Pure Mathematics 理论数学, 2024, 14(2), 669-681
Published Online February 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.142067

Hans汉斯

Logistic模型分析脑中风影响因素

褚明阳

南京审计大学统计与数据科学学院，江苏 南京

收稿日期：2024年1月5日；录用日期：2024年1月31日；发布日期：2024年2月29日

摘　要

脑中风是一种急性脑血管疾病，是由于脑部血管突然破裂或因血管阻塞导致血液不能流入大脑而引起脑组织损伤的一组疾病。城乡合计脑中风已成为我国第一位死亡原因，也是中国成年人残疾的首要原因。其具有发病率高、死亡率高和致残率高的特点。本文通过对于脑中风数据的描述性分析，建立Logistic回归模型，得到了患脑中风的五个显著的重要因素：年龄、高血压、心脏病、工作类型和平均血糖水平，并得到了经验回归方程。

关键词

广义线性模型，Logistic模型，脑中风

Logistic Model Was Used to Analyze the Influencing Factors of Cerebral Apoplexy

Mingyang Chu

School of Statistics and Data Science, Nanjing Audit University, Nanjing Jiangsu

Received: Jan. 5th, 2024; accepted: Jan. 31st, 2024; published: Feb. 29th, 2024

Abstract

Stroke is an acute cerebrovascular disease, which is a group of diseases caused by the sudden rupture of blood vessels in the brain or the blockage of blood vessels that prevents blood from flowing into the brain. Combined urban and rural cerebral apoplexy has become the first cause of death in China and the leading cause of adult disability in China. It has the characteristics of high morbidity, high mortality and high disability rate. In this paper, through descriptive analysis of cerebral stroke data, Logistic regression model was established, and five significant factors of cerebral stroke were obtained: age, hypertension, heart disease, job type and average blood sugar level, and empirical regression equation was obtained.

文章引用: 褚明阳. Logistic 模型分析脑中风影响因素[J]. 理论数学, 2024, 14(2): 669-681.
DOI: 10.12677/pm.2024.142067

Keywords

Generalized Linear Models, Logistic Model, Cerebral Apoplexy

1. 研究背景

脑中风又称脑卒中，是一种急性脑血管疾病。脑中风是由脑部血管破裂或是血管堵塞导致血液无法流入大脑而引起的一种脑组织损伤疾病，其包括两种：缺血性以及出血性卒中。其中，缺血性卒中的发病比例略高于出血性卒中的发病比例，大概占总发病比例的 60%~70%。常见的缺血性卒中多发生在年龄 40 周岁以上的老年人，颈内动脉和椎动脉闭塞常常是引发老年人缺血性卒中的主要原因。相较于缺血性卒中而言，出血性卒中的危害更加明显，死亡率远高于缺血性卒中。

目前，脑中风已经一跃成为我国致死原因第一位，也是我国成年人残疾的首要原因。对于脑中风的治疗，目前而言预防是最为主要的措施，而高血压是导致脑中风的重要因素之一，因此对于血压的观测应是脑中风患者的重中之重。

关于脑中风的研究，目前国内已有许多学者发表了许多文献研究。刘艳娇[1]通过临床流行病学方法研究了肥胖人痰湿体质与脑中风的相关性并证明了肥胖人痰湿体质是引发脑中风的相关因素之一。赵孔华与张沁园[2]通过对缺血性脑中风的研究分析出其发病机理以及治疗原则。杜恩[3]通过老年脑中风患者观察组与对照组的比较说明了早期康复治疗对于老年脑中风偏瘫患者的重要性。甘勇、杨婷婷以及刘建新等人[4]通过对脑中风的研究，分析出了影响脑中风的主要因素。本文选取了该篇文献内的相关指标来使用 Logistic 回归模型得到几个重要影响因素。

2. 广义线性模型

2.1. 模型介绍

在数据分析的过程中，很多分析方法和模型往往要求目标变量(数据)服从某些假设如正态分布、方差齐次等。一般来说，如果数据不能服从这些假设，那么采用对应的方法或模型获得的结果往往不可信。例如，我们经常使用的经典模型，即形如 $y = kx + b$ 的一般线性模型就要求数据(目标变量)必须满足正态分布和残差的方差齐次。然而，在实际科研工作中，很多数据往往不能满足以上条件。这种情况就要求我们寻找一种没有以上假设的方法来替代存在假设的模型如：一般线性模型。此时就产生了广义线性模型(GLM)。

在统计学中，广义线性模型(GLM)是普通线性回归的灵活概括，它允许响应变量具有除正态分布以外的误差分布模型。GLM 通过允许线性模型通过链接函数与响应变量相关，并允许每种度量的方差大小成为其预测值的函数，从而推广线性回归。广义线性模型由 John Nelder 和 Robert Wedderburn 制定，作为统一各种其他统计模型的一种方式，包括线性回归，逻辑回归和泊松回归。他们提出了一种迭代重加权的最小二乘方法，用于模型参数的最大似然估计。最大似然估计仍然很流行，并且是许多统计计算包上的缺省方法。其他方法，包括贝叶斯方法和最小二乘拟合方差稳定响应，已经开发出来。

线性回归模型主要适用于因变量为连续性(特别是服从正态分布)的随机变量的情况，GLM 通过一个已知的连接函数将因变量的数学期望与自变量的线性函数连接起来，并将因变量的分布由正态分布推广到广义指数分布族，GLM 可以处理因变量为一些离散型、连续性随机变量的回归问题。

在线性回归模型中，设因变量 Y 与自变量 $X_1,\cdots,X_p$ 的 n 组观测值：$\left(Y_i;X_{i1},\cdots,X_{ip}\right),i=1,\cdots,n$，其有两个基本要素：$Y_1,\cdots,Y_n$ 相互独立，服从同方差的正态分布 $N\left(\mu_i,\sigma^2\right)$；$\mu_i=E\left(Y_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j$，通常取 $X_{i1}=1,i=1,\cdots,n$。

