科技经济导刊 2019，27（28）
Technology and Economic Guide

基于 Logistic 回归的胃癌预测研究

孙明娟
（延安大学数学与计算机学院，陕西 延安 716000）

[摘要] 胃癌是世界范围内最常见的恶性肿瘤疾病之一，发病率居我国恶性肿瘤第二位，死亡率排在前三位。因此，探究引发胃癌疾病的致病因素以及建立合理有效的疾病诊断模型对个人及医疗机构就显得尤为重要。建立 Logistic 胃癌预测模型对胃癌数据进行分析，采用逐步回归分析法进行变量筛选；对模型参数的估计采用极大似然法，得到患胃癌概率的 Logistic 回归方程，并进行预测。通过对比预测精度和误分类精度的值可以发现，Logistic 胃癌预测模型适用于对胃癌进行预测研究。

[关键词] 极大似然估计；Logistic 回归；预测
[中图分类号] O212 **[文献标识码]** A **[文章编号]** 2096-1995(2019)28-0126-02

随着人民生活水平的不断提升，工作和生活的压力随之而来。工作上的熬夜加班应酬以及生活中的暴饮暴食不规律作息也是屡见不鲜，由此引发大大小小的胃部疾病，也引起了越来越多人的关注与重视，探究引发胃癌疾病的致病因素，建立合理有效的疾病诊断模型对个人及医疗机构是非常重要的。

本文通过建立 Logistic 回归模型，对标准化后的胃癌数据进行分析，采用逐步回归法选取解释变量，对回归模型参数的估计选用极大似然法，得到了预测患胃癌概率的 Logistic 回归方程。

1 Logistic 回归及逐步回归法

1.1 Logistic 回归

对 p 元线性回归模型

$$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon,E_\varepsilon=0,D_\varepsilon=\sigma^2, \tag{1}$$

其中未知参数 $\beta_0,\beta_1,\cdots\beta_p$ 称为偏回归系数，显然有

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \tag{2}$$

式 (2) 称为 y 对 $x_1,x_2,\cdots,x_p$ 的回归函数。

当因变量是一个二元变量，只取 0 与 1 两个值时，$E(y)=P\{y=1\}=p$ 是因变量，对其做 Logit 变换，得

$$\ln\frac{p}{1-p}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \tag{3}$$

称为 Logistic 线性回归。

极大似然估计就是选取 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 的估计值 $\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$ 使得式 (3) 达到极大。由此得到的 Logistic 回归模型为：

$$\hat{p}_i=\frac{e^{\hat{\beta}_0+\sum_{j=1}^{p}x_{ij}\hat{\beta}_j}}{1+e^{\hat{\beta}_0+\sum_{j=1}^{p}x_{ij}\hat{\beta}_j}} \tag{4}$$

1.2 变量选取方法

回归自变量的选取是建立回归模型的一个关键问题所在。在对一个实际问题进行建模时我们往往最开始遇到的问题就是对自变量的筛选。逐步回归分析法是将自变量一个一个选入而后进行回归分析。对已经选入的变量进行显著性检验。将其中显著性较低的自变量剔除。当从选取自变量或者在回归方程中剔除一个自变量是逐步回归分析的一步。将这个步骤反复执行，一直到既没有显著的自变量选入回归方程，也没有不显著自变量从回归方程中剔除为止。

2 基于 Logistic 回归的胃癌预测模型

2.1 数据的分析和处理

本文收集了某一地区是否患胃癌的数据。该数据集共有 400 条是否患胃癌的记录，每一条是否患病记录都包含个人情况属性变量和个人标签变量。个人情况属性变量包含了有关人类生活习惯、年龄、性别、遗传、内在因素五个方面的 11 项指标。分别是是否过度饮酒、是否经常吸烟、是否长期食用烟熏，盐腌和霉变食物、是否过度肥胖、精神状况、睡眠质量、是否经常加班熬夜、是否坚持锻炼身体、是否有胃癌或食管癌家族史、性别、年龄。个人标签变量是对个人健康的定义，有“健康”和“疾病”两种。

