Operations Research and Fuzziology 运筹与模糊学, 2020, 10(4), 321-329
Published Online November 2020 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2020.104034

糖尿病并发症的关键因素分析与风险预测

张　立，范馨月*，魏斐斐

贵州大学数学与统计学院，贵州　贵阳
Email: *1432177413@qq.com

收稿日期：2020年11月4日；录用日期：2020年11月19日；发布日期：2020年11月26日

摘　要

通过分析检查者各属性情况，研究对其患糖尿病概率的影响并制定相应的风险指标。利用检查者数据作为研究数据，用各属性的相关系数值来分析选取可能影响检查者存活概率的属性，并对所选取的属性分别进行单个属性的一元logistic回归和多个属性的多元logistic回归，并加入标准化后的HbA1c值联合患病概率考虑风险系数。多元logistic回归分析显示，性别、年龄γ-谷氨酰基转移酶，血清白蛋白，钙属性较大影响检查者患病概率的因素。认为检查者的患病概率并不是随机的，而是受检查者多个属性所影响的。

关键词

糖尿病并发症，关键因素，logistic回归，风险预测

Key Factors Analysis and Risk Prediction of Diabetic Complications

Li Zhang, Xinyue Fan*, Feifei Wei

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
Email: *1432177413@qq.com

Received: Nov. 4th, 2020; accepted: Nov. 19th, 2020; published: Nov. 26th, 2020

Abstract

To study the influence on the probability of diabetes mellitus (DM) of examinees by analyzing the attributes of examinees, using the data of examiners as the research data, the correlation coefficient value of each attribute was used to analyze and select the attributes that may affect the survival probability of examiners, and the single logistic regression of single attribute and multiple logistic regression of multiple attributes were carried out for the selected attributes, and the

*通讯作者。

文章引用: 张立, 范馨月, 魏斐斐. 糖尿病并发症的关键因素分析与风险预测[J]. 运筹与模糊学, 2020, 10(4): 321-329.
DOI: 10.12677/orf.2020.104034

standardized HbA1c value was added to consider the risk coefficient. Multiple logistic regression analysis shows that sex, age, y-glutamyltransferase, serum albumin, calcium attribute have a great influence on the probability of disease. It is not random, but influenced by multiple attributes of examinees.

Keywords

Diabetic Complications, Key Factors, Logistic Regression, Risk Prediction

1. 引言

糖尿病是一组由多病因引起的以慢性高血糖为特征的终身性代谢性疾病。长期血糖增高，大血管、微血管受损并危及心、脑、肾、周围神经、眼睛、足等，据世界卫生组织统计，糖尿病并发症高达 100 多种，是目前已知并发症最多的一种疾病。糖尿病死亡者有一半以上是心脑血管所致，10%是肾病变所致。因糖尿病截肢的检查者是非糖尿病的 10~20 倍。临床数据显示，糖尿病发病后 10 年左右，将有 30%~40%的检查者至少会发生一种并发症，且并发症一旦产生，药物治疗很难逆转，因此强调尽早预防糖尿病并发症。关于糖尿病并发症它常常与多种危险因素相关，如遗传、性别、年龄、社会、肥胖、生活方式、血脂、环境、血糖、糖化血红蛋白等等，医学上关于一般恒定糖尿病并发症的风险指标为 FPG、2h-PG、糖化血红蛋白(HbA1c)三类[1]。

虽然在是否患有糖尿病并发症的问题上有一些幸运因素，但有些群体比其他群体，如女性、年老者等更有可能患上糖尿病并发症，所以是否患上糖尿病并发症其实并非随机，而是基于一些背景的。针对糖料病患者临床研究公布数据，包含检查者的基本信息表、诊断表、检查表、医嘱表、费用表、生化检查表、糖化检查表、尿常规检查表等临床科研数据集，每个表通过用户 ID 进行关联。可以尝试根据检查者各因素(性别、年龄、各类检查数据)为自变量，建立合适的广义线性模型，分析人的患有糖尿病并发症情况与其关键因素的关系。

