寻路前方，影动未来

——基于全球电影数据的票房回归分析

俞奕涵 余萌曦 秦浩然

2022.12

一、研究背景及研究目的

随着物质生活的不断丰富，人们对精神世界的追求也不断增强。电影作为艺术的直接表达形式，在当今时代也已经成为了人们的生活中不可分离的一部分。但随着电影产业的不断扩大，人们发现高票房和高口碑并不等价，甚至部分评分堪忧的电影依然能够取得高票房，而部分制作精良、引人深思的电影却无法获得相应的回报。电影制作方认为，这是由于大众对于具有深意的电影不感兴趣，而更倾向于爆米花式电影；而观众却认为是电影制作方在宣发投入上不到位，或是电影标题与电影内容不符，如评分颇高的电影《闪光少女》就因为电影标题让观众认为是烂片，导致票房仅为 3900 万。此外，电影也是传播文化和价值观的重要途径，但近年来我国电影不仅没有产出类似迪士尼、漫威这样的知名 IP，更没有像韩国一样诞生《寄生虫》这类夺得奥斯卡名利双收的佳作。

为了避免优秀的电影成为沧海遗珠，促进电影市场良性发展，提高我国电影产业向外输送文化和价值观的作用，本次研究将深入探索电影票房的影响因素，并根据模型提出相应建议。

二、数据介绍

（一）原始数据获取

本实验中所使用的原始数据来自 Kaggle 网站的 Movie Industry 数据集，该数据集收纳了 1981 年至 2020 年发布的超过 7000 部电影的各项有关数据，数据涵盖的内容包括：电影标题、电影评级、电影类型、电影发布时间、电影制作地区、电影评分及评分人数、电影预算、电影票房、电影主演和电影时长等。考虑到数据量的大小和模型预测的时效性，我们选取了发布时间自 2017 年至 2019 年的共 376 条电影数据，构建了原始数据集。

（二）原始数据处理

在进行后续分析之前，我们对原始数据进行了初步处理。出于研究目的考虑，我们决定将电影票房作为因变量处理，由于电影票房数值过大，我们将其取对数，将对数电影票房作为因变量 y。接下来，我们从原始数据中选取了可能对票房产生影响的多个离散型变量和连续型变量。而对于其他变量，如电影标题、电影主演和电影公司等，我们考虑到，由于电影放映时期的电影公司、电影主演等实力难以衡量，我们没有将电影公司、电影主演等变量考虑在内。而对电影标题，我们经过分析后认为，其对电影票房产生影响的可能性较大，因此考虑从电影标题数据中设计新的自变量。在确定处理思路后，详细的处理过程如下：

1. 标题数据处理

由于电影的标题名称在数据分析过程中难以作为量化指标，而我们希望能从标题名称的一些性质中提取与电影票房相关的信息，我们在去除停用词后，绘制了标题名称的词云图，如图 1 所示，以便分析其性质。

图 1 标题名称词云图

从词云图中我们容易发现，“man”、“day”、“you”和“night”等在

口语中使用较多的单词在标题中出现得最为频繁，而在口语中出现较少、在标题中出现较多的如“love”、“evil”、“dead”和“movie”等单词，相对更集中地出现在某一特定类型的电影中，如“evil”在犯罪片中出现更多、“dead”在恐怖片中出现更多等。因此，我们认为标题词汇含义包含的信息可能与电影类型包含的信息具有较强的相关性。在已经获取了电影类型数据的前提下，我们经过对标题的分析，不选择从标题含义中提取出自变量，而是从标题的长度和特殊符号两个方面设计自变量。

一方面，标题长度可能影响观众的观影欲望。考虑到两个单词之间一般以空格或标点符号作为间隔，我们通过 Excel 中的函数计算出电影标题中空格符和标点符号的个数，再将这个数加上 1，设计为连续型自变量。

另一方面，标题中特殊符号的出现与否可能作用于电影对观众的吸引力。我们通过网络信息收集和日常观影经验，筛选出“!”“:”等 20 种常见特殊符号，利用 Excel 中的函数判断标题中是否存在以上特殊符号，将其设计为 0-1 自变量。

2.连续型数据处理

考虑到变量可能对电影票房的影响，我们选取了电影评分、评分人数、电影预算和电影时长作为连续型自变量，并对其进行处理。

首先，考虑到电影评分、电影时长的量级与电影预算、评分人数的量级相差过大，以及后续处理中可能遇到的异方差问题，我们对电影预算、评分人数的数据取对数，设置为新的自变量：对数电影预算和对数电影评分人数。

其次，由于电影时长数据以分钟为基本单位，与电影评分仍相差两个量级，我们决定转换电影时长的单位，将基本单位设置为小时，从而得到以 1 小时即 60 分钟为基本单位的电影时长自变量。

最后，数据集中的连续型数据存在一部分空缺值，在数据集中将其删除。

3.离散型数据处理

经过考虑，我们选取了电影评级、电影类型、电影发布年份、电影发布月份和电影制作地区作为离散型自变量，并对其进行处理。

一方面，由于数据集中电影制作地区为北美洲的数据量占比超过 70%，数据量差异过大，我们将制作地区为除北美洲外其他大洲的数据水平合并，设置为“非北美洲”水平，从而把电影制作地区设计为 0-1 自变量。

另一方面，在进行回归分析之前，将每个离散型变量中数据量最大的水平设置为基准组。

（三）选取自变量展示

最终我们选取的自变量如表 1 所示：

表 1 自变量展示

变量类型	变量名	水平数
连续型	标题长度（len）	无
离散型	标题是否有符号（symbol）	是（1）否（0）
离散型	电影评级（rating）	R、PG、PG-13
离散型	电影类型（genre）	动作、冒险、动画、传记、喜剧、犯罪、剧情、恐怖
离散型	电影发布年份（year）	2017、2018、2019
离散型	电影发布月份（released.month）	1 月~12 月
连续型	电影评分（score）	无
连续型	电影评分人数（votes）	无
离散型	电影制作地区（region）	北美洲（1）非北美洲（0）
连续型	电影成本（budget）	无
连续型	电影时长（runtime）	无

三. 描述性分析

（一）因变量：票房

票房（gross）是模型的因变量，为连续型，首先通过直方图查看票房的分布情况，如图 2 左侧所示：

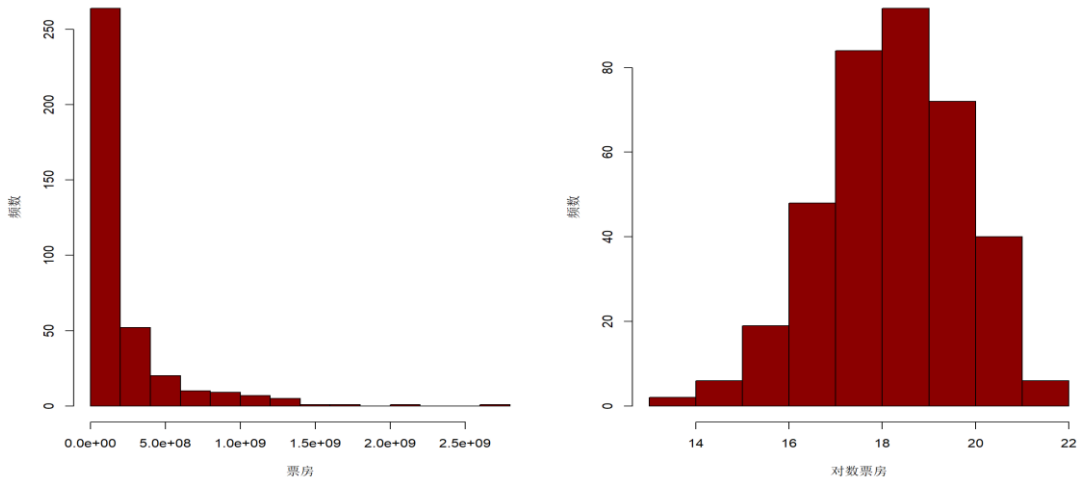

图 2 票房分布情况

可以发现，在我们的数据中，票房的值及其取值范围均较大，两端分布频数有非常显著的差异。为了便于数据的处理，我们对其作对数变换以得到对数票房。再作直方图对对数票房进行分析，得到图 2 右侧。我们发现，对数变换以后得到的票房数据分布更加均匀，且有近似于均值为 18.5 的正态分布的趋势。因此在后续数据建模中，因变量使用对数票房是符合预期的。

