Statistics and Application 统计学与应用, 2022, 11(4), 722-731
Published Online August 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.114076

Hans汉斯

基于机器学习方法的肺癌病人生存时间报告

李竺珊

云南财经大学，云南 昆明

收稿日期：2022年7月9日；录用日期：2022年7月20日；发布日期：2022年8月2日

摘　要

肺癌作为发病率及死亡率极高的恶性肿瘤之一，给病人也给家庭带来了无可弥补的巨大的精神压力和经济压力。本文基于R软件自带的survival-lung数据集，利用传统的线性模型、决策树、随机森林以及岭回归四种模型，研究了影响肺癌患者生存时间的因素，对医学工作者制定有效、合理的医疗方案具有现实意义。研究表明：第一，在相关关系研究中，肺癌患者的生存时间与年龄、心电图表现分、最近六个月减重、膳食中消耗的卡路里、医疗机构数等都具有较强的相关关系；第二，通过性别对比研究生存时间，得出女性肺癌患者生存时间高于男性肺癌生存时间；第三，通过生存时间曲线图得出，肺癌患者的生存时间与生存概率呈负相关关系。最后，根据研究结果提出相应的延长肺癌患者生存时间的建议。

关键词

肺癌，生存时间，Kaplan-Meier生存图

Survival Time Report of Lung Cancer Patients Based on Machine Learning Method

Zhushan Li

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Jul. 9th, 2022; accepted: Jul. 20th, 2022; published: Aug. 2nd, 2022

Abstract

As one of the malignant tumors with high incidence rate and mortality, lung cancer has brought irreparable mental and economic pressure to patients and families. Based on the survival-lung data set of R software, this paper studies the factors affecting the survival time of lung cancer patients by using the traditional linear model, decision tree, random forest and ridge regression models, which has practical significance for medical workers to formulate effective and reasonable medical plans. The results show that: First, in the correlation study, the survival time of lung can-

文章引用: 李竺珊. 基于机器学习方法的肺癌病人生存时间报告[J]. 统计学与应用, 2022, 11(4): 722-731.
DOI: 10.12677/sa.2022.114076

cer patients has a strong correlation with age, ECG score, wt.loss, calories consumed in diet, and the number of medical institutions; Second, the survival time of female lung cancer patients is higher than that of male lung cancer patients through gender comparison; Thirdly, the survival time curve shows that the survival time of lung cancer patients is negatively correlated with the survival probability. Finally, according to the results of the study, some suggestions were put forward to prolong the survival time of lung cancer patients.

Keywords

Lung Cancer, Survival Time, Kaplan-Meier Survival Map

1. 引言

1.1. 研究背景

肺癌作为中国乃至全球报告的最大恶性肿瘤之一，不仅对人类健康甚至对人类生命造成了严重的威胁。根据我国癌症中心发布的全国癌症报告数据显示，肺癌每年新发病例高达七十万，而生存率不足 20%，病死率极高。导致肺癌确诊的原因尚不明确，且每个患者肺癌确诊后的生存时间也有所不同，大量资料表明，长期吸烟人员确诊肺癌的概率比不吸烟人员确诊的概率高 10~20 倍。并且患肺癌的几率与开始吸烟的年龄有关，越小开始吸烟，患肺癌的几率就越高[1]。还有学者研究发现，肺癌发病及致死率在所有癌症患者中是最高的，且男性发病及致死率高于女性[2]。

相关研究显示，肺癌病人的生存时间与三个重要因素有关：

1) 病人心理因素。大多数情况下，癌症在晚期才会被发现，肺癌细胞很可能不止存在于肺部，还扩散于其他身体器官。此时，家人的陪伴和鼓励给癌症病人一个健康、良好的心态，能够帮助病人积极配合治疗方案，能够在很大程度上缓解病情。不乏有报道癌症病人积极面对生活、放松心情，来对抗病魔的案例。所谓积郁成疾，想必心态对身体健康的重要性也就不言而喻了。

