Hans Journal of Data Mining 数据挖掘, 2016, 6(3), 106-115
Published Online July 2016 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2016.63013

Hans汉斯

Analysis of Factors Influencing Hypertension in Different Ethnic Groups

Mengting Hu, Yu Fei*

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Yunnan Kunming
Email: 825634143@qq.com, *feiyukm@aliyun.com

Received: Aug. 1st, 2016; accepted: Aug. 21st, 2016; published: Aug. 24th, 2016

Abstract

Recently, it is more likely to have high blood pressure, and complications related to high blood pressure are dangerous. Complications of hypertension have gradually become one of the killers of modern health. In this paper, we use the United States health survey data in UCI database for analysis and processing. We dealt with each factor of people from different races using Logit Classification and the Random Forest Classification, and obtained the following conclusions: Regardless of race, age had significant effects on high blood pressure; For different ethnic groups, influence of other factors on hypertension is different.

Keywords

Hypertension, Logit Classification, Random Forest Classification

影响不同种族患高血压的因素分析

胡梦婷，费　宇*

云南财经大学统计与数学学院，云南　昆明
Email: 825634143@qq.com, *feiyukm@aliyun.com

收稿日期：2016年8月1日；录用日期：2016年8月21日；发布日期：2016年8月24日

*通讯作者。

文章引用: 胡梦婷, 费宇. 影响不同种族患高血压的因素分析[J]. 数据挖掘, 2016, 6(3): 106-115.
http://dx.doi.org/10.12677/hjdm.2016.63013

摘　要

由于现在人们得高血压的几率越来越大，而高血压所引起的并发症十分危险，高血压各种并发症病逐渐成为了现代健康杀手之一。本文选用UCI数据库中的美国健康普查数据进行分析处理。本文通过对不同种族的人的各因素用logit分类以及随机森林分类进行分析，得到了以下结论：不论种族，年龄对高血压都有显著影响；对于不同种族，其他的因素对高血压的影响程度不同。

关键词

高血压，Logit分类，随机森林分类

1. 引言

正常人的血压随内外环境变化在一定范围内波动。在整体人群，血压水平随年龄逐渐升高，以收缩压更为明显，但 50 岁后舒张压呈现下降趋势，脉压也随之加大。所谓高血压，就是以人体内血压异常增高为主要特征的一种心脑血管疾病。当一般人的舒张血压高于 90 mmHg，收缩血压高于 140 mmHg 时即可判定出现高血压症状。高血压已经成为全球十大危险因素之一，每年可导致七百多万人死亡，约占全世界死亡人数的 13% [1]。

国内外研究表明，高血压是一种多因子疾病，致病因素多，各因素间关系复杂，是高血压研究的重要特点[2]。我国高血压的现状十分不容乐观，2002 年全国高血压抽样调查结果显示我国 18 岁以上城市人群高血压的患病率为 19.3%。按照人口比例可得出，我国患高血压的人数已经过亿。2002 年中国居民营养与健康状况调查显示，按照经济发展水平，将城市分为大城市和中小城市。将农村分为一至四类农村，进一步分析发现大城市和中小城市高血压患病率分别为 20.4%和 18.8%，一至四类农村分别达到 21.0%、19.0%、20.2%和 12.6% [3]。我国高血压的特点是“三高三低”，患病率高、增长趋势高、危害性高，同时知晓率低(30.2%患者知道自己患有高血压)、治疗率低(24.7%高血压患者接受治疗)、控制率低(6.1%的高血压患者血压控制达标)。

高血压病的早期，仅有全身小动脉痉挛，而血壁没有明显器质性改变，因此及时治疗，高血压病完全可以治愈或被控制。若血压持续增高多年不降，动脉壁由于长期缺氧、营养不良，动脉内膜通透性增高，内膜及中层有血浆蛋白渗出，渗入壁管的血浆蛋白逐渐凝固发生透明样变，血管壁因透明变性而发生硬化。硬化的小动脉管壁日渐增厚而失去弹性，管腔逐渐狭窄甚至闭塞，从而导致血压特别是舒张压的持续性升高。最常见的六种严重危害的后果如下：冠心病、脑血管病、高血压心脏病、高血压脑病、慢性肾功能衰竭、高血压危现。