（二）解释性变量

1.len(标题长度)

标题长度指电影标题的单词个数，如名为 Life 的电影的标题长度为 1。将标题长度为 1 的电影标题定义为短标题，长度大于等于 4 的标题定义为长标题，其余定义为适中标题。再绘制频数分布直方图和箱线图如图 3 所示，发现大部分电影标题长度适中，而长标题电影的对数票房中位数高于其余两类。

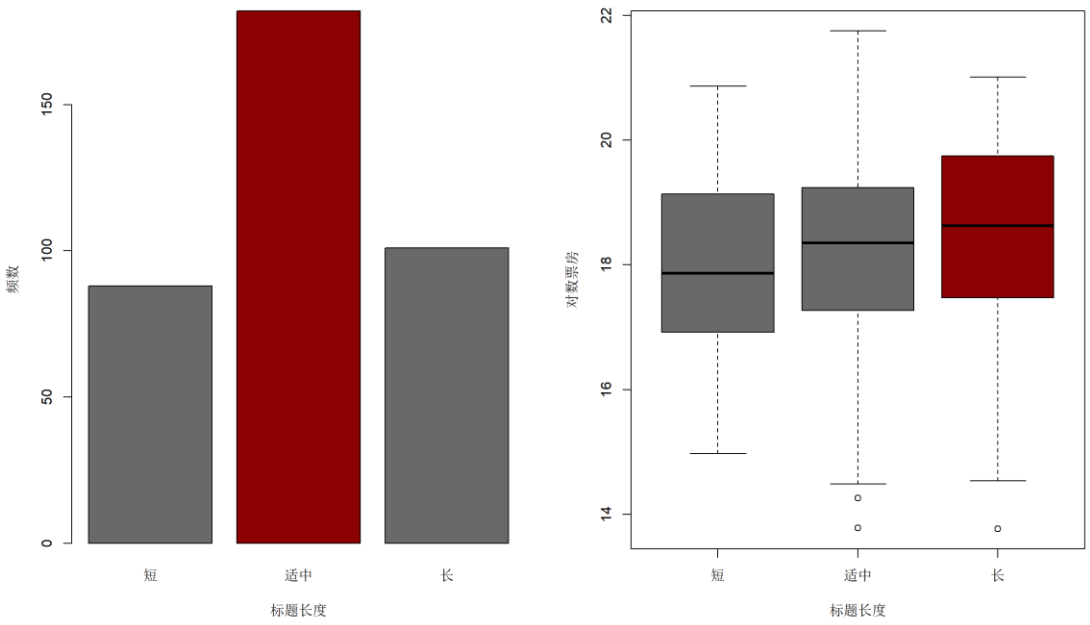

图 3 标题长度分布

2.symbol(标题是否含符号)

symbol 为 0-1 变量，若电影标题含有非字符符号，如“!”“#”“$”或“%”等，则值为 1，否则为 0。绘制频数直方图和箱线图如图 4 所示。观察发现，大部分电影标题不含符号，但标题含符号的电影的对数票房中位数更高。

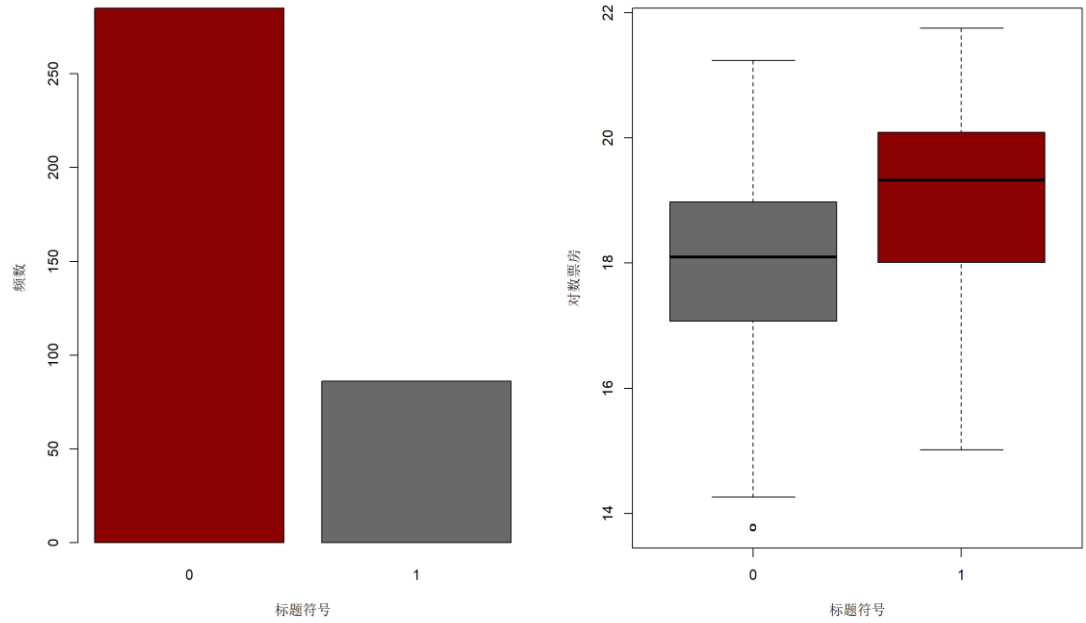

图 4　标题是否有符号分布

3.rating(电影评级)

电影评级指按照电影分级制度将电影进行限制性分类，包含 PG，PG-13 和 R 三类，画出频数直方图和箱线图，如图 5 所示。观察发现 PG 级别的电影较少，但 R 级的电影对数票房中位数较低。

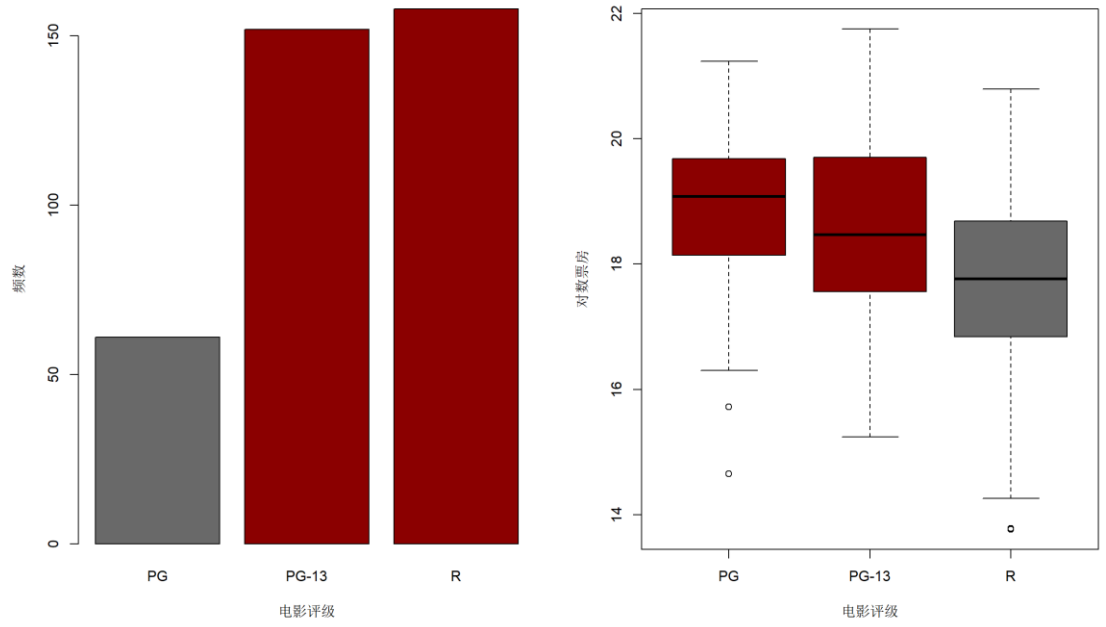

图 5 电影评级分布

4.genre(电影类型)

电影类型一共分为动画(Animation)、动作(Action)、冒险(Adventure)、恐怖(Horror)、犯罪(Crime)、戏剧(Drama)、喜剧(Comedy)和传记(Biography)八类。由于电影类型是离散型变量，因此用箱线图和直方图来得到它的分布。

首先得到每一种类型的对数年度票房均值，按其降序排列得到直方图 6，如图所示，可以看出动画(Animation)、动作(Action)、冒险(Adventure)、恐怖(Horror)的平均对数票房明显高于其他四种类型的平均对数票房。

