电影票房影响因素探究与预测

——基于线性回归模型与神经网络模型

吴张璇　2020110540

费璐瑶　2020110731

辛柏赢　2020111753

牟思宜　2020111082

2022 年 12 月 23 日

摘 要

21 世纪以来，随着电影行业的发展与成熟，电影行业逐渐成为文娱产业的重要一环。本文基于 TMDB 网页上所提供的电影数据，历年奥斯卡奖项的提名信息以及爬取了 boxoffice.com 网页上提供的电影上映后的荧幕数的相关数据影响电影票房的相关因素进行了探究。在对数据描述性分析后，根据变量间的线性关系，建立了 R 方为 0.76，解释性较高的多元线性回归模型。

在建立了全模型之后，基于模型诊断结果，模型可能存在异方差的问题，于是利用加权最小二乘法对其进行优化，使得随机误差项满足同方差假定。接着根据 AIC 准则进行逐步回归，得到了拟合效果较好，R 方超过 0.9 的回归模型，并对影响电影票房。针对数据正态性不佳的特点，使用广义线性模型 Lasso 回归对模型进行改进和优化，实现变量选择和参数估计。

为了找到高票房电影之间的共性特征，并对高票房电影进行预测，我们将票房分为高、中、低三个档次，将原数据分为 90%与 10%，划分出测试集和训练集，通过 Anova 分析后进行了有序多分类 Logistic 回归。该模型预测效果达到 84.6%，预测效果较好。最后建立了神经网络模型再次提高了预测的准确度，达到了 88.7%。

关键词：电影票房 线性回归 加权最小二乘法 Lasso 回归 Logistic 回归 神经网络

目 录

一、 研究背景与研究意义

近年来随着人们生活水平的改善，人们对于精神世界的追求越来越高，文娱产业的发展也越来越受到重视。电影作为文娱产业的重要支柱，是不同国家与地区间文化交流与传达的主要途径之一。电影票房不仅体现出该国甚至世界范围内的观众的审美以及价值取向，也一定程度上体现了电影技术水平的发展。

近十年来，全球电影票房保持着增长态势，电影娱乐成为了文娱产业的重要支柱，也是体验经济时代的重要消费方式。据《电影蓝皮书：全球电影产业发展报告（2019）》发布的数据显示，2018 年全球电影票房达 411 亿美元，创下历史新高，较 2017 年增长约 1.48%。

电影市场存在着丰厚的投资回报，对于投资者来说具有很大的潜力。但随着互联网的普及与发展，人们的价值取向与娱乐偏好呈现多元化的态势，使得电影投资的风险增加。因此，电影的投资方出于利益考虑，在选择投资的电影时，对于观众的审美和价值取向十分看重，希望由此降低投资风险，提高投资效率。

网络信息与数据挖掘技术领域飞速发展，使得利用神经网络等模型来预测电影票房成为可能，有利于捕捉大众喜好，提高投资收益率。同时，通过电影票房影响因素的研究使我们对整个电影行业的发展有更明晰的了解，掌握电影市场未来的发展趋势。**如何结合电影的题材、演职人员、预算、是否原创等信息对电影票房进行预测，对于爆款电影的挖掘具有重要意义**。

二、 国内外文献综述

2.1 国外研究综述

国外的电影产业发展已久，对于电影票房的研究较为深入与完善。国外最早有关电影票房的研究起于 1980 年，多以美国的早期电影数据为例，对电影票房的影响因素进行探究。

美国的经济学家 Litman 和 Kohl(1989)最早确立了电影票房影响因素研究的基本模型，并提出运用传播学方法与经济学方法来进行研究与预测。在此基础之上，Sochay(1994)进一步对自变量进行细分，并提出市场集中度、营销等新的自变量，通过更加全面的数据对票房影响因素与预测进行研究。后期也有诸多学者对该问题进行进一步探索，并得出一些结论。**De Vany 和 Walls(1999)和 Bagella 和 Becchetti(1999)发现电影票房分布的偏度系数较大，因此对电影票当进行挡档位划分并换用 Logit 模型进行分析**。Lee 和 Chang(2009)则运用贝叶斯网络，进一步丰富了研究的方法。G. 等(2022)运用随机森林回归模型，提供了较优的预算、运行时间、明星影响力和预期人气，具有实践意义。

2.2 国内研究综述

国内对电影票房的研究起步较晚，并且许多是基于艺术学等理论来进行的定性研究，定量研究的数量较少，但近年来也在逐渐上升。**王铮和许敏(2013)选用 2007-2012 年中国的 554 部电影作为样本，运用 OLS 方法和 Logit 模型进行回归，发现明星与导演存在显著的票房效应**。郑坚和周尚波(2014) 提出一种基于反馈神经网络的电影票房预测模型，进行具有针对性的改进后得到较好的预测效果与分类性能。池建宇(2016)发现明星导演的影响力甚至超过了明星演员对票房的影响。**苏永华和王哲平(2021)运用多元线性回归与 BP 神经网络算法对国内电影网络口碑对票房的影响进行了探索，发现网络口碑有显著正向影响，但是在预测模型中的作用很小**。国内研究往往以中国的电影市场作为研究主题，研究结果对于国内电影的投资、执导、选角、宣传等多方面具有指导作用。

三、 数据预处理与描述性分析

3.1 数据获取与变量说明

本文数据来源为 **Kaggle 公司提供的 TMDB 5000 Movie Dataset 以及 The Oscar Award, 1927 - 2020。并且通过爬虫技术从 boxoffice.com 网站爬取了 1953 部电影上映时的荧幕数 Movie Theater。**

Movie dataset 数据集共有 4083 个数据集共包括 4083 部电影的基本信息。时间跨度为 1916—2016，记录了电影的语言、演职人员、内容简介、宣传语、预算、票房、评分、热度等等。The Oscar Award 则记录了 1927-2020 年的奥斯卡获奖信息。

数据集的主要信息如表 1 所示：

表 1-1 Movie dataset 主要信息

变量类型	变量名称		变量说明
使用量	电影票房	revenue	该电影上映后的总票房
电影投入	电影预算	budget	该电影的发行预算
电影内容	电影类型	genres	按重要性排序的电影的所有类型(如 action comedy 等)
	电影热度	popularity	网站上该电影的点击率
	是否原创	original	电影原创则 original=1，改编电影为 0
演职人员	演员	cast	出演该电影的演员，按重要程度排序
	导演	director	该电影的导演
	发行公司	production_companies	该电影的发行公司
时间因素	发行时间	release_date	该电影上映的时间月份日期

	电影时长	runtime	电影放映时长（分钟）
电影知名度	电影评分	vote_average	该电影的平均评分
	电影评分人数	vote_count	该电影在 IMDB 网页被评分的次数
无关因素	电影主页	homepage	电影的网址主页
	电影 id	id	该电影在 IMDB 网页的编码
字符型变量	电影标题	title	电影的标题
	电影简介	overview	电影的简介
	宣传语	agline	电影的宣传语

表 1-2 The Oscar Award 主要信息

变量名称	变量说明
年份（year）	奥斯卡奖所在年份
奖项类别（category）	该奖项的类别，如 actor,actress
姓名（name）	提名者姓名
电影名（film）	提名者所对应的电影
是否得奖（winner）	若只是提名而并未获奖则为 false，若获奖则为 true

表 1-3 Movie Theater 主要信息

变量名称	变量说明
电影名（film）	电影名称
电影上映后的放映银幕数（theater）	荧幕数

3.2 数据预处理

3.2.1 无关因素筛选

在 TMDB 5000 Movie Dataset 中，电影 Id 为该电影在 IMDB 网页的编码，homepage 为电影网址，所能提供的与电影相关的信息较少，故列为无关因素，不考虑这两个变量 。