2. 数据处理与属性选择

关于糖尿病并发症数据来源于 2019 年人口健康共享杯大赛[2]发布的原始数据，它给出了检查者的各类具体数据，数据集包含基本信息表、诊断表、检查表、医嘱表、费用表、生化检查表、糖化检查表、尿常规检查表等，每个表通过用户 ID 进行关联，剔除无效数据和大量缺失数据样本，如铁和不饱和铁结合力两类属性有效数据为 23 组，477 组空白数据，考虑删除此类属性，通过初步人工检索选择所有可能影响糖尿病并发症出现的数据，包括检查者的 patient_id、性别、年龄、血清血蛋白、丙氨酸氨基转移酶等二十四类属性。利用 python3.7 对所有表格的原始数据做一个大致的显示，最终确定发现样本数据中总共有 500 名检查者，其中除了年龄数据与性别数据外的属性数据均不全，例如：

- 天冬氨酸氨基转移酶属性只有 402 名检查者有记录
- 葡萄糖属性只有 445 名检查者有记录

再观察具体数据的数值情况，利用 describe 函数得到数值型数据的分布情况，其中有 2 个属性(性别、诊断情况)为非数值型数据，故无法显示，详见表 1 数值型属性数据分布表。

Table 1. Distribution of original data of numerical attributes
表 1. 数值型属性原始数据部分分布表

Attribute	患病	年龄	丙氨酸氨基转移酶	天冬氨酸氨基转移酶	总蛋白
count	499	499	402	401	326
mean	0.136273	56.785571	109.411443	107.558853	64.704294
std	0.343422	12.897655	455.682158	652.442121	9.677023
min	0	0	4.5	4.2	32.3
25%	0	50	13.9	14.4	58.85
50%	0	58	24.2	22.8	66.25
75%	0	66	49.5	43.6	71.3
max	1	84	5514.2	10,897.1	90.9

从表 1 观察部分属性值，在 mean 字段处，约有 13.6%的人患有糖尿病并发症，平均年龄从表 1 观察部分属性值，在 mean 字段处，约有 13.6%的人患有糖尿病并发症，平均年龄在 57 岁，但在这里的显示均是忽略了缺失值后的数值。现将所有的原始数据进行处理：在 57 岁，但在这里的显示均是忽略了缺失值后的数值。现将所有的原始数据进行处理：

- 针对所有只有部分检查者有记录的属性，将缺失值由其各自属性的平均值代替。
- 针对每一位检查者均有其特定的 ID (Passenger ID)，可剔除申请序号这个数值型属性。
- 正常情况下，检查者的信息和检查情况表之间是有直接关系的，均由检查者 ID 进行关联，我们针对本次分析样本中的 500 个检查者信息重新做一个顺序排列(1~500)。
- 将性别属性转换为数值型属性，女性为 0，男性为 1。
- 将诊断属性转换为数值型属性(患病)，诊断患有糖尿病及并发症为 1，其他情况均记为 0。

将以上工作完成后，得到 500 组检查者的二十四类属性情况(性别，年龄，患病，丙氨酸氨基转移酶(0~40 U/L)，天冬氨酸氨基转移酶(0~40 U/L)，总蛋白(55~80 g/L)，血清白蛋白(35~50 g/L)，总胆红素(0~21 umol/L)，直接胆红素(0~8.6 umol/L)，碱性磷酸酶(0~130 U/L)，尿素(1.8~7.5 mmol/L)，γ-谷氨酰基转移酶(0~50 U/L)，肌酐(30~110 umol/L)，葡萄糖(3.4~6.1 mmol/L)，甘油三酯(0.4~1.7 mmol/L)，血清尿酸(104~444 umol/L)，总胆固醇(3.1~5.7 mmol/L)，肌酸激酶(2~200 U/L)，乳酸脱氢酶(40~250 U/L)，钙(2.09~2.54 mmol/L)，钠(130~150 mmol/L)，钾(3.5~5.5 mmol/L)，氯化物(94~110 mmol/L)，无机磷(0.89~1.6 mmol/L)，镁(0.6~1.4 mmol/L))，将这二十四类属性做一个线性相关性的热力图，见图 1。