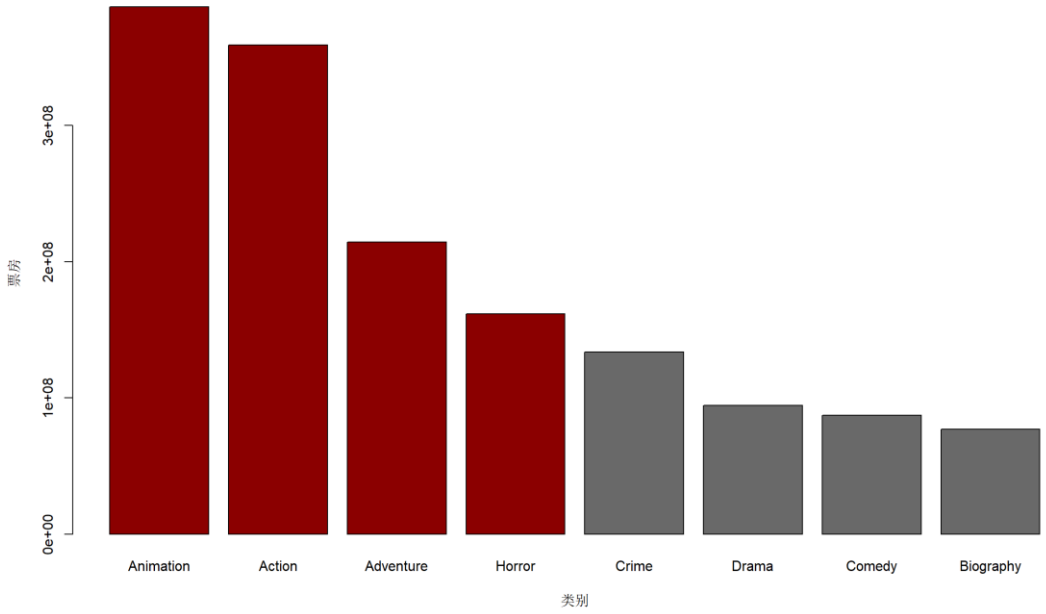

图 6 电影类型分布直方图

再用箱线图对电影类型进行数据分析，按每种类型对数年度票房的中位数排序得到如图的箱线图，箱线图中箱体的水平宽度表示该类型的电影样本数量。观察发现，样本中类型为动作(Action)的电影较多，犯罪(Crime)和恐怖(Horror)电影最少。此外我们发现，动画(Animation)、动作(Action)、冒险(Adventure)、恐怖(Horror)类型的电影的对数票房中位数高于相对其他类型，也就是说电影类型确实能够影响对数年度票房。

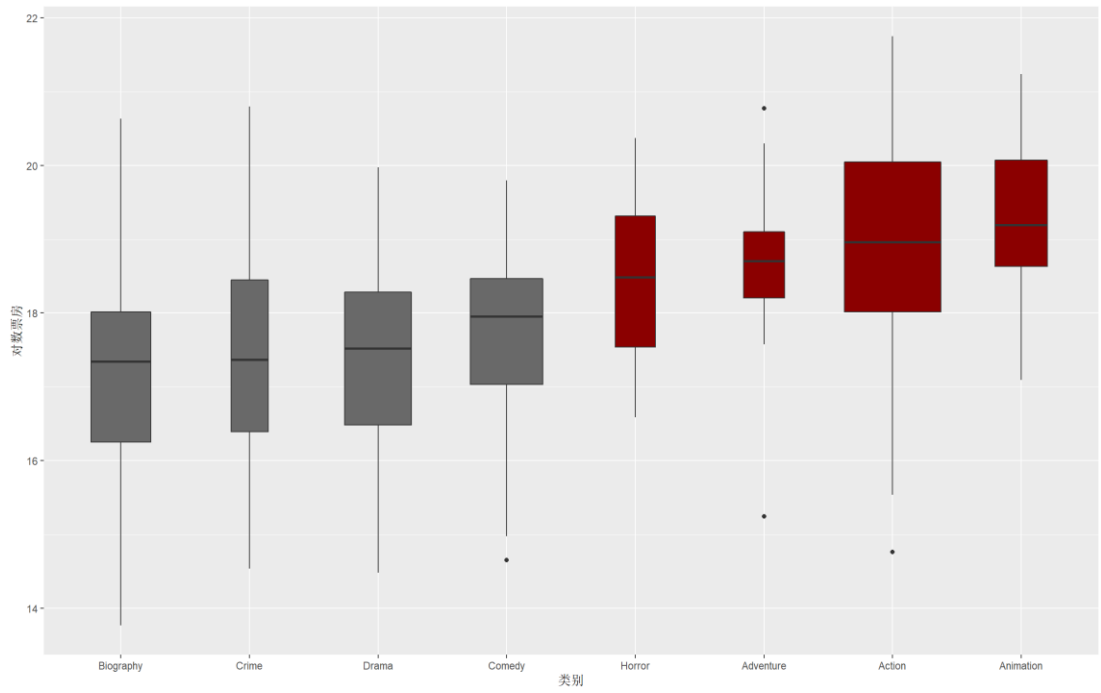

图 7 电影类型分布箱线图

5.year(电影发布年份)

本研究选取了 2017，2018，2019 三个年度上映的电影，绘制频数直方图和箱线图，如图 8 所示。观察得出，2019 年上映的电影数量略少，但对数票房中位数较高。

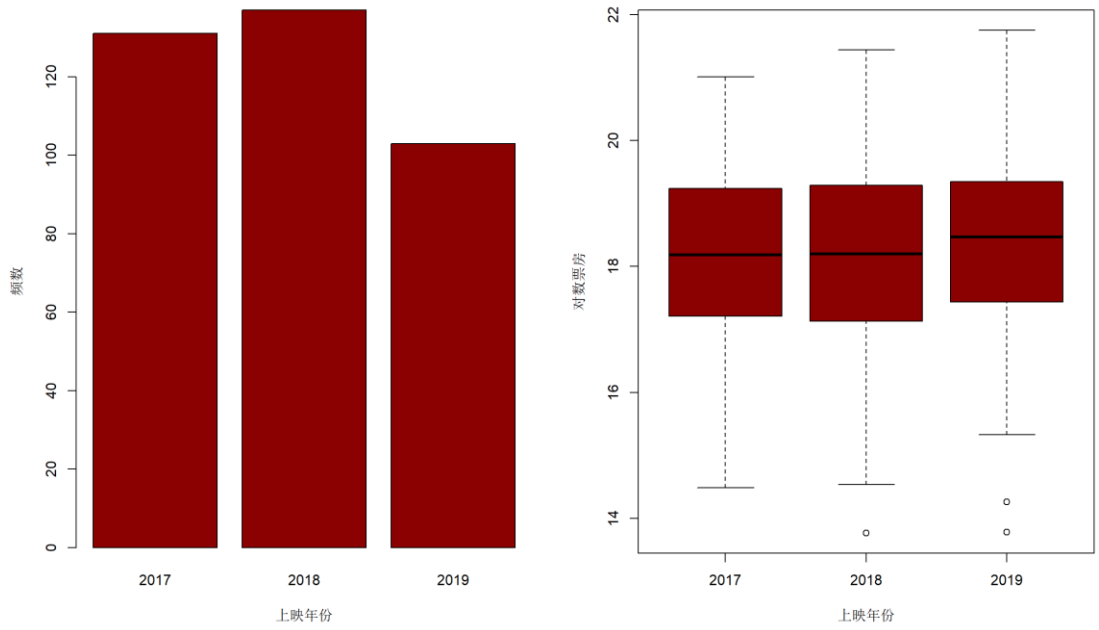

图 8 电影发布年份分布

6. released.month(电影发布月份)

电影发布月份指电影第一次在影院上映的月份，用箱线图对电影发布月份进行数据分析。箱线图中箱体的水平宽度表示该月上映的电影样本数量。观察发现，样本中各月上映的电影数量较为均匀，无明显不同。而七月上映的电影对数票房中位数明显高于其他月份上映的电影。