2) 病人饮食习惯。在采取治疗的情况下，癌症病人死亡原因可能并非仅因为癌症，很可能是治疗阶段营养不均衡。由于不能正常吃喝，身体所需的微量元素没有及时补充，每天蔬菜水果也需要一定量的摄入。

3) 治疗方法。治疗的目的在于缓解病情，减轻痛苦，延长寿命。盲目统一的治疗方案反而会给病人及家属带来更大的痛苦。因此，需要根据每个病人的不同情况，制定有针对性、合理、有效的治疗方案。

1.2. 研究目的及意义

本文计划使用 R 软件自带的 survival-lung 数据对肺癌病人生存时间进行分析。通过考察年龄、性别及其他变量对肺癌患者生存时间的影响，得出相应的结论，从而针对不同的患者制定高效的治疗方案，延长肺癌患者生存时间。

2. 数据介绍

本案例分析数据来自 R 软件自带 survival 数据集中的 lung 数据。由于肺癌病人生存时间与生活环境、

自身状况因素都有比较密切的关系[3]。基于以上考虑，本文将该数据中所有变量纳入对肺癌病人生存时间与性别、膳食中消耗的卡路里、过去六个月的体重下降等的研究来考察生存时间长度。该数据一共 228 条记录，10 个指标。各变量指标具体描述如表 1 所示：

Table 1. Specific values of variables
表 1. 变量的具体取值

变量类型	变量名	详细说明	备注
因变量	Time： 生存时间	肺癌病人生存时间	定量变量
自变量	Status： 审查状况	1 为审查；2 为未审查	定性变量
	Age： 年龄	肺癌病人年龄	定量变量
	Sex： 性别	1 为男性；2 为女性	定性变量
	Ph.ecog： 心电图的表现评分	0 为无症状；1 为有症状但可行走； 2 为在床上 < 每天 50%；3 为在 床上 > 每天 50%；4 为卧床	定性变量
	Ph.karno： Karnofsky 表演评分		定量变量
	Pat.karno： 病人的卡诺夫斯基表现评分		定量变量
	Meal.cal： 膳食中消耗的卡路里		定量变量
	Wt.loss： 最近六个月的减肥		定量变量
	Inst： 机构数量		定量变量