考虑以检查者的患病情况(序号 4)为因变量，除去编号序列，观察图 1，表 2 中剩余的二十四个自变量分别与因变量之间的相关性。根据相关系数背景知识[3]，其中检查者的性别，总蛋白，血清白蛋白，血清尿酸，钙五个属性与患病之间有一般的正相关(相关系数分别为 0.086、0.096、0.13、0.13、0.12)，检查者的年龄，总胆红素，直接胆红素，γ-谷氨酰基转移酶与患病属性之间有一般的负相关(相关系数为−0.28、−0.096、−0.13、−0.100)，暂定性别，总蛋白，血清白蛋白，血清尿酸，钙，年龄，总胆红素，直接胆红素，γ-谷氨酰基转移酶这九个属性为自变量。但注意到检查者的总蛋白，血清白蛋白，钙之间有较强的正相关，总胆红素，直接胆红素，γ-谷氨酰基转移酶之间有较强的正相关，一般不宜将相关性较强的变量同时作为自变量进行建模，本次建模保留钙，γ-谷氨酰基转移酶两个属性，删除与之相关的总蛋白，血清白蛋白，总胆红素，直接胆红素四个属性，最终选择性别，年龄，γ-谷氨酰基转移酶，血清白蛋白，钙五个属性作为自变量。