在广义线性模型中，将 Y_i 的分布由正态分布推广到指数族分布；将 μ_i 推广到它的一个单调、可微函数(称为连接函数) $\eta_i=g\left(\mu_i\right)$，使得 $\eta_i=g\left(\mu_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j,i=1,\cdots,n$。

称随机变量 $Y_1,\cdots,Y_n$ 满足广义线性模型，如果：$Y_1,\cdots,Y_n$ 相互独立且服从指数族分布；对于某单调、可微的连接函数 $g(\bullet)$，有 $g\left(\mu_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j,i=1,\cdots,n$，其中 $\mu_i=E\left(Y_i\right),i=1,\cdots,n$。指数族分布包含诸如二项分布、Poisson 分布、正态分布、Γ 分布等常见的离散型和连续性分布。对指数族分布的定义为，若 Y (一元) 概率分布 (离散) 或概率密度 (连续) 有如下形式：$f\left(y;\theta,\phi\right)=\exp\left(\frac{y\theta-b\left(\theta\right)}{a\left(\phi\right)}+c\left(y,\phi\right)\right)$，其中 $a(\cdot),b(\cdot),c(\cdot,\cdot)$ 为已知连续函数，θ 称为自然参数(表示位置)，ϕ 称为散度参数(表示尺度)，则称 Y 服从指数族分布。通常，θ 与 $E\left(Y\right)=\mu$ 有关，是我们所感兴趣的参数，ϕ 与 $Var\left(Y\right)=\sigma^2$ 有关，通常作为多余参数。

一般连接函数是，连接函数 $g(\cdot)$ 是将自变量第 i 组观测值的线性组合 $\sum_{j=1}^{p}X_{ij}\beta_j$ 与 $\mu_i=E\left(Y_i\right)$ 联系起来的函数，即 $g\left(\mu_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j,i=1,\cdots,n$。常见的连接函数有：对数函数，$g\left(\mu\right)=\ln\mu,\left(\mu>0\right)$；Logit 或 Logistic 函数：$g\left(\mu\right)=\ln\left(\mu/\left(1-\mu\right)\right),\left(0<\mu<1\right)$；Probit 函数或 Gauss 函数：$g\left(\mu\right)=\Phi^{-1}\left(\mu\right),\left(0<\mu<1\right)$；重对数函数：$g\left(\mu\right)=\ln\left(-\ln\mu\right),\left(0<\mu<1\right)$；互补重对数函数：$g\left(\mu\right)=\ln\left(-\ln\left(1-\mu\right)\right),\left(0<\mu<1\right)$。

典型连接函数是随机变量 Y 服从指数族分布，典型连接函数 $g\left(\mu\right)$ 满足 $g\left(E\left(Y\right)\right)=g\left(\mu\right)=\theta$。此时，$b'\left(\theta\right)=E\left(Y\right)=\mu=g^{-1}\left(\theta\right)$，$b''\left(\theta\right)=V\left(\mu\right)=\frac{\mathrm{d}\left(b'\left(\theta\right)\right)}{\mathrm{d}\theta}=\frac{\mathrm{d}\left(b'\left(\theta\right)\right)}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}\theta}=\left(\frac{\mathrm{d}\theta}{\mathrm{d}\mu}\right)^{-1}=\frac{1}{g'\left(\mu\right)}$。

当 $Y\sim N\left(\mu,\sigma^2\right)$：$\theta=\mu$，典型连接：$g\left(\mu\right)=\theta=\mu$；$Y\sim P\left(\mu\right)$：$\theta=\ln\mu$，典型连接：$g\left(\mu\right)=\theta=\ln\mu$；$Y\sim B\left(m,\mu\right)/m$：$\theta=\ln\left(\frac{\mu}{1-\mu}\right)$，典型连接：$g\left(\mu\right)=\theta=\ln\left(\frac{\mu}{1-\mu}\right)$，即 Logit (或 Logistic)函数；$Y\sim\Gamma\left(\alpha,\mu\right)/\alpha$：$\theta=1/\mu$，典型连接：$g\left(\mu\right)=\theta=1/\mu$。

设 $\left(Y_i;X_{i1},\cdots,X_{ip}\right),i=1,\cdots,n$ 为因变量 Y 和自变量 $X_1,\cdots,X_p$ 的观测值，若：$Y_1,\cdots,Y_n$ 相互独立，且对每个 i，Y_i 服从指数族分布：$Y_i\sim f\left(y;\theta_i,\phi_i\right)=\exp\left(\frac{y\theta_i-b\left(\theta_i\right)}{a\left(\phi_i\right)}+c\left(y,\phi_i\right)\right)$；$g\left(\mu_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j$ 其中 $\mu_i=E\left(Y_i\right),i=1,\cdots,n$，则称 Y 与 $X_1,\cdots,X_p$ 服从广义线性模型(GLM)。

广义线性模型有十分广泛的应用背景，例如：研究人类某种疾病的发病率与人的性别、年龄、家庭经济情况、职业、自然环境情况等的关系，我们可以用典型连接函数创建 Logistic 模型，通过检验 β_j 是否显著为 0，了解哪些因素是影响该疾病发病率的主要因素，各因素影响发病率的强度等等。还可以研

究某地区在某时段内矿难发生次数与矿山类型、企业类型、企业管理水平、企业规模、安全资金投入等等因素的关系，我们可以建立 Poisson 模型或者对数线性模型来分析。

2.2. GLM 的极大似然估计

设有数据：$\left(Y_i;X_{i1},\cdots,X_{ip}\right),i=1,\cdots,n$，$Y_i$ 服从参数为 θ_i 和 ϕ_i 的指数族分布，其概率分布或密度：$f\left(y;\theta_i,\phi_i\right)=\exp\left(\dfrac{y\theta_i-b\left(\theta_i\right)}{a\left(\phi_i\right)}+c\left(y,\phi_i\right)\right)$。设 $\mu_i=E\left(Y_i\right)$，选定连接函数 $g\left(\cdot\right)$，考虑广义线性模型：$g\left(\mu_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j,i=1,\cdots,n$。

