Advances in Applied Mathematics 应用数学进展, 2021, 10(3), 666-673
Published Online March 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.103072

Hans汉斯

基于路径模型的肺癌死亡率影响因素研究

肖　霏，滕文珍，徐汝屹，刘啟竹，张汝阳

曲阜师范大学统计学院，山东　济宁
Email: x_feidem@163.com

收稿日期：2021年2月8日；录用日期：2021年3月8日；发布日期：2021年3月15日

摘　要

为研究影响肺癌死亡率的因素，本文基于美国数据(DATA USA)中的相关数据作为样本，建立以肺癌死亡率因变量的路径模型。首先通过相关分析确定纳入模型的解释变量，运用SPSS软件分析所获得的样本数据，通过逐步回归法建立回归方程，得到相关系数和回归系数，并绘制路径图。分析结果表明：抽烟是引发肺癌的主要原因之一，此外石油，煤和内燃机等燃烧后和沥青公路尘埃产生的苯等有害物质也是导致肺癌的重要因素，因此应该提倡不吸烟，并加强城市环境卫生工作。

关键词

肺癌死亡率，路径模型，相关分析，逐步回归法

Study on Influencing Factors of Lung Cancer Mortality Based on Path Model

Fei Xiao, Wenzhen Teng, Ruyi Xu, Qizhu Liu, Ruyang Zhang

School of Statistics, Qufu Normal University, Qufu Shandong
Email: x_feidem@163.com

Received: Feb. 8th, 2021; accepted: Mar. 8th, 2021; published: Mar. 15th, 2021

Abstract

In order to study the factors affecting lung cancer mortality, this paper selected relevant DATA from Data USA as samples to establish a path model based on dependent variables of lung cancer mortality. Firstly, the explanatory variables included in the model were determined by correlation analysis, and the sample data obtained were analyzed by SPSS software. The regression equation was established by step-by-step regression method, and the correlation coefficient and regression

文章引用: 肖霏, 滕文珍, 徐汝屹, 刘啟竹, 张汝阳. 基于路径模型的肺癌死亡率影响因素研究[J]. 应用数学进展, 2021, 10(3): 666-673. DOI: 10.12677/aam.2021.103072

coefficient were obtained, and the path diagram was drawn. The analysis results show that smoking is one of the main causes of lung cancer. In addition, benzene and other harmful substances produced by the combustion of petroleum, coal, internal combustion engines and asphalt road dust are also important factors of lung cancer. Therefore, we should advocate non-smoking and strengthen the urban environmental health work.

Keywords

Lung Cancer Mortality, Path Model, Correlation Analysis, Stepwise Regression

1. 引言

在 1950~1999 年期间，美国人口数量和肺癌死亡率之间的相关系数达到 0.92，如果把肺癌死亡率在人口数量上回归，数据会非常接近回归线[1]。然而，肺癌不具传染性，人口数量不会影响肺癌死亡率，得到这样无理的结论是因为我们错误地把相关性当成了因果性。因果性会导致相关性，反之则不成立。统计学把两个或者多个看起来高度相关，但却不存在因果性的现象叫做伪相关性。伪相关的成因有：1. 存在其他变量同时对这两个事件造成直接或者间接的影响，但我们未意识到它的存在；2. 样本量过小导致的偶然现象。伪相关不代表是错误的，只要不是纯粹巧合，那么就值得分析与思考。

周脉耕等定量地分析人口老化及危险因素对肺癌死亡率的影响[2]。冯丹等利用灰色关联分析探讨该市大气污染指标与肺癌死亡率的关系[3]。张国钦等拟合烟草消费和大气污染与肺癌死亡率关系的分布滞后模型，根据模型分析烟草消费和大气污染对肺癌死亡率的影响[4]。本文旨在通过相关分析、建立路径模型，初步探究影响肺癌死亡率的潜在变量，并结合实际情况提出降低肺癌死亡率的合理建议。

2. 理论依据

2.1. 相关分析

相关分析是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。对总体中确实具有联系的标志进行分析，其主体是对总体中具有因果关系标志的分析。它是描述客观事物相互间关系的密切程度并用适当的统计指标表示出来的过程。