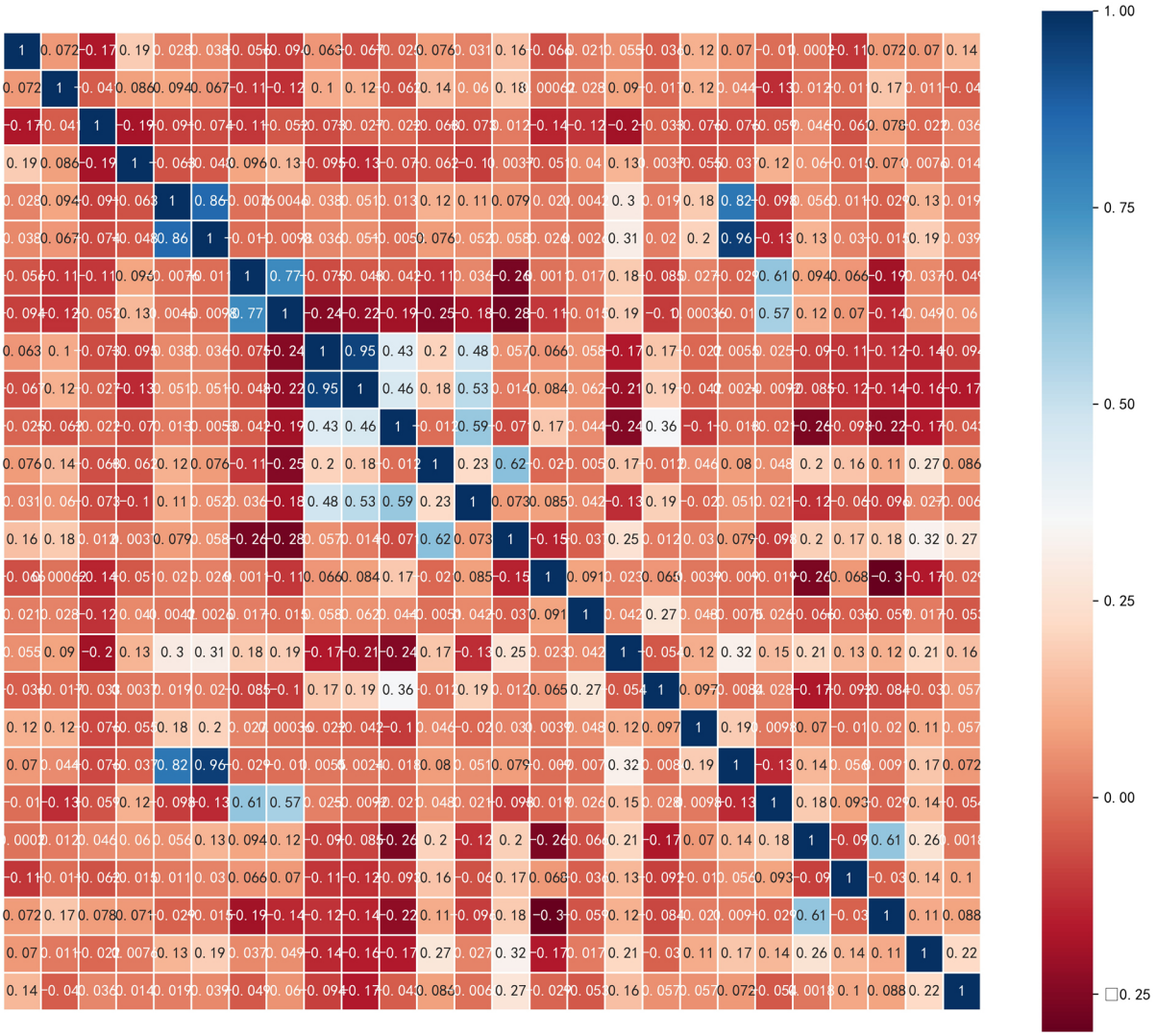

Figure 1. Thermodynamic diagram of linear correlation of attributes
图 1. 属性线性相关性的热力图

Table 2. Correlation between various attributes and disease attributes
表 2. 各属性与患病属性的相关性表

属性	相关性	属性	相关性	属性	相关性
性别	0.086	碱性磷酸酶	−0.070	肌酸激酶	−0.05
年龄	−0.28	尿素	−0.062	乳酸脱氢酶	−0.03
丙氨酸氨基转移酶	−0.063	γ-谷氨酰基转移酶	−0.100	钙	0.12
天冬氨酸氨基转移酶	−0.04	肌酐	−0.0037	钠	−0.060
总蛋白	0.096	葡萄糖	−0.051	钾	−0.016
血清白蛋白	0.13	甘油三酯	0	氯化物	0.07
总胆红素	−0.096	血清尿酸	0.13	无机磷	−0.007
直接胆红素	−0.13	总胆固醇	−0.0037	镁	−0.007

在进行模型回归之前，先简单观察检查者性别属性和年龄属性与患病情况之间的关系，分析数据之间的关系：

- 性别与患病间的关系

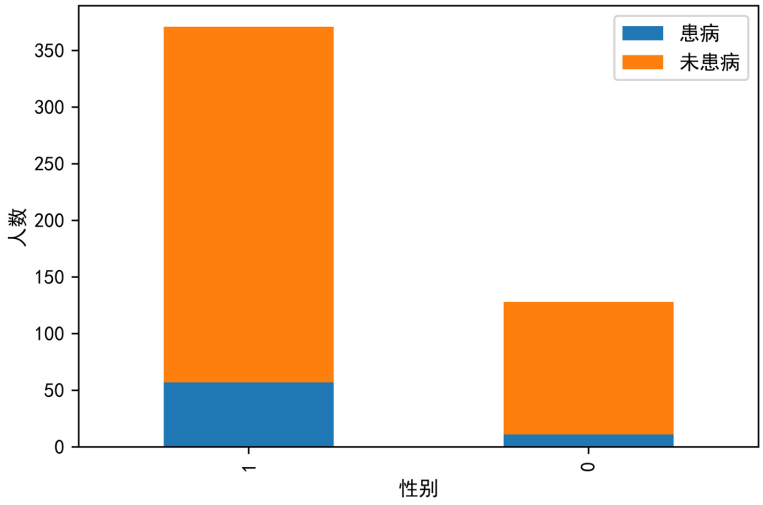

Figure 2. Relationship between gender and disease
图 2. 性别与患病关系图

利用 python 软件绘出 Titanic 事件中男女存活的分布情况，由图 2 可见。图 2 左的柱状图表示男性患病的数量情况，图 2 右的柱状图表示女性患病的数量情况，明显男性比女性的患病人数数量更多，从这里可以看出性别对是否患病情况有一定的影响，而且结合现实发现，男性相比校与女性而言更容易患病。

