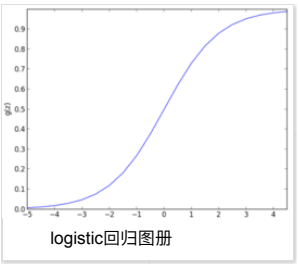
logistic回归

本词条由"科普中国"百科科学词条编写与应用工作项目 审核。

logistic回归又称logistic回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此因变量就为是否胃癌，值为"是"或"否"，自变量就可以包括很多了，如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的。然后通过logistic回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。



logistic回归图册

中文名	logistic回归	分　类	计算机 数学
外文名	logistic regressive	用　途	预测 判别
定　义	线性回归模型	领　域	数据挖掘 疾病诊断 经济预测

目录

概念

编辑

logistic回归是一种广义线性回归（generalized linear model），因此与多重线性回归分析有很多相同之处。它们的模型形式基本上相同，都具有 w'x+b，其中w和b是待求参数，其区别在于他们的因变量不同，多重线性回归直接将w'x+b作为因变量，即y=w'x+b，而logistic回归则通过函数L将w'x+b对应一个隐状态p，p =L(w'x+b),然后根据p 与1-p的大小决定因变量的值。如果L是logistic函数，就是logistic回归，如果L是多项式函数就是多项式回归。[1]

logistic回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释，多类可以使用softmax方法进行处理。实际中最为常用的就是二分类的logistic回归。[1]

Logistic回归模型的适用条件

1 因变量为二分类的分类变量或某事件的发生率，并且是数值型变量。但是需要注意，重复计数现象指标不适用于Logistic回归。

2 残差和因变量都要服从二项分布。二项分布对应的是分类变量，所以不是正态分布，进而不是用最小二乘法，而是最大似然法来解决方程估计和检验问题。

3 自变量和Logistic概率是线性关系

4 各观测对象间相互独立。[2]

原理：如果直接将线性回归的模型扣到Logistic回归中，会造成方程二边取值区间不同和普遍的非直线关系。因为Logistic中因变量为二分类变量，某个概率作为方程的因变量估计值取值范围为0-1，但是，方程右边取值范围是无穷大或者无穷小。所以，才引入Logistic回归。[2]

Logistic回归实质：发生概率除以没有发生概率再取对数。就是这个不太繁琐的变换改变了取值区间的矛盾和因变量自变量间的曲线关系。究其原因，是发生和未发生的概率成为了比值，这个比值就是一个缓冲，将取值范围扩大，再进行对数变换，整个因变量改变。不仅如此，这种变换往往使得因变量和自变量之间呈线性关系，这是根据大量实践而总结。所以，Logistic回归从根本上解决因变量要不是连续变量怎么办的问题。还有，Logistic应用广泛的原因是许多现实问题跟它的模型吻合。例如一件事情是否发生跟其他数值型自变量的关系。[2]

注意：如果自变量为字符型，就需要进行重新编码。一般如果自变量有三个水平就非常难对付，所以，如果自变量有更多水平就太复杂。这里只讨论自变量只有三个水平。非常麻烦，需要再设二个新变量。共有三个变量，第一个变量编码1为高水平，其他水平为0。第二个变量编码1为中间水平，0为其他水平。第三个变量，所有水平都为0。实在是麻烦，而且不容易理解。最好不要这样做，也就是，最好自变量都为连续变量。[2]

spss操作：进入Logistic回归主对话框，通用操作不赘述。

发现没有自变量这个说法，只有协变量，其实协变量就是自变量。旁边的块就是可以设置很多模型。

"方法"栏：这个根据词语理解不容易明白，需要说明。

共有7种方法。但是都是有规律可寻的。

"向前"和"向后"：向前是事先用一步一步的方法筛选自变量，也就是先设立门槛。称作"前"。而向后，是先把所有的自变量都进来，然后再筛选自变量。也就是先不设置门槛，等进来了再一个一个淘汰。

"LR"和"Wald",LR指的是极大偏似然估计的似然比统计量概率值，有一点长。但是其中重要的词语就是似然。

“进入”就是所有自变量都进来，不进行任何筛选

将所有的关键词组合在一起就是7种方法，分别是“进入”“向前LR”“向前Wald”“向后LR”“向后Wald”“向后条件”“向前条件”

下一步：一旦选定协变量，也就是自变量，“分类”按钮就会被激活。其中，当选择完分类协变量以后，“更改对比”选项组就会被激活。一共有7种更改对比的方法。

“指示符”和“偏差”，都是选择最后一个和第一个个案作为对比标准，也就是这二种方法能够激活“参考类别”栏。“指示符”是默认选项。“偏差”表示分类变量每个水平和总平均值进行对比，总平均值的上下界就是“最后一个”和“第一个”在“参考类别”的设置。