求参数 $\beta_1,\cdots,\beta_p$ 的最大似然估计，$Y_1,\cdots,Y_n$ 的对数似然函数为：

$$\ln L\left(\beta_1,\cdots,\beta_p\right)=\ln\left(\prod_{i=1}^{n}f\left(Y_i;\theta_i,\phi_i\right)\right)=\sum_{i=1}^{n}\left(\frac{Y_i\theta_i-b\left(\theta_i\right)}{a\left(\phi_i\right)}+c\left(Y_i,\phi_i\right)\right)$$

由于 θ_i 通过 μ_i 与 $\beta_1,\cdots,\beta_p$ 相联系，而 $c\left(Y_i,\phi_i\right)$ 与 $\beta_1,\cdots,\beta_p$ 无关，所以 $\beta_1,\cdots,\beta_p$ 的最大似然估计即是下列方程组的解：

$$\frac{\partial\ln L\left(\beta_1,\cdots,\beta_p\right)}{\partial\beta_r}=\frac{\partial}{\partial\beta_r}\sum_{i=1}^{n}\left(\frac{Y_i\theta_i-b\left(\theta_i\right)}{a\left(\phi_i\right)}\right)=0,r=1,\cdots,p$$

$$\mu_i=b'\left(\theta_i\right),\ V\left(\mu_i\right)=b''\left(\theta_i\right),\ g\left(\mu_i\right)=\sum_{j=1}^{p}X_{ij}\beta_j,\ \mu_i=g^{-1}\left(\sum_{j=1}^{p}X_{ij}\beta_j\right)。$$

利用链式求导法则得：

$$\frac{\partial\theta_i}{\partial\beta_r}=\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_r}=\left(\frac{\partial\mu_i}{\partial\theta_i}\right)^{-1}\frac{\partial\mu_i}{\partial\beta_r}=\frac{1}{b''\left(\theta_i\right)}\frac{\partial\mu_i}{\partial g\left(\mu_i\right)}\frac{\partial g\left(\mu_i\right)}{\partial\beta_r}=\frac{1}{b''\left(\theta_i\right)}\frac{1}{g'\left(\mu_i\right)}X_{ir}=\frac{X_{ir}}{V\left(\mu_i\right)g'\left(\mu_i\right)}$$

$$\frac{\partial b\left(\theta_i\right)}{\partial\beta_r}=\frac{\partial b\left(\theta_i\right)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta_r}=b'\left(\theta_i\right)\frac{\partial\theta_i}{\partial\beta_r}=\frac{\mu_iX_{ir}}{V\left(\mu_i\right)g'\left(\mu_i\right)}$$

代入似然方程得 $\sum_{i=1}^{n}\dfrac{\left(Y_i-\mu_i\right)X_{ir}}{a\left(\phi_i\right)V\left(\mu_i\right)g'\left(\mu_i\right)}=0$，其中 $\mu_i=g^{-1}\left(\sum_{j=1}^{p}X_{ij}\beta_j\right)$。对很多有重要应用背景的指数族分布，有 $a\left(\phi_i\right)=a_i\phi$，$a_i$ 已知。这时似然方程为 $\sum_{i=1}^{n}\dfrac{\left(Y_i-\mu_i\right)X_{ir}}{a_iV\left(\mu_i\right)g'\left(\mu_i\right)}=0$。

讨论一般情形下的似然方程的迭代最小二乘解。$\sum_{i=1}^{n}\dfrac{\left(Y_i-\mu_i\right)X_{ir}}{a\left(\mu_i\right)V\left(\mu_i\right)g'\left(\mu_i\right)}=0$。在 $a\left(\phi_i\right)=a_i\phi$ 的情况下，似然方程为 $\sum_{i=1}^{n}\dfrac{\left(Y_i-\mu_i\right)X_{ir}}{a_iV\left(\mu_i\right)g'\left(\mu_i\right)}=0$。令 $Z_i=g\left(\mu_i\right)+\left(Y_i-\mu_i\right)g'\left(\mu_i\right)$，则

$E\left(Z_i\right)=E\left(g\left(\mu_i\right)+\left(Y_i-\mu_i\right)g'\left(\mu_i\right)\right)=g\left(\mu_i\right)$，由于 $Var\left(Y_i\right)=V\left(\mu_i\right)a\left(\phi_i\right)=a_i\phi V\left(\mu_i\right)$，因此

$$Var\left(Z_i\right)=Var\left(Y_i\right)\left(g'\left(\mu_i\right)\right)^2=a_i\phi V\left(\mu_i\right)\left(g'\left(\mu_i\right)\right)^2=\tilde{a}_i\phi。$$

从而有：

$$\sum_{i=1}^{n}\frac{Z_i-g\left(\mu_i\right)}{\tilde{a}_i}X_{ir}=\sum_{i=1}^{n}\frac{\left(Y_i-\mu_i\right)g'\left(\mu_i\right)}{a_iV\left(\mu_i\right)\left(g'\left(\mu_i\right)\right)^2}X_{ir}=\sum_{i=1}^{n}\frac{Y_i-\mu_i}{a_iV\left(\mu_i\right)}\frac{X_{ir}}{g'\left(\mu_i\right)}$$

从而似然方程变为 $\sum_{i=1}^{n}\frac{Z_i-g(\mu_i)}{\tilde{a}_i}X_{ir}=0$， $E(Z_i)=\sum_{j=1}^{p}X_{ij}\beta_j$。

若 Z_i 已知，则 $\beta_1,\cdots,\beta_p$ 的加权最小二乘估计为 $\hat{\beta}=\left(X^{\mathrm{T}}WX\right)^{-1}X^{\mathrm{T}}WZ$，其中，$W=Diag(w_1,\cdots,w_n)$，$w_i=\frac{1}{\tilde{a}_i}=\frac{1}{a_iV(\mu_i)\left(g'(\mu_i)\right)^2}$，$Z=(Z_1,\cdots,Z_n)^{\mathrm{T}}$。