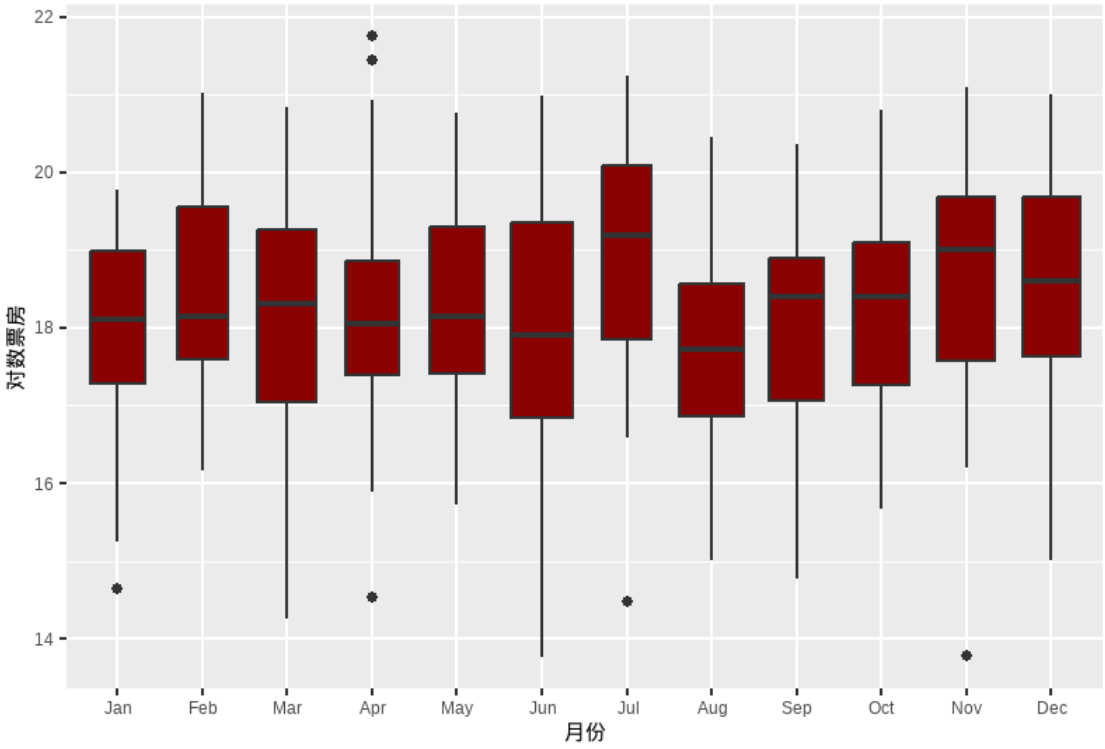

图 9 电影发布月份分布

7. region(电影制作地区)

电影制作地区指电影出版方所属地区，包含除南极洲的六大洲，但由于北美地区电影显著多于其他洲，故将非北美地区的所有电影制作地区定义为非北美。绘制箱线图如图 10 所示，且箱体的水平宽度表示该地区制作的电影样本数量。发现北美地区制作的电影数量较多，且对数票房中位数也较高。

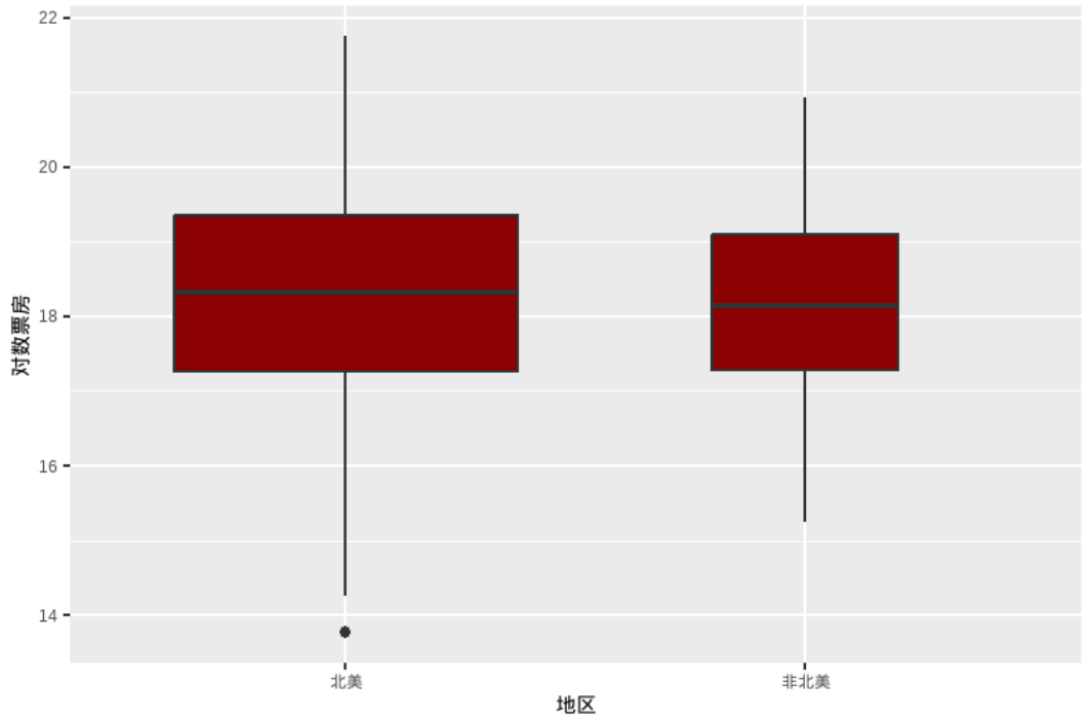

图 10 电影发布月份分布

8. score(电影评分)

电影评分为 IMDB 网站上该电影的分数，为十分制的连续型变量。绘制评分与对数票房的散点图和频数分布图如图 11 所示。发现评分与对数票房相关关系不明晰，各分数段电影呈现近似正态分布。

