亚洲临床医学杂志 · 第 02 卷 · 第 04 期 · 2019 年 10 月
DOI: https://doi.org/10.26549/yzlcyxzz.v2i4.2258

Analysis of Etiology of Breast Cancer Based on Data Mining Method

Liwei Li　Pengyan Ren

China Academy of Information and Communications Technology, Beijing, 100037, China

Abstract

In order to study the causes of female breast cancer, data mining methods such as factor analysis were used to analyze the 13 selected influencing factors by collecting and processing breast cancer data and some socio-economic data from 173 countries on seven continents. The study found that the social factors, stress factors, external environment factors all play an inducing role in the prevalence of female breast cancer, among which social factors play the most important role, indicating that with the improvement of regional economic development, women live under more pressure and are more likely to suffer from breast cancer. At the same time, environmental pollution will also increase the prevalence of breast cancer in women.

Keywords

breast cancer; cluster analysis; factor analysis; data mining

基于数据挖掘方法的乳腺癌病因分析

李立委　任鹏燕

中国信息通信研究院，中国 · 北京 100037

摘　要

为研究女性乳腺癌诱因，通过收集处理七大洲包含 173 个国家的乳腺癌数据与部分社会经济数据，对选取的 13 个影响因素运用因子分析等数据挖掘方法进行分析。研究发现，社会发达因子、压力因子、外部环境因子对女性乳腺癌患病率均起诱发作用，其中社会因子的作用最大，说明随着地区经济发展水平的提高，女性生活压力加大，更容易患得乳腺癌，同时环境污染也会增加女性乳腺癌患病率的提升。

关键词

乳腺癌；聚类分析；因子分析；数据挖掘

1 理论基础及变量说明

1.1 因子分析

作为一种常用的降维方法，因子分析通过研究众多变量之间的内部依赖关系，提取公共因子，用以表示原有数据的基本结构，并且利用这些公因子反映变量的重要信息，由于这些假想变量是不可观测的潜在变量，故称为因子。

在进行因子分析时，首先对数据进行标准化处理，然后估计因子载荷矩阵，具体公式如下所示：

$$\begin{cases} Z_1 = a_{11}F_1 + a_{12}F_2 + ... + a_{1p}F_p + C_1U_1 \\ Z_2 = a_{21}F_1 + a_{22}F_2 + ... + a_{2p}F_p + C_2U_2 \\ \cdots \\ Z_m = a_{m1}F_1 + a_{m2}F_2 + ... + a_{mp}F_p + C_mU_m \end{cases} \quad (1)$$

其中 Z1、Z2、…、ZM 为原始变量，F1、F2、…、FP 为公共因子，表示为矩阵形式为：

$$Z=A \cdot F+C \cdot U \quad (2)$$

A 为因子载荷矩阵，一般采用主成分法进行估计，随后对 A 进行正交变换，从而对因子的意义进行解释。最后，通过因子得分函数，可以计算原有的每个解释变量在每个公共因子上的得分，从而解决公共因子不可观测的问题。

1.2 多元线性回归

在实际问题分析中，因变量的影响因素众多，因此需要引入多个自变量进行量化分析，而在回归分析中，如果有两个或两个以上的自变量，就称为多元回归，多元线性回归多借助最小二乘法进行参数估计，并通过相关统计量判定参数

亚洲临床医学杂志 · 第 02 卷 · 第 04 期 · 2019 年 10 月
DOI: https://doi.org/10.26549/yzlcyxzz.v2i4.2258

的显著性大小。

本文使用的数据集主要包含 173 个国家的数据，其中乳腺癌数据来自 ARC 国际癌症研究机构，部分社会经济数据来源于世界银行数据库。其中，解释变量包含：人均收入、酒消费量、入伍率、CO_2 排放量、女性就业率、艾滋死亡率、网络使用率、寿命、政治得分、人均居民用电量、自杀率、就业率，以及城镇率，分别用 X_1−X_{13} 表示，被解释变量为新增乳腺癌患病案例。