实际上，Z 是未知的，可用迭代加权 LS 法：

1) 给定 $\mu_1,\cdots,\mu_n$ 的一组初值 $\mu_1^{(0)},\cdots,\mu_n^{(0)}$，比如 $\mu_i^{(0)}=Y_i$

2) 计算 $g(\mu_i)$ 的初值 $\eta_i^{(0)}=g\left(\mu_i^{(0)}\right)$

3) 由 $Z_i=g(\mu_i)+(Y_i-\mu_i)g'(\mu_i)$，计算 Z_i，w_i 的初值：

$$Z_i^{(0)}=\eta_i^{(0)}+\left(Y_i-\mu_i^{(0)}\right)g'\left(\mu_i^{(0)}\right),$$

$$w_i^{(0)}=\frac{1}{a_iV\left(\mu_i^{(0)}\right)\left(g'\left(\mu_i^{(0)}\right)\right)^2},$$

$$Z^{(0)}=\left(Z_1^{(0)},\cdots,Z_n^{(0)}\right)^{\mathrm{T}},$$

$$W^{(0)}=Diag\left(w_1^{(0)},\cdots,w_n^{(0)}\right)。$$

计算 $\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 的第一次估计：$\hat{\beta}^{(1)}=\left(X^{\mathrm{T}}W^{(0)}X\right)^{-1}X^{\mathrm{T}}W^{(0)}Z^{(0)}$

4) 令 $\eta^{(1)}=\left(\eta_1^{(1)},\cdots,\eta_n^{(1)}\right)^{\mathrm{T}}=X\hat{\beta}^{(1)}$，这里 $\eta_i^{(1)}=g\left(\mu_i^{(1)}\right)=\sum_{j=1}^{p}X_{ij}\hat{\beta}_j^{(1)}$，$\mu_i^{(1)}=g^{-1}\left(\eta_i^{(1)}\right)$。由 $\mu_1^{(1)},\cdots,\mu_n^{(1)}$，根据 3)计算得 $Z_i^{(1)}=\eta_i^{(1)}+\left(Y_i-\mu_i^{(1)}\right)g'\left(\mu_i^{(1)}\right)$，$w_i^{(1)}=\frac{1}{a_iV\left(\mu_i^{(1)}\right)\left(g'\left(\mu_i^{(1)}\right)\right)^2}$，由此求得 β 的第二次迭代估计：

$$\hat{\beta}^{(2)}=\left(X^{\mathrm{T}}W^{(1)}X\right)^{-1}X^{\mathrm{T}}W^{(1)}Z^{(1)}$$

5) 重复上述步骤得 β 的迭代估计式：$\hat{\beta}^{(m+1)}=\left(X^{\mathrm{T}}W^{(m)}X\right)^{-1}X^{\mathrm{T}}W^{(m)}Z^{(m)}$，直到 $\hat{\beta}^{(m+1)}$ 收敛。

3. Logistic 模型

3.1. 模型

Logistic 模型最早在 20 世纪四五十年代由 Berkson，Dyke 和 Patterson 等人使用过。当因变量 Y 是 0~1 变量(二值变量)时，即 Y 表示分两类的类别，用取值 1 和 0 表示，将取值 1 称为成功，我们关心的时成功的概率 $p=P(Y=1)$。这是一个[0, 1]区间内的值。如果把 Y 当作一般因变量做线性回归，会给出不合理的结果，比如负值，另外线性回归假定误差项为正态分布在这里也不适用。

如果 Y 是 m 次试验中成功的次数，这时可以设 Y 服从二项分布 $B(m,p)$，关心的是成功概率 p，也可以归入相同的模型。

为此考虑对应于二项分布的广义线性回归模型：

$$Y\sim B(m,p),$$

$$g(p)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$$

其中 m 为试验次数，Y 为 m 次试验中成功次数，p 为给定 $x_1,\cdots,x_k$ 条件下的成功概率。一般取联系函数 $g(p)$

为 Logit 函数 $\text{logit}(p)=\ln\left(p/(1-p)\right)$。此模型称为 Logistic 模型， $p/(1-p)$ 是成功概率与失败概率的比值，称为发生比。 $\ln\left(p/(1-p)\right)$ 称为对数发生比。

Logistic 模型有如下的模型假定：因变量表示成败型结果，为零一变量或者已知试验次数中成功次数；各个观测独立；如果某个观测的试验次数为 m，成功概率为 p，则因变量方差为 $mp(1\text{-}p)$，方差不等于常数值，与期望值 mp 有关系；对数发生比 $\ln\left(p/(1-p)\right)$ 与自变量之间为线性关系。

3.2. 最大似然估计

考虑观测的因变量是二值因变量情形。数据为 (x_i,y_i)，一个观测对似然函数的贡献为：$p^{y_i}(1-p_i)^{1-y_i}$，其中 $p_i=\text{logit}^{-1}(a+bx_i)=\dfrac{e^{a+bx_i}}{1+e^{a+bx_i}}$，对数似然函数为：

$$\begin{aligned} l(a,b) &= \sum_{i=1}^{n}\left\{y_i\ln p_i+(1-y_i)\ln(1-p_i)\right\} \\ &= \sum_{i=1}^{n}\left\{y_i\ln\frac{p_i}{1-p_i}+\ln(1-p_i)\right\} \\ &= \sum_{i=1}^{n}\left\{y_i(a+bx_i)-\ln\left(1+e^{a+bx_i}\right)\right\} \end{aligned}$$