3. 实证分析

3.1. 各种方法的实证性分析

3.1.1. 数据的描述性分析

1) 因变量的描述性分析

从图 1 中可以看出，因变量生存时间总体呈偏正态分布的特征，其峰值出现在 200 附近，呈右偏趋势，则说明该数据存在极大值，会导致均值向右偏。

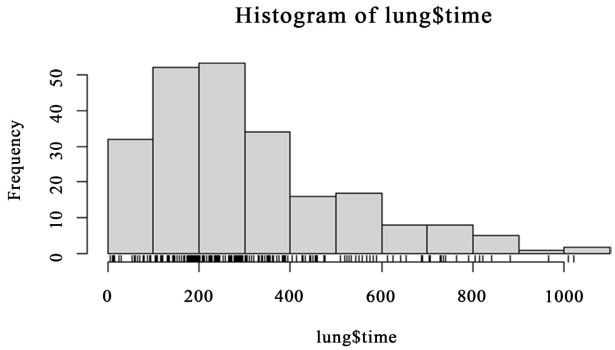

Figure 1. Histogram of dependent variable
图 1. 因变量直方图

2) 变量之间的散点图矩阵

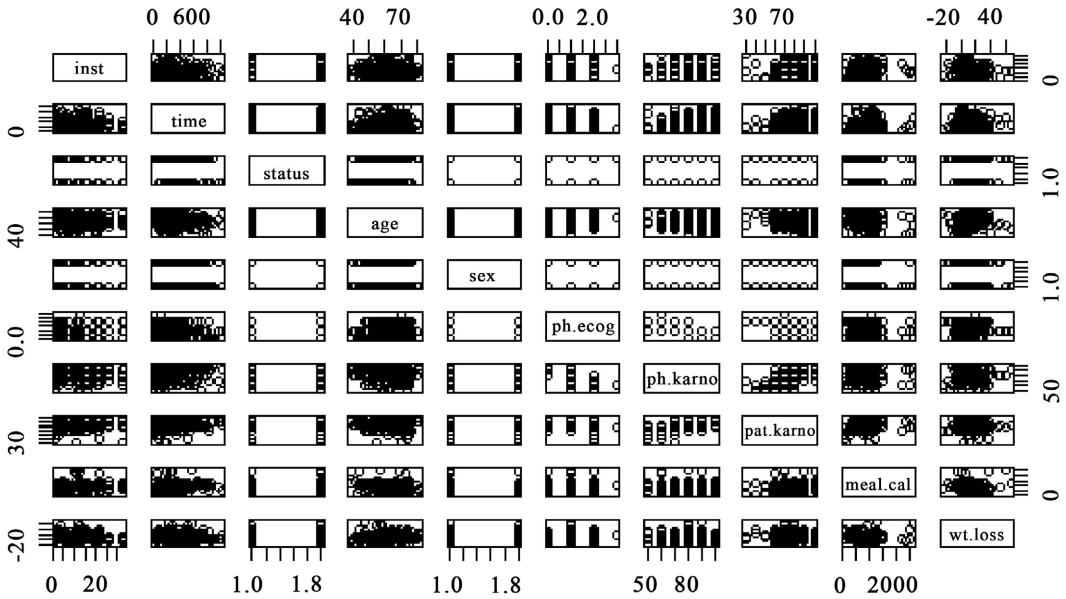

Figure 2. Scatter matrix between variables
图 2. 变量之间的散点图矩阵

从图 2 中可以看出，inst (机构数)与 time (生存时间)存在相关关系；time (生存时间)与 age (年龄)、ph.ecog (心电图表现分)、pat.larno (卡诺夫斯基表现评分)、meal.cal (膳食中消耗的卡路里)以及 wt.loss (最近六个月减肥)均存在相关关系；age (年龄)与 meal.cal (膳食中消耗的卡路里)以及 wt.loss (最近六个月减肥)存在相关关系；meal.cal (膳食中消耗的卡路里)与 wt.loss (最近六个月减肥)存在相关关系。

3) 变量的箱线图

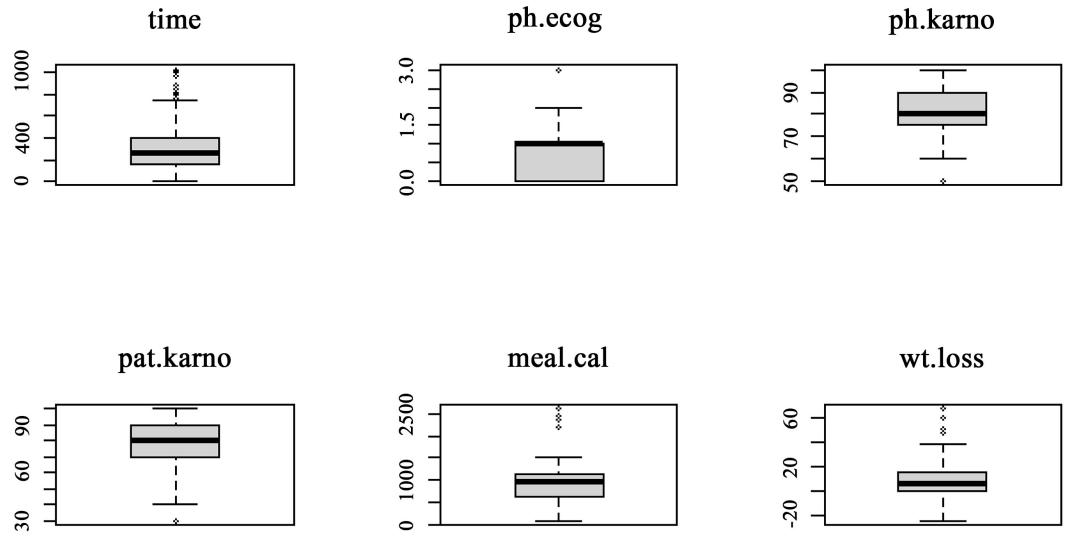

Figure 3. Box diagram of variables
图 3. 变量的箱线图

从图 3 中可以看出，生存时间、心电图的表现评分、Karnofsky 表演评分、病人的卡诺夫斯基表现评分、膳食中消耗的卡路里、过去六个月的体重下降这几个变量均存在异常点。

