Advances in Applied Mathematics 应用数学进展, 2022, 11(1), 108-115
Published Online January 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.111016

Hans汉斯

基于因子分析的二元Logistic回归对糖尿病预测的研究

徐　娇

长安大学理学院，陕西　西安

收稿日期：2021年12月13日；录用日期：2022年1月3日；发布日期：2022年1月18日

摘　要

目前全球糖尿病患者数量庞大，且数据在持续增长，有近一半人不知晓自己已经患病。为了尽早诊断及时治疗防控，本文建立了一个糖尿病初步诊断模型。本文选取UCI机器学习库中的糖尿病数据集，利用SPSS软件进行数据分析。首先利用因子分析方法对影响因素数据进行降维处理，得到公共因子和其对应数值，然后根据公共因子建立二元Logistic回归模型，根据模型可以预测是否患病。

关键词

糖尿病，因子分析，二元Logistic回归

Prediction of Diabetes by Binary Logistic Regression Based on Factor Analysis

Jiao Xu

School of Sciences, Chang'an University, Xi'an Shaanxi

Received: Dec. 13th, 2021; accepted: Jan. 3rd, 2022; published: Jan. 18th, 2022

Abstract

At present the number of diabetics worldwide is huge and growing, and nearly half of people don't know that they are sick. In order to diagnose and treat diabetes as soon as possible, this article establishes a preliminary diagnosis model of diabetes. We select diabetes data set from UCI machine learning library and use SPSS software for data analysis. First, the factor analysis method is used to reduce the dimension of the influencing factor data, and the common factors and their corresponding values are obtained. Then, a binary Logistic regression model is established according to the common factors, and the disease can be predicted according to the model.

文章引用: 徐娇. 基于因子分析的二元 Logistic 回归对糖尿病预测的研究[J]. 应用数学进展, 2022, 11(1): 108-115.
DOI: 10.12677/aam.2022.111016

Keywords

Diabetes, Factor Analysis, Binary Logistic Regression

1. 绪论

根据国际糖尿病联盟(IDF) 2019 年发布的第九版全球糖尿病概览，数据显示截止 2019 年，在 20 岁到 79 岁的人群中，共有约 4.63 亿糖尿病患者，其中中国糖尿病患者数居于首位，总人数约为 1.164 亿人，占全球 1/4。第二为印度，约有 7700 万糖尿病患者，第三为美国，有 3100 万糖尿病患者。根据联盟估算，中美两国花费在糖尿病相关的健康支出分别为 2946 亿美元与 1090 亿美元，而每年在中国因糖尿病而导致的死亡人数约为 83.4 万。全球约 4.63 亿 20~79 岁成人患有糖尿病，即每 11 个人中就有一个糖尿病患者，并且数据在持续快速增长。依照目前的趋势，预计到 2045 年将有 7 亿人患糖尿病。

其实，糖尿病本身并不可怕，而糖尿病带来的危害，主要来自糖尿病并发症对我们身体造成的损伤。糖尿病的并发症分为急性和慢性两种，急性并发症一般来势汹汹，在及时有效地处理后即可去无痕迹，而糖尿病的慢性并发症则发病隐匿、难以察觉，早期发现尚可进行纠正，一旦发展为中晚期，则会对身体造成不可逆转的巨大伤害。早期诊断出糖尿病，不间断的适当治疗能让患者远离糖尿病相关的致命并发症。但令人堪忧的是，我国的糖尿病患者的知晓率偏低，有 46.5%的患者未被诊断，不知晓已经患病。如何根据各项指标进行初步诊断，及时治疗防控十分重要。

对 1 型糖尿病而言，目前尚未出现安全有效的预防手段，治愈 1 型糖尿病目前还无法实现，但延迟其发生发展已经成为可能。大部分的 2 型糖尿病患者都在中低收入国家，对于 2 型糖尿病需要全社会都行动起来，采取综合性的预防措施。这些措施不仅针对 2 型糖尿病，同时也能降低其他慢性非传染病的风险。本文旨在建立一个糖尿病风险评估体系，预测个人在未来患有糖尿病的风险。