需要用数值迭代算法求最大似然估计。若样本为满足模型的随机样本，最大似然估计是相合估计，渐近有效估计，具有渐近正态分布。得到最大似然估计 $(\hat{a},\hat{b})$ 后，一般用信息阵的逆矩阵估计其协方差阵，信息阵为：

$$I(a,b)=-E\left[\frac{\partial^2 l(a,b)}{\partial(a,b)\partial(a,b)^T}\right]$$

在用数值迭代算法计算最大似然估计时，一般会得到最大值点处的对数似然函数的海色阵，加上负号并求逆矩阵，就可以作为参数协方差阵估计。有了参数估计的协方差阵估计，就得到了参数估计的标准误差。设参数估计为 $\hat{\theta}$，估计的方差开平方根作为其抽样分布标准差估计，称为标准误差，记为 $SE(\hat{\theta})$。令 $Z=\dfrac{\hat{\theta}}{SE(\hat{\theta})}$，可以用 Z 统计量检验 $H_0:\theta=0$，在 H_0 成立，大样本且满足适当正则性条件时近似服从标准正态分布。也可用 Z^2 近似服从 $\chi^2(1)$ 分布进行检验。这样对回归系数进行的检验称为 Wald 类型的检验。此检验对中小样本有一定缺陷，参数的实际数值绝对值较大时，容易错判成不显著。另一种可用的方法是利用偏差统计量的方法。

4. 数据选取、预处理与描述

根据世界卫生组织(WHO)的数据，中风是全球第二大死亡原因，约占总死亡人数的 11%。从 kaggle 上选取脑中风数据集，共有 5110 条数据，包含 id、性别等共 12 个指标。脑中风数据集包含指标如表 1 所示。

首先 BMI 中存在的缺失值删除。将性别指标中的男性改为 1，女性改为 0，由于只存在一例其他性别，并且女性有 2994 人，男性有 2115 人，因此将其设为人数较多女性，这样不至于使误差过大。婚姻状况上，未婚改为 0，曾今结过婚改为 1。工作类型上，孩子、政府工作、从不工作、私人、自由职业分别对应设置为 1~5。居住类型上，城市改为 1，农村改为 0。在吸烟状况上，曾经吸烟、从不吸烟、吸烟、不明分别对应设置为 0~3。

Table 1. Indicator specification
表 1. 指标说明

变量类型	变量名称	变量说明
自变量	id	唯一标识符
	性别	男、女或其他
	年龄	患者年龄
	高血压	0：没有高血压 1：有高血压
	心脏病	0：没有心脏病 1：有心脏病
	曾经结过婚	“不”或“是”
	工作类型	孩子、政府工作、从不工作、私人、自由职业
	居住类型	农村、城市
	平均血糖水平	血糖水平
	体重指数	身体质量指数
	吸烟状况	曾经吸烟、从不吸烟、吸烟、不明
因变量	中风	0：没有中风 1：有中风

查看处理后数据中的脑中风情况，如图 1 所示。

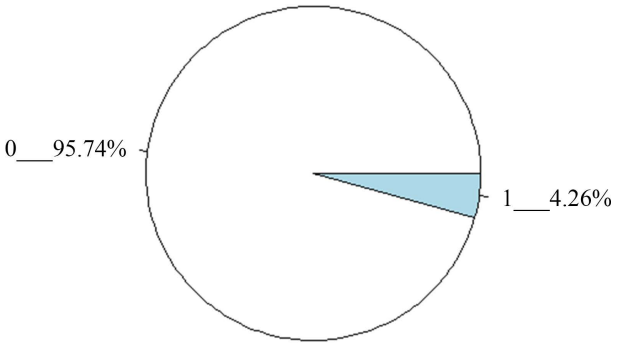

Figure 1. Cerebral apoplexy ratio
图 1. 脑中风比例

在 4909 条数据中，脑中风的人数有 209 人，占比为 4.26%。没有脑中风的有 4700 人，占比为 95，74%。在分性别查看患脑中风的情况，如图 2 所示。

对于选取的 2898 名女性中，患脑中风的人数为 120 人，占比为 4.14%。选取的男性人数为 2011 人，患脑中风的人数为 89 人，占比为 4.43%。可以看出，男性、女性患脑中风的比例相差并不是很大，男性要稍微高一点。

再按年龄来看脑中风情况，如图 3 所示。

我们可以看出，脑中风现象大约是从 30 岁开始少量出现，并且随着年龄段的增加，脑中风的人数也逐渐增加，50~80 岁之间为脑中风的高发年龄段，同时，脑中风率也有增高的趋势。