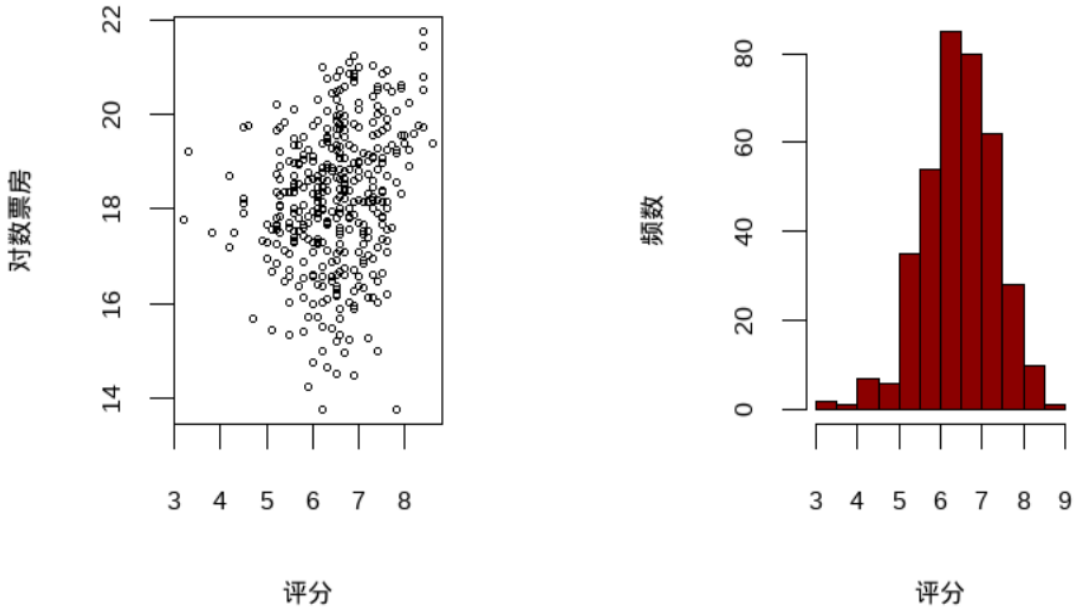

图 11 电影评分分布散点图和直方图

将评分小于等于 5 的电影定义为低分电影，评分大于 7 的电影定义为高分

电影，其余定义为中等分数电影，绘制箱线图如图 12 所示，且箱体宽度代表此分数等级的电影数量。观察发现，中等分数电影最多，平均评分高的电影的对数票房中位数较高。

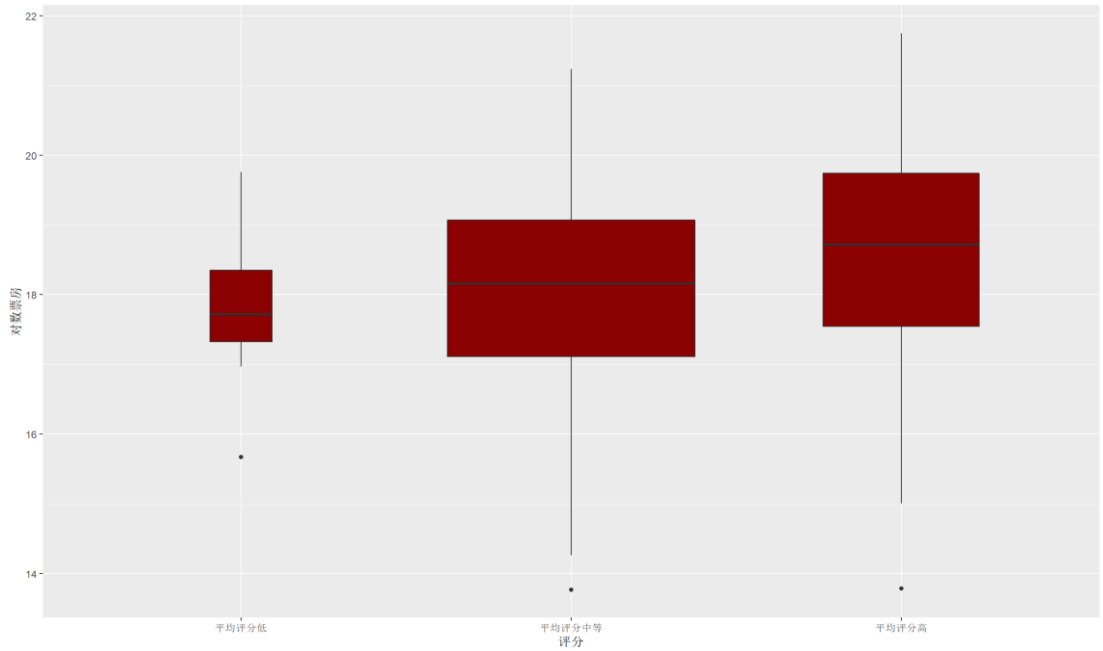

图 12 电影评分分布箱线图

9. logvotes(对数电影评分人数)

电影评分人数为 IMDB 网站上给该电影评分的观众数量，为连续型变量，且数值较大，故做对数变换处理，并绘制对数评分人数与对数票房的散点图和频数分布图如图 13 所示。发现对数评分人数与对数票房呈正相关，对数评分人数分布呈现近似正态分布。

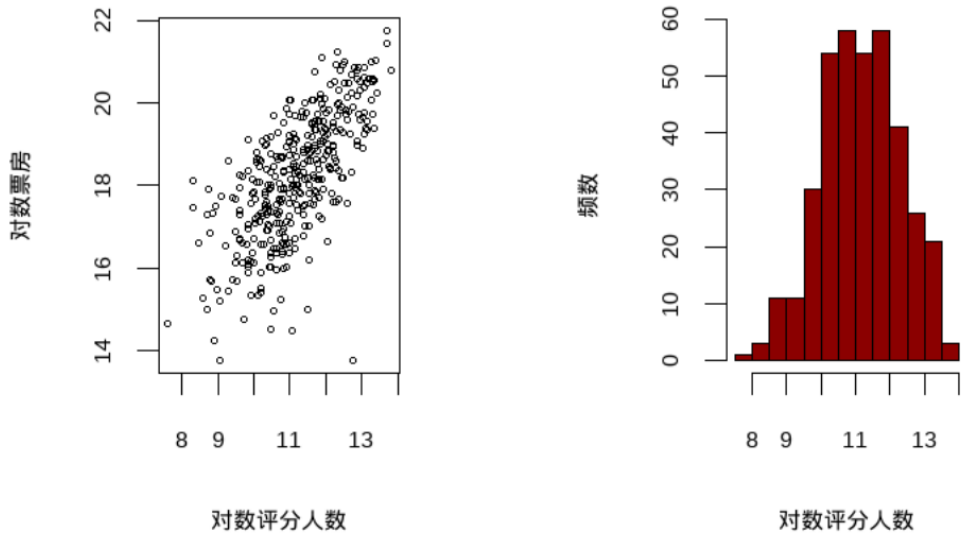

图 13 电影对数评分人数分布散点图和直方图

10. logbudget(对数电影成本)

电影成本为该电影制作方公布的包含拍摄、宣传等在内的成本，为连续型变量，且数值较大，故做对数变换。绘制对数成本与对数票房的散点图和频数分布图如图 14 所示。发现对数成本与对数票房呈正相关，对数成本分布呈现近似正态分布。

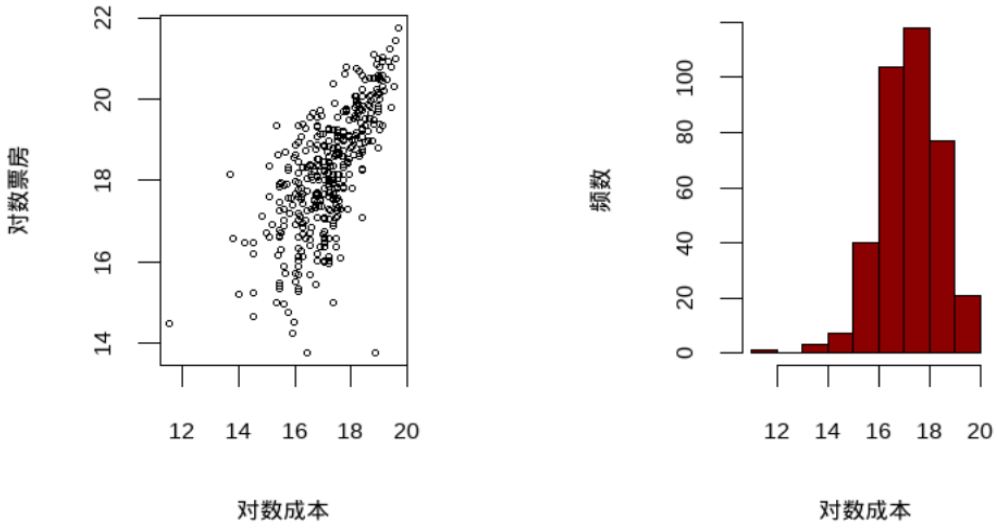

图 14 电影对数成本分布散点图和直方图

11. runtime(电影时长)

电影时长值是包含片头片尾在内的电影总时长，以分钟为单位，绘制电影时长与对数票房的散点图和频数分布图如图 15 所示。发现对电影时长与对数票房关系不明晰，大多数电影时长在 120 分钟左右。

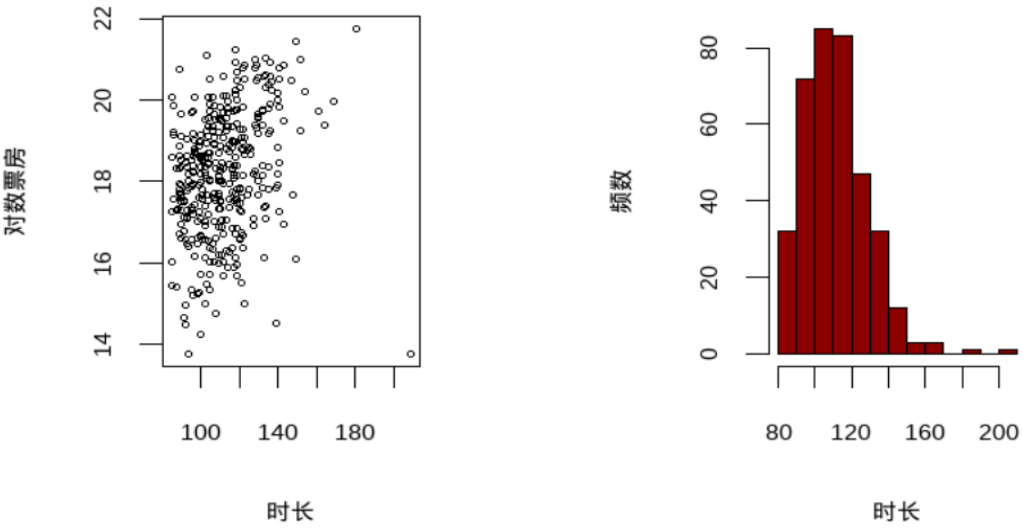

图 15 电影时长分布散点图和直方图

将电影时长小于等于 90 的定义为短电影，大于 120 的定义为长电影，其余定义为中等时长电影，绘制箱线图如图 16 所示。箱体长度为此类电影数量。发现中等时长电影最多，长电影的对数票房中位数较高。