目前大量学者对糖尿病预测模型进行了深入研究，构建了许多预测模型。这些预测模型根据不同应用场景而建立，对于推动糖尿病的快速诊断进而及早干预防治具有重要作用。回归模型是一种研究因变量与自变量的预测性模型，综合考虑各种可能的危险因素，通过多元回归模型预测未来一定时间内糖尿病的发病概率[1]。

本文的研究方法：以糖尿病患病风险为研究对象，选择影响糖尿病预测的因素为目标变量，建立 Logistic 回归模型。本文选取来自 UCI 机器学习库中的“Pima Indians Diabetes”数据集。其中的预测变量包括怀孕次数、血糖、血压、皮褶厚度、胰岛素水平、BMI、糖尿病谱系功能、年龄。首先对预测指标进行因子分析，对数据进行降维处理，剔除数据间的相关性，确定最终指标；然后根据最终指标建立 Logistic 回归模型；最后对模型的预测效果进行检验，并结合研究结果对预防糖尿病提出相关建议。

2. 基本理论

2.1. 因子分析

因子分析是一种降维、简化数据的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个“抽象”的变量来表示其基本的数据结构。这几个抽象的变量被称作“因

子”，能反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而因子一般是不可观测的潜在变量。因子分析中的公共因子是不可直接观测但又客观存在的共同影响因素；每一个变量都可以表示成公共因子的线性函数与特殊因子之和，它的数学模型可表示为[2]：

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ a_{21} & \cdots & a_{2p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pm} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix} \tag{2.1}$$

即 $X = AF + \varepsilon$。其中 $X = \left(x_1, x_2, \cdots, x_p\right)^{\mathrm{T}}$ 是可观测的 p 维随机向量，每个分量代表一个指标或者向量。$F = \left(F_1, F_2, \cdots, F_m\right)^{\mathrm{T}}$ 中的 $F_1, F_2, \cdots, F_m$ 为 $m\left(m \le p\right)$ 个公因子变量，是各个原观测变量的表达式中都出现的因子，是相互独立的不可观测的理论变量。矩阵 A 称为因子载荷矩阵，a_{ij} 称为因子载荷，表示第 i 个原有变量和第 j 个公共因子变量的相关系数，a_{ij} 越大说明公共因子 F_j 和原有变量 X 的相关性越强。ε 为特殊因子，表示原有变量不能被公共因子变量所解释的部分，相当于多元线性回归分析中的残差部分。

因子分析利用了降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，根据相关性的大小把原始变量分组，使得同组内的变量之间相关性高，而不同组的变量间的相关性较低[3]。每组变量代表一个基本结构，并用一个不可观测的综合变量表示，这个基本结构就称为公共因子。抓住这些主要的因子就可以帮助我们对复杂的问题进行分析和解释。

2.2. Logistic 回归模型

2.2.1. 因变量为定性变量的回归模型

1) 定性变量：因变量只取两结果，当 $y = 0$ 时表示事件未发生，$y = 1$ 时表示事件发生。考虑简单的线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.2}$$

$$E\left(y_i\right) = \beta_0 + \beta_1 x_i \tag{2.3}$$

由于 y_i 是 0~1 型伯努利随机变量，得到如下概率

$$P\left(y_i = 1\right) = \pi_i$$

$$P\left(y_i = 0\right) = 1 - \pi_i$$

根据离散型随机变量期望定义，得

$$E\left(y_i\right) = 1\left(\pi_i\right) + 0\left(1 - \pi_i\right) = \pi_i \tag{2.4}$$