再按高血压分类来看脑中风情况。如图 4 所示。

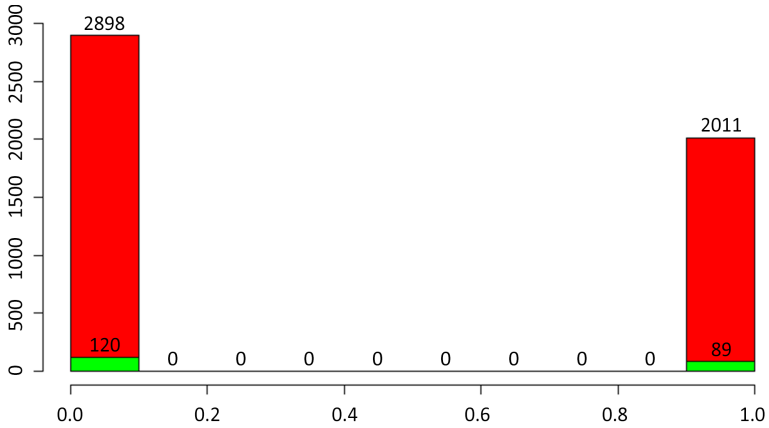

Figure 2. Number of cerebral strokes by sex
图 2. 按性别分类脑中风的人数

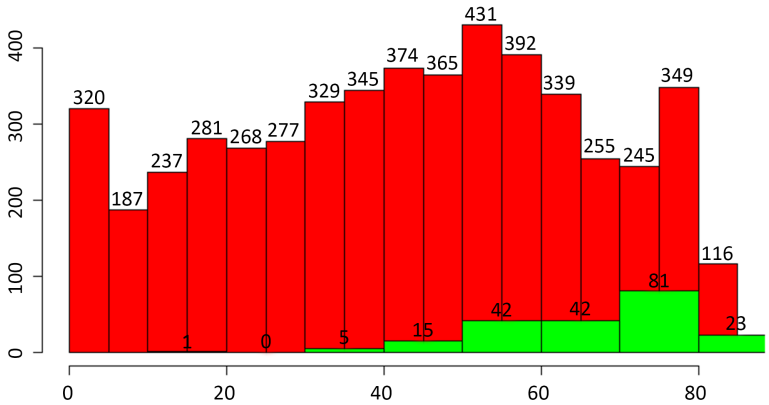

Figure 3. Cerebral stroke by age
图 3. 按年龄分类的脑中风情况

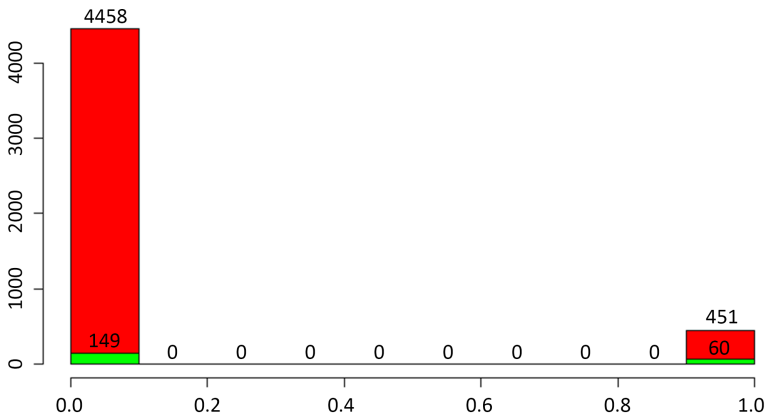

Figure 4. Cerebral stroke classified by hypertension
图 4. 按高血压分类的脑中风情况

不患有高血压的 4458 人中，有 149 人患了脑中风，比例为 3.34%，患有高血压的 451 人中，有 60 人患了脑中风，比例为 13.3%。可以看出，患有高血压的人群得脑中风的比例比不患高血压的人得脑中风的比例高很多，高了约有 10%。

再按心脏病分类来看脑中风情况。如图 5 所示。

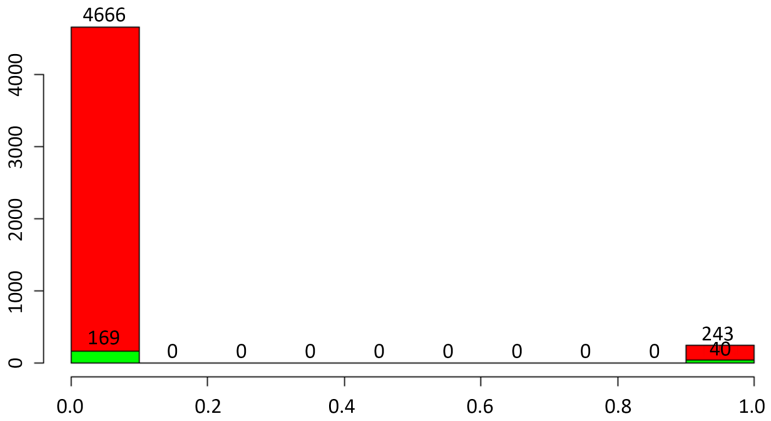

Figure 5. Cerebral stroke classified by heart disease
图 5. 按心脏病分类的脑中风情况

不患心脏病的 4666 人中，有 169 人患了脑中风，比例为 3.62%，患心脏病的 243 人中，有 40 人患了脑中风，比例为 16.46%。可以看出，患有心脏病的人群得脑中风得比例比不患心脏病的人得脑中风的比例高很多，高了约有 13%。

再按结婚状况分类来看脑中风情况。如图 6 所示。

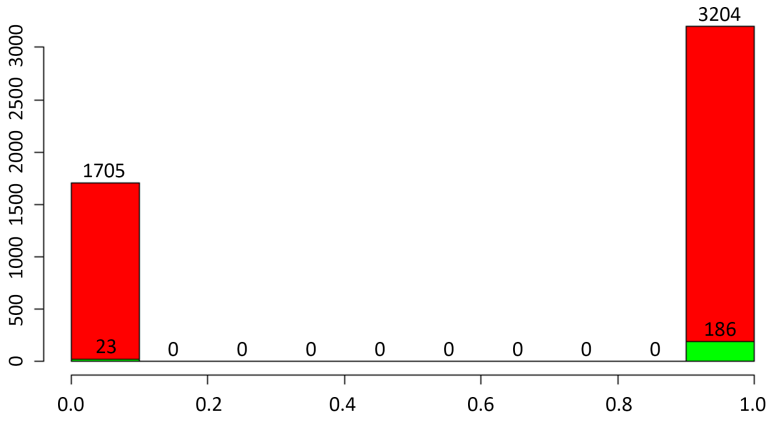

Figure 6. Cerebral stroke classified by marital status
图 6. 按结婚情况分类的脑中风情况

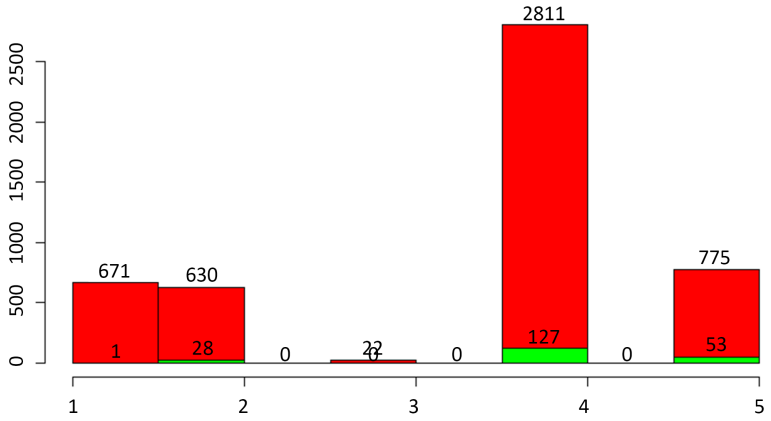

Figure 7. Cerebral stroke classified by type of work
图 7. 按工作类型分类的脑中风情况

在未婚的 1705 人中，只有 23 人患了闹中风，比例为 1.35%，曾经结过婚的 3204 人中，有 186 人患了脑中风，比例为 5.8%。可以看出，结过婚的人患脑中风的比例要比未婚的人高，这可能主要是和年龄有关，曾经结过婚的人一般年龄较大，而未婚的人年龄较小。