3.2.2 时间因素处理

对电影票房而言，时间也是较为重要的一项指标，由上映日期可以得出上映的年份、月份、日期三个变量。其中该数据集所对应的年份为 1916——2016，跨度较大。而 20 世纪的电影仍停留在默片与简易动画，电影技术和电影市场均未成熟，并且考虑到通胀因素，若选取电影时间跨度过大会使得后续回归分析的残差偏大，于是我们选取了 2000——2016 年的电影数据。

年份因素主要从通胀和该年度的电影市场状态两方面影响票房，但是年份因素作为数值型变量或是分类变量均难以处理，并且考虑到我们已经缩小了电影年份跨度，电影市场状态的信息与本文研究方向——探究与电影相关的哪些因素得以影响票房有所差距，而年份所代表的信息更接近于市场信息，将其看作系统误差。

通过对数据的观察，可以看出不同月份电影票房存在差异，故将上映月份看作具有代表性的时间因素，纳入变量选择中。而上映的具体日期关联性则不大，将其舍去。

3.2.3 缺失值与异常值处理

电影票房、预算、演职人员、时长均存在少量缺失值，将有缺失值的数据筛去。对于电影票房而言，发现有 20 部电影票房小于 10000 美元，数值过小，且占比不到 1%，删去对数据集不会产生较大影响，故将其删去。对于电影而言，预算为 0 与事实相违背，应当是数据集爬取时出现错误或者是统计时出现误差，若涵盖在数

据集中会使得系统误差偏大，故将预算为 0 的电影删去。

3.2.4 演职人员信息处理

出演该电影的演员以及该电影的导演与电影内容相关，且热门演员具有票房吸引力，我们采用获得奥斯卡奖的提名次数来衡量该部电影演员、导演的热门程度和专业程度。我们通过 The Oscar Award, 1927 - 2020 数据集找出该电影上映之前其演职人员有无获得奥斯卡的提名，将演员获提名的次数相加作为演员对于票房的影响程度，将导演获得提名的次数相加，作为导演对于票房的影响程度。

3.2.5 预处理后变量说明

表 2 预处理后变量概览

变量类型	变量名称		变量说明
使用量	电影票房	revenue	该电影上映后的总票房
电影投入	电影预算	budget	该电影的发行预算
电影内容	电影类型	genres	按重要性排序的电影的所有类型 (如 action comedy 等)
	是否原创	original	电影原创则 original=1，改编电影为 0
	电影评分	vote_average	该电影的平均评分
演职人员	演员	cast	该电影的演员获奥斯卡奖提名数
	导演	director	该电影的导演获奥斯卡奖提名数
	发行公司	production_companies	该电影的发行公司
时间因素	发行时间	month	该电影上映的时间月份
	电影时长	runtime	电影放映时长分钟）
电影知名度	电影热度	popularity	网站上该电影的点击率
	评分人数	vote_count	该电影在 IMDB 网页被评分的次数
发行因素	荧幕数	theatre	电影上映后的荧幕数
字符型变量	电影标题	Title	电影的标题

电影简介	Overview	电影的简介
宣传语	Tagline	电影的宣传语

至此为止，已完成数据的清洗与变换操作，后续的模型建立与求解均建立在此数据预处理的基础上。

3.3 描述性分析

3.3.1 电影投入信息

从预算的柱状图可看出，电影投入分布呈现右偏，即存在部分电影有较大预算，从而拉高了电影整体的均值，而大部分电影预算小于平均值。由于电影预算为连续型变量，通过散点图大致可看出电影预算与票房之间的关系趋势。可看出当 log（budget）大于 15 时，预算与票房呈现较强的正相关性。而当 log(budget)小于 15 时，呈现的分布则较为分散，出现低预算高票房和高预算低票房的现象，这与电影市场出现“黑马”，“爆冷”的现象相符合。

同时，我们也可以看出 2000 年之后电影的预算和票房随年份变量，并未呈现明显的数值变化，这也说明前面我们将年份的信息纳入系统误差是可以接受的。

图 1 电影预算直方图

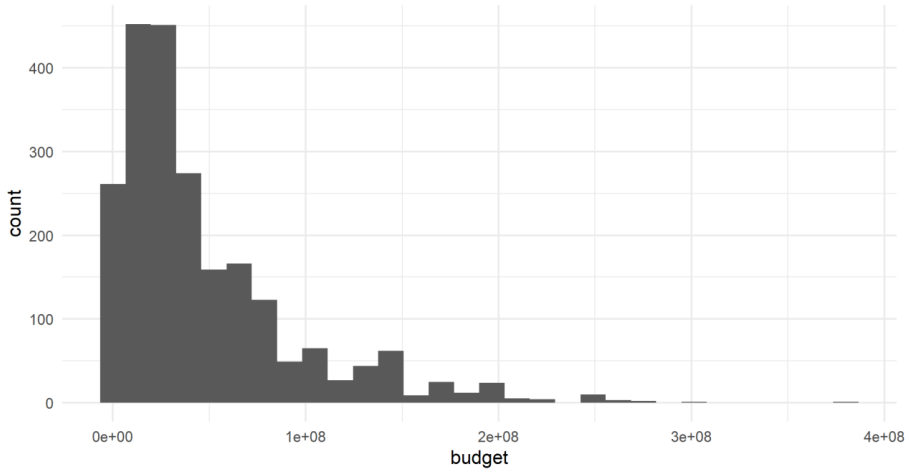
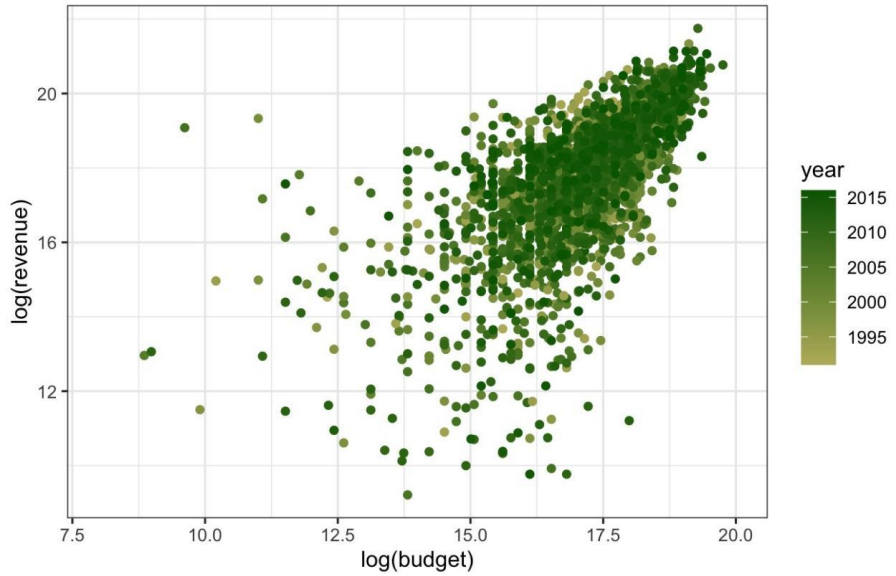


