0	-0.001
体格	(x21) 体重	0.01	0.01	0.01
	(x22) 腰围	0.19	0.14	0.17
	(x23) 基础代谢	0.25	0.25	0.25
	(x24) 身体脂肪率	0.11	0.10	0.11
	(x25) 内脏脂肪指数	0.03	0.05	0
体脂指标	(x26) 甘油三酯	0.88	0.2	0.77
	(x27) 高密度脂蛋白胆固醇	-0.84	-0.05	-0.53
	(x28)APOB	0.13	0.05	0.13
	(x29)LPA	-0.014	0	-0.016
	(x30) 葡萄糖	0.2	0	0
	(x31)UA	0.03	0	0.03
	(x32) 血肌酐	0.02	0.06	0.02
	(x33) 超敏 C 反应蛋白	0.35	0.05	0.4
	(x35) 微量白蛋白	0.08	0.08	0.08

3) 综合考虑影响因素数量和强度, 可以发现, 体格方面的影响因素对血压的影响最为重要, 其次是体脂指标及生活方式和行为方面, 再次为疾病家族史、年龄及文化程度, 最后是婚姻状况及收入水平.

4) 查阅文献资料, 发现许多专家学者的研究成果均支持本文综合系统研究所得结果, 验证了本文的科学性及合理性.

参考文献

[1] 胡世红, 贾卫鸿, 韦春凌. 成年人腰围与血压, 血脂及血糖关系 [J]. 中国慢性病预防与控制, 2007, 15(5): 459-461.

[2] 韩春姬, 俞星, 申红梅等. 延吉市朝鲜族和汉族老年人血压与体内脂肪的关系及差异 [J]. 中国临床康复, 2005, 9(15): 20-22.

[3] Puddey I B, Beilin L J. Alcohol is bad for blood pressure[J]. Clinical and Experimental Pharmacology and Physiology, 2006, 33(9): 847-852.

[4] Snodgrass J J, Leonard W R, Sorensen M V, et al. The influence of basal metabolic rate on blood pressure among indigenous Siberians[J]. American journal of physical anthropology, 2008, 137(2): 145-155.

[5] 李新立, 倪春辉, 王震震. 不同年龄健康人血压和动脉顺应性关系的研究 [J]. 南京医科大学学报: 自然科学版, 2003, 23(3): 255-256.

[6] 郝唯蔚, 赵春华, 武权, 等. 不同血压水平与血肌酐相关性的研究 [J]. 中国临床保健杂志, 2009, 12(3): 238-239.

[7] Munger R G, Prineas R J, Gomez-Marin O. Persistent elevation of blood pressure among children with a family history of hypertension: the Minneapolis Children's Blood Pressure Study[J]. Journal of hypertension, 1988, 6(8): 647-653.

[8] 李东光, 张翠莉. 高血压危险因素的研究 [J]. 中国慢性病预防与控制, 1995, 3(4): 147-148.

[9] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

[10] Fan J, Li R.Variable selection via nonconcave penalized likelihood and its oracle properties[J].Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

[11] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[12] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67.

[13] Huang J, Ma S, Xie H, et al. A Group Bridge approach for variable selection[J]. Biometrika, 2009, 96(2): 339-355.

[14] Oda E, Kawai R. Low-density lipoprotein (LDL) cholesterol is cross-sectionally associated with preclinical chronic kidney disease (CKD) in Japanese men[J]. Internal medicine (Tokyo, Japan), 2009, 49(8): 713-719.

[15] 孔灵芝, 方圻, 王文, 等. 中国高血压防治指南 [J]. 2005.

[16] 张梅, 姜勇, 汪媛, 等. 中国成人腰围, 体质指数与高血压关系 [J]. 中国公共卫生, 2009, 25(6): 693-695.

An Empirical Analysis of Influencing Factors of Blood Pressure Based on Group Bridge

LI Zhong-qiu, ZHANG Ru-fei, LU Ya-jun

(Renmin University of China, School of Statistics, Beijing 100872, China)

Abstract: Hypertension is a common cardiovascular disease. In-depth research on factors causing significant changes in blood pressure for play an important significance in the prevention of hypertension and its complications. In this paper, based on 11,624 sample data, Group Bridge was used for analyzing blood pressure and eight groups and a total of 35 factors of blood pressure including: age, educational level, etc. The results showed: Group Bridge can provide effective sparse fitting results; fitting results of the six groups selected 18 factors, including 15 factors have positive influence to blood pressure and 3 factors which negative influence to blood pressure. Considering the number and intensity of factors, we can find physical aspects have most influence on blood pressure, the second body fat index and lifestyle and behavior, thirdly, again as a family history of the disease, age and education level, and the final was marital status and income level.

Keywords: blood pressure; group bridge; influencing factors