基于多元线性回归的空腹血糖影响因素分析方法

张福旺 苑会娟
(哈尔滨理工大学测控技术与通信工程学院 哈尔滨 150080)

摘 要 通过分析空腹血糖影响因素的关系,提出了一种基于多元线性回归分析的空腹血糖影响因素分析方法。首先,收集影响空腹血糖的主要因素数据,包括血清总胆固醇、甘油三酯、空腹胰岛素、糖化血红蛋白;然后,通过散点图对这些影响因素进行分析和确定,利用收集到的数据构建基于最小二乘法的多元线性回归模型,并通过逐步回归分析得到修正后的模型;最后,运用此模型确定了影响空腹血糖的关键因素,以为糖尿病患者的平时饮食给予指导以及为医生的临床治疗提供参考。

关键词 多元线性回归,空腹血糖,最小二乘法,逐步回归
中图法分类号 O212.4 **文献标识码** A

Analysis of Factors Influencing Fasting Plasma Glucose Based on Multiple Linear Regression

ZHANG Fu-wang YUAN Hui-juan
(Institute of Measurement and Control Technology and Communication Engineering, Harbin University of Science and Technology, Harbin 150080, China)

Abstract A kind of fasting plasma glucose factors analysis method based on multiple linear regression analysis was proposed by analyzing the relationship of influencing factors of fasting plasma glucose. Firstly, the data of major factors influencing fasting plasma glucose is collected, including serum total cholesterol, triglyceride, fasting insulin, and glycated hemoglobin. Later, these influencing factors are analyzed and determined through scatter diagram. The multiple linear regression model based on least square method is constructed by the collected data. Meanwhile, through stepwise regression, the revised model is otained. At last, this model is applied to determine the key factors that affect fasting plasma glucose, so as to give diet guidance for diabetic patients, and provide reference for the clinical treatment of doctors.

Keywords Multiple linear regression, Fasting plasma glucose, Least-square method, Stepwise regression

1 引言

糖尿病在全世界成为了一种慢性终身性疾病,现如今已经演变成严重影响人们公共卫生的问题。目前尚不了解糖尿病的病因和发病机制,这种疾病主要由高血糖维持其基本的生化变化[1]。进入 21 世纪后,糖尿病发病率飞速上涨,糖尿病患者的健康日益受到关注[2]。血糖,特别是空腹血糖,是诊断糖尿病的唯一标准。通过分析空腹血糖变化的影响因素可以加深对空腹血糖与糖尿病之间关系的了解。

目前,国内关于空腹血糖的影响因素的实证研究逐渐增多。文献[3]探讨了空腹血糖受损的相关影响因素,结果表明空腹血糖受损患者已开始聚集心血管危险因素,且他们的发病率与年龄、超重肥胖、血压、血脂代谢异常、糖尿病家族史相关。文献[4]探讨了正常空腹血糖与心血管疾病及其危险因素的相关性,结果显示正常空腹血糖的不同水平结合传统的心血管疾病危险因素能够预测心血管疾病发生的可能性。文献[5]探讨了不同空腹血糖水平诊断糖尿病的敏感度和特异度,寻找空腹血糖诊断糖尿病的合理截点,结果表明空腹血糖≥6.13 mmol·L^{-1}诊断糖尿病的敏感度优于空腹血糖≥7.0 mmol·L^{-1},且能降低漏诊率,适于作为筛查早期糖尿病的空腹血糖截点。

本文从空腹血糖本身存在的密切关系出发,结合多元线性回归分析模型在处理由多因素共同影响的现实生活问题上的高效性,建立了基于多元线性回归的空腹血糖分析,用于对影响空腹血糖的因素进行分析。与以往的关于空腹血糖的影响因素分析方法相比,该方法具有模型操作简练、分析结果准确、解释力强的特点,且已广泛应用于各种模型分析[6-9],在实际应用中效果显著。

2 多元线性回归

回归分析分为单变量回归分析与多元回归分析[10]。在现实生活中,对因变量产生作用的单一变量较少,其常常是被多个自变量综合影响。比如,影响人才流动意向的变量不但有工作环境满意度,还包括工作压力、职业发展规划、晋升机制和奖惩制度满意程度等因素。对于多元回归的分析,若两个或两个以上的自变量与因变量的关系为线性,则称其为多元线性回归。在现实生活中,多元线性回归在解决由多个因素影响的问题中得到了广泛应用。例如,文献[11]根据 2006—2015 连续 10 年的历史数据对北京等 7 个不同类型城市的商品房销售率与房屋均价、人口可支配收入、人口密度等 3 个