3.1.2. 线性回归分析

线性回归模型诊断结果如图 4 所示：

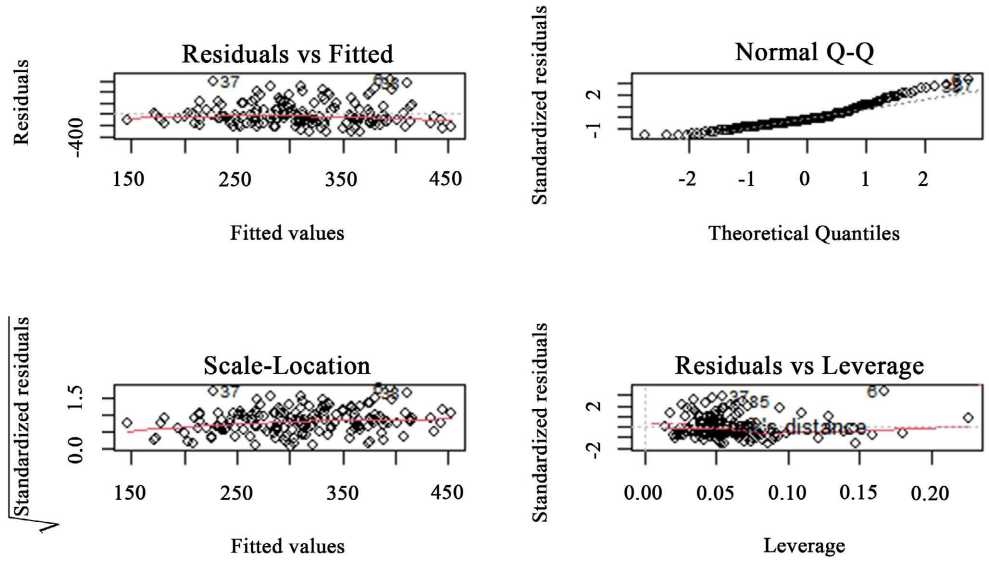

Figure 4. Model diagnosis of linear regression
图 4. 线性回归的模型诊断

从图 4 中可以看出，线性回归模型不存在异方差现象，残差服从正态分布。其中，Residuals vs Fitted 图是预测值与残差的关系图，根据该图可以判断该 lung 数据集的线性关系以及方差齐性；Normal Q-Q 图是残差 QQ 图，根据该图判断残差服从正态分布；Scale-Location 图是用来检验等方差假设的，根据该图可以判断模型的方差齐性；Residuals vs Leverage 图是杠杆和残差的关系图，用于检查数据是否存在非常极端的点。

线性回归是利用回归分析来明确自变量与因变量之间的相互影响关系，通常影响因变量 y 的因素存在许多个，假设存在有 $x_1, x_2, \cdots, x_k$， k 个自变量，则自变量与因变量之间的关系可以被线性表示为：

$$y = \beta_0 + \beta_1 x_2 + \cdots + \beta_k x_k + \varepsilon$$

下列表 2 为线性回归系数表：

Table 2. Linear regression coefficient
表 2. 线性回归系数表

变量	inst	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss	常数项
系数	0.64	−39.89	0.26	48.37	−84.21	−3.31	1.5	0.04	1.77	470.72

