影响血糖水平因素的偏最小二乘回归分析

中国科学院文献情报中心医务室(100190) 杨红宁
哈尔滨医科大学 武海滨 李 叶

偏最小二乘 (partial least squares, PLS) 回归分析方法是近年来应实际需要而产生和发展的一种具有广泛适用性的多元统计分析方法,它集多元线性回归分析、典型相关分析和主成分分析的基本功能于一体,将建模预测类型的数据分析方法与非模式的数据认识性分析方法有机地结合在一起,该方法能够在自变量存在严重多重相关性的条件下进行回归建模,更易于分析与因变量相关的因素,使自变量的回归系数更容易解释。本研究将此方法应用于某科研单位的体检资料,并与传统线性回归方法进行比较,研究与血糖值有关联的血浆生化及血压指标,反映在变量间存在多重相关时偏最小二乘回归分析方法的优势。

资料与方法

数据来源于我国科研机构 2007年人员体检资料,共包括 432个观察对象,被检测者抽血前 3天避免高脂饮食,空腹 12～14小时,清晨静止状态下抽取静脉血,测量甘油三酯 (TG)、总胆固醇 (TC)、高密度脂蛋白中胆固醇含量 (HDL - C)、低密度脂蛋白中胆固醇含量 (LDL - C),同时测量收缩压 (SBP)和舒张压 (DBP)。各项生化指标均采用日产希森美康 CHEM IX - 180型全自动生化分析及其配套试剂进行测定,3小时内完成操作。同时采用多元线性回归分析方法和偏最小二乘回归分析方法对数据资料进行分析,以进行对比,统计软件使用 SAS 9.1。

偏最小二乘法的主要特点是,在尽可能提取包含自变量更多信息成分的基础上,保证提取成分与因变量间具有最大相关性,当数据的观测数量较少,有缺失数据或存在多重相关性时,该方法仍可以对因变量进行较好的预测。每个自变量对因变量影响的大小可以用变量投影重要性 (variable importance for projection, VIP) 表示,VIP表示每个自变量在回归模型拟合中对于其他自变量和因变量的价值,VIP值越大则相应自变量对因变量解释的相关性越强,因此,VIP可以作为进行自变量筛选的重要指标。Wold in Umetrics (1995)建议当 VIP的值小于 0.8时可以认为相对较小。

结果与分析

1. 线性回归方法的分析结果

以血糖值为因变量,其他测量指标为自变量,对资料进行多元线性回归分析,结果见表 1。

表 1 多元线性回归分析结果

变量	回归系数	标准误	t值	P值	方差膨胀因子 (VIF)
常数项	2.7379	0.6538	4.19	<0.0001	0
SBP	0.0076	0.0056	1.38	0.1691	2.1551
DBP	0.0031	0.0107	0.29	0.7708	2.0076
TG	0.1997	0.0714	2.80	0.0054	1.2417
TC	-0.1504	0.1899	-0.79	0.4288	8.9556
HDL - C	-0.0745	0.1861	-0.40	0.6890	1.5122
LDL - C	0.4119	0.2443	1.69	0.0925	8.81171

回归模型具有统计学意义 ($F = 6.65$, $P < 0.0001$),表 1显示作为自变量的 6个变量中只有一个变量 TG具有统计学意义 ($P = 0.0054$);另外,各变量的方差膨胀因子 (variables influation factor, VIF)均较大 (VIF > 1),提示这些变量之间存在多重共线性。对所有自变量采用逐步选择法进行变量筛选,结果见表 2。

表 2显示,经过逐步选择法对变量进行筛选,SBP、TG和 LDL - C三个变量进入了模型,具有统计学意义 ($P < 0.05$)。

表 2 逐步选择法对变量的筛选结果

变量	回归系数	标准误	t值	P值
常数项	2.4092	0.4385	5.49	<0.0001
SBP	0.0088	0.0041	2.16	0.0316
TG	0.2105	0.0678	3.10	0.0020
LDL - C	0.2308	0.0915	2.52	0.0120

2. 偏最小二乘回归方法的分析结果

以血糖值为因变量,其他测量值为自变量,利用偏最小二乘回归方法对资料进行回归分析,结果见表 3。

(下转第 79页)

参 考 文 献

1. 中华医学会肝脏病学分会脂肪肝和酒精性肝病学组. 非酒精性脂肪肝病治疗指南. 2006: 1-3.
2. Fan-J-G, Saibara-T, Chitturi-S, et al What are the risk factors and settings for non-alcoholic fatty liver disease in Asia-Pacific? J-Gastroenterol, 2007, 22(6): 794-800.
3. Kojima-S, Watanabe-N, Numata-M, et al Increase in the prevalence of fatty liver in Japan over the past 12 years: analysis of clinical background. J-Gastroenterol-Hepatol, 2003, 38(10): 954-61.
4. Amarapurkar-D, Kamani-P, Patel-N, et al Prevalence of non-alcoholic fatty liver disease: population based study. Ann-Hepatol, 2007, 6(3): 161-3.

(上接第 77页)

表 3 偏最小二乘回归分析结果

变量	回归系数	标准化回归系数	变量投影重要性 (VIP)
常数项	2.0087	0	—
SBP	0.0063	0.07694	1.6807
DBP	0.0099	0.06078	1.3278
TG	0.1204	0.08719	1.9046
TC	0.0965	0.07054	1.5409
HDL - C	- 0.1306	- 0.04004	0.8746
LDL - C	0.1551	0.08741	1.9094

表 3显示,除 HDL - C的标准化回归系数相对较小(- 0.04004)外,其余各自变量的标准化回归系数相差不大。自变量中只有 HDL - C的 VIP值相对较小(VIP = 0.8746)。这表明除 HDL - C对回归模型的贡献相对较小外,其他的变量在回归模型中都有较大的贡献。

讨 论

应用多元线性回归和偏最小二乘回归,分析结果存在着较大的差异,这种差异主要由变量之间的多重相关性引起。当变量之间存在相关性时,由于变量之间效应的重叠,一些变量的效应掩盖了另一些变量的效应,所以在传统的线性回归模型中,某些与因变量有关联的自变量可能变得不显著。如表 1,由于变量之间的多重相关作用,只有 TG具有统计学意义。应用逐步选择法对自变量进行筛选后,虽有一定改善,但仍可能漏掉一些重要变量,而偏最小二乘回归对数据信息进行分解和筛选,可以有效地提取重要变量,剔除多重相关信息和无解释意义信息的干扰,使得有价值的自变量信息均可显现出来,如表 3中 SBP、DBP、TG、TC和 LDL - C对模型均有较大贡献。说明偏最小二乘回归具有传统线性回归方法不具备的许多优点,在存在共线情况下其结论可能更为可靠。

参 考 文 献

1. 邓念武,徐晖. 单因变量的偏最小二乘回归模型及应用. 武汉大学学报, 2001, 34(2): 14-16.
2. Zhang XA, Tian P. modeling and analysis of post - purchase intentions using partial least squares. Journal of the Chinese Institute of IndustrialEngineers, 2004, 21(1): 68-74.
3. 秦涛,林志娟,陈景武. 偏最小二乘回归原理、分析步骤及程序. 数理医药学杂志, 2007, 20(4): 450-451.
4. 王惠文. 偏最小二乘回归方法及其应用. 第 1版. 北京:国防工业出版社, 1999, 1 - 11.