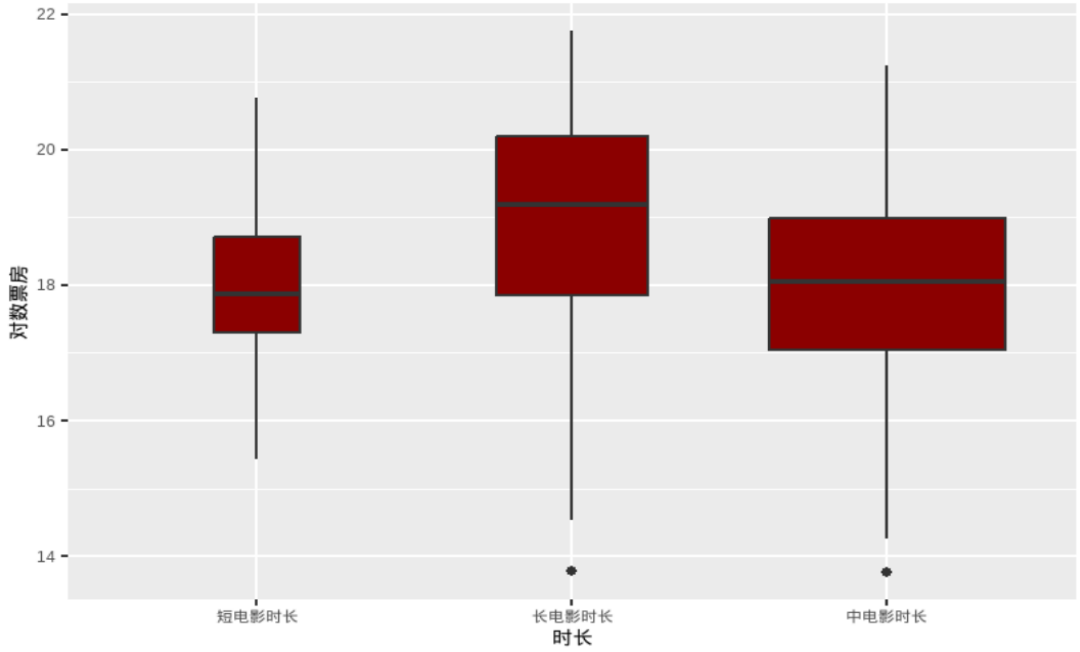

图 16 电影时长分布箱线图

四、数据建模

（一）全模型分析

1.全模型建模

由于因变量为连续型数据，自变量既有连续型又有离散型数据，我们使用多因素协方差分析建模。由描述性统计分析可知，标题长度（len），标题是否有符号（symbol），电影评级（rating），电影类型（genre），电影发布月份（released.month），电影评分（score），对数电影评分人数（logvotes），对数电影预算（logbudget）和电影时长（runtime）与对数电影票房（loggross）有较大关系，电影发布年份（year）和电影地区（region）与对数电影票房（loggross）关系较小。

考虑到模型完整性，我们首先运用全模型分析。方差分析结果如表 2：

表 2 全模型方差分析

变量	p 值
标题长度	0.7029627

标题是否有符号	0.4639723
电影评级	7.096e-13 ***
电影类型	0.0009224 ***
电影发布年份	0.1264223
电影发布月份	0.9532034
电影评分	0.0042991 **
电影评分人数	< 2.2e-16 ***
电影制作地区	0.2968838
电影预算	1.023e-09 ***
电影时长	0.290204

Residual standard error: 0.748 on 341 degrees of freedom

Multiple R-squared: 0.8126, Adjusted R-squared: 0.8033

F-statistic: 39.95 on 29 and 341 DF, p-value: < 2.2e-16

2.模型检验

为了确保模型分析结果的合理性，我们首先对模型的正态性以及同方差假设进行检验，得到的结果如图 17 所示：

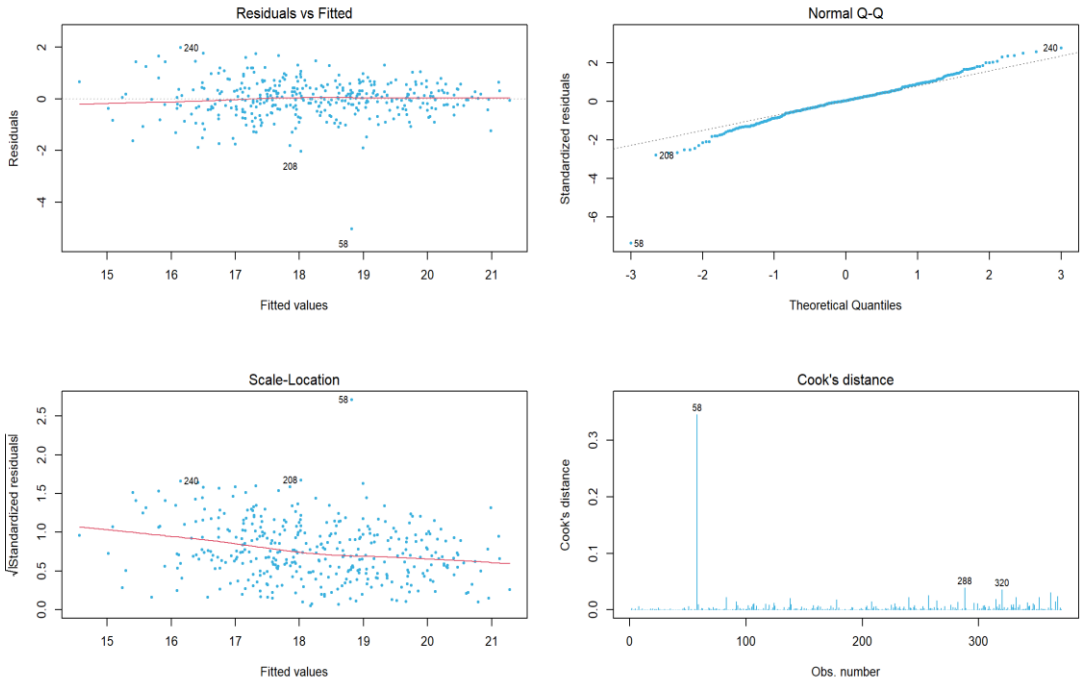

图 17 模型诊断图

利用 ncvTest 函数检验模型的异方差性，得到 p 值为 3.3725e-05，因此拒绝原假设，并且认为模型具有异方差。

表 3 ncvTest 检验指标

Chisquare	Df	p
17.19538	1	3.37E-05

然后，我们利用 dwtest 函数对误差进行自相关假设，得到 DW 值为 1.9946，p 值为 0.4519，接受原假设，认为误差不具有自相关性。

表 4 dwtest 检验指标

DW	p-value
1.9946	0.4519

此外，我们利用 vif 函数和 kappa 函数进行了多重共线性检验，所得结果如表 5、表 6 所示：

表 5 vif 检验指标

变量	GVIF	Df	GVIF^(1/(2*Df))
标题长度	1.398125	1	1.182423
标题是否有符号	1.393802	1	1.180594
电影评级	3.18258	2	1.335657
电影类型	8.319234	7	1.163376
电影发布年份	1.196656	2	1.045905
电影发布月份	2.0871	11	1.03401
电影评分	2.417177	1	1.554727
电影评分人数	3.18257	1	1.783976
电影制作地区	1.091362	1	1.044683
电影预算	3.489362	1	1.867983
电影时长	2.423078	1	1.556624