根据表 2 结果，肺癌病人生存时间的线性表达式为：

$$\begin{aligned}\text{time} = {} & 0.64 * \text{inst} - 39.89 * \text{status} + 0.26 * \text{age} + 48.37 * \text{sex} - 84.21 * \text{ph.ecog} - 3.31 * \text{ph.karno} \\ & + 1.5 * \text{pat.karno} + 0.04 * \text{meal.cal} + 1.77 * \text{wt.loss} + 470.72\end{aligned}$$

3.1.3. 决策树回归分析

决策树模型是一种简单直观的树形结构，广泛运用于分类以及回归的情况。一个完整的决策树由一个根节点、有向边以及若干非叶节点和叶子结点组成。

根节点(Root Node)是决策树开始决策的起点。通过每个内部节点问题对决策树进行分支，这一过程就是对数据的某个特性进行测试，根据测试结果，将数据分配到内部节点的某个分支，再通过下一个测试分配到下一个分支。通过所有的内部节点后，就到达最终的叶子节点，叶子节点也就是最终结果，显示的百分比为期望值。中间的分支可以无限分配，以达到最优的决策结果。

利用 R 语言的 rpart.plot 包我们得到决策树，如图 5 所示。

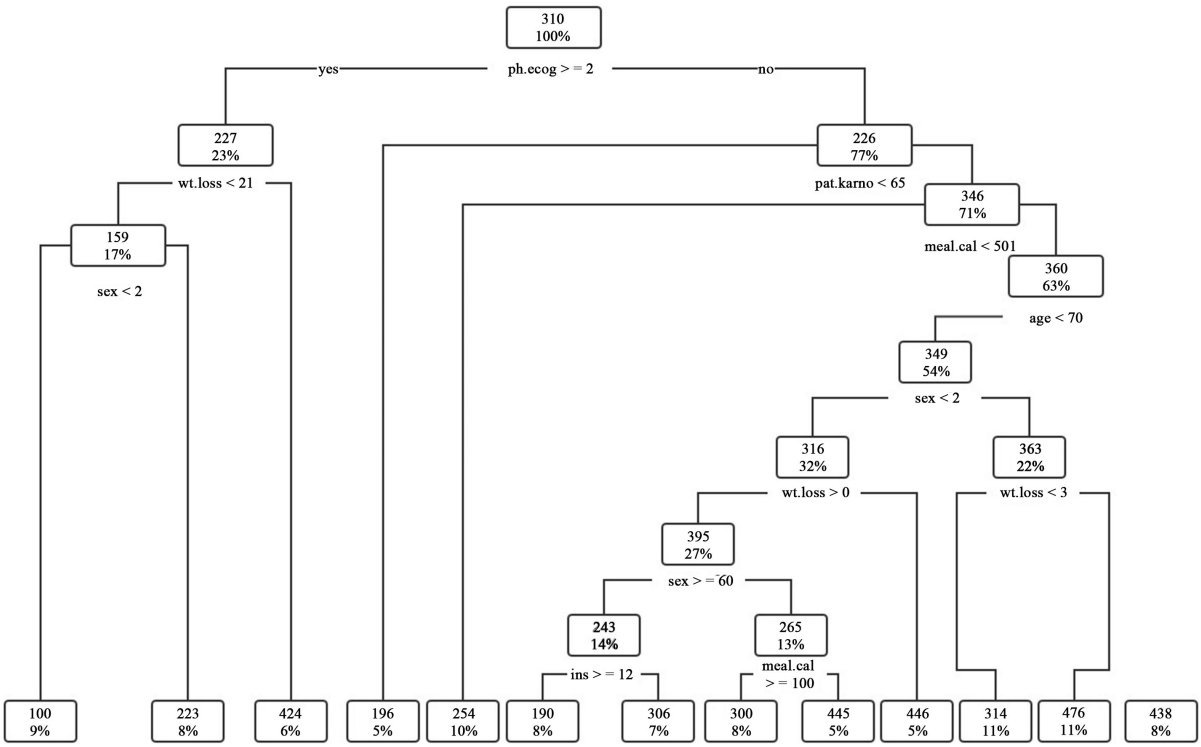

Figure 5. Regression tree
图 5. 回归树

从图 5 结果来看，回归树共有十四次分割，根节点分割变量为 ph.ecog (心电图表现分)，其分割点为 2，值越大表示病情越严重(0 表示无症状；1 表示有症状但可以行走；2 表示在床上的时间小于 50%；3 表示在床上的时间大于 50%；4 表示卧床)大于等于该值的分到左边，小于该值的分到右边。左边分支为 wt.loss (最近六个月减肥)，以 21 磅为分割点，大于等于 21 磅的到右边分支，最终结果为生存时间 424 天，期望值为 6%。小于 21 磅的部分继续通过 sex (性别)进行分割，最终结果为生存时间为 100 天的男性占比 9%，生存时间为 223 天的女性，期望值为 8%。

心电图表现分右边分支为 pat.karno (卡诺夫斯基表现评分)小于 65 分的患者生存时间为 196 天，期望值为 5%；表现评分大于等于 65 分，meal.cal (膳食中消耗的卡路里)小于 501 卡的患者生存时间为 254 天，期望值为 10%；膳食中消耗卡路里大于 501 卡，年龄 60~70 岁之间的男性患者，且最近六个月体重有所减轻，周围医疗机构数大于等于 12 的生存时间为 190 天，期望值为 8%；医疗机构数少于 12 的患者生存时间为 306 天，期望值为 7%。