再按工作类型分类来看脑中风情况。如图 7 所示。

可以看到，各个工作类型的人数分别为 671，630，22，2811，775，他们的脑中风情况分别为 1，28，0，127，53，比例分别为 0.15%，4.44%，0，4.51%，6.84%。可以看到，小孩与从不工作患脑中风的比例最低，几乎为 0；自由职业者患脑中风的概率最高，约有 7%；政府工作与私人工作的患脑中风的比例相差不大，都约为 4.5%。

再按居住地类型分类来看脑中风情况。如图 8 所示。

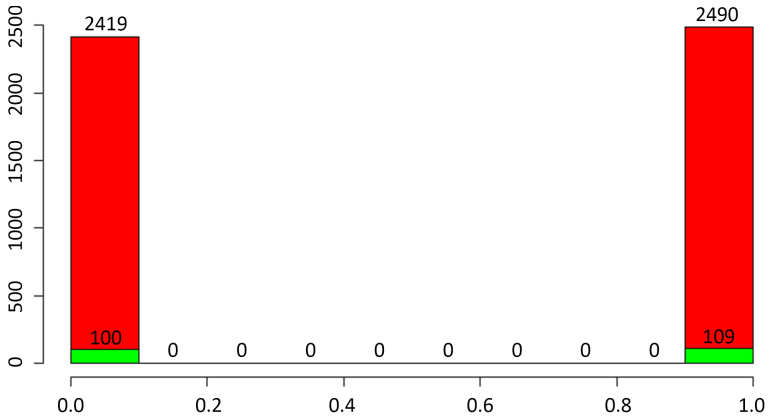

Figure 8. Cerebral stroke classified by type of residence
图 8. 按居住地类型分类的脑中风情况

可以看到，居住在城市和农村的人都差不多，得脑中风得人数与比例也都差不多。

再按血糖水平来看脑中风情况。如图 9 所示。

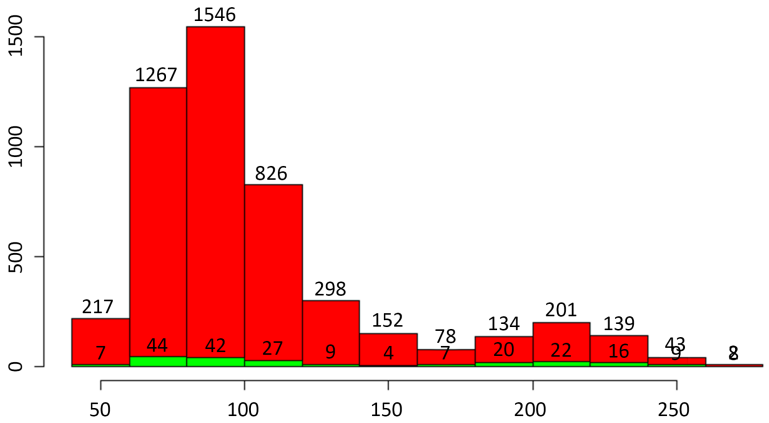

Figure 9. Cerebral stroke classified by blood sugar level
图 9. 按血糖水平分类的脑中风情况

可以看到，血糖水平基本都集中在 50~150 之间，约有 4306 人，患脑中风的有 133 人，占比为 3.08%，血糖水平 150 以上约有 603 人，患脑中风的有 74 人，占比为 12.27%。可以看出，血糖在 170~250 之间脑中风率较高。

最后是 BMI 与吸烟情况，对于 BMI，各区间患脑中风的比例都约为 3%~5%，相差不大。各吸烟情况得脑中风的比例也都相差不大，在 6%左右。

5. 模型建立与结果

首先我们通过相关系数查看各自变量之间的线性关系。如图 10 所示。

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type
gender	1.000000000	-0.03014922	0.021863236	0.082982904	-0.036138008	-0.0729422033	-0.0041782084
age	-0.030149218	1.00000000	0.274424873	0.257122776	0.680781652	0.5381240067	0.0109481144
hypertension	0.021863236	0.27442487	1.000000000	0.115990991	0.162406260	0.1246547061	-0.0010741462
heart_disease	0.082982904	0.25712278	0.115990991	1.000000000	0.111245121	0.0921448190	-0.0023617439
ever_married	-0.036138008	0.68078165	0.162406260	0.111245121	1.000000000	0.4259143556	0.0049891711
work_type	-0.072942203	0.53812401	0.124654706	0.092144819	0.425914356	1.0000000000	-0.0008827106
Residence_type	-0.004178208	0.01094811	-0.001074146	-0.002361744	0.004989171	-0.0008827106	1.0000000000
avg_glucose_level	0.053007822	0.23583816	0.180542699	0.154525119	0.151377377	0.0924897366	-0.0076165420
bmi	-0.026019865	0.33339800	0.167810584	0.041357443	0.341694652	0.3414343947	-0.0001224412
smoking_status	0.039494024	-0.38667582	-0.132831660	-0.071396924	-0.310702330	-0.3444032458	0.0027191093

	avg_glucose_level	bmi	smoking_status
gender	0.053007822	-0.0260198650	0.039494024
age	0.235838155	0.3333979952	-0.386675819
hypertension	0.180542699	0.1678105844	-0.132831660
heart_disease	0.154525119	0.0413574429	-0.071396924
ever_married	0.151377377	0.3416946516	-0.310702330
work_type	0.092489737	0.3414343947	-0.344403246
Residence_type	-0.007616542	-0.0001224412	0.002719109
avg_glucose_level	1.000000000	0.1755021761	-0.108983692
bmi	0.175502176	1.0000000000	-0.235739765
smoking_status	-0.108983692	-0.2357397646	1.000000000