2.2. 路径分析

路径分析的主要目的是检验一个假想的因果模型的准确和可靠程度，测量变量间因果关系的强弱，回答下述问题：1) 模型中两变量 xj 与 xi 间是否存在相关关系；2) 若存在相关关系，则进一步研究两者间是否有因果关系；3) 若 xj 影响 xi，那么 xj 是直接影响 xi，还是通过中介变量间接影响或两种情况都有；4) 直接影响与间接影响两者大小如何。

2.3. 逐步回归

逐步回归将变量一个一个的引入，每引入一个变量后都要进行 F 检验，并对已选入的变量要进行逐个的 t 检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其删除，以确保回归方程只

包含显著的变量。反复执行此过程，直至既无显著的自变量选入回归方程，也无不显著的自变量被剔除回归方程为止。

3. 模型构建与结果分析

3.1. 研究假设

通过阅读文献确定影响肺癌死亡率的主要因素，可从吸烟、空气质量等方面展开讨论[5]，同时结合1820年-1924年美国完成了第一次和第二次工业革命，工业污染加剧，本文初步选取空气污染指数(AQI)、人均工业烟尘(吨/人)、烟民数量、森林覆盖率、在职医生数量和政府医疗资金投入六个指标。做出如下假设：

1) 空气污染指数(AQI)、人均工业烟尘(吨/人)、人均工业废水排放总量(吨/人)、烟民数量与肺癌死亡率呈正相关，森林覆盖率与肺癌死亡率呈负相关；

2) 以在职医生数量和政府医疗资金为指标刻画一个时期的医疗水平，与肺癌死亡率呈负相关；

3) 空气污染指数(AQI)与人均工业烟尘(吨/人)呈正相关；

4) 烟民数量增加，一定程度上会增加酸性气体污染，空气污染指数越高，表示空气污染越严重,故烟民数量与空气污染指数呈正相关；

5) 绿色植被有助于改善空气，吸附大气中的烟尘，故空气污染指数(AQI)、人均工业烟尘(吨/人)与森林覆盖率呈负相关。

通过上面的假设，做出初始路径图，如图1：

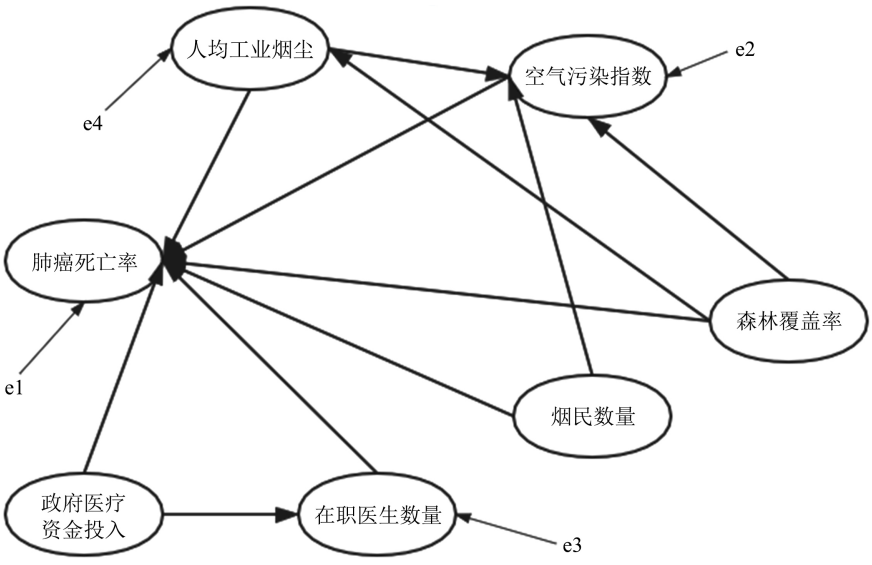

Figure 1. Initial path diagram
图 1. 初始路径图

3.2. 符号规定

为便于论述以及表示模型，我们特规定如下符号表，见表1。

3.3. 数据预处理

考虑数据的可获得性和真实性，本文选取了美国数据(DATAUSA)中的相关数据作为样本。由于指标

的量级不同，当各指标间的水平相差很大时，若直接利用原始指标值进行分析，就会突出数值水平较高的指标在综合分析中的作用，降低数值水平较低指标的作用。首先对数据进行标准化处理。