年龄在 60 岁以下的男性患者，且最近六个月体重有所减轻，膳食中消耗的卡路里大于等于 1100 卡的生存时间为 300 天，期望值为 8%；膳食中消耗卡路里小于 1100 卡的患者生存时间为 445 天，期望值为 5%。70 岁以下的男性患者，且最近六个月体重没有减轻的生存时间为 446 天，期望值为 5%。

年龄 70 岁以下的女性患者，最近六个月体重减轻在 3 磅以内的生存时间为 314 天，期望值为 11%，是生存时间的众数；最近六个月体重减轻 3 磅以上的女性患者生存时间为 476 天，期望值为 11%。

最后一条分支为心电图评分在 2 分以内卡诺夫斯基表现评分 65 分以上，膳食中消耗的卡路里在 501 卡以上，且年龄在 70 岁以上的患者生存时间为 438 天，期望值为 8%。

3.1.4. 随机森林回归

随机森林是一种集训练与测试的一种分类器，顾名思义由多棵相互独立的决策树构成，且最终模型是在众多决策树共同影响作用下形成的[4]。

在 R 语言中使用 importance( )函数可以浏览各个变量的重要性：

Table 3. Importance measurement of random forest for each variable
表 3. 随机森林对各个变量的重要性度量

随机森林对各个变量的重要性度量		
	%IncMSE	IncNodePurity
inst	−856.9933	762578.2
status	739.7872	241015.0
age	−333.6646	1014770.4
sex	1587.2545	230568.6
ph.ecog	4536.3915	369351.2
ph.karno	4365.2048	844147.5
pat.karno	2166.9551	601802.2
meal.cal	533.1502	1071148.3
wt.loss	642.5195	1072908.8

表 3 中第一列数字是从替换该变量而致精确度平均递减的角度来衡量变量重要性，基于当一个给定的变量被排除在模型之外时，预测样本的准确性的平均减小值。第二列数字则是从以该变量为拆分变量所造成的均方误差的平均递减的角度来衡量变量重要性。

其中，Karnofsky 表演评分以及病人的卡诺夫斯基表演评分是目前最重要的两个变量。

将上表结果可视化展示，如图 6 所示。

3.1.5. 岭回归

岭回归方法是一种改良之后的 OLS (最小二乘估计法)。是适用于存在共线性问题的有偏估计，要求自变量 $x_1, x_2, \cdots, x_k$ 至少有一个是定量变量。通过检验自变量与因变量之间的关系，分析岭迹图，确定岭参数[3]。