图 2 电影预算与票房散点图



3.3.2 电影内容信息

3.3.2.1 电影类型

我们首先统计了电影的不同类型出现的频次，发现 Drama Comedy Action Adventure Horror 为电影市场上 70%的电影类型，随后我们绘制了票房与电影类型分类的箱线图。从中位数可以看出，不同电影类型之间的电影票房存在明显差异，说明我们将电影类型选入影响票房的因素是可以接受的。其中 Adventure 有最大的中位数，Drama 有最小的中位数。Action 的中位数虽然小于 Adventure 但是其最大值大于 Adventure，说明该类型电影的潜力更大。Comedy，Drama，Horror 有较多的低异常值。从箱线图可以看出电影票房并不稳定，并容易出现极端低的异常值。

图 3 各类型电影分布饼图

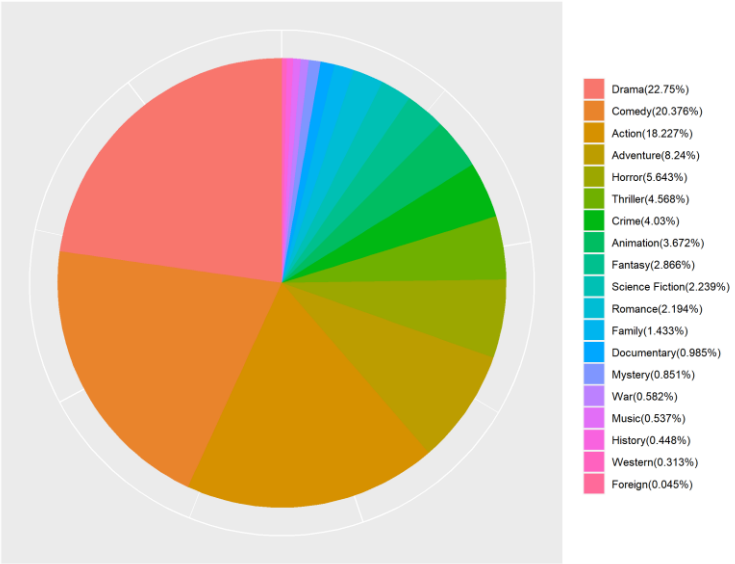
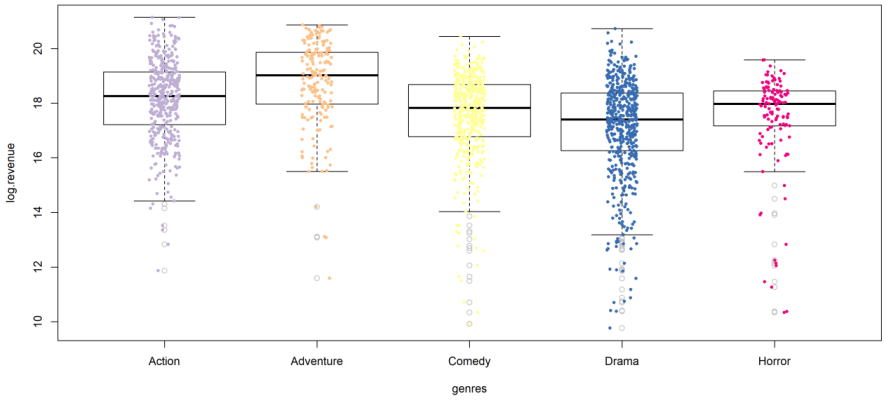


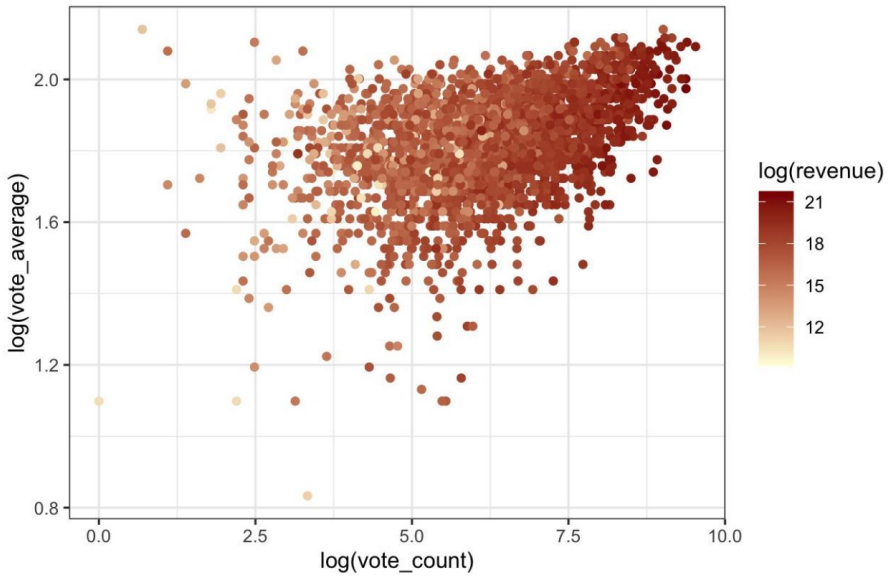
图 4 各类型电影票房分布直方图



3.3.2.2 电影热度

由下图不同颜色所代表的不同票房的散点的分布可以看出，高热度电影往往也是高票房电影，而低票房电影的热度也较低。所以电影热度可以看作是影响电影票房的因素之一；且热度高电影的评分也通常会更高。

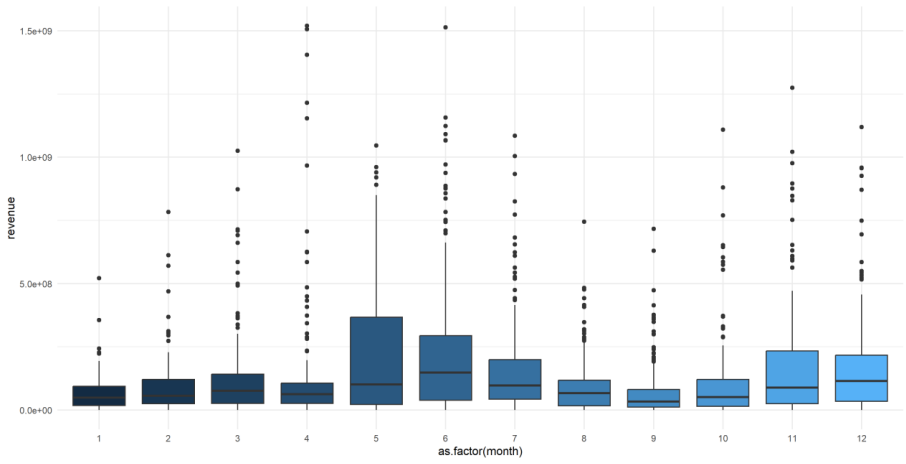
图 5 电影热度与电影评分散点图



3.3.2.3 时间因素

从下图可以看出，不同月份发行电影的票房数存在差异，将月份纳入影响电影票房的因素是可行的。其中 5 月、6 月、7 月、12 月的票房中位数较大，这与电影档期中夏季档、冬季档相对应，符合实际。并且每个月份都有一系列高票房电影的存在，从箱线图可以看出电影票房的分布并不稳定。