表 6 kappa 检验指标

kappa
267.6061

因此，虽然从 vif 检验来看，模型多重共线性相对符合要求，但从 kappa 检验来说，模型依然存在较强的多重共线性问题，后续处理中仍需对模型加以改进。

3.BOX-COX 变换

为了改善数据的非正态性与异方差性，我们对因变量进行 BOX-COX 变换，寻找 BOX-COX 变换中最优的 λ 值。λ 与对数似然函数的关系如图 18 所示：

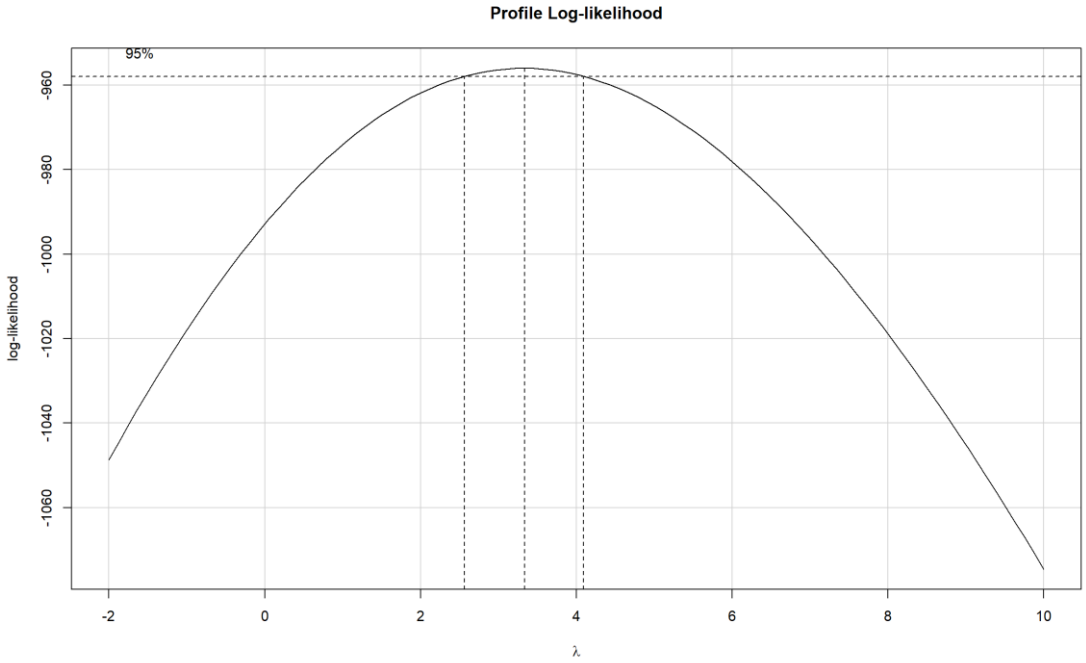

图 18 λ与似然函数关系图

从图中可以发现 λ 存在唯一的最优值。我们利用 R 找到了使似然函数达到最 大的 λ=3.333333，并以此对因变量进行相应变换。我们将 BOX-COX 变换后的因变量代入模型进行分析，得到的残差图如图 19 所示:

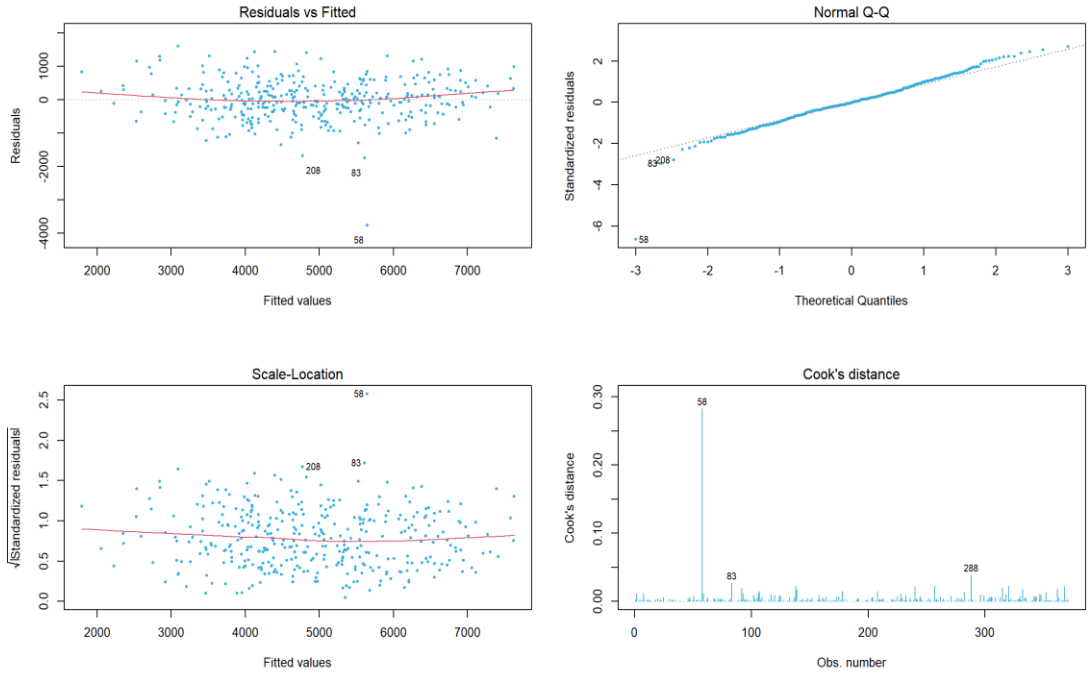

图 19 模型诊断图

利用 ncvTest 再次检验模型异方差，得到 p 值为 0.5752，因此接受原假设，并且认为模型具有同方差。

表 7 ncvTest 检验指标

Chisquare	Df	p
0.3140615	1	0.5752

（二）模型选择和处理

其次，我们考虑解决模型的多重共线性问题，并将采用逐步回归的方法，分别依据 AIC、BIC 准则选出最优模型。首先，对 BOX-COX 变换后的模型计算其 AIC、BIC 值，所得结果如表 8 所示：

表 8 原模型 AIC、BIC 值计算

AIC	BIC
5851.199	5972.601

接着，基于 AIC、BIC 准则分别选出最优模型，计算新模型的 AIC、BIC 值，并对其进行方差分析，结果如表 9、表 10 所示：

表 9 选模型 AIC、BIC 值计算

指标	AIC 最优模型	BIC 最优模型
AIC	5830.407	5852.129
BIC	5900.899	5883.459

表 10 选模型变量和显著性

AIC 最优模型变量	显著性	BIC 最优模型变量	显著性
标题是否有符号		标题是否有符号	*
电影评级	***	电影评级	***
电影评分	**	电影评分	***
电影评分人数	***	电影评分人数	***
电影预算	***	电影预算	***
电影类型	***		
电影发布年份	.		
电影制作地区			

通过对比可以认为，AIC、BIC 准则所选出的最优模型都优于原模型，并且两个模型的区别在于是否选择电影发布年份、电影制作地区和电影类型作为自变量。我们发现，两个模型的调整后的 kappa 值如表 11 所示：

表 11 选模型 kappa 值

指标	AIC 最优模型	BIC 最优模型
kappa	187.5031	163.2567

可以看出，AIC 所选出的最优模型的多重共线性强于 BIC 所选出的最优模型。因此，我们使用 BIC 模型所选择的变量进行后续的回归分析。由 kappa 值可以看出，由 BIC 准则筛选后的模型一定程度上改善了多重共线性问题，但并未完全解决。考虑到所选数据可能存在由于量纲不同、自身变异或者数值相差较大所引起的误差，我们对 BIC 模型中的数据进行中心化和标准化处理，并在处理后重新建立线性模型。模型的残差图和各项检验指标统计值如下：