2.2 Logistic 胃癌预测模型

2.2.1 变量筛选

运用 SPSS 软件对原始数据集进行逐步回归后筛选出的变量列于表 1：

表 1 Logistic 回归变量选择结果

输入／移去的变量[a]

模型	输入的变量	移去的变量	方法
1	是否过度肥胖	.	步进（准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。
2	是否经常吸烟	.	步进（准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。
3	是否经常酗酒	.	步进（准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。
4	是否经常吃烟熏等不健康食物	.	步进（准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。
5	精神状况	.	步进（准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。
6	是否遗传	.	步进（准则：F-to-enter 的概率 <= .050，F-to-remove 的概率 >= .100）。

a. 因变量：是否患病

从表 1 可看出，筛选出的变量有：是否过度肥胖（x_1），是否经常吸烟（x_2），是否经常喝酒（x_3），是否经常吃烟熏等不健康食物（x_4），精神状况（x_5），是否遗传（x_9）。

项目来源： 陕西省教育厅自然科学基金（18JK0877）；延安大学自然科学基金（YD2015-10）。

2.2.2 模型建立

以“是否患病”作为因变量 y，y=1 表示“患病”，y=0 表示“健康”，以 $x_1, x_2, x_3, x_4, x_5, x_9$ 为自变量，运用 SPSS 软件对 Logistic 回归模型参数进行极大似然估计，结果列于表 2：

表 2 Logistic 回归参数估计结果

参数估计

是否患病[a]		B	标准误	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95% 下限	上限
健康	截距	6.044	.773	61.202	1	.000			
	是否经常酗酒	-3.551	.784	20.499	1	.000	.029	.006	.133
	是否经常吸烟	-3.555	.709	25.128	1	.000	.029	.007	.115
	是否经常吃烟熏等不健康食物	-1.740	.618	7.926	1	.005	.176	.052	.589
	是否过度肥胖	-3.369	.805	17.496	1	.000	.034	.007	.167
	精神状况	-1.196	.312	14.737	1	.000	.302	.164	.557
	是否遗传	-4.042	1.642	6.060	1	.014	.018	.001	.439

a. 参考类别是: 患病。

可以得到具体的回归模型如下：

$$p = \frac{e^{(6.044-3.551x_1-3.555x_2-1.740x_8-3.369x_4-1.196x_5-4.042x_9)}}{1+e^{(6.044-3.551x_1-3.555x_2-1.740x_8-3.369x_4-1.196x_5-4.042x_9)}} \quad (5)$$

2.2.3 模型的预测

将上面筛选出的变量作为预测变量，以是否患病作为因变量。利用 SPSS 软件可以得出测试数据每一条数据的预测概率。

统计的结果为：当 y_i = 0（即”健康”）时，预测概率小于 0.5 的数据有 336 条，大于 0.5 的数据有 24 条；当 y_i = 1（即”疾病”）时，预测概率小于 0.5 的数据有 14 条，大于 0.5 的数据有 346 条。

综上所述，即就是将“疾病”预测成“健康”的比例是 93.3%，将“健康”预测成“疾病”的比例是 6.7%。将“疾病”预测成“疾病”的比例是 96.1%，将“疾病”预测成“健康”的比例是 3.9%。总的预测精度为 89.4%。误分类错误为 10.6%。

表 3 Logistic 回归预测精度

真实 \ 预测	健康	疾病	预测精度	误分类错误
健康	336	24	93.3%	6.7%
疾病	14	346	96.1%	3.9%

3 结语

本文采用 Logistic 回归对数据进行建模，采用传统的逐步回归方法进行变量选择，从回归预测精度可以看出，模型的拟合效果比较好，说明 Logistic 回归适用于对胃癌的预测。

【参考文献】

[1] 张婷婷 .Logistic 回归及其相关方法在个人信用评分中的应用 [D]. 太原：太原理工大学，2017.

[2] 高慧璇，张继平，李忠，等 . 应用多元统计分析 [M]. 北京：北京大学出版社，1995：166-169.

[3] 谭宏卫，曾捷 .Logistic 回归模型的影响分析 [J]. 数理统计与管理，2013，32(03)：476-485.