- 年龄(Age)与存活间的关系

根据图 1 的热力图显示，发现年龄与存活之间的线性相关性为−0.28，这可以视作为具有一定的相关性。而且结合生活实际来看，年龄是会影响存活的情况的，不同年龄段的人存活概率是不一样的，可如果仅根据−0.28 的相关性条件来看，那么年纪越小患病概率越大，这比较实际来说是不可观的，考虑将年龄做一个分段观察。

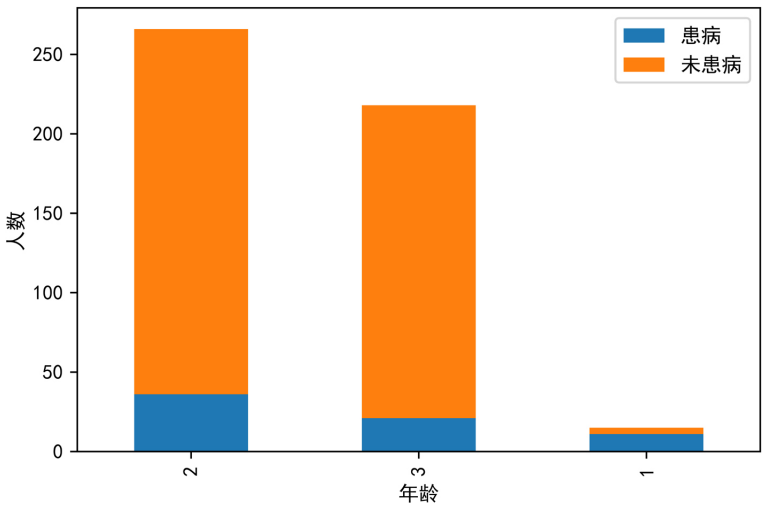

Figure 3. Distribution of examinees in different age groups
图 3. 不同年龄段的检查者情况分布图

将年龄段分为三个等级(0~29, 30~59, 60~100)，绘出不同年龄段的检查者获救情况分布情况，由图3可见，发现年龄分布在第一等级和第二等级上的检查者患病概率更大，而且关于患病人群逐渐年轻化。在人们的印象中，糖尿病是中老年人的“专利”，但通过查阅发现，过去20年间，我国15岁以下儿童发病率增加了近4倍，多在35~40岁之后发病的2型糖尿病也逐渐在青少年中产生。