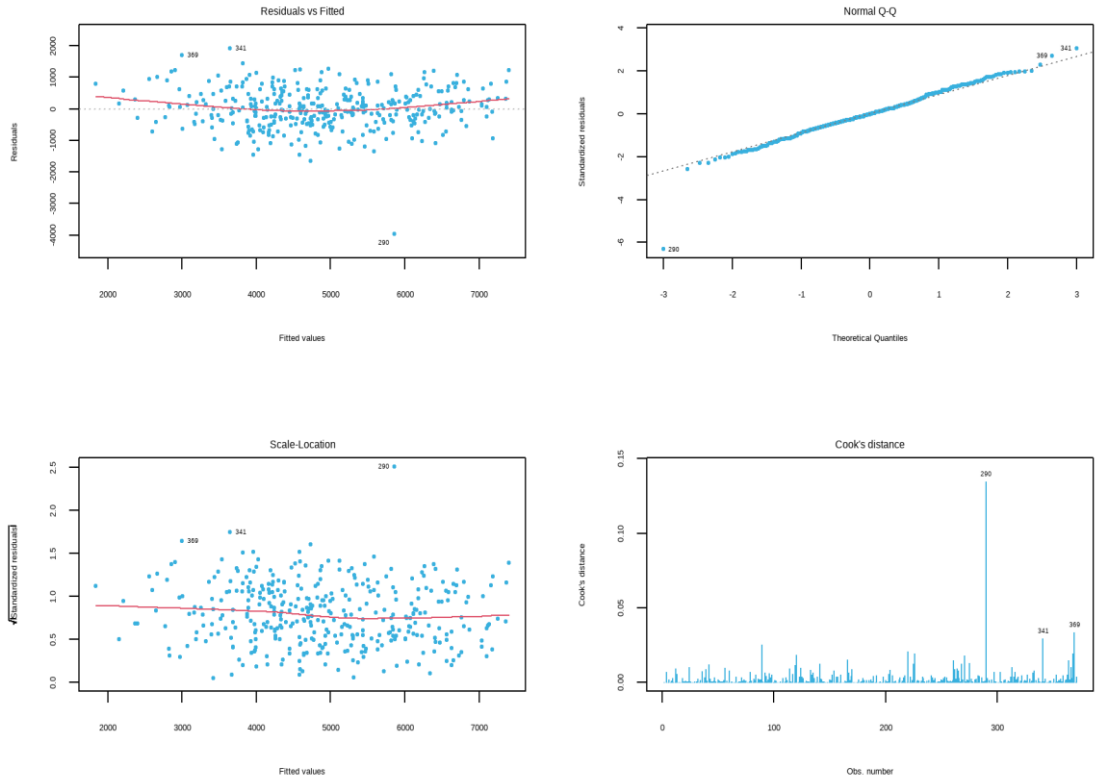

图 20 模型诊断图

表 12 检验指标

检验名称	检验指标	指标取值	结论
ncvTest	p-value	0.50682	满足同方差假设
dwTest	p-value	0.1402	自相关性较低
vif	max(GVIF^(1/(2*Df)))	1.587254	多重共线性较低
kappa	kappa	7.197476	多重共线性较低
R^2	Adjusted R-squared	0.8019	拟合度较高

通过右下角的 Cook 距离图，我们找到 Cook 距离最大的强影响点“290”，并将其删除。从而，经过一系列处理后，我们得到的回归模型已经通过了各项检验，满足了同方差和正态分布假设，不存在较强的复共线性和自相关性，且拟合程度相比原模型得到了提高，保证了后续分析和结果的准确性。最终模型展示如表 13：

表 13 模型展示

变量名	系数估计值	标准差	p 值
(Intercept)	-0.35084	0.04242	2.53e-15 ***
symbol1	0.15956	0.06317	0.012 *
ratingPG	0.98854	0.08428	<2e-16 ***
ratingPG-13	0.36933	0.05766	4.64e-10 ***
score	-0.16959	0.03252	3.09e-07 ***
logvotes	0.70718	0.04027	<2e-16 ***
logbudget	0.26041	0.03453	3.73e-13 ***

Residual standard error: 0.488 on 364 degrees of freedom

Multiple R-squared: 0.8107, Adjusted R-squared: 0.8079

F-statistic: 198.3 on 6 and 364 DF, p-value: < 2.2e-16

（三）主成分分析

我们采用主成分分析法，进一步降低模型的多重共线性。首先将 AIC 模型中的变量选出，将分类变量转化为哑变量，得到变量如下：

① 哑变量：Symbol_0（标题无符号）、Symbol_0（标题有符号）、PG-13、PG、R；

② 连续型数值变量：score（评分）、logvotes（对数评分人数）、logbudget（对数成本）。

将这些数据标准化后，先进行主成分个数的判断。图 21 展示了基于观测特征值的碎石检验（×）与由 100 个随机数据推导出的特征值均值（-）。由于同时高于均值和大于 1 的特征值检验只有 4 个，因此选择 4 个主成分以保留数据集的大部分信息。

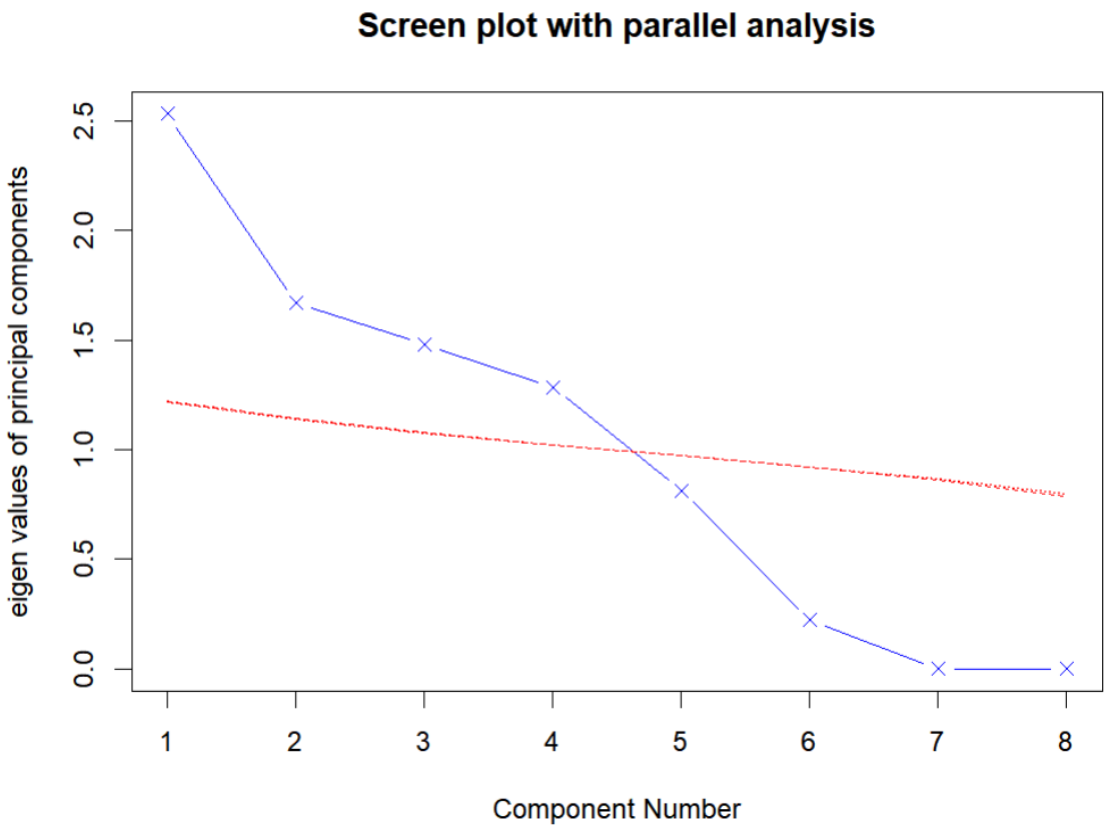

图 21 碎石图

接下来使用方差极大旋转的主成分分析法（尽可能对成分去噪），提取出 4 个成分如表 14、表 15 所示：

表 14 主成分分析

变量名	RC2	RC1	RC3	RC4	h2	u2	com
Symbol_0	-0.99	-0.05	-0.1	-0.02	0.99	0.01	1
Symbol_1	0.99	0.05	0.1	0.02	0.99	0.01	1
PG	0.03	0.01	-0.02	0.99	0.99	0.014	1
PG-13	0.04	0.91	0.01	-0.39	0.98	0.018	1.4
R	-0.07	-0.91	0.01	-0.36	0.97	0.033	1.3
score	-0.04	-0.12	0.78	0.06	0.63	0.374	1.1
logvotes	0.14	0.04	0.9	-0.23	0.88	0.124	1.2
logbudget	0.28	0.34	0.56	0.21	0.55	0.451	2.5

表 15 主成分分析指标

指标	RC2	RC1	RC3	RC4
SS loadings	2.06	1.8	1.74	1.36
Proportion Var	0.26	0.23	0.22	0.17
Cumulative Var	0.26	0.48	0.7	0.87
Proportion Explained	0.3	0.26	0.25	0.2
Cumulative Proportion	0.3	0.55	0.8	1

可以发现，在主成分分析中 h2（主成分对变量的方差解释度）基本都较好。4 个成分（由于方差旋转没有保留单个主成分方差最大化性质，故称成分）解释了所有变量 87%的方差。各成分解释情况如下：

表 16 成分解释情况

RC1	Symbol_0, Symbol_1
RC2	PG-13, R
RC3	Score, logvotes, logbudget
RC4	PG

利用