“简单”也能激活“参考类别”设置。表示对分类变量各个水平和第一个水平或者最后一个水平的均值进行比较。

“差值”对分类变量各个水平都和前面的水平进行作差比较。第一个水平除外，因为不能作差。

“Helmert”跟“差值”正好相反。是每一个水平和后面水平进行作差比较。最后一个水平除外。仍然是因为不能做差。

“重复”表示对分类变量各个水平进行重复对比。

“多项式”对每一个水平按分类变量顺序进行趋势分析，常用的趋势分析方法有线性，二次式。[2]

主要用途 编辑

寻找危险因素

正如上面所说的寻找某一疾病的危险因素等。

预测

如果已经建立了logistic回归模型，则可以根据模型，预测在不同的自变量情况下，发生某病或某种情况的概率有多大。

判别

实际上跟预测有些类似，也是根据logistic模型，判断某人属于某病或属于某种情况的概率有多大，也就是看一下这个人有多大的可能性是属于某病。

这是logistic回归最常用的三个用途，实际中的logistic回归用途是极为广泛的，logistic回归几乎已经成了流行病学和医学中最常用的分析方法，因为它与多重线性回归相比有很多的优势，以后会对该方法进行详细的阐述。实际上有很多其他分类方法，只不过Logistic回归是最成功也是应用最广的。[1]

案例分析 编辑

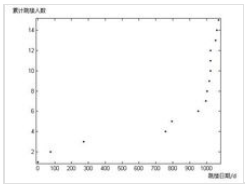
关于富士康跳楼曲线的Logistic回归分析。

首先找出所有富士康员工自杀的日期：

列出如下表格：（以07年6月18号，第一例自杀案例为原点，至今（10年5月25日）1072天）

自杀时间x/d	0	75	272	758	794	950	997	1003	1015
1023	1024	1024	1053	1051	1072				
累计自杀人数y	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15				

在MATLAB中容易做出散点图：



可见这是一个指数增长的曲线。

其增长曲线与对数增长很接近。

对其做指数函数拟合：

General model Exp2:

f(x) = a*exp(b*x) + c*exp(d*x)

Coefficients (with 95% confidence bounds):

a = 7.569e-007 (-6.561e-006, 8.075e-006)

b = 0.01529 (0.006473, 0.0241)

c = 1.782 (0.5788, 2.984)

d = 0.001075 (2.37e-005, 0.002125)

Goodness of fit:

SSE: 8.846

R-square: 0.9684

Adjusted R-square: 0.9598

RMSE: 0.8968

因此，和很多流行病分析一样，该曲线很有可能呈S型。对于该曲线的分析，使用Logistic回归。

首先假设Logis(B,x)=F(x),之中B为参数数组，则由经验和可能的微分方程关系，回归曲线应该为

S（x）=m*Logis（B,x+t）/（n+Logis（B,x+t））格式

由于当Logis（B,x）较小时S(x)=Logis（B,x），则可以认为f（x）的参数可以直接引入S（x）作为一种近似，而对于m，n的确定，以1为间隔，画出m*n=40*20的所有曲线，

选出其中最吻合的的一条（m=22 n=20 t=50）：[1]

其他信息

编辑

由此可以见，富士康的跳楼人数最终会稳定在在22人左右，仍然不会超过全国平均跳楼率。

对此曲线的分析，借鉴微生物生长曲线的方法，将其分为：

缓慢期，对数期，稳定期，衰亡期

缓慢期，富士康员工虽然受到很大的工作压力，可是其自身的心理并没有崩溃，因此跳楼这种事件发生频率很少，而且呈线性关系，说明没有跳楼者受到别的跳楼者的影响。

对数期，富士康员工由于受到工厂巨大的工作压力，以及来自社会各方的压力，甚至加上上级的欺压，心理防线渐渐崩溃，无处发泄。而一旦有想不开者跳楼，则为其提供了一个发泄的模板，这种情况下，很容易有相同经历的员工受到跳楼者的影响，从而一个接一个的跳楼自杀。目前的富士康正处于此时期。

稳定期，由于社会、媒体各方面的关注以及社会、广大人民对工厂的压力，工厂不得不做出改变，员工的心理压力渐渐得到释放，从而员工跳楼轻生频率会很快下降。[1]

词条图册

更多图册

参考资料

1. logistic回归 . 人文网[引用日期2017-01-07]

2. 78logistic 回归与线性回归的比较 . 三亿文库[引用日期2017-01-07]

词条标签： 科学百科信息科学分类