图 6 各月份电影票房箱线图



3.3.2.4 字符型变量

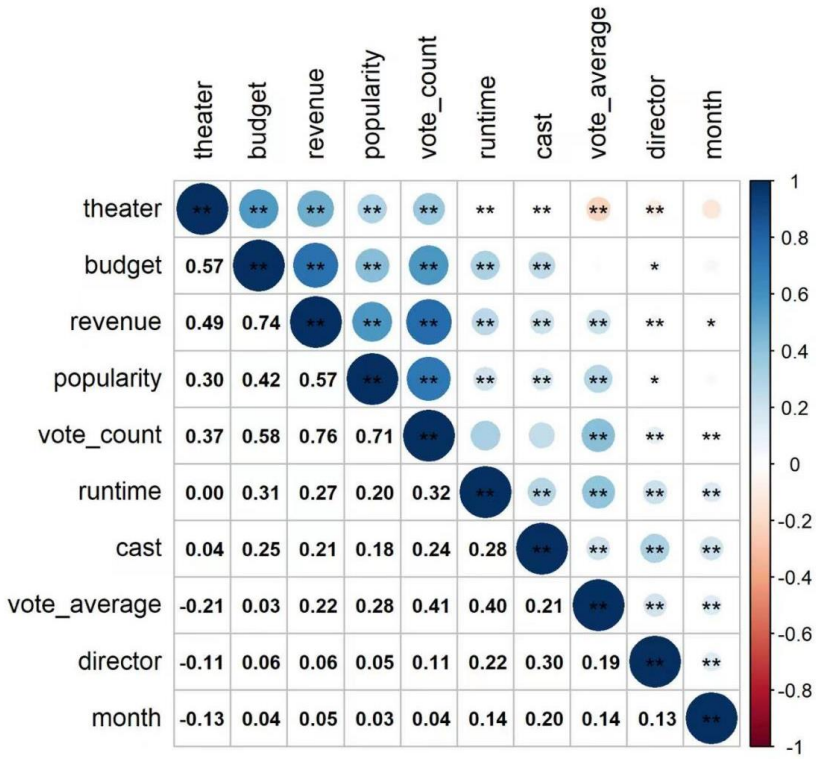
下图左图为提取电影的所有关键词所绘制的词云图。从该图中可以看出电影在选择自己的关键词时常常出现其导演、性别、年龄的信息，并对电影类型和人物关系有了初步描绘。高词频字符：based，根据观察，为”based on novel”,代表电影是否原创这一信息。右二为电影的内容概要绘制的词云图，其与电影的剧情及意义联系更为紧密。

图 7 电影关键词及内容概要词云图

3.4 变量间相关性分析

通过散点图的绘制，我们发现电影热度与评分人数存在一定的相关性，为了衡量变量之间的相关性，我们对模型中的连续型变量，计算其斯皮尔曼相关系数，并绘制热力图。由图可以发现 revenue 与所选取的变量均存在一定的相关性，说明变量选取是合理的。popularity 与 vote_count、budget 之间存在一定的相关性。Popularity 与 vote_count 之间的相关系数为 0.73，与 budget 之间的相关系数为 0.45，runtime 与 vote_average 有一定的相关性，相关系数为 0.39。Budget 与 theater 之间存在一定相关性，相关系数为 0.57。由热力图可以看出，在后续的处理中，我们可能需要考虑变量之间的相关性对模型的干扰。

图 8 连续型变量相关性热力图



四、 模型建立

4.1 多元线性回归模型全模型

由于描述性分析中提及的各因素同电影票房的关系，为进一步测算其具体数值关系，本文选用多元线性回归模型测量各因素对电影票房的影响。

首先，本文将票房收入（revenue）作为因变量，将电影预算（budget）、电影热度（popularity）等连续性变量和电影类型（genres）、是否原创（original）等名义变量生成的哑变量加入模型中。

各变量基本统计量如下表所示：

表 3 连续型变量基本统计量

	mean	std	min	max	N
revenue	139767233.6	194782282.6	10018	1506249360	1822
runtime	109.83	18.51	41	338	1822
vote_average	6.27	0.80	2.3	8.3	1822
vote_count	1169.97	1499.94	3	13752	1822
theater	2066.08	1399.51	0	4468	1822
cast	4.65	5.46	0	39	1822
director	0.56	1.70	0	18	1822
popularity	33.70	41.91	0.056	875.58	1822
budget	48721555.09	49357435.46	8000	380000000	1822

本文设定模型如下：

$$revenue = \beta_0 + \beta_1 * budget + \beta_2 * popularity + \beta_3 * vote\_average$$
$$+\beta_4 * vote_{count} + \beta_5 * cast + \beta_6 * director + \beta_7 * theater$$
$$+\beta_8 * original + \beta_9 * competition + \sum_k \beta_k Genres_k + \varepsilon$$

在前文的相关性分析中，可以发现电影预算和电影热度等变量间的相关性较强，即存在相关性较强的变量。为避免自变量间的多重共线性对模型的合理性的影响，

在构建多元线性回归模型时，先计算了自变量的方差膨胀因子。结果表明，所有变量的 vif 均小于 5，由此可以认为自变量间不存在多重共线性，可以进一步进行回归分析。

多元线性回归的全模型部分结果如下表所示（完整结果将放在附录中），模型的 R^2 为 0.76，被认为是拟合效果较好且可以接受的模型。

表 4 线性回归全模型回归结果

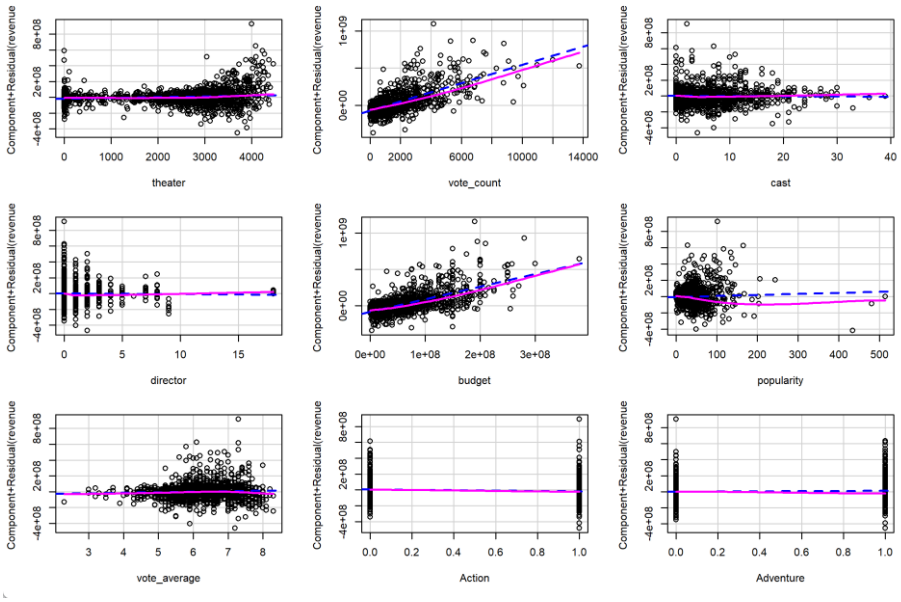
	Dependent variable:
	revenue
theater	9,736.801***
	t = 4.474
vote_count	61,648.900***
	t = 21.007
Action	-23,825,032.000***
	t = -3.722
Animation	49,490,971.000***
	t = 4.227
ScienceFiction	-49,514,637.000***
	t = -6.417
cast	-299,639.200
	t = -0.646
director	-861,722.600
	t = -0.614
budget	1.744***
	t = 23.405
Constant	-79,731,576.000***
	t = -2.970
Observations	1,822

R^2	0.764
Adjusted R^2	0.761
Residual Std. Error	94,738,769.000 (df = 1794)
F Statistic	215.349^{***} (df = 27; 1794)
Note:	*p<0.1 **p<0.05 ***p<0.01