[4] 何晓群，刘文卿，易丹辉 . 应用回归分析 [M]. 北京：中国人们大学出版社，2000：153-158.

[5]Lee Saro，Jeon Seong Woo，Oh Kwan-Young，Lee Moung-Jin. The spatial prediction of landslide susceptibility applying artificial neural network and logistic regression models： A case study of Inje，Korea[J]. Open Geosciences，2016，8(1)：106-122.

[6] 丁澍，王艳 . 高职院校课堂教学质量影响因素研究——基于 Lasso-logistic 回归模型 [J]. 数理统计与管理，2017，36(06)：1039-1048.

[7] 陆兵焱，陈友龙，李映颖 . 基于 SPSS 对试飞数据进行的相关性分析 [J]. 科技信息，2009(15)：487-442.

[8] 袁鹏 .R 语言在统计学教学中的探讨 [J]. 科技展望，2016，26(07)：226.

[9] 麦鸿坤，肖坚红，吴熙辰，等 . 基于 R 语言的负荷预测 ARIMA 模型并行化研究 [J]. 电网技术，2015，39(11)：3216-3220.

（上接 P133）到相关的科研机构深造、学习；聘请相关的专家、学者为图书馆的管理人员讲学、座谈及交流沟通；提供给本图书馆管理人员之间互相学习的机会等。计算机技术和网络信息技术的广泛应用，高校图书馆的管理人员要有意识的加强在网络信息技术方面的学习。例如网络管理、信息资源的采集及提炼等方面的专业技能，不断提升自己的综合能力。总之，高校图书馆也要高度重视本馆管理人员的综合能力提升，通过开展一系列的计算机技能和网络技能等的专业培训和学习活动，切实增强管理人员的管理能力，打造一支知识全面、专业知识丰富、业务能力一流、创新能力较强的图书馆管理队伍，有力推动高校图书馆管理工作的更快、更高效的全面发展 [2]。

3.3 创新高校图书馆管理服务方式

在信息化的时代发展背景下，高校图书馆一方面要注重传承传统的图书馆管理和服务内容、服务方式；另一方面还要依据时代的发展变化要求，做好服务创新。其一，要积极开展网络化服务体系的构建工作。当前在网络信息技术的助推下，网上图书馆以其强大的便利性，受到了众多读者的追捧，发展相当迅猛。高校图书馆是网上图书馆的主力军，这些网上高校图书馆可以为广大读者提供多种服务。例如查询服务、数字化资源、相关链接、在线咨询及特色馆藏等 [3]。其二，提供网络上的导航服务。网上图书馆可以为读者提供丰富的信息资源，为了查找方便，建立网上导航服务能够满足许多读者的查询需要，可以更便捷地找到自己所需要的网站、图书馆等终端，节省了时间，提高了效率。

3.4 不断创新完善馆藏资源

丰富而完善的馆藏资源，是吸引读者的有效方法之一。其一，不断的扩充纸质图书的数量，针对学校学生专业的特点，注意添置新图书的均衡性。其二，多征求各专业师生的意见和建议，满足他们多元化、个性化的信息资源需求 [4]。其三，积极构建高校图书馆间的数据库系统共享平台，提升信息资源的利用效率，满足读者的阅读需求。

4 结语

信息化时代的到来，给高校图书馆的管理工作带来了机遇和挑战。高校图书馆只有顺应时代的发展，不断的改革和创新，才能在发展中不被淘汰，满足更多读者的需求，发挥更大的作用。

【参考文献】

[1] 韩建明 . 如何进行高校图书馆管理工作的创新改革 [J]. 课程教育研究 ,2019(26):244.

[2] 宋成城 . 网络环境下现代高校图书馆管理工作的创新策略探索 [J]. 产业与科技论坛 ,2018,17(18):240-241.

[3] 张旭 . 刍议我国高校图书馆管理工作的创新 [J]. 现代经济信息 ,2017(11):122.

[4] 罗玉琴 . 对高校图书馆读者服务创新路径的几点思考 [J]. 才智 ,2017(31):167.