因此，控制和防治高血压以及高血压并发症刻不容缓。在医学上，认为高血压的病因有大概以下四种，分别是：习惯因素遗传(大约半数高血压患者有家族史)；环境因素；年龄(发病率有随着年龄增长而增高的趋势，40 岁以上发病率高)；其他(肥胖者发病率高；避孕药；睡眠呼吸暂停低通气综合症)。

目前国内外对高血压危险因素的研究已经很全面，不仅使用了传统的统计学方法，还使用了机器学习的方法，比如 Logistic 回归、分类树回归、BP 神经网络，而且都对它们的准确性进行比较，但是结果各不相同[4]-[6]。

2006 年傅传喜等分别利用 Logistic 回归和分类树分析对高血压危险因素进行分析得出高血压的主要危险因素为年龄、血脂以及肥胖，同时得到分类树分析较 Logisic 回归分析分类效果好[7]。

2010 年杨洋用 BP 人工神经网络对辽宁省彰武县农村人群进行患病预测，并与 Logistic 回归模型进行比较，利用 ROC 曲线(receiver operator characteristic curve)评价人工神经网络模型的预测性能[8]。

因此本文将通过对美国全国健康和营养调查的数据进行研究，利用 Logistic 分类和随机森林来探讨影响不同种族患高血压病的因素。

2. 数据来源及数据描述

本数据来源 http://www.umass.edu/statdata/statdata/stat-logistic.html。本数据共有 17,030 个观测对象，16 个变量，有 10,472 个缺失值。表 1 是变量清单。

其中，在变量清单中的前四个变量(受访者号码，伪初级抽样单位，伪阶层，统计权重)在本文的研究中是没有实际作用的，因此删去这四个变量。由于原始数据中给出的变量名不方便直接使用，因此我在表 1 的最后一列将所要用的变量名进行了重新命名。

Table 1. Variable list
表 1. 变量清单

变量名	变量名含义	取值/单位	新变量名
SEQN	受访者号码		
SDPPSU6	伪初级抽样单位	1, 2	
SDPSTRA6	伪阶层	01~49	
WTPFHX6	统计权重	225.93~139744.9	
HSAGEIR	年龄	岁	A
HSSEX	性别	0 表示女性	B(0)
		1 表示男性	B(1)
DMARACER	种族	1 表示白人	
		2 表示黑人	
		3 表示其他	
BMPWTLBS	体重	英镑	C
BMPHTIN	身高	英寸	D
PEPMNK1R	平均收缩血压	mmHg	
PEPMNK5R	平均舒张血压	mmHg	E
HAR1	受访者吸烟 > 100 支香烟	1 表示有	F(1)
		2 表示没有	F(2)
HAR3	受访者现在是否抽烟	1 表示抽烟	
		2 表示不抽烟	
SMOKE	现在受访者抽烟状况	1 = if HAR1 = 2	G(1)
		2 = if HAR1 = 1 & HAR3 = 2	G(2)
		3 = if HAR1 = 1 & HAR3 = 1	G(3)
TCP	血清胆固醇	mg/100ml	H
HBP	高血压	如果平均收缩血压 <= 140，记为 0	
		如果平均收缩血压 > 140，记为 1	

由于该数据变量 HAR3 的缺失值占了该变量的一半多，无法通过 R 语言中的 missForest 函数进行很好的弥补，因此本文舍去这个变量。同时，这个数据的种族包括白人、黑人以及其他，本文将根据这个将数据分为三个小数据(nhanes_1 表示白人，nchanes_2 表示黑人，nchanes_3 表示其他)。