4.1.1 模型诊断

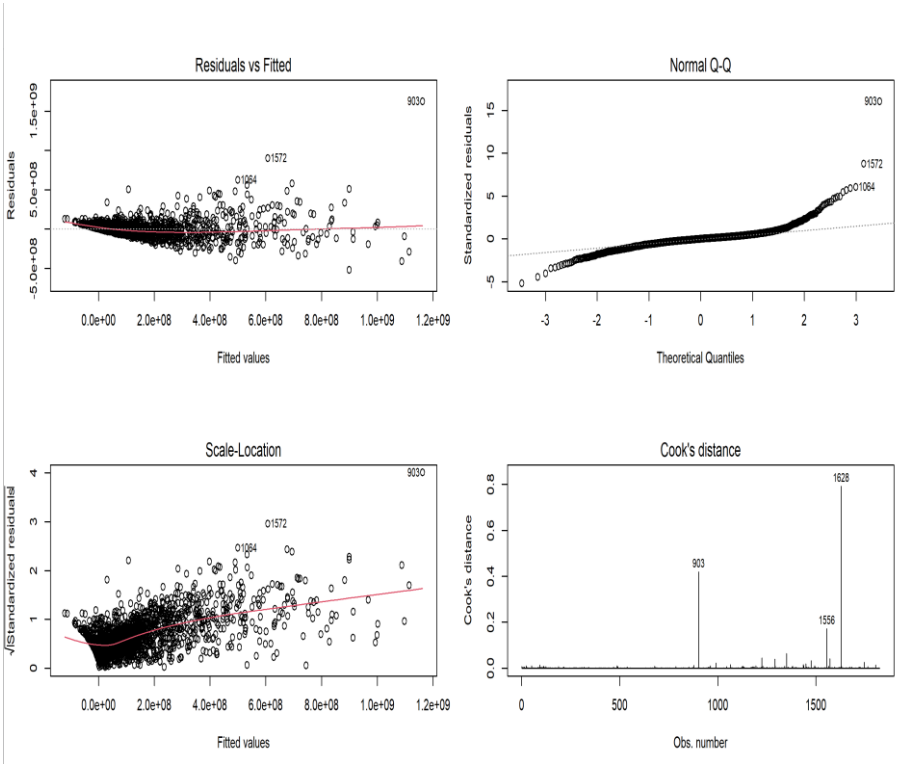
对于多元线性回归模型结果，我们可以发现其拟合效果较好，且由下图观测与主要变量的关系可以发现，主要变量同票房近似可成线性关系，模型选择较为合理。

图 9 主要变量和票房间散点图



于是我们对模型进行进一步的检验，在 R 中输出模型的 Q-Q 图、Cook 距离图等，由左上图发现残差项与预测值几乎没有关联，在 0 的上下随机分布；由 Q-Q 图发现数据大致呈正态分布，但存在部分异常点；而根据 Scale-Location 图，发现该模型的残差标准差随拟合值变大而增加，因此该模型可能存在异方差。

图 10 模型诊断图



因此，在删除了部分异常值后，对模型进行进一步检验。首先对模型进行 Durbin-Watson 检验，得到 p 值大于 0.1，即在 10%水平下，残差具有独立性。其次，又对模型进行了同方差假设检验，通过 ncvTest 命令对该线性模型进行同方差假设检验，发现 p=0.000；同样使用 Goldfeld-Quandt 检验也得到相似结论。检验显著表明模型存在异方差性，残差方差随拟合值水平变化而变化。

4. 2 模型优化-加权最小二乘法

由于使用普通最小二乘法的多元线性回归全模型的残差项不满足 Guass-Markov 假设，考虑使用加权最小二乘法对模型进行优化，即根据变异程度的大小赋予不同的权重，以实现解决了异方差的问题，使得加权后回归直线的残差平方和最小。

在模型设定时，为了选择最优的模型，以 AIC 准则使用逐步回归法进行变量的筛选和选择。AIC = -2*log(模型的极大似然函数) + 2(模型的独立参数个数)。选择 AIC 最小的参数个数及选择最优的留在模型中的变量。

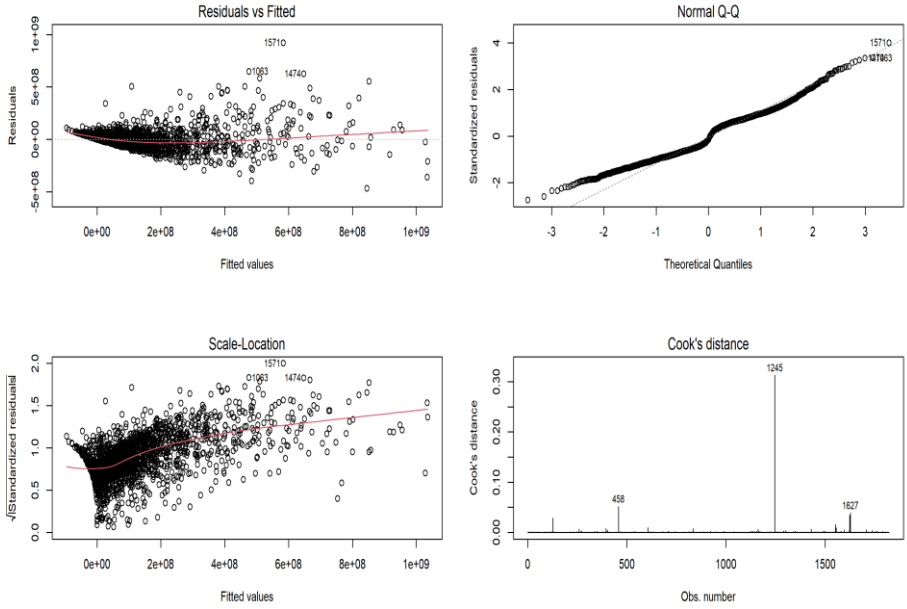
由此，逐步回归法得到的加权最小二乘法模型部分结果如下表所示。

表 5 加权最小二乘法逐步线性回归结果（部分）

	Dependent variable:
	revenue
theater	9,038.983***
	t = 15.238
vote_count	58,022.550***
	t = 42.333
budget	1.687***
	t = 71.893
popularity	222,566.900***
	t = 3.918
vote_average	6,012,287.000***
	t = 7.643
Action	-18,896,692.000***
	t = -11.560
original	11,365,528.000***
	t = 4.403
Constant	-74,013,162.000***
	t = -13.469
Observations	1,822
R2	0.993
Adjusted R2	0.993
Residual Std. Error	7,656.942 (df = 1798)
F Statistic	10,754.210*** (df = 23; 1798)
Note:	*p<0.1 **p<0.05 ***p<0.01

根据回归结果，我们发现加权最小二乘法下的多元线性回归模型的 R^2 超过 0.9，可以认为在最小 AIC 准则下得到的最优模型几乎涵盖了所有电影票房的所有信息，拟合效果非常好。且根据下图的模型诊断图来看，残差的正态性等指标也得到了有效的提升。该模型具有较高的可信度和研究意义。