所以 $E\left(y_i\right) = \pi_i = \beta_0 + \beta_1 x_i$。

2) 误差项

对取值为 0 或 1 的因变量，误差项 $\varepsilon_i = y_i - \left(\beta_0 + \beta_1 x_i\right)$ 只能取两值

$$y_i = 1,\ \varepsilon_i = 1 - \left(\beta_0 + \beta_1 x_i\right) = 1 - \pi_i$$

$$y_i = 0,\ \varepsilon_i = -\left(\beta_0 + \beta_1 x_i\right) = -\pi_i$$

误差项是两点型离散分布，所以不能假设其是正态误差回归模型。

零均值异方差：误差项为零均值，其方差不相等

$$D\left(\varepsilon_i\right) = D\left(y_i\right) = \pi_i\left(1 - \pi_i\right) = \left(\beta_0 + \beta_1 x_i\right)\left(1 - \beta_0 - \beta_1 x_i\right) \tag{2.5}$$

若用多元线性回归方程分析因变量与自变量之间的定量关系

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m \tag{2.6}$$

3) 等式左边 y 取 0 或 1，等式右边可取任意实数，左右两边取值范围不对应。因此不能采用多元线性回归进行因变量为定性变量的拟合。

2.2.2. Logistic 回归模型

Logistic 函数的形式为[4]

$$f(x)=\frac{e^x}{1+e^x}=\frac{1}{1+e^{-x}} \tag{2.7}$$

其自变量的取值范围是$(-\infty,+\infty)$，函数值的取值范围为$(0,1)$。

因变量 y 本身只取0,1两离散值，不适于作为回归模型中的因变量，令

$$\pi_i=f(x_i)=\frac{1}{1+\exp(-(\beta_0+\beta_1x_i))} \tag{2.8}$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_1x_i \tag{2.9}$$

其中 π_i 是随机变量 y 取 1 的概率，其值在[0, 1]区间内连续变化，因此可用 π_i 代替 y 作为因变量。

设 y 是 0~1 型变量，n 组观测数据为$(x_{i1},\cdots,x_{ip},y_i)$，其中 $y_1,y_2,\cdots,y_n$ 是取值 0 或 1 的随机变量，

$$E(y_i)=\pi_i=f(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}) \tag{2.10}$$

Logistic 回归模型[5]

$$\pi_i=f(x_i)=\frac{e^{\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}}}{1+e^{\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}}} \tag{2.11}$$

于是 y_i 是均值为 $\pi_i=f(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})$ 的 0~1 型随机变量，概率函数为

$$P(y_i=1)=\pi_i$$

$$P(y_i=0)=1-\pi_i$$

可以把 y_i 的随机概率定义为

$$P(y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i},\ y_i=0,1;i=1,\cdots,n \tag{2.12}$$

于是 $y_1,y_2,\cdots,y_n$ 的似然函数为

$$L=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{2.13}$$

似然函数取对数，得

$$\ln L=\sum_{i=1}^{n}\left[y_i\ln\pi_i+(1-y_i)\ln(1-\pi_i)\right]=\sum_{i=1}^{n}\left[y_i\ln\frac{\pi_i}{1-\pi_i}+\ln(1-\pi_i)\right] \tag{2.14}$$

将式(2.14)带入得

$$\ln L=\sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})-\ln(1+\exp(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}))\right] \tag{2.15}$$

最大似然估计得到 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$。

3. 糖尿病预测模型

3.1. 数据处理

本文所取的数据是来自 UCI 机器学习库中的“Pima Indians Diabetes”，最初来自国家糖尿病消化肾脏疾病研究所。数据集中包含 768 位女性的相关数据，数据集由多个医学预测变量和一个目标变量组成。预测变量包括患者的怀孕次数、血糖、血压、皮褶厚度、胰岛素水平、BMI、糖尿病谱系功能、年龄，变量说明见表 1。数据集的目标是基于数据集中包含的某些诊断测量来诊断性的预测患者是否患有糖尿病。

Table 1. Variable declaration
表 1. 变量说明

X_1	怀孕次数
X_2	血糖
X_3	血压
X_4	皮褶厚度
X_5	胰岛素
X_6	BMI (身体指数) = (身高/体重)^2)
X_7	糖尿病谱系功能，根据家族史评估
X_8	年龄