交叉验证选择岭回归的参数 λ 的过程中最优模型所对应的均方误差的运动轨迹，直观地看出最佳 λ 的大致取值，如图 7 所示：

随机森林的特征选择

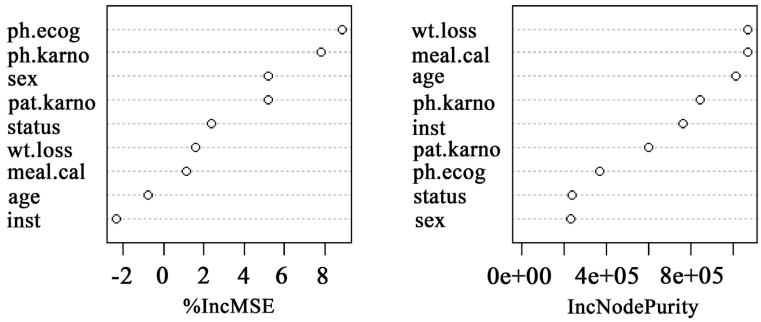

Figure 6. Feature selection of random forest
图 6. 随机森林的特征选择

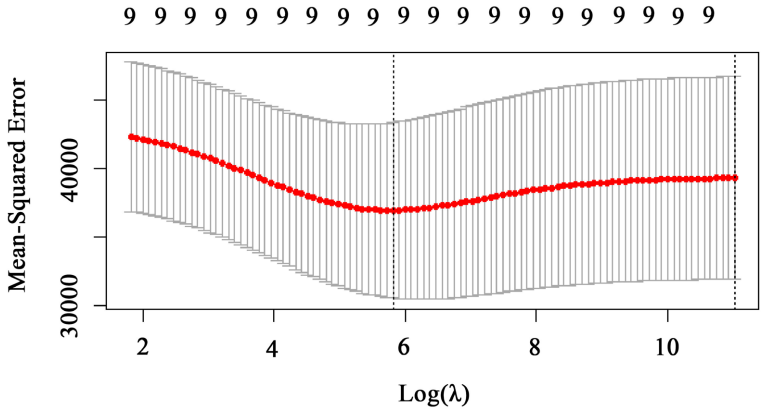

Figure 7. Trajectory of parameters and mean square error of ridge regression
图 7. 岭回归的参数与均方误差的运动轨迹

bestlam = 336.6244

从上述输出结果可知，使得交叉验证误差最小的 λ 为 336.6244。基于整个数据集，使用交叉验证所得的 λ 值重新拟合岭回归模型，得到的模型系数估计情况，如表 4 所示：

Table 4. Linear regression fitting results
表 4. 线性回归拟合结果

变量	inst	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss	常数项
系数	0.18	−22.97	0.26	17.51	−16.12	0.09	0.67	0.01	0.4	278.52

根据表 4 可知，从而重新拟合得到的肺癌病人生存时间回归模型表达式为：

$$\text{time} = 0.18 * \text{inst} - 22.97 * \text{status} + 0.26 * \text{sex} - 16.12 * \text{ph.ecog} + 0.09 * \text{ph.karno} + 0.67 * \text{pat.karno} + 0.01 * \text{meal.cal} + 0.4 * \text{wt.loss} + 278.52$$

3.1.6. 按性别计算生存时间

从表 5 可知，男性肺癌患者平均生存时间为 339 天，下分位时间为 223 天，上分位时间为 353 天；

女性肺癌患者平均生存时间为 460 天，下分位时间为 345 天，上分位时间为 641 天。从而得出女性平均生存时间高于男性，即与男性肺癌患者相比，女性肺癌患者更具有生存优势。

Table 5. Survival time by sex
表 5. 按性别计算生存时间

	records	n.max	n.start	events	rmean	Se (rmean)	median	0.95LCL	0.95UCL
Sex = 1	103	103	103	82	339.19	26.51	284	223	353
Sex = 2	64	64	64	38	460.24	42.51	426	345	641