Table 1. The symbol table
表 1. 符号表

y	肺癌死亡率(%)
x1	空气污染指数(AQI)
x2	人均工业烟尘(吨/人)
x3	烟民数量(人)
x4	森林覆盖率(%)
x5	在职医生数量(人)
x6	政府医疗资金投入(千万)

3.4. 数据预处理

3.4.1. 相关分析

为进一步考察变量之间的相关性，需要对它们进行相关分析，从结果显示中可见肺癌死亡率与空气污染指数、人均工业烟尘、烟民数量、森林覆盖率、在职医生数量和政府医疗资金投入相关性较为显著，空气污染指数、人均工业烟尘和森林覆盖率关系较为显著，在职医生数量和政府医疗资金投入关系较为显著。

3.4.2. 建立各相关因素间的回归方程

运用逐步回归分析法建立回归方程：1) 因变量是肺癌死亡率，自变量是空气污染指数、人均工业烟尘、烟民数量、森林覆盖率、在职医生数量和政府医疗资金投入 6 个因素的回归方程；2) 因变量是肺癌死亡率，自变量是空气污染指数、烟民数量和政府医疗资金投入 3 个因素的回归方程；3) 因变量是空气污染指数为因变量，自变量是烟民数量的回归方程；4) 因变量是人均工业烟尘为因变量，自变量是空气污染指数和森林覆盖率 2 个因素的回归方程；5) 因变量是在职医生数量为因变量，自变量是以烟民数量、政府医疗资金投入的回归方程。其中对第二个回归方程的逐步回归系数表，如表 2。

Table 2. Regression analysis results using lung cancer mortality as the dependent variable
表 2. 以肺癌死亡率为因变量的回归分析结果

模型	系数*			t	Sig.
	非标准化系数		标准系数		
	B	标准误差			
(常量)	−2916E−45	0.096		0.000	1.000
空气污染指数	0.266	0.138	0.266	1.921	0.001
烟民数量	0.329	0.129	0.259	1.642	0.005
政府医疗资金投入	0.313	0.117	0.237	1.973	0.002

*因变量：Zscane (肺癌死亡率)。

结果显示，以肺癌死亡率为因变量，以空气污染指数、烟民数量和政府医疗资金投入 3 个因素为自变量的回归方程的 $R^2 = 0.509$，F 值为 17.998，对应的 P 值为 0.001，说明该回归方程通过了 F 检验。

由表 2 可知，在 0.1 的显著性水平下，对以肺癌死亡率为因变量，空气污染指数、烟民数量和政府医疗资金投入 3 个因素为自变量的回归方程的各自变量的回归系数的 t 检验对应的 P 值都显著，说明该回归方程的各回归系数都通过了检验。

其他四个回归方程的分析过程同上，将所有回归分析结果汇总为表 3。

Table 3. Summary table of regression analysis results
表 3. 回归分析结果汇总表

因变量 / 自变量	空气污染指数		在职医生数量		人均工业烟尘		肺癌死亡率	
	β	P	β	P	β	P	β	P
空气污染指数					1.105	0.000	1.921	0.001
人均工业烟尘								
0 烟民数量	0.025	0.001	0.006	0.827			1.642	0.005
森林覆盖率					0.005	0.003		
在职医生数量								
政府医疗资金投入			0.561	0.001			1.973	0.002

3.4.3. 修改并完成路径图的绘制

通过表 4，对初始路径图进行修改，修改后的路径图如图 2 所示。

Table 4. Table of coefficients after normalization of data
表 4. 标准化数据后的系数表

自变量	应变量	相关系数	路径系数	决定系数
空气污染指数	肺癌死亡率	0.581	0.233	0.135
烟民数量	肺癌死亡率	0.663	0.421	0.279
政府医疗资金投入	肺癌死亡率	−0.345	−0.220	0.076
森林覆盖率	空气污染指数	−0.827	−0.827	0.684
烟民数量	空气污染指数	0.329	0.329	0.108
人均工业烟尘	空气污染指数	0.705	0.690	0.486