张福旺(1990—),男,硕士生,主要研究方向为计算机视觉、自然语言处理、数理统计,E-mail:3537064032@qq.com(通信作者);**苑会娟**(1963—),女,教授,硕士生导师,主要研究方向为传感技术、3D 测量技术,E-mail:huijuany@126.com。

因素建立多元线性回归模型，分别确定了总体线性相关的显著性和各个变量影响的显著性，以识别影响决策变量的关键因素，为在制定房地产调控政策时选择主攻方向提供了一个值得借鉴的思路。文献[12]基于多元线性回归模型，分析了流域内不同子流域的土地利用类型和降水量变化对各子流域径流量变化的影响，得出了不同子流域降水量与各土地利用类型对流域内径流量变化的驱动性存在较大差异，在进行水利工程规划设计时需进行单独讨论与分析的结论。

本文中，对于需要进行分析的因变量 y 与自变量 x_1，x_2，…，x_m（$m\geqslant 2$），其对应的多元线性回归分析模型为：

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\varepsilon \tag{1}$$

其中，$\varepsilon\sim N(0,\delta^2)$为随机误差，$\beta_1$，$\beta_2$，…，$\beta_m$ 是未知参数，β_0 为回归常数，β_1，β_2，…，β_m 为回归系数。

3 模型的建立

3.1 样本的选取

本文收集了 27 名糖尿病患者的各项指标数据作为测试集，如表 1 所列。其中包括血清总胆固醇、甘油三酯、空腹胰岛素以及糖化血红蛋白数据。因为本文主要考虑糖尿病的综合影响因素，故选取以上几项因素作为影响因素。例如，空腹胰岛素值偏高，则空腹血糖值就应该下降。

表 1 27 名糖尿病患者的指标

1	X_1	X_2	X_3	X_4	Y
1	5.68	1.90	4.53	8.2	11.2
2	3.79	1.64	7.32	6.9	8.8
3	6.02	3.56	6.95	10.8	12.3
4	4.85	1.07	5.88	8.3	11.6
5	4.60	2.32	4.05	7.5	13.4
6	6.05	0.64	1.42	13.6	18.3
7	4.90	8.50	12.60	8.5	11.1
8	7.08	3.00	6.75	11.5	12.1
9	3.85	2.11	16.28	7.9	9.6
10	4.65	0.63	6.59	7.1	8.4
11	4.59	1.97	3.61	8.7	9.3
12	4.29	1.97	6.61	7.8	10.6
13	7.97	1.93	7.57	9.9	8.4
14	6.19	1.18	1.42	6.9	9.6
15	6.13	2.06	10.35	10.5	10.9
16	5.71	1.78	8.53	8.0	10.1
17	6.40	2.40	4.53	10.3	14.8
18	6.06	3.67	12.79	7.1	9.1
19	5.09	1.03	2.53	8.9	10.8
20	6.13	1.71	5.28	9.9	10.2
21	5.78	3.36	2.96	8.0	13.6
22	5.43	1.13	4.31	11.3	14.9
23	6.50	6.21	3.47	12.3	16.0
24	7.98	7.92	3.37	9.8	13.2
25	11.54	10.89	1.20	10.5	20.0
26	5.84	0.92	8.61	6.4	13.3
27	3.84	1.20	6.45	9.6	10.4

3.2 自变量的选取

散点图是描述变量之间关系的直观方式，从图中可以看出变量之间的关系和强度之间的关系。因此，散点图是最有效和最简单的相关分析工具之一。如果它们之间存在线性关系，则通过线性回归可以进一步阐明它们之间的函数关系。

通过散点图可以有效地选出相对应的自变量。从残差图可以明显地看出数据是否出现异常，便于对数据进行处理。假设 X_1 为血清总胆固醇数据，X_2 为甘油三酯数据，X_3 为空腹胰岛素数据，X_4 为糖化血红蛋白数据，X_1-X_4 为自变量；Y 为空腹血糖数据，它们为因变量。图 1 为数据的残差图，空腹血糖与各影响因素的散点图如图 2 所示。