2 实证分析

2.1 缺失值处理

由于部分国家数据存在缺失，在进行分析之前，利用 spss 软件对现有数据进行分析，发现变量人均收入、艾滋死亡率、人均用油量、人均居民用电量，均存在数据缺失，总缺失数据量达到 190，其中人均用油量数据缺失值个数高达 112，占总数据缺失值的 64.7%，故需要对缺失值进行处理。

因此，对于遗失值 X_1，X_6，X_9，X_{11}，应当进行遗漏值处理，当前常用的处理方法包括回归法和 EM 法，但是在处理前首先应该进行缺失值分析，判断是完全随机缺失 MCAR 还是随机缺失 MAR。在处理时，分别采用了 EM 方法和回归方法，但是在使用 EM 方法时，25 迭代中的 EM 算法收敛失败，故采用回归方法进行分析和处理。独立方差 t 测试结果，给出了影响其他定量变量的遗失值模式，计算发现当 X_9 存在时 X_1 的均值为 12752.2，缺失时均值为 4072.17，因此可以看出 X_9 的缺失对于 X_1 的影响较大，故 X_9 的缺失不是完全随机缺失。从回归协方差和回归相关性也可以看出，我们应该拒绝缺失值为完全随机缺失假设，并利用回归法进行插值处理。

2.2 基本统计量分析

在进行相应分析之前，首先对被解释变量进行基本的统计分析，对其分布和基本统计特性进行研究。首先分析其基本统计量，可以看出新增乳腺癌病例数的均值为 37.4029/十万人，方差为 515.195，方差和标准差较大，数据较离散。说明被解释变量在各个不同变量分组之间的差异性较大。通过做被解释变量与其他几个变量之间的柱状图，分析分组之间的数据值差异。

从七大洲新增乳腺癌患病数分布图可以看出，非洲、AS 亚洲、EE 西欧、WE 东欧、LATAM 拉丁美洲、NORAM 北美洲、OC 大洋洲，亚非地区的新增乳腺癌病例数相对较低，西欧和北美地区的新增乳腺癌病例数较高。西欧和北美的发病率是亚非的约 3.5 倍，东欧地区的病例数约为亚非地区发病率的 2.5 倍，大洋洲的病例数约为亚非地区病例数的两倍。根据数据显示，可以初步判断，经济发展水平较高的地区新增乳腺癌病例数水平高于经济发展水平较低的地区。

通过不同洲际之间和乳腺癌患病数之间的频数分布柱状图可以看出，各大洲间如果乳腺癌潜在患者出现在经济欠发达的地区，他们就不会花费巨额医疗费去医治，而如果这一情况发生在高收入群体中，这一疾病是有很大希望治愈的；其次，不同大洲用于人民医疗的财政支出也有很大区别，用于人民重大医疗报销的费用越多，新增乳腺癌病例的比例就会减少；还有，不同大洲的饮食习惯和作息时间也是不一样的。沿袭传统健康的饮食模式和作息时间会使人身体的各项机能保持稳定，大大减少新增乳腺癌病例的比例，而不健康的饮食规律和不正常的作息时间会使身体机能紊乱，使得很多人变得肥胖、亚健康，增加了乳腺癌患病率。针对这些定性的描述，本文对相关特征变量进行数据处理，定量的分析乳腺癌的诱发因素。

2.3 因子分析

由于解释变量之间存在较为严重的多重共线性，结合多元统计学的方法，故采用因子分析法提取公因子，探测数据的基本结构，同时有效地消除多重共线性的问题，继而利用公因子对数据进行处理，从而采用逐步回归法分析乳腺癌发病率的影响因素，KMO 值越接近 1 说明越适合做因子分析，此例中该值达到 0.688，说明适合利用因子分析法对数据进行处理。Bartlett 球形度检验原假设为相关系数矩阵是单位阵，sig 值小于 0.05，说明拒绝原假设，即变量之间存在相关关系，适合做因子分析。