本文选择 HBP 作为因变量，由于 HBP 是完全由平均收缩血压(PEPMNK1R)决定的，因此本文将删除 PEPMNK1R 这个变量。HBP 是一个定性变量，分为 0 和 1 这两个水平。由于自变量中也有定性变量，本文将考虑用 logistic 分类和随机森林分类来对数据进行拟合，从中选择出最好的模型，并进行预测。

3. 实证研究

3.1. 模型原理及形式

1) 建立二分类 Logit 模型

在研究本文采用对于一个有两个结果的随机试验，实验的两个可能结果分别是有高血压和没有高血压，这也就是最简单的概率模型就是伯努利实验，该实验假定成功的概率为 p 失败的概率为 $1-p$。二项分布就是由多次伯努利实验导出的。在实际生活中，有各种不同的因素干扰随机实验结果。那么成功和失败的概率就不是固定的，而是其他变量的一个函数。

假定自变量向量为 X，那么一个简单的函数为公式(1)：

$$\ln\frac{p}{1-p}=X^{\mathrm{T}}\beta \quad (P\text{ 为患高血压病的概率}) \tag{1}$$

2) 建立随机森林模型

在机器学习中，随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定。

随机森林的原理为：从所有(n 个)观测值中抽取 n 个观测值作为自助法样本(bootstrap sample)，也就是说，等概率地放回抽取和原数据同样样本量的样本，然后根据这个新样本建造一个(分类)决策树，在建造树的过程中并不用所有的变量当候选拆分变量，而是随机地挑选部分变量来竞争拆分变量，这样，不仅仅是每棵树所用的数据是随机抽取的，而且每个节点的拆分变量的选择都是随机的；不断重复上一步骤直到建成的决策树个数等于指定的数目为止(这里用的程序包 randomForest 中的 randomForest()函数的默认值为 500 棵树)；如果来了新的数据，每棵树给出一个预测值，然后所有的树(默认 500 棵树)用简单多数投票来决定其因变量的预测值。

3.2. 模型结果分析

1) 对种族为白人的数据进行模型结果分析

① Logit 参数估计结果分析

对数据 nhanes_1 用 Logit 方法进行建模分析，之前介绍过，选择 HBP 作为因变量，其余变量作为自变量。我们用软件 R 中的 glm()函数进行分析，得到了表 2 中的结果。

表 2 表示各因素分类中的变量对高血压影响的参数估计结果。

从结果中可以看出，在 0.05 的显著性水平下，种族为白人的时候，只有性别、受访者吸烟 > 100 支香烟 = 2、现在受访者抽烟状况 = 2 这三种因素不显著(现在受访者抽烟状况 = 3 缺失)，其余的因素对高血压都是有显著影响的。由于年龄的系数为正，也就是说随着年龄的增长对高血压的影响会越来越大；性别为负数可以看出男性比女性患高血压的几率要小一点；体重的系数为正数可以看出越胖的人得高血压的可能性要稍微大一点点；身高的系数为负数表示身高越高的人得高血压的风险就越低；血清胆固醇对患高血压稍微有点影响，血清胆固醇高的话得高血压的概率要稍微大一点。

Table 2. Effect of various factors on hypertension (white)
表 2. 各因素对高血压的影响(白人)

变量名	系数	标准误	Z 值	P 值
截距	−12.590	0.830	−15.178	0.000
年龄(A)	0.104	0.003	39.574	0.000
性别 = 1 (B(1))	−0.159	0.089	−1.777	0.076
体重(C)	0.004	0.001	3.893	0.000
身高(D)	−0.084	0.012	−6.781	0.000
平均舒张血压(E)	0.124	0.004	31.845	0.000
受访者吸烟 > 100 支香烟 = 2 (F(2))	−0.161	0.914	−1.760	0.078
现在受访者抽烟状况 = 2 (G(2))	−0.125	0.093	−1.338	0.181
现在受访者抽烟状况 = 3 (G(3))	NA	NA	NA	NA
血清胆固醇(H)	0.003	0.001	4.205	0.000