图 11 加权最小二乘法模型诊断图

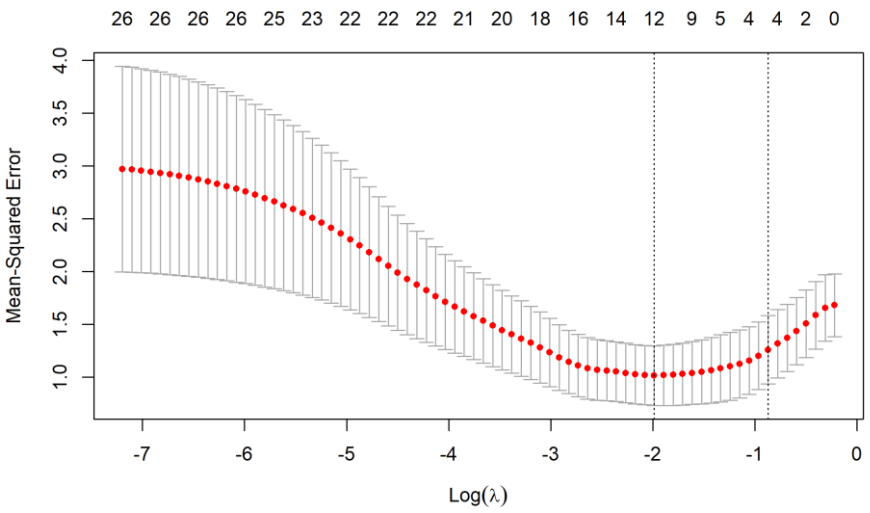


4. 3 模型优化-Lasso 回归

由于模型不满足基本最小二乘的线性回归的假定，本文转而选用了 Lasso 回归这一广义线性模型，对变量进行选择。Lasso 回归在优化函数中增加了一个偏置项，以自变量减少共线性的影响，从而减少模型方差 。Lasso 回归使用正则化方法，使用绝对值偏差作为正则化项。

本文通过五折交叉验证选取λ拟合 Lasso 回归。首先生成λ取值序列，基于给定λ序列拟合Lasso回归模型，利用五折交叉验证验证用最小MSE准则得到最优的取值，记录模型拟合出的参数。

图 12 MSE 和 Log(λ)取值关系图



根据以上结果，我们选用λ=0.173，固定此λ计算得到 Lasso 回归的模型选择及参数估计。具体结果如下表所示。

表 6 Lasso 回归结果

	Dependent variable:
	revenue
popularity	0.523
vote_average	0.278
month	0.01
Comedy	0.021
Crime	-0.085
Drama	-0.022
ScienceFiction	-0.104
Thriller	-0.111
cast	0.005
director	0.063
Observations	1,822
R2	0.993

4.4 线性回归结论

从回归结果看，可以得到如下结论：

1. 电影票房收入与电影投入的关系紧密，电影预算的增加能有效提高电影的票房收入；电影的排片增加也能正向促进电影票房收入的提高。

2. 电影票房收入与电影内容的各个方面有关。原创电影的票房收入显著高于改编电影等非原创电影；电影热度的提升和电影口碑的好转也能显著增加电影票房收入。

3. 在电影类型方面，冒险类、动画片、家庭伦理、恐怖、音乐、浪漫等类型的电影更为卖座，而动作、悬疑、犯罪、战争、惊悚等类型的电影则相对在票房上表现不佳。

五、 票房预测模型

5.1 票房分类

为进一步探讨高票房电影的成因，寻找其内部共性，并对电影进行分类预测。本文考虑将电影按照票房进行分类后再使用 Logistic 进行有序多分类回归。因此，在因变量的选择上将电影按票房收入分为三个档次，即为高票房、中票房、低票房电影。其中，高票房电影为票房收入大于 1 亿美元的电影，低票房电影为票房收入小于 1 千万美元的电影，其他电影则归为中票房电影。在本数据集中，高票房电影有 715 部，中票房电影有 860 部，低票房电影有 247 部。

将电影按票房分成三类后，对数据集中的连续型变量分组进行描述性分析，发现组间存在明显差异。对主要变量分组可视化有如下结果。

图 13 按票房分类的电影预算直方图

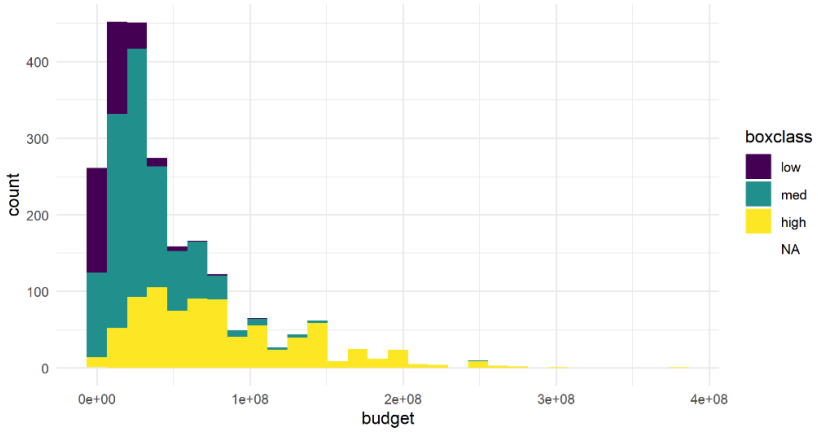
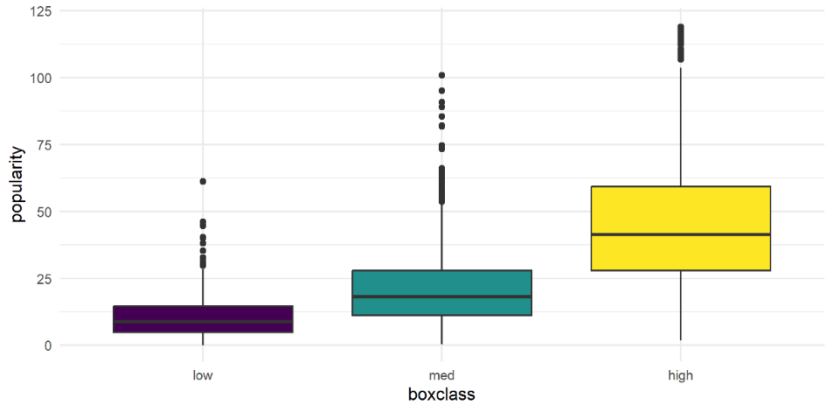


图 14 按票房分类的电影热度箱线图



5.2 有序多分类 Logistic 回归模型

5.2.1 模型建立

由于票房等级间的有序关系，考虑使用有序多分类 Logistic 回归。首先将所有变量加入模型中，对数据进行平行性检验（Brant-Wald 检验），结果显示各变量的 p 值都大于 0.05，可以使用有序逻辑回归。其次，构建只有截距项的模型，与全模型进行比较，使用 Anova 分析，p 值为 0.00，模型具有统计学意义。综上，可以在该数据基础上使用有序多分类 Logistic 回归对电影票房等级进行预测。

同样使用逐步回归法对变量进行筛选，结果所选的变量与多元线性回归结果相似，部分结果如下表所示。

表 7 有序多分类 Logistic 回归结果（部分）

	Dependent variable:
	revenue
theater	0.001***
	t = 13.218
cast	0.034***
	t = 2.743
director	0.116***
	t = 3.138
vote_average	0.549***
	t = 6.665
original	0.410***
	t = 16.557
vote_count	0.001***
	t = 7.934
budget	0.799***
	t = 24.867
popularity	0.025***
	t = 3.902
Romance	0.263*
	t = 1.736
Observations	1,822
Note:	*p<0.1 **p<0.05 ***p<0.01

5.2.2 模型评价及预测效果

对本文中建立的有序多分类 Logistic 回归模型，使用在多分类问题上更为通用的一致性指数评价其预测准确度。在本模型中，一致性指数 C-Index 为 0.874，可以认为模型对电影票房的分类具有较好的预测能力。