- 年龄、性别和患病间的关系

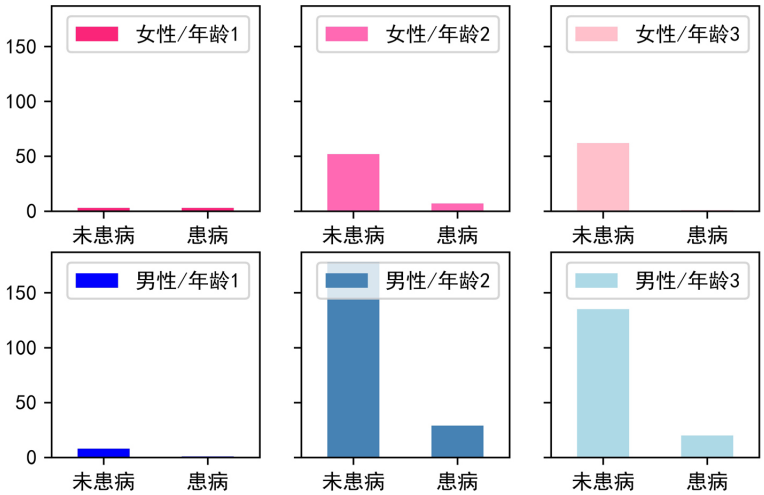

Figure 4. Distribution of age-related diseases in men and women
图4. 男女不用年龄患病分布图

绘出三类不同年龄等级的男女患病分布情况，由图4可见。图4上的三幅小图分别表示女性在年龄段1、年龄段2、年龄段3下患病与否的数量情况，明显处于年龄段1和年龄段2的女性比年龄段3患病的人数更多；图4下的三幅小图分别表示男性在一年龄段1、年龄段2、年龄段3下患病与否的数量情况，明显处于年龄段1比年龄段2和年龄段3患病的概率更大。从这里可以看出不同性别下的年龄段不同对患病情况有一定的影响，针对女性而言，应加大关注年龄段1和年龄段2的人群，然而针对男性而言，年龄段3患病的概率仍旧较大需要时刻保持警醒，而且患病群体的年轻化针对男性和女性都较明显。

3. 模型回归与预测

现将检查者的多个属性(性别，年龄，γ-谷氨酰基转移酶，血清白蛋白，钙)对其患病概率的关系进行分析。对于患病情况的处理，将引入一个二值因变量：

$$Y=\begin{cases}1, & \text{患病}\\ 0, & \text{未患病}\end{cases}$$

针对因变量为二分类数值型变量的情况，因此可以拟合Logistic模型[4] [5] [6]进行分析，文中将从单个自变量研究出发，逐渐拟合多元情形。开始建立检查者患病的概率与检查者性别的一元Logistic模型。

设$\pi(\varpi_1)$表示的是性别为ϖ_1的检查者患病的概率，建立Logistic模型：

$$\ln\left(\frac{\pi(\varpi_1)}{1-\pi(\varpi_1)}\right)=\beta_0+\beta_1\varpi_1 \tag{1}$$

其中，$\pi(\varpi_1)$为分类变量：

$$\varpi_1=\begin{cases}0, & sex=male\\ 1, & sex=female\end{cases}$$

利用 SPSS 进行 logistic 回归，可得拟合的模型：

$$\ln\left(\frac{\bar{\pi}(\varpi_1)}{1-\bar{\pi}(\varpi_1)}\right)=\bar{\beta}_0+\bar{\beta}_1\varpi_1=-2.364+0.658\omega_1$$

或

$$\bar{\pi}(\varpi_1)=\frac{\exp(-2.364+0.658\omega_1)}{1+\exp(-2.364+0.658\omega_1)}$$

其中 $\bar{\beta}_1=0.658$ 的标准差为 $s(\bar{\beta}_1)=0.067$，检验 $\beta_1=0$ 的 wald 统计量的值为 156.204，p 值小于 0.0001，因此回归关系高度显著，利用回归得到的 logistic 模型进行预测得到准确率为 86.40%。

由于 $\bar{\beta}_1=0.658>0$，说明检查者是否患病的概率随检查者的性别是有关系的，当 $\omega_1=1$ 即检查者性别为男性时，患病的概率更高，这与之前研究性别(sex)与生存间的关系的结论是一致的。

关于性别属性它是一个二值变量，现在选择研究年龄这个单个属性，年龄属性是一个离散数值变量，开始建立检查者患病的概率与检查者年龄的一元 Logistic 模型。

设 $\pi(\varpi_2)$ 表示的是年龄为 ϖ_2 的检查者患病的概率，建立 Logistic 模型：

$$\ln\left(\frac{\pi(\varpi_2)}{1-\pi(\varpi_2)}\right)=\beta_0+\beta_1\varpi_2 \tag{2}$$

利用 SPSS 进行 logistic 回归，可得拟合的模型：

$$\ln\left(\frac{\bar{\pi}(\varpi_2)}{1-\bar{\pi}(\varpi_2)}\right)=\bar{\beta}_0+\bar{\beta}_1\varpi_2=-0.055+1.103\omega_2$$