通过以上的路径图，可以得到每个自变量对肺癌死亡率的直接效应和间接效应，其中直接效应指的是自变量直接影响肺癌死亡率的路径系数，间接效应指的是自变量通过其他变量间接地影响肺癌死亡率，其大小等于各路径系数的乘积。

由图 2 可知，烟民数量除了对肺癌死亡率有直接效应外，还通过对空气污染指数的影响产生了对肺癌死亡率的间接影响。具体如下：

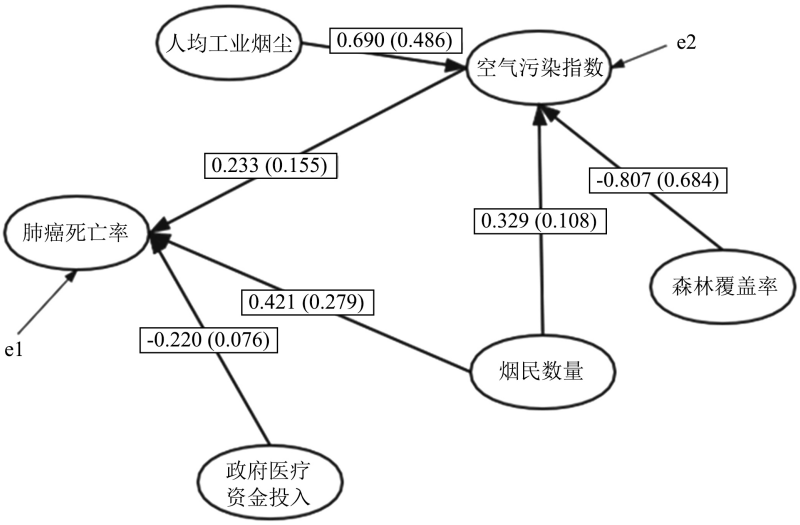

Figure 2. Path diagram
图 2. 路径图

烟民数量对肺癌死亡率的直接效应为 0.421；烟民数量通过空气污染指数对肺癌死亡率的间接影响为：0.329 × 0.233 = 0.076657；于是，烟民数量对肺癌死亡率的总效应为直接效应与间接效应之和：0.421 + 0.076657 = 0.55649。

同理得到人均工业烟尘通过空气污染指数对肺癌死亡率的间接效应为：0.690 × 0.233 = 0.16077；

空气污染指数对肺癌死亡率的直接效应为 0.233；

政府医疗资金投入对肺癌死亡率的直接效应为 0.220；

森林覆盖率通过空气污染指数对肺癌死亡率的间接效应为 0.827 × 0.233 = 0.192691。

4. 模型的评价与改进

观察肺癌死亡率的路径模型，我们发现烟民数量对空气污染指数影响的路径系数超越预期。同时，如人均工业烟尘和森林覆盖率等对肺癌死亡率影响较大的变量未通过显著性检验，模型存在明显的缺陷。由于采用逐步回归法消除了变量间的多重共线性，于是我们考虑是否是数据标准化带来的问题。

现将未经标准化的原始数据整理导入 SPSS，并重复上述统计分析过程，依次做数据的相关分析、逐步回归和显著性检验后，得到新的各因变量与自变量之间的相关系数和回归系数，见表 5。

Table 5. Table of coefficients with unstandardized data
表 5. 未经标准化数据的系数表

自变量	应变量	相关系数	路径系数	决定系数
空气污染指数	肺癌死亡率	0.416	0.233	0.097
烟民数量	肺癌死亡率	0.647	0.421	0.272
政府医疗资金投入	肺癌死亡率	−0.214	−0.220	0.047
森林覆盖率	肺癌死亡率	−0.773	−0.612	0.473
人均工业烟尘	肺癌死亡率	0.476	0.329	0.157
森林覆盖率	空气污染指数	−0.815	−0.782	0.637
人均工业烟尘	空气污染指数	0.844	0.690	0.582

对比原系数表，我们可以直观地发现部分相关系数和回归系数发生了较大的改变，制作新的路径图如图 3。其中各路径及相应路径系数较符合我们的预期，说明之前的路径模型的确存在数据标准化带来的问题。