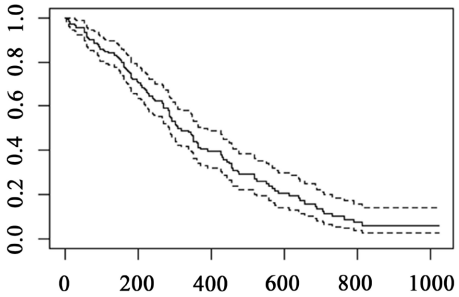

Figure 8. Kaplan-Meier survival char
图 8. Kaplan-Meier 生存图

Kaplan-Meier 是估计生存时间和概率最常用的非参数方法。图 8 中横坐标表示生存时间，纵坐标表示对应的生存概率。两条虚线内的区域表示置信区间。根据该研究，对应图 8 所示，生存时间与生存概率之间呈反比例关系。在起点时刻，生存概率为 100%；生存时间为 200 天的概率大概是 60%；生存时间为 400 天的概率大概是 35%。

4. 结论

4.1. 结果分析

本文为了研究 R 自带的数据集 survival-lung 中影响肺癌病人存活时间的主要影响因素，选取了传统的线性模型、决策树、随机森林、岭回归四种模型，分析影响肺癌病人存活时间的主要因素。最终分析出，影响肺癌病人存活时间的主要因素是心电图表现评分、Karnofsky 表现评分、膳食中消耗的卡路里。在此数据集上表现较好的模型是传统的线性模型以及随机森林。

研究结果显示：

1) 在相关关系研究中，肺癌患者的生存时间与年龄、心电图表现分、最近六个月减重、膳食中消耗的卡路里、医疗机构数等都具有较强的相关关系；

2) 通过性别对比研究生存时间，得出女性肺癌患者生存时间高于男性肺癌生存时间[4]，女性肺癌患者比男性肺癌患者更具有生存优势；

3) 通过生存时间曲线图得出，肺癌患者的生存时间与生存概率呈负相关关系。

4.2. 相关建议

从我国乃至世界肺癌发病、致死的数据报告来看，肺癌患者的数量每年以几十万的速度增长，不仅严重地危害了肺癌患者的生命健康，更是对家庭造成了不可弥补的伤害[5]。传统的统一治疗方案存在很

大的局限性，难以根据患者实际情况达到高效的治疗效果。

第一，大量提倡居民定期检查。政府部门应该对居民定期体检进行大力宣传，通过预防或者提早发现肿瘤，及时发现，尽早治疗；已经确诊的患者进行详细检查，明确病情[4]。

第二，制定个性化治疗方案。每个患者的具体情况有所不同，根据具体情况制定治疗方案，控制病情，延长生存期。

第三，减少吸烟。吸烟作为肺癌的头号致病因素，有研究表明吸烟者肺癌发病率是不吸烟者的 10~20 倍。因此，减少吸烟、戒烟是降低肺癌发病率的重要因素。

第四，加强体育锻炼。大部分疾病是能够通过良好的生活方式来预防的。

致　谢

感谢学校老师对我的严格要求，在学习方面的指导以及生活方面的帮助。感谢各位同学们不厌其烦地为我解答学术问题。

参考文献

[1] 黄天姿, 王素珍, 袁媛, 王莎莎. 影响肺癌病人生存期的相关因素分析[J]. 护理研究, 2016, 30(32): 4033-4036.

[2] 庄巧蕙. 基于改进随机森林算法的研究与应用[D]: [硕士学位论文]. 泉州: 华侨大学, 2019.

[3] 王纯杰, 温男, 马元嘉. 基于岭回归和 Lasso 回归的螺纹钢期货价格实证分析[J]. 吉林师范大学学报(自然科学版), 2020, 41(1): 36-41.

[4] 肖永红, 张翠敏, 张娟, 阎子海. 肺癌病人生存时间与生存质量的相关分析[J]. 现代预防医学, 2009(20): 3823-3825.

[5] 唐海涛. 基于 LASSO 的肝癌预后风险基因的数据挖掘分析[D]: [硕士学位论文]. 兰州: 兰州大学, 2019