数据清理的过程中，需要考虑以下几个方面：

1) 重复或者无关的数据；
2) 错误标识的数据或者多次出现相同的标识；
3) 缺失或空的数据点；
4) 异常值。

该数据为标准数据库，所以无重复以及无关数据，并且经检验没有空缺数据点。由于血压、血糖水平、身体指数一般不可能为 0，0 为异常数据点，所以过滤掉血压、血糖和身体指数各特征中为 0 值的行。处理后剩余有效数据 724 个。正常人的皮褶厚度一般不会小于 10 mm，但在该数据集中该参数共出现 227 次 0 值。因为出现大量异常值，所以我们暂时弃用该特征。

3.2. 因子分析

1) 因子分析适用性检验

利用 Spss 软件对糖尿病数据进行 KMO 和 Bartlett 球度适用性检验，结果如表 2 所示，一般认为 KMO 度量值若大于 0.5，则可以进行因子分析。且显著性 $p = 0$，说明原有变量之间存在一定的关联性，具备进行因子分析的条件。

2) 提取公因子

对数据进行因子分析，通过主成分分析法进行主成分的提取。在特征值为 1 的原则下，保留三个主因子，即 7 个变量归为 3 类。减少了运算量，但归类会造成信息损失，保留下来的信息为 64.493%，所损失的信息量较大，所以增加一个公共因子，使得损失信息量在可接受的范围内。下方差解释表 3，显示各主成分包含了各个原始变量总方差的情况，改进后保留信息 77.419%。

Table 2. KMO and Bartlett test
表 2. KMO 和巴特利特检验

KMO 取样适切性量数		0.593
巴特利特球形度检验	近似卡方	952.822
	自由度	28
	显著性	0.000

Table 3. Total variance explained
表 3. 总方差解释

成分	总计	初始特征值方差百分比	累积%	总计	载荷平方和方差百分比	累积%	总计	旋转载荷平方和方差百分比	累积%
1	1.980	28.288	28.288	1.980	28.288	28.288	1.709	24.418	24.418
2	1.533	21.905	50.194	1.533	21.905	50.194	1.355	19.359	43.777
3	1.001	14.299	64.493	1.001	14.299	64.493	1.309	18.703	62.480
4	0.881	12.587	77.080	0.881	12.587	77.080	1.022	14.600	77.080
5	0.654	9.339	86.419						
6	0.544	7.776	94.194						
7	0.406	5.806	100.000						

3) 公共因子命名

通过提取出来的 4 个公共因子，进行最大方差正交旋转，对原始因子载荷矩阵进行旋转，得到方差最大正交旋转矩阵，如表 4 所示。

Table 4. Rotated factor and Score matrix
表 4. 旋转因子矩阵与得分矩阵

变量	指标	旋转因子矩阵				因子得分矩阵			
		1	2	3	4	1	2	3	4
X_1	怀孕次数	0.847	−0.019	0.009	−0.002	0.520	−0.029	−0.117	0.045
X_2	年龄	0.869	0.078	0.114	0.012	0.517	0.030	−0.049	0.036
X_3	胰岛素	−0.156	0.856	−0.036	0.124	−0.097	0.675	−0.150	−0.032
X_4	血糖	0.279	0.749	0.232	0.002	0.125	0.562	0.033	−0.130
X_5	身体指数	−0.147	0.208	0.799	0.174	−0.206	0.011	0.651	0.091
X_6	血压	0.335	−0.045	0.774	−0.109	0.084	−0.146	0.613	−0.128
X_7	糖尿病谱系功能	0.018	0.099	0.051	0.982	0.052	−0.105	−0.029	0.992