将原数据按照 90%和 10%的比例随机划分为训练集和测试集，在训练集上使建立逐步回归法得到的有序多分类 Logistic 最优模型，用该模型对测试集上的观测使用 predict()函数进行预测，取预测概率最高的值作为对该条观测的预测值，并计算预测正确的个数及正确率。

最终该模型取得了良好的预测效果，在测试集上的预测正确率高达 84.6%，认为该模型能较好地对电影票房实现预测。下为预测结果混淆矩阵，具体预测结果见附录。

表 8 测试集预测结果混淆矩阵

预测 实际	Low	Med	High
Low	11	3	0
Med	5	77	15
High	0	5	66

5.3 神经网络模型

5.3.1 神经网络模型介绍

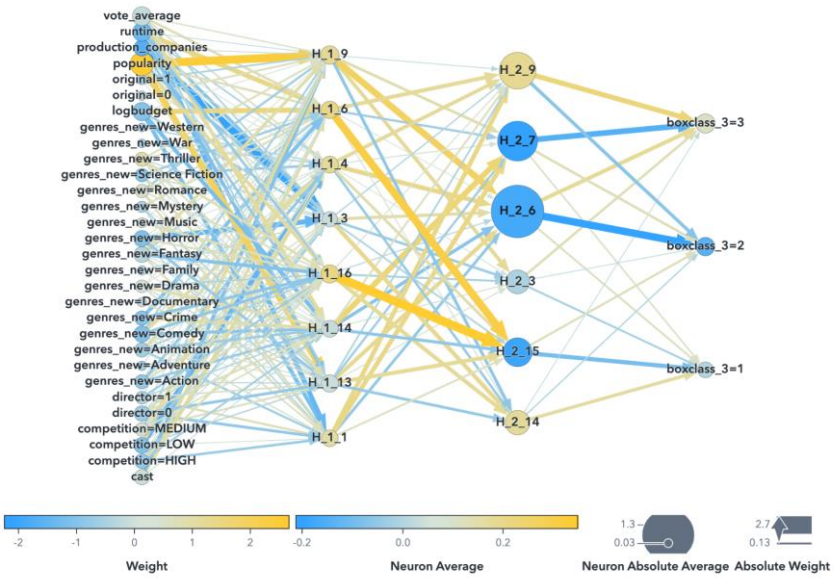
神经网络是一种由生物学启发而来的统计学技术，其对于非常复杂的非线性函数建模具有很强的构建能力。神经网络一般由若干节点层组成，其中包括一个输入层，一个或多个隐藏层和一个输出层。每一层中又包含了若干的节点，或称为神经元。层与层之间的神经元按一定权重与阈值彼此相互连接，当节点的输出通过一个激活函数计算的结果高于一个给定的阈值，该节点则将被激活，并将数据传递至下一层

的与之相连的神经元去（否则则将不被激活）。神经网络学习是有监督学习的一种，通过给定的训练目标计算损失函数或成本函数，并按照设定的学习率与优化策略（如梯度下降等）通过不断的迭代更新各部分权重，从而不断学习减小成本函数使其最终收敛至最小值，最终提高其预测的准确性。

5.3.2 神经网络模型的构建及结果

本报告中的神经网络是通过 SAS Viya 实现的，具体代码详见附录。构造的神经网络示意图如下图所示：

图 15 神经网络示意图



在本模型中，最终采用 2 层隐藏层，每层各 16 个神经节点，激活函数选择为 tanh。这样的选择是在设置 0～3 层隐藏层，每层上限 50 个神经元的条件下通过计算得到的最优化结果。神经网络中的其余各详细参数如下图所示：

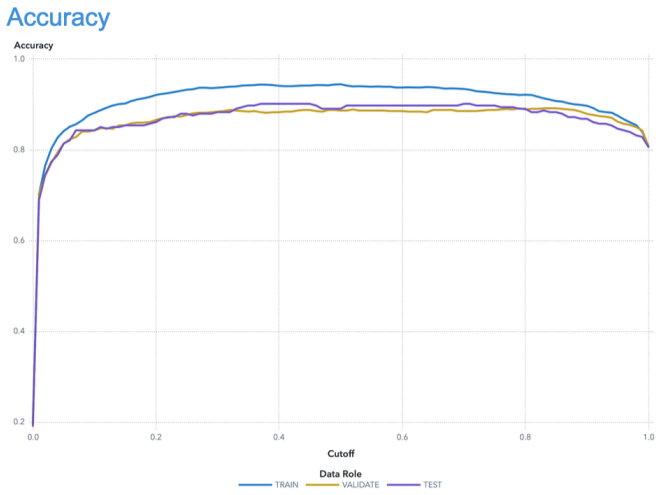
表 9 神经网络模型参数

Model	Neural Net
Number of Nodes	66
Number of Input Nodes	31
Number of Output Nodes	3
Number of Hidden Nodes	32

Number of Hidden Layers	2
Number of Weight Parameters	800
Number of Bias Parameters	35
Architecture	MLP
Seed for Initial Weight	12345
Optimization Technique	SGD
Number of Neural Nets	1

最终其预测准确度随 cutoff 的关系如图：

图 16 预测准确度与 cutoff 的关系



可见在 0.5 的判断水平下，测试集的准确度达到 0.89，训练集为 0.945，验证集为 0.887，其预测性相比较传统线性回归模型得到大幅提升。

5. 4 神经网络模型与 Logistic 回归模型的比较

对比模型结果，在验证集上神经网络模型的准确度为 0.887，而线性回归模型的预测准确度为 0.846，从准确度上看，神经网络模型优于线性回归模型，但是从可解释性来看，随机森林模型虽然可以起到很好的预测结果，但是其运行的过程更接近于“黑箱”，并不能对每个变量的效应做出很好的说明。而传统回归模型可以对模型中的每个变量的效应都给出教精确的解释。

总之，如果目标是对电影票房进行预测，那么我们优先考虑神经网络模型，而若是想要对影响电影票房的每个因素的效应都给出一定的解释，那么我们则优先考虑线性回归模型。在建模的过程中，交叉表中的频数存在较小的情况，对$\tau-\gamma$系数的适用性有偏差。

六、 结论

本文利用线性回归模型，结合加权最小二乘估计与 AIC 原则下的逐步回归法及使用 Lasso 回归对模型进行了优化。在进行票房预测的过程中，采取了 Logistic 回归与神经网络的方法，并将两种方法进行了对比。

首先通过描述性统计初步探究了可能影响电影票房的因素，接着通过热力图探究了变量之间的相关性，基于VIF值确定变量之间并无相关性后建立线性回归模型。线性回归模型结果显示，电影投入，电影排片，电影热度和口碑对电影票房有显著影响，并且电影的原创性与电影类型也会影响票房，最终 R 方为 0.76，说明拟合效果较好。对于出现的异方差的问题，采取了加权最小二乘估计进行优化。接着对票房进行高中低三分类，通过 logistic 回归模型与神经网络模型对票房类型分别进行了预测，其中 Logistic 回归模型预测准确度接近 85%，神经网络模型预测准确度接近 90%。

最终，我们发现电影投入，电影排片，电影热度均为提升电影票房需要投入的要素。从电影内容上看，电影的原创性能增加电影票房，高质量电影也往往能获得高票房。同时，电影类型的选择也较为重要，悬疑、惊悚等类型的电影的票房表现相对不佳。

我们还发现，在进行票房预测的过程中，虽然神经网络能够提升票房的预测准确度，但是它难以具体解释各个变量的效应，而 Logistic 回归模型则能够更准确地给出各个变量的效应。

七、 参考文献

[1] Bagella M, Becchetti L. The Determinants of Motion Picture Box Office Performance: Evidence from Movies Produced in Italy[J]. Journal of Cultural Economics, 1999, 23(4): 237-256.