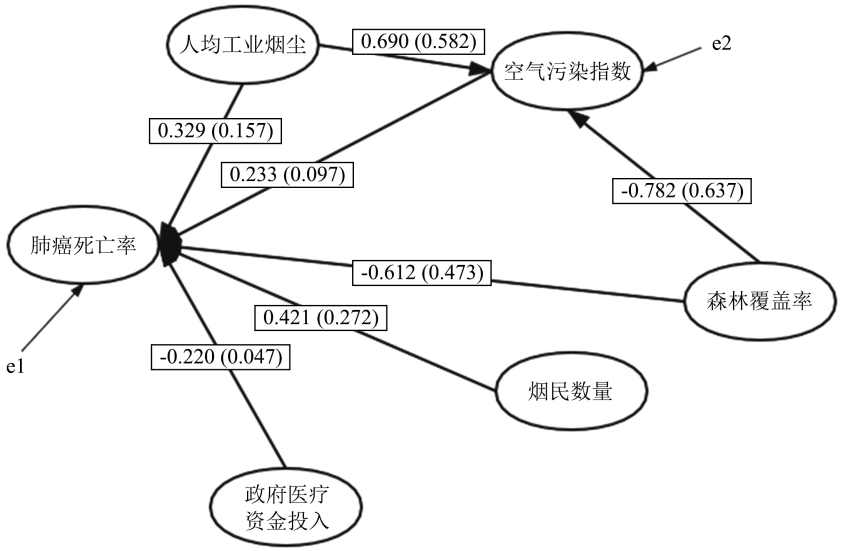

Figure 3. Modified path diagram
图 3. 修正后路径图

5. 结果与讨论

5.1. 结构与建议

根据以上的分析结果可以得出以下初步结论：人均工业烟尘、空气污染指数、森林覆盖率、烟民数量和政府医疗资金投入与肺癌死亡率呈显著的相关关系，而在职医生数量对肺癌死亡率有一定的影响，但是不显著，这与先前的假设有一定的差别。另外，烟民数量的总效应最高，开始吸烟的年龄越小，患肺癌的几率越高。此外，吸烟不仅直接影响本人的身体健康，还对周围人群的健康产生不良影响，导致被动吸烟者肺癌患病率明显增加。城市居民肺癌的发病率比农村高，这可能与城市大气污染和烟尘中含有致癌物质有关。因此应该提倡不吸烟，并加强城市环境卫生工作。1950~1999 年期间，美国先后经历了第一次和第二次工业革命，工业和交通得到发展的同时，石油，煤和内燃机等燃烧后和沥青公路尘埃产生的苯等有害物质污染大气。基于结果，给出以下建议：避免接触与肺癌发生有关的因素，不吸烟或及早戒烟，同时避免吸入二手烟，减少接触烟尘、车尾气等刺激性有害气体。

5.2. 不足与展望

本文统计分析的部分结论仍不完全符合预期的假设，这或许与选取变量有关。受专业知识限制，无法准确全面地确定影响肺癌死亡率的因素。此外，部分原本将纳入模型的影响因素因无法找到可靠的数据来源而被迫删除或选取其他因素替代，对最终的结果产生一定的影响。

参考文献

[1] Freedman, D.A. 统计模型[M]. 吴喜之, 译. 北京: 机械工业出版社, 2010.

[2] 周脉耕, 王黎君, 黄正京, 等. 人口老化及危险因素改变对肺癌死亡率的影响[J]. 中国卫生统计, 2002, 19(3):

161-161.

[3] 冯丹, 徐桂永, 赵连伟, 等. 肺癌死亡率与大气污染关系的灰色关联分析[J]. 数理医药学杂志, 2001, 14(4): 364-365.

[4] 张国钦, 王宁, 王涛, 等. 北京市城区居民烟草消费和大气污染对肺癌死亡率的影响[J]. 环境与健康杂志, 2009(8): 666-669.

[5] 李兰曼, 魏玮. 肺癌流行病学和危险因素研究进展[J]. 肿瘤研究与临床, 2018, 30(12): 875-879.