而由总方差解释结果中可以看出，只有前四个因子的特征值大于 1，并且前四个特征值能够解释数据总特征值的 69.756%，因此提取前四个因子作为主因子代表数据样本进行分析。确定主因子个数后，根据成分得分系数矩阵计算的因子得分对该因子进行定义解释。作为计算因子得分的依据，系数矩阵中每个因子只有少数几个指标的载荷较大，因此根据载荷值的大小，将 13 个解释变量分为以下四类：人均收入、网络使用率、人均寿命在第一个因子上载荷较大，可以命名第一个因子为社会发达因子；入伍率、女性就业率、就业率在第二个因子上载荷较大，可以命名为就业因子；酒精消费量、

政治得分、人均居民用电量、自杀率在第三个因子上载荷较大，可以称为压力因子；CO_2 排放量、艾滋死亡率在第四个因子的载荷较大，可以称为外部环境因子。

2.4 多元线性回归进行分析

基于各变量在相关因子的得分情况，对数据进行处理，得到 i_1、i_2、i_3、i_4 四个主因子作为解释变量对女性乳腺癌发病率进行解释，利用多元线性回归法进行分析，研究各因子对因变量的解释情况以及显著性水平大小。均以通过 1% 的显著性水平拒绝原假设，各变量的系数较为显著，模型的拟合优度为 0.697，调整拟合优度为 0.69，F 统计量通过 1% 的显著性水平，说明模型的解释程度较高，且 DW 值为 1.86 接近 2，说明模型残差不存在自相关，故多元线性回归法最终确定的模型较为合理。

社会发达因子、压力因子、外部环境因子对女性乳腺癌患病率均起诱发作用，其中社会因子的作用最大，说明随着地区经济发展水平的提高，女性更容易患得乳腺癌。分析其原因，可能是居住在经济发达地区的女性生活压力较大，长期处于焦虑状态而诱发乳腺癌，而压力因子的系数为正也说明了这一点；就业因子系数为负，可能的原因在于随着就业率的提高，人们的生活水平得到提升，能够降低生活压力，同时收入的提高也能使女性定期进行体检，提前预防乳腺癌的发生；[1] 外部环境因子的系数为正，而构成该因子最重要的一个因素是的排放量，说明 CO_2 排放量的增加对女性乳腺癌的发病率起正向促进作用，即污染是引发患病的一个重要因素。

3 结语

本文首先分析不同洲际之间、城镇化率和乳腺癌患病数之间的频数分布柱状图，通过直观的图形表述，可以看出城市发展水平高，经济发展程度高的发达经济体，乳腺癌发病率较高，随后通过一系列数据处理，消除多重共线性，添加遗失值，利用逐步回归法进行定量分析，发现社会发达因子、压力因子、外部环境因子对女性乳腺癌患病率均起诱发作用，其中社会因子的作用最大，说明随着地区经济发展水平的提高，女性更容易患得乳腺癌。分析其原因，可能是居住在经济发达地区的女性生活压力较大，长期处于焦虑状态而诱发乳腺癌，而压力因子的系数为正也说明了这一点；就业因子系数为负，可能的原因在于随着就业率的提高，人们的生活水平得到提升，能够降低生活压力，同时收入的提高也能使女性定期进行体检，提前预防乳腺癌的发生；外部环境因子的系数为正，而构成该因子最重要的一个因素是 CO_2 的排放量，说明 CO_2 排放量的增加对女性乳腺癌的发病率起正向促进作用，即污染是引发患病的一个重要因素。

综合分析结果，可以看出经济发达地区的女性患有乳腺癌的几率较大。因此，政府及相关部门应该重视并制定相关政策，减少污染排放，同时作为个人而言，女性应该找到适合自己的排压方式，以减轻工作生活中的压力，同时注意保持良好的饮食和锻炼习惯，拥有一个健康的身体，减低乳腺癌的发病率。从政府政策外部因素和个人减压锻炼身体内部因素出发，相信女性居民中新增乳腺癌案例数将会得到有效地减少。

参考文献

[1] 张嘉庆 , 王殊 , 乔新民 . 乳腺癌的现状和远景 [J]. 中华外科杂志 ,2002,(03):161-163.