根据旋转后的成分矩阵，可将 4 个公共因子进行命名。第一个因子 Z_1 在年龄和怀孕次数指标上具有较大载荷，第二个因子 Z_2 在胰岛素、和血糖指标上具有较大载荷，第三个因子 Z_3 在身体质量、血压上具有较大载荷，第四个因子 Z_4 在糖尿病谱系功能上具有较大载荷。可以发现，Z_1 所对应的评价指标是间接影响数据，Z_2 所对应的评价指标是血糖相关数据，Z_3 所对应的是其他身体数据，Z_4 代表糖尿病谱系。分别命名为间接因素、胰岛功能、身体素质、糖尿病谱系。

3.3. 二元 Logistic 回归

1) 霍斯默－莱梅肖检验

原假设 H_0：模型与观测值能很好拟合。结果如表 5 所示，$p = 0.279 > 0.05$，接受原假设，该回归模型可以较好拟合数据。

Table 5. Hosmer-Lemeshow test
表 5. 霍斯默－莱梅肖检验

卡方	自由度	显著性
9.808	8	0.279

2) 准确度如表 6，准确度 75.3%，说明模型预测较为准确。

Table 6. Forecast precision
表 6. 预测准确度

		预测是否患病		正确百分比
		0	1	
实际是否患病	0	413	62	86.9
	1	117	134	53.3
总体百分比				75.3

3) 由表 7，显著性 P 值均为 0，表示间接因素、胰岛功能和身体素质对于糖尿病诊断具有十分显著的影响。其影响程度由高到低排序如下：胰岛功能 > 身体素质 > 间接因素 > 糖尿病谱系。

Table 7. Logistic regression model
表 7. Logistic 回归分析

	B	标准误差	瓦尔德	自由度	显著性	Exp (B)	EXP (B)的 95%置信区间	
							下限	上限
间接因素	0.655	0.091	51.667	1	0.000	1.925	1.610	2.301
胰岛功能	0.869	0.100	75.364	1	0.000	2.386	1.960	2.903
身体素质	0.679	0.097	49.180	1	0.000	1.971	1.631	2.383
糖尿病谱系	0.388	0.093	17.329	1	0.000	1.474	1.228	1.769
常量	−0.828	0.095	76.333	1	0.000	0.437		

4) 由多因素的回归分析，建立二元 Logistic 回归方程

$$\text{Logit}P = -0.828 + 0.655z_1 + 0.869z_2 + 0.679z_3 + 0.388z_4$$

其中，$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$。

4. 总结

根据本文得到的模型，可以对是否患有糖尿病进行初步诊断。且从回归方程可以看出，在诊断过程中，胰岛功能和身体素质对是否患有糖尿病影响最为明显。因此，为预防糖尿病，需要有良好的生活习惯，强健体魄，规律饮食，从而保证自己身体处于良好的状态。同时糖尿病还受年龄和怀孕次数影响，而年龄与怀孕次数呈正相关，年龄越大，怀孕次数越多。因此，随着年龄增长，要更加注重锻炼和摄糖量。同时糖尿病会受到遗传因素影响，所以如果家中有糖尿病患者，那就要格外注意，从小预防。

本文的主要目的是利用 Logistic 回归模型对糖尿病患病风险进行预测，同时结合因子分析的思想，对数据进行降维处理。总体来说，用于初步诊断，本文得到的模型预测效果较好，准确率达到了 75.3%。其中胰岛功能、身体素质对诊断结果影响较大。

参考文献

[1] 苏天培. 基于 XGBoost 的糖尿病风险预测[J]. 科技视界, 2019(2): 160-161.

[2] 何晓群. 多元统计分析[M]. 北京: 中国人民大学出版社, 2004.

[3] 李丹. 基于因子分析的我国商业银行经营绩效评价研究[D]: [硕士学位论文]. 上海: 东华大学, 2013.

[4] 王济川, 郭志刚. Logistic 回归模型:方法与应用[M]. 北京: 高等教育出版社, 2001.

[5] 蒋雁. Logistic 回归及其在上市公司信用风险度量中的应用[D]: [硕士学位论文]. 大连: 大连理工大学, 2019.