表 2 的参数表达式为公式(2)：

$$Logit\,HBP = \ln\frac{p}{1-p} = -12.590 + 0.104*A - 0.159*B(1) + 0.004*C - 0.084*D + 0.124*E - 0.161*F(2) - 0.125*G(2) + 0.003*H \quad (2)$$

用公式(2)得到的模型进行预测，我们可以得到用 logit 方法的误判率为 0.142 (在数据 nhanes_1 中对于变量HBP，将0误判给1的有504个，将1误判给0的有1024个，正确判断的有9239，误判率 $=\frac{\text{误判个数}}{\text{总个数}}$)，因此这个误判率稍微有点高。下面我们来看随机森林的结果。

② 随机森林分析结果

跟据吴喜之[9]的介绍，我们可以利用 R 软件的 randomForest()函数构建一个随机森林分类模型，并且画出各因素对高血压影响的重要性示意图(如图 1)。

从图 1 中我们可以看出在种族为白人的时候，对高血压影响最为显著的因素是年龄(HSAGEIR)和平均舒张血压(PEPMNK5R)，且年龄的重要性要高于平均舒张压。而且这两个都是呈正相关影响，其余的几个因素对高血压的影响都不是特别显著。简单来说，年龄越大，越容易得高血压；平均舒张血压值越高，也越容易得高血压。抽烟对得高血压并没有显著影响，在某些情况下，抽烟甚至会对高血压有反作用。

通过所建立的随机森林模型对数据进行预测，分类结果如下表 3 (行是真实值，列是预测值)，并得到误判率为 0.006，跟 Logit 方法相比减少了很多。也就是说，用随机森林得到的结果要比用 Logit 方法得到的结果准确率更高。

2) 对种族为黑人的数据进行模型结果分析

① Logit 参数估计结果分析

仿照前面的建模方法对数据 nhanes_2 构建模型，并用软件 R 进行运算，得到了表 4 中的结果。

表 4 表示各因素分类中的变量对高血压影响的参数估计结果。

从结果中可以看出，在 0.05 的显著性水平下，种族为黑人的时候，只有身高和血清胆固醇这两种因素不显著(现在受访者抽烟状况 = 3 缺失)，其余的因素对高血压都是有显著影响的。也就是说，身高、血清胆固醇对高血压几乎没有影响。由于年龄的系数为正，也就是说随着年龄的增长对高血压的影响会

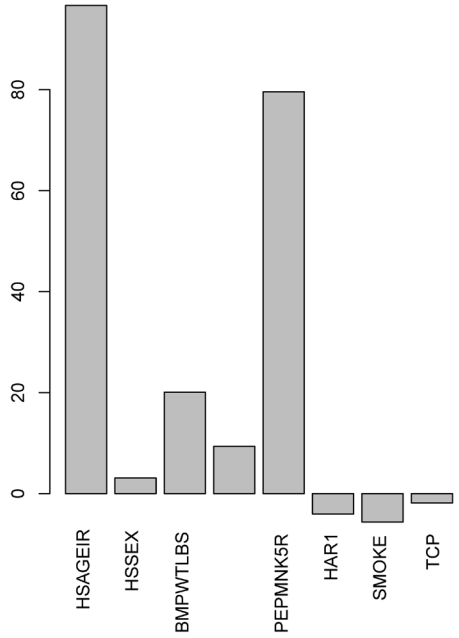

Figure 1. The importance of each factor to the influence of hypertension (white)
图 1. 各因素对高血压影响的重要性(白人)

Table 3. Classification results of nhanes_1 data in random forest
表 3. 随机森林对 nhanes_1 数据的分类结果

	HBP = 0	HBP = 1
HBP = 0	8447	0
HBP = 1	65	2255

Table 4. Effect of various factors on hypertension (black)
表 4. 各因素对高血压的影响(黑人)