Figure 10. Correlation coefficient of each variable
图 10. 各变量相关系数

可以看出，曾经结婚与年龄的相关系数最大，为 0.68，还没有达到 0.8，不是很强。对于连续型变量来说，年龄越大，血糖的水平较高的人数越多，但总体仍是血糖水平较低的人数多，相关系数为 0.23。年龄和 BMI 之间存在弱相关关系，BMI 和血糖水平之间关系不大。再根据之前的描述性统计，工作类型为 1、3 的为一个水平，2、4 的分为一个水平，5 为一个水平，所以将它们分为 3 类。

对其建立广义线性模型中的 Logistic 回归，结果如图 11 所示。

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.039538   0.647136 -10.878  < 2e-16 ***
A3[, 2]      0.004368   0.152907   0.029 0.977211
A3[, 3]      0.071987   0.006024  11.951  < 2e-16 ***
A3[, 4]      0.540369   0.174580   3.095 0.001966 **
A3[, 5]      0.390926   0.206030   1.897 0.057772 .
A3[, 6]     -0.098003   0.243659  -0.402 0.687527
A3[, 7]     -0.395133   0.175689  -2.249 0.024509 *
A3[, 8]      0.020100   0.149505   0.134 0.893050
A3[, 9]      0.004621   0.001294   3.572 0.000355 ***
A3[, 10]     0.003931   0.011682   0.337 0.736485
A3[, 11]    -0.029984   0.073921  -0.406 0.685023
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1369.0  on 4898  degrees of freedom
AIC: 1391

Number of Fisher Scoring iterations: 8
```

Figure 11. Regression result of Logistic
图 11. Logistic 回归结果

年龄、高血压、心脏病、工作类型和平均血糖水平都通过了显著性检验，与之前的描述性统计相符。在未通过显著性检验的变量中，只有曾经结婚与我们的描述性统计量不符。

由此我们可以得到初步的经验回归方程为：

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=-7.04+0.004x_1+0.07x_2+0.54x_3+0.39x_4-0.098x_5$$
$$-0.4x_6+0.02x_7+0.005x_8+0.004x_9-0.03x_{10}$$

继续进行逐步回归。逐步回归的结果建议我们选取年龄、高血压、心脏病、工作类型和平均血糖水平。对这五个变量和脑中风进行 Logistic 回归。结果如图 12 所示。

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.021571   0.478208 -14.683  < 2e-16 ***
A3[, 3]      0.071527   0.005831  12.267  < 2e-16 ***
A3[, 4]      0.551985   0.173338   3.184 0.001450 **
A3[, 5]      0.390191   0.204057   1.912 0.055855 .
A3[, 7]     -0.393813   0.175603  -2.243 0.024921 *
A3[, 9]      0.004732   0.001256   3.767 0.000165 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1728.4  on 4908  degrees of freedom
Residual deviance: 1369.4  on 4903  degrees of freedom
AIC: 1381.4

Number of Fisher Scoring iterations: 7
```

Figure 12. Regression result of Logistic after selecting the variable
图 12. 选择变量后的 Logistic 回归结果

所有回归系数在 0.1 的显著性水平下都通过了检验，并且 AIC 值较之前有所降低。但是心脏病得系数还是不够好，这可能是由于样本中患有心脏病得人比例太少了，仅有 243 人，而且是两点分布，总人数却有 4909 人，相较于其他指标，人数是比较少的。我们可以得到经验回归方程：

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=-7.022+0.07x_2+0.056x_3+0.39x_4-0.39x_6+0.005x_8$$

$$P(Y=1)=\frac{\exp\left(-7.022+0.07x_2+0.056x_3+0.39x_4-0.39x_6+0.005x_8\right)}{1+\exp\left(-7.022+0.07x_2+0.056x_3+0.39x_4-0.39x_6+0.005x_8\right)}$$

我们可以看出，随着年龄的增加、平均血糖水平的升高，患脑中风的概率也增加，患高血压与心脏病也会增加患脑中风的概率，对于三个水平的工作类型，即孩子、从不工作；私人、政府工作；自由职业，患脑中风的概率也随之降低。

以一个 50 岁，不患高血压，职业为政府工作，血糖水平为 120 的人比较其是否患心脏病得脑中风得概率。如果其没有心脏病，则患脑中风得概率为 3.5%，反之，患中风得概率为 2.4%，有了显著得降低。

6. 结论

年龄、高血压、心脏病、工作类型、平均血糖水平对于判断患脑中风的概率有着显著的影响，年龄是我们不可改变的，工作类型由于生活需要也是不能随意改变的，但是高血压、心脏病、平均血糖水平却可以靠我们自己来改变。在日常生活中多锻炼，多吃一些清淡食物，养成良好的生活习惯，可以有效降低患脑中风的概率。

参考文献

[1] 刘艳骄. 肥胖人痰湿体质与脑中风的相关性研究[J]. 河北中医学院学报, 1996(3): 13-17. https://doi.org/10.16370/j.cnki.13-1214/r.1996.03.007

[2] 赵孔华, 张沁园. 缺血性脑中风的发病机理探讨[J]. 河南中医, 2006(7): 7-9.

https://doi.org/10.16367/j.issn.1003-5028.2006.07.002

[3] 杜恩. 老年脑中风偏瘫患者的早期康复应用研究[J]. 成都医学院学报, 2013, 8(1): 65-67.

[4] 甘勇, 杨婷婷, 刘建新, 等. 国内外脑卒中流行趋势及影响因素研究进展[J]. 中国预防医学杂志, 2019, 20(2): 139-144. https://doi.org/10.16506/j.1009-6639.2019.02.013