或

$$\bar{\pi}(\varpi_2)=\frac{\exp(-0.055+1.103\omega_2)}{1+\exp(-0.055+1.103\omega_2)}$$

其中 $\bar{\beta}_1=-0.055$ 的标准差为 $s(\bar{\beta}_1)=0.001$，检验 $\beta_1=0$ 的 wald 统计量的值为 146.357，p 值小于 0.0001，因此回归关系高度显著，利用回归得到的 logistic 模型进行预测得到准确率为 87.6%。

由于 $\bar{\beta}_1=-0.055>0$，说明检查者是否存活下来的概率随检查者的年龄的增加而减少，即患糖尿病的患者逐渐年轻化，一改人们之前认为年老者更易患糖尿病的想法，随着当代生活节奏的加快，工作压力的增大，患有糖尿病的年轻人也越来越多，这与之前研究年龄等级和患病间的关系的结论是类似的。

下面以检查者的性别，年龄，γ-谷氨酰基转移酶，血清白蛋白，钙为自变量拟合多元的 logistic 模型分析检查者的属性对其患病概率的影响。设 $\pi(x)$ 为设 $x=(\varpi_1,\varpi_2,\varpi_3,\varpi_4,\varpi_5)$ 下检查者存活下来的概率，建立如下的 Logistic 模型：

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1\varpi_1+\beta_2\varpi_2+\beta_3\varpi_3+\beta_4\varpi_4+\beta_5\varpi_5 \tag{3}$$

利用 SPSS 将整理后的数据进行 logistic 回归分析，得到(3)的参数估计以及其显著性检验结果，见表 3。

对于性别，γ-谷氨酰基转移酶，钙属性，$\bar{\beta}_1=0.607>0$，$\bar{\beta}_3=0.257>0$，$\bar{\beta}_5=0.009>0$，并且在 $\alpha=0.05$ 下显著非零，检查者的性别，γ-谷氨酰基转移酶，钙对其患病概率有一个正的影响，年龄，血清白蛋白对其患病概率有一个负的影响，说明检查者的性别为男性时对其患病的概率更大，检查者患病年龄逐渐趋于年轻化，血清白蛋白太低加大患病概率。

Table 3. Parameter estimation and significance test of the model
表 3. 模型的参数估计以及其显著性检验表

参数	自由度	参数估计值	标准差	Wald 卡方值	p 值
β_0	1	1.597	0.423	14.232	0.0001
β_1	1	0.607	0.187	195.290	$2*10^{-55}$
β_2	1	−1.149	0.002	0.207	0.0001
β_3	1	0.257	0.126	64.383	$1*10^{-15}$
β_4	1	−0.071	0.262	14.378	0.0001
β_5	1	0.009	0.242	14.378	0.0001

4. 风险指标

一般针对于检查者是否患有糖尿病三个较为重要的风险指标[7]：FPG、2h-PG、糖化血红蛋白(HbA1c)。一般来说 FPG、2h-PG 主要是常规监测中运用较多，HbA1c 用来评估出糖尿病患者近 90 天的血糖平均水平具有一定的科学性，糖化血红蛋白正常值小于 6.0，如果糖化血红蛋白大于等于 6.5 时，即高度怀疑其存在糖尿病的可能；需要行 75 克无水葡萄糖的糖耐量试验，如果血糖超出正常范围内，即可诊断为糖尿病；糖化血红蛋白化验会受到血红蛋白值的影响，如果存在贫血，糖化血红蛋白也会偏低；这时可以进一步化验糖化白蛋白以明确目前病情。