变量名	系数	标准误	Z 值	P 值
截距	−16.229	1.479	−10.976	0.000
年龄(A)	0.101	0.004	23.226	0.000
性别 = 1 (B(1))	−0.445	0.149	−2.983	0.003
体重(C)	0.003	0.001	2.282	0.022
身高(D)	−0.031	0.021	−1.469	0.142
平均舒张血压(E)	0.143	0.006	23.678	0.000
受访者吸烟 > 100 支香烟 = 2 (F(2))	−0.415	0.127	−3.274	0.001
现在受访者抽烟状况 = 2 (G(2))	−0.493	0.146	−3.370	0.001
现在受访者抽烟状况 = 3 (G(3))	NA	NA	NA	NA
血清胆固醇(H)	0.001	0.001	1.190	0.234

越来越大；性别的系数为负数可以看出男性比女性患高血压的几率要小一点；体重的系数为正数可以看出越胖的黑人得高血压的可能性要稍微大一点点；平均舒张血压对高血压是正相关；抽烟的各种系数均为负数，也就是说不抽烟会对高血压产生负影响。

表 4 的参数表达式为公式(3)：

$$Logit\,HBP = \ln\frac{p}{1-p} = -16.229 + 0.101*A - 0.445*B(1) + 0.003*C - 0.031*D + 0.143*E - 0.415*F(2) - 0.493*G(2) + 0.001*H \quad (3)$$

同时我们可以得到用 Logit 方法的误判率为 0.124，这个误判率稍微有点高了。下面我们来看随机森林的结果。

② 随机森林分析结果

图 2 是通过随机森林得到的各因素对高血压影响的重要性示意图。

从图 2 中我们可以看出在种族为黑人的时候，对高血压影响最为显著的因素是平均舒张血压(PEPMNK5R)和年龄(HSAGEIR)，且平均舒张压的重要性要高于年龄。而且这两个都是呈正相关影响，其余的几个因素对高血压的影响都不是特别显著。简单来说，年龄越大，越容易得高血压；平均舒张血压值越高，也越容易得高血压。抽烟对得高血压并没有显著影响。

通过所建立的随机森林模型对数据进行预测，分类结果如表 5 (行是真实值，列是预测值)，并得到误判率为 0.003，跟 Logit 方法相比减少了很多。

3) 对种族为其他的数据进行模型结果分析

① Logit 参数估计结果分析

仿照前面的建模方法对数据 nhanes_3 构建模型，，我们用软件 R 中的 glm()函数进行分析，得到了表 6 中的结果。

表 6 表示各因素分类中的变量对高血压影响的参数估计结果。

从结果中可以看出，在 0.05 的显著性水平下，其他种族的情况下，只有性别(男)、受访者吸烟 > 100 支香烟 = 2 和血清胆固醇这三种因素不显著(现在受访者抽烟状况 = 3 缺失)，其余的因素对高血压都是有较为显著影响的。由于年龄的系数为正，也就是说随着年龄的增长对高血压的影响会越来越大；体重的系数为正数可以看出越胖的人得高血压的可能性要稍微大一点点；身高的系数为负数表示越高的人得高血压的可能性越小；平均舒张血压对高血压是正相关；抽烟的各种系数均为负数，也就是说不抽烟会对高血压产生负影响。

表 6 的参数表达式为公式(3)：

$$Logit\,HBP = \ln\frac{p}{1-p} = -5.784 + 0.109*A + 0.897*B(1) + 0.012*C - 0.194*D + 0.121*E - 0.571*F(2) - 1.442*G(2) - 0.002*H$$

同时我们可以得到用 Logit 方法的误判率为 0.088。

② 随机森林分析结果

图 3 是通过随机森林得到的各因素对高血压影响的重要性示意图。

从图 3 中我们可以看出在种族为其他的时候，对高血压影响最为显著的因素是年龄(HSAGEIR)和平均舒张血压(PEPMNK5R)，且年龄的重要性要高于平均舒张压。而且这两个都是呈正相关影响，其余的几个因素对高血压的影响都不是特别显著。简单来说，年龄越大，越容易得高血压；平均舒张血压值越高，也越容易得高血压。现在不抽烟对得高血压并没有显著影响。