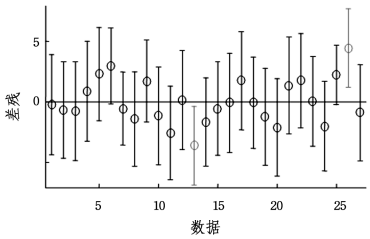
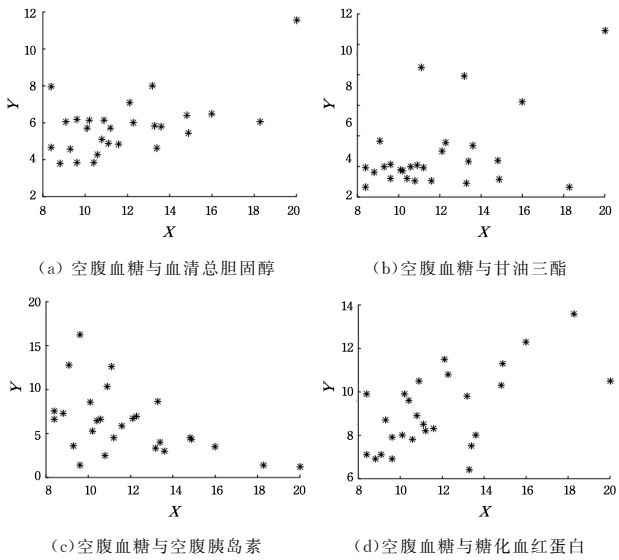


图 1 数据残差图



(a) 空腹血糖与血清总胆固醇　(b)空腹血糖与甘油三酯

(c)空腹血糖与空腹胰岛素　(d)空腹血糖与糖化血红蛋白

图 2 空腹血糖与各影响因素散点图

通过对图 2 的分析可知，血清总胆固醇、甘油三酯、糖化血红蛋白与空腹血糖之间均呈现出正向线性相关的关系，而空腹胰岛素与空腹血糖呈现出负向线性相关的关系。本文中，糖化血红蛋白与空腹血糖之间的正向线性相关关系尤其显著，血清总胆固醇与空腹血糖之间的线性相关关系次之，而甘油三酯与空腹血糖之间的线性相关性一般，空腹胰岛素与空腹血糖之间的负向线性相关关系也尤其显著。从图 1 可以明显看出，第 13 点和第 26 点是异常的，在这里将其剔除。

3.3 利用最小二乘法建立多重回归模型

通过多元线性回归对以上 4 个变量建立模型，得到的结果如表 2、表 3 所列。观察表 2 可以看出，$h=0$，$h_1=0$，$P_1=1.0000$，由 jbtest 检验、ttest 检验可知残差服从均值为 0 的正态分布。将 25 个数据从小到大排列，去掉中间的 5 个数据，由观测值 $f<F(6,6)$可知没有异方差。由 $du<DW<4-du=2.479$ 可知不存在残差自相关。

表 2 统计量数值

统计量	统计量值
h	0
P	0.3546
h_1	0
P_1	1.0000
f	2.7569
DW	1.9860
$F(6,6)$	4.284
dl	0.832
du	1.521

表 3　回归系数的最小二乘法估计结果

系数	系数估计值	标准误差	t	P
常量	3.6366	2.4408	1.490	0.1518
X_1	0.2393	0.3568	0.671	0.5101
X_2	0.3198	0.1889	1.693	0.1059
X_3	−0.2447	0.1048	−2.335	0.0301
X_4	0.8185	0.2078	3.939	0.0008

从表 3 可得到模型各自变量的回归系数估计值、标准误差、t 检验值以及 P 值信息，结合以上信息，得到最小二乘法下的回归模型：

$$Y=3.6366+0.2398X_1+0.3198X_2-0.2447X_3+0.8185X_4 \quad (2)$$

4　利用逐步回归法建立多元线性回归模型

从表 3 可以看出，空腹血糖与血清总胆固醇、甘油三酯、空腹胰岛素和糖化血红蛋白呈显著线性关系，即显示出总体的显著性，但是这并不意味着空腹血糖与各自变量的关系都显著，要确定每个自变量对空腹血糖的影响是否显著，需要对每个自变量进行检验。通过对 P 值的分析可知，血清总胆固醇、甘油三酯对空腹血糖的影响不显著；空腹胰岛素、糖化血红蛋白对空腹血糖的影响显著。因此，如果选择所有的自变量来建立回归模型，效果将不理想。

作为多元线性回归分析的相关分析方法——逐步回归分析，其在多元线性回归模型的基础上建立最优或者最合适的回归模型，使得对所研究的变量的分析及其依赖关系得到了进一步深入了解。逐步回归的主旨思想在于对于显著性的变量保留在模型中，以此同时剔除无显著性变量，得到所求的最优回归模型，使得可以有效避免模型的多重共线性。

在逐步回归分析中，对因变量 Y 和自变量 X 建立的多元线性回归模型进行模型和各个自变量的假设检验。当所建立的模型不显著时，该回归模型的线性关系不成立；模型中的任一自变量若对因变量不显著，则对其进行剔除，对于对因变量显著的自变量则将其筛选出来，在不包括该自变量的情况下，重新建立多元线性回归模型，从而得到最优的回归模型。