[2] De Vany A, Walls W D. Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office?[J]. Journal of Cultural Economics, 1999, 23(4): 285-318.

[3] G. V, R. V, S. L, et al. Movie Box-Office Success Prediction Using Machine Learning[A]. 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)[C]. 2022: 1-6.

[4] Lee K J, Chang W. Bayesian belief network for box-office performance: A case study on Korean movies[J]. Expert Systems with Applications, 2009, 36(1): 280-291.

[5] Litman B R, Kohl L S. Predicting financial success of motion pictures: The '80s experience[J]. Journal of Media Economics, 1989, 2(2): 35-50.

[6] Sochay S. Predicting the Performance of Motion Pictures[J]. Journal of Media Economics, 1994, 7(4): 1-20.

[7] 池建宇. 演员与导演谁更重要——中国电影票房明星效应的实证研究[J]. 新闻界, 2016(21): 36-41.

[8] 苏永华,王哲平. 网络口碑对票房表现的影响研究——基于人工神经网络的电影票房预测[J]. 电影文学, 2021(21): 3-14.

[9] 王铮,许敏. 电影票房的影响因素分析——基于Logit模型的研究[J]. 经济问题探索, 2013(11): 96-102.

[10] 郑坚,周尚波. 基于神经网络的电影票房预测建模[J]. 计算机应用, 2014(03): 742-748.

八、 附录

8.1 多元线性回归全模型回归结果

	Dependent variable:
	revenue
theater	7,865.874***
	t = 3.317
vote_count	65,838.780***
	t = 22.808
cast	-776,717.000
	t = -1.539
director	-417,003.100
	t = -0.272
budget	1.748***
	t = 21.516
popularity	196,005.300**
	t = 2.371
vote_average	2,695,644.000
	t = 0.656
Action	-24,650,308.000***
	t = -3.530
Adventure	10,815,295.000
	t = 1.418
Animation	51,564,503.000***
	t = 4.036
Comedy	-1,491,104.000
	t = -0.229
Crime	-11,291,437.000
	t = -1.519
Drama	-5,699,476.000
	t = -0.890
Family	14,378,200.000
	t = 1.375
Fantasy	5,276,924.000
	t = 0.591
Foreign	25,704,219.000
	t = 0.247

History	-2,349,024.000
	t = -0.170
Horror	9,857,931.000
	t = 1.025
Music	17,316,628.000
	t = 1.214
Mystery	-16,129,356.000*
	t = -1.719
Romance	9,550,945.000
	t = 1.355
ScienceFiction	-48,156,408.000***
	t = -5.721
Thriller	-5,949,491.000
	t = -0.904
War	-19,317,944.000
	t = -1.273
Western	-97,443,172.000***
	t = -3.931
competition	-2,723,652.000
	t = -0.358
original	14,353,054.000
	t = 1.377
Constant	-59,359,054.000**
	t = -2.027
Observations	1,825
R^2	0.749
Adjusted R^2	0.745
Residual Std. Error	103,465,007.000 (df = 1797)
F Statistic	198.518*** (df = 27; 1797)
Note:	*p**p***p<0.01

8.2 加权最小二乘法下的逐步回归多元线性回归结果

	Dependent variable:
	revenue
theater	9,038.983***

	t = 15.238
vote_count	58,022.550***
	t = 42.333
cast	-233,684.400*
	t = -1.784
budget	1.687***
	t = 71.893
popularity	222,566.900***
	t = 3.918
vote_average	6,012,287.000***
	t = 7.643
Action	-18,896,692.000***
	t = -11.560
Adventure	6,909,858.000***
	t = 3.164
Animation	47,551,720.000***
	t = 9.101
Crime	-12,878,886.000***
	t = -8.537
Drama	-7,018,078.000***
	t = -4.722
Family	14,877,168.000***
	t = 4.608
Foreign	20,518,186.000***
	t = 10.154
Horror	9,084,948.000***
	t = 4.506
Music	17,226,549.000***
	t = 4.670
Mystery	-13,648,641.000***
	t = -5.443
Romance	6,392,919.000***
	t = 3.749
ScienceFiction	-48,212,685.000***
	t = -22.145
Thriller	-4,520,932.000***
	t = -3.330
War	-20,361,864.000***
	t = -5.990
Western	-85,344,627.000***
	t = -7.795

competition	-3,416,933.000**
	t = -2.032
original	11,365,528.000***
	t = 4.403
Constant	-74,013,162.000***
	t = -13.469
Observations	1,822
R2	0.993
Adjusted R2	0.993
Residual Std. Error	7,656.942 (df = 1798)
F Statistic	10,754.210*** (df = 23; 1798)
Note:	*p**p***p<0.01

8.3 逐步回归的 Logisitic 回归模型结果

	Dependent variable:
	boxclass
Crime	-0.338**
	t = -2.416
Drama	-0.481***
	t = -3.564
Family	0.541***
	t = 6.948
History	0.976***
	t = 50.277
Horror	0.431***
	t = 2.814
Music	0.800***
	t = 85.599
Romance	0.263*
	t = 1.736
ScienceFiction	-0.822***
	t = -6.326
Thriller	-0.377***
	t = -2.926
Western	-1.151***

	t = -287.574
theater	0.001***
	t = 13.218
cast	0.034***
	t = 2.743
director	0.116***
	t = 3.138
vote_average	0.549***
	t = 6.665
original	0.410***
	t = 16.557
vote_count	0.001***
	t = 7.934
budget	0.799***
	t = 24.867
popularity	0.025***
	t = 3.902
Observations	1,822
Note:	*p**p***p<0.01

8. 4 Logistic 预测结果

boxclass	pred
med	high
high	med
high	med
med	med
med	med
high	high
high	high
high	high
high	high
med	med
high	med
high	high
high	high
med	med
high	high
med	med
med	med
low	low
med	med
high	med
med	med

med	med
high	high
high	high
med	med
med	med
high	high
med	med
high	high
med	med
high	high
low	med
high	high
high	high
low	low
high	high
med	med
med	med
high	med
med	med
high	med
high	high
med	med
med	med
high	high
med	med
high	high
med	med
med	med
med	med
med	med
high	high
high	high
high	med
med	med
high	high
med	med
med	med
low	low
med	med
med	med
high	high
med	med
med	high
med	med
high	med
low	med
high	med
high	high
high	high
med	med
med	med
med	med

high	high
high	med
high	high
med	med
med	med
med	med
high	high
med	med
med	med
low	low
med	med
med	med
high	high
med	med
med	med
med	low
med	med
high	med
med	med
low	low
med	med
low	med
high	high
med	med
med	med
med	low
med	med
med	high
high	high
high	high
med	med
med	med
high	high
high	high
med	med
med	med
high	high
high	high
high	high
med	med
high	high
med	med
med	med
med	med
high	med
high	high
high	high
high	med
low	low
med	med
high	high
low	low

med	med
med	med
high	high
high	high
high	high
high	high
high	high
med	med
low	med
med	med
low	low
high	high
high	high
med	med
high	high
high	high
low	low
low	low
high	high
med	med
med	high
med	med
high	high
high	high
med	med
med	med
med	med
low	med
high	high
high	high
high	high
high	med
med	med
med	med
high	high
med	high
med	med
high	high
med	med
med	low
high	high
high	high
high	high
high	high
med	med
med	med
high	high
high	high
high	high
high	high
med	med
high	med

high	high
med	med
low	low
med	med
med	med