此时我们可以得到随机森林的误判率为 0。

综上，我们可以得到随机森林分类模型都比 Logit 分类准确率要高。但是随机森林无法给出一个准确的公式，而 Logit 方法可以给出。

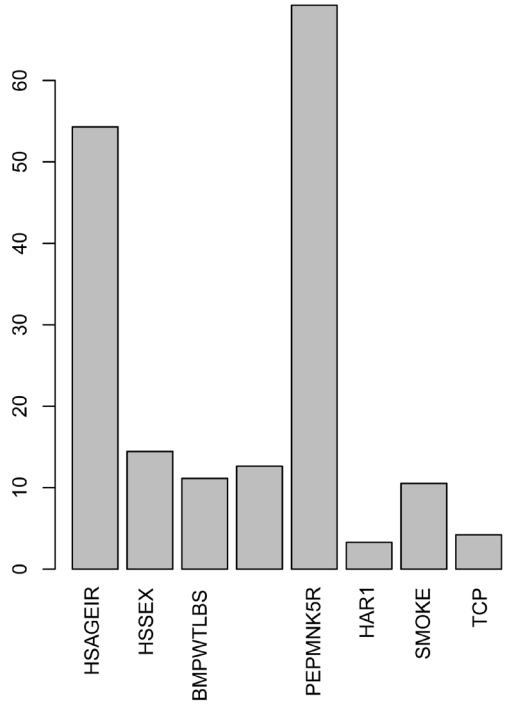

Figure 2. The importance of each factor to the influence of hypertension (black)
图 2. 各因素对高血压影响的重要性(黑人)

Table 5. Classification results of nhanes_2 data in random forest
表 5. 随机森林对 nhanes_2 数据的分类结果

	HBP = 0	HBP = 1
HBP = 0	3507	0
HBP = 1	13	845

Table 6. Effect of various factors on hypertension (other)
表 6. 各因素对高血压的影响(其他)

变量名	系数	标准误	Z 值	P 值
截距	−5.784	4.701	−1.230	0.219
年龄(A)	0.109	0.013	8.129	0.000
性别 = 1 (B(1))	0.897	0.518	1.731	0.083
体重(C)	0.012	0.006	2.127	0.033
身高(D)	−0.194	0.077	−2.514	0.012
平均舒张血压(E)	0.121	0.020	6.014	0.000
受访者吸烟 > 100 支香烟 = 2 (F(2))	−0.571	0.423	−1.348	0.178
现在受访者抽烟状况 = 2 (G(2))	−1.442	0.509	−2.833	0.005
现在受访者抽烟状况 = 3 (G(3))	NA	NA	NA	NA
血清胆固醇(H)	−0.002	0.004	−0.580	0.562

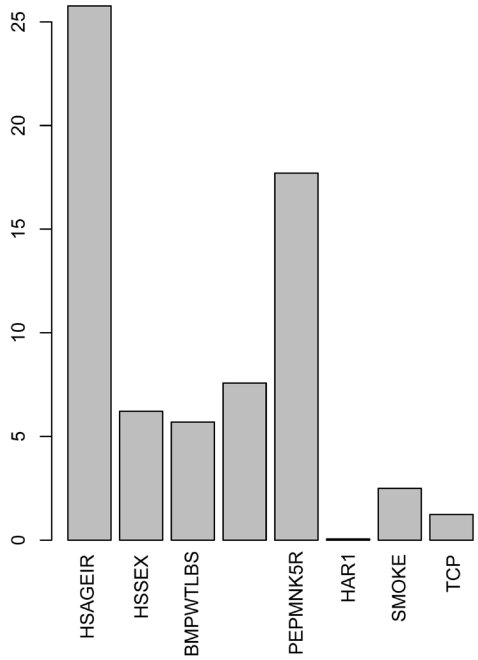

Figure 3. The importance of each factor to the influence of hypertension (other)
图 3. 各因素对高血压影响的重要性(其他)