对于已经建立好的多元线性回归模型，基于 Akaike 信息统计，通过逐步回归分析选择最小的 AIC 信息统计量，达到删除自变量的目的，在多元线性回归模型的基础上，建立最优或者最合适的回归模型。

表 4　逐步回归统计量

变量	AIC	变量	AIC
Start	28.68	Step	27.23
X_1	27.232	None	27.232
X_2	30.026	X_2	35.511
X_3	32.700	X_3	34.316
X_4	41.029	X_4	41.716

从表 4 可以看出，对所有自变量使用回归模型时，AIC 统计值为 28.68。如果自变量 X_1 被剔除，则 AIC 统计值为 27.232；如果自变量 X_2 被剔除，则 AIC 统计值为 30.026。由于剔除 X_1 可以得到最小化的 AIC 统计量，因此将自变量 X_1 剔除并且进行下一轮逐步回归分析。在下一轮中，无论哪个自变量被删除，AIC 统计量的值都会增加，从而终止计算并获得最优的回归模型。

表 5　逐步回归分析后的统计结果

系数	系数估计值	标准误差	t	P
常量	4.52599	2.02201	2.238	0.0362
X_2	0.41295	0.12636	3.268	0.0037
X_3	−0.27754	0.09149	−3.033	0.0063
X_4	0.86222	0.19473	4.428	0.0002

从表 5 得到模型各自变量的回归系数估计值、标准误差、t 检验值以及 P 值信息，结合以上信息可以看出，回归系数的显著性水平得到极大的提高，且各项检验均表明各个自变量对因变量的影响尤其显著，效果理想，分析准确。因此，得到最优的回归模型：

$$Y=4.52599+0.41295X_2-0.27754X_3+0.86222X_4 \quad (3)$$

结束语　本文通过对影响空腹血糖的因素进行分析，建立了基于血清总胆固醇、甘油三酯、空腹胰岛素以及糖化血红蛋白 4 个因素的多元线性回归模型来分析影响空腹血糖的因素。由于所选取的各要素之间与因变量是否存在显著性是未知的，因此通过逐步回归分析方法，基于 Akaike 信息统计，参考 AIC 信息统计量，选取合适的自变量对模型进行优化，得到了最优的回归模型。通过最优回归模型可知，空腹血糖和甘油三酯、空腹胰岛素、糖化血红蛋白密切相关。其中，甘油三酯、糖化血红蛋白与空腹血糖呈正相关；空腹胰岛素与空腹血糖呈负相关。因此加深了对数据间客观关系的了解，使变量间的规律得到了进一步的揭示，在现实生活中具有相应的价值和重要的理论意义。

参 考 文 献

[1] BRAMLAGE P，BINZ C，GITT A K，et al. Diabetes treatment patters and goal achievement in primary diabetes care-study protocol and patent characteristics at baseline[J]. Cardiovasc Diabetol，2010，9(53)：1-14.

[2] CURRY A. Diabetes and dementia. Dose type 2 care also bolster brain function[J]. Diabetes Forecast，2010，63(9)：64-66.

[3] 邹海洪. 空腹血糖受损相关影响因素分析[J]. 海南医学，2009，20(9)：18-20.

[4] 李兵强，李广平，袁如玉，等. 正常空腹血糖与心血管疾病危险因素的相关性研究[J]. 中国全科医学，2011，14(1B)：136-139.

[5] 马辰星，许颖，康向辉，等. 空腹血糖诊断糖尿病截点的探讨与评价[J]. 中国全科医学，2014，17(17)：1943-1945.

[6] 王勇，黄国兴，彭道刚. 带反馈的多元线性回归法在电力负荷预测中的应用[J]. 计算机应用与软件，2008，25(1)：82-84.

[7] 李军成，陈国华，石小芳. 基于灰色多元线性回归的粮食产量预测[J]. 安徽农业科学，2010，38(16)：8281-8282.

[8] 周晨，冯宇东，肖匡心，等. 基于多元线性回归模型的东北地区需水量分析[J]. 数学的实践与认识，2014，44(1)：118-123.

[9] 周永生，肖玉欢，黄润生. 基于多元线性回归的广西粮食产量预测[J]. 南方农业学报，2011，42(9)：1165-1167.

[10] UYANLK G K，GULER N. A Study on Multiple Linear Regression Analysis[J]. Procedia-Social and Behavioral Sciences，2013，106(106)：234-240.

[11] 谈元伟，季红蕾，刘桂兰. 基于多元线性回归的城市房地产销售率影响要素分析[J]. 南通职业大学学报，2017，31(3)：60-63.

[12] 张竞月. 基于多元线性回归模型的滦河流域径流量变化驱动因子分析[J]. 水利水电技术，2017，48(11)：43-47.