因此可以利用 HbA1c 指标参照训练好的 logistic 模型做一个风险指标系数[8]来衡量患者是否患有糖尿病的风险：

$$L = \mu_1 \pi(\varpi) + \mu_2 H \tag{4}$$

(3)中的 H 为标准化后的 HbA1c 值，$\pi(\varpi)$ 为某患者的各检查值确定的患病概率，给定两类参考值各自权重，综合考虑确定风险指标。其中 $\mu_1 = 0.7$，$\mu_2 = 0.3$，因此确定的最终风险指标系数：

$$L = 0.7\pi(\varpi) + 0.3H \tag{5}$$

由于 $\pi(\varpi)$ 为某患者的各检查值确定的患病概率，H 为标准化后的 HbA1c 值，他们的值均在 0~1 之间，因此由其得到的风险指标系数控制在 0~1 之间。其中，当 L 接近 1 是，患病风险显示最高，应及时检查治疗，当 L 接近 0 是，患病风险显示最低。

5. 结论

相比较于单个属性的一元 logistic 回归而言，多个属性的多元 logistic 回归的准确率有所下降，其准确率只达到了 79.1%，仍旧不能够较好地判定检查者的患病概率情况。寻找其中的原因，有两个限制讨论。

首先，关于样本的问题。在官网上所得到的样本并非是全部检查者的数据，真实情况一共有 500 条有效数据，如果这 500 名检查者是从所有检查者中随机选出，根据中心极限定理[9] [10] [11]，样本足够大，分析结论具有代表性，如果不是随机抽取，那么分析的结果就不可靠。

其次，官网给出的检查者属性数据并不完全，可能还有其它影响检查者存活概率的情况，比如说生育情况是否影响检查者的患病概率，检查者的饮食情况也会影响检查者的患病概率，检查者的职业是否影响患病概率，当考虑加入更多影响检查者存活概率的属性后，那么得到的多元 logistic 回归模型的准确率必定有很大的提升。

基金项目

本论文得到国家自然科学基金项目(11961008)；贵州省科技计划项目(黔科合基础[2019]1122 号)；贵州大学线上线下混合式课程建设项目(XJG202060)资助。

参考文献

[1] 邹小伟, 刘海燕, 主编. 苏州工业园区年鉴[M]. 苏州: 古吴轩出版社, 2019: 113.

[2] 2019 年人口健康共享杯大赛数据库[EB/OL]. https://www.kaggle.com/c/titanic/data

[3] David S. Moore 统计学的世界[M]. 北京: 中信出版社, 2003.

[4] 罗登菊, 戴家佳, 罗兴甸. 随机效应模型的复合分位数回归估计[J]. 贵州大学学报(自然科学版), 2019, 36(2): 96-100, 108.

[5] 高波, 王小乐, 李肖瑛, 赵静. 基于 Logistic 回归模型的航天器健康状态评估方法[C]//中国卫星导航系统管理办公室学术交流中心. 第十一届中国卫星导航年会论文集——S08 测试评估技术. 中国卫星导航系统管理办公室学术交流中心: 中科北斗汇(北京)科技有限公司, 2020: 6.

[6] 张彩云. 儿童病毒性脑炎急性期继发癫痫的药物控制及预后不良相关危险因素的 Logistic 回归分析[J]. 临床医学, 2020, 40(10): 66-68.

[7] 陈莹, 杨彩哲, 王良宸, 肖黎, 张妲, 王晨蕊, 陈红梅, 王璐宁. 趾臂指数对糖尿病足患者心脑血管事件发生风险的预测价值研究[J]. 中国全科医学, 2020, 23(34): 4332-4336+4348.

[8] 郭金生, 张生成. 一类具有垂直传染的非线性发生率的 SIS 传染病模型的稳定性分析[J]. 贵州大学学报(自然科学版), 2017, 34(3): 6-9.

[9] 孟祥飞, 王瑛, 李超, 亓尧, 孙赟. 独立不同分布不确定变量中心极限定理证明及其应用[J]. 上海交通大学学报, 2019, 53(10): 1230-1237.

[10] Ajay, J. and Yu, F.Y. (2020) Central Limit Theorems for Coupled Particle Filters. *Advances in Applied Probability*, **52**, 942-1001. https://doi.org/10.1017/apr.2020.27

[11] Dudley, R.M. (2013) Uniform Central Limit Theorems. Cambridge University Press, Cambridge.