4. 结论

在 Logit 分类方法下，我们可以看到不同种族的显著性因素不同，但是年龄、体重、平均舒张血压不论在什么种族下对高血压都有显著影响，且都是正相关。而对于性别 = 1 (即性别为男性)的时候，只有在种族为黑人的时候显著，而且是负相关，虽然在白人和其他的情况下，该因素并不显著，但是两个的系数一个为正一个为负。也就是说对于不同种族来说，男女得高血压的几率并不相同(白人和黑人男性更不容易得高血压，而其他的时候男性更容易得高血压)。对于身高这个因素，虽然在种族为黑人的时候不显著，但是对所有的种族而言，都是身高越高的人越不容易得高血压。而对于有关于抽烟的两个因素(受访者吸烟 > 100 支香烟 = 2 和现在受访者抽烟状况 = 2)，虽然在各个种族下并不是都是显著的，但是它们的系数都是很小的负数，也就是说不抽烟对得高血压是负相关的。而血清胆固醇在种族为白人的时候是显著的，而在剩下两种情况是不显著的，其中在白人和黑人的时候系数为正数，在其他的时候为负数，但是这三个数的绝对值都非常小，我们甚至可以近似为 0。

简而言之，就是年龄、体重、平均舒张血压对于所有的人都是影响高血压的正因素，不抽烟对高血压都是负相关。而在不同的种族下，其余不同的因素的影响都不完全相同。

在随机森林分类方法下，年龄和平均舒张血压都是影响高血压的最重要的因素。其中，在种族为白人和其他的情况下，年龄对高血压的影响比平均舒张血压的影响要更为重要一点，而在种族为黑人的情况下，平均舒张血压对高血压的影响比年龄的影响要更为重要一点。抽烟对高血压并没有显著影响。

综上所述，年龄和平均舒张血压对于每个人而言都是影响高血压的重要因素，且体重也是。但是在不同的种族下，各个因素的影响程度不同。而抽烟对高血压的影响并不显著。

基金项目

1) 国家自然科学基金项目“广义估计方程(GEE)框架下的回归诊断：基于均值和协方差结构同时拟

合的研究”(11561071)。

2) 云南省哲学社会科学研究基地 2015 年重点项目“云南社会经济可持续发展竞争力指标体系研究”(JD2015ZD20)。

参考文献 (References)

[1] WHO (2002) Reducing Risks Promoting Healthy Life. World Health Organization, Geneva, 1.

[2] 孙振球. 医学统计学[M]. 北京: 人民卫生出版社, 2007: 333-341.

[3] 李英华. 高血压的现状与流行[J]. 中华心血管病杂志, 2004(7): 456.

[4] Tian, J.Y., Cheng, Q., Song, X.M., *et al.* (2006) Birthweight and Risk of Type-2diabetes, Abdominal Obesity and Hypertension among Chinese Adults. *European Journal of Endocrinology*, **155**, 601-607. http://dx.doi.org/10.1530/eje.1.02265

[5] Ning, G., Su, J., Li, Y., *et al.* (2006) Artificial Neural Network Based Model for Cardiovascular Risk Stratification in Hypertension. *Medical and Biological Engineering and Computing*, **44**, 202-208. http://dx.doi.org/10.1007/s11517-006-0028-2

[6] Ture, M., Kurt, I., Yavuz, E. and Kurum, T. (2005) Comparison of Multiple Prediction Models for Hypertension (Neural Networks, Logistic Regression and Flexible Discriminant Analyses). *Anadolu Kardiyoloji Dergisi*, **5**, 24-28.

[7] 傅传喜, 马文军, 梁建华. 高血压危险因素 logistic 回归与分类树分析[J]. 中华疾病控制杂志, 2006, 10(3): 652-952.

[8] 杨洋. 利用人工神经网络模型预测原发性高血压的研究[D]: [硕士学位论文]. 北京: 中国医科大学, 2010.

[9] 吴喜之. 复杂数据统计方法: 基于 R 的应用(第 2 版) [M]. 北京: 中国人民大学出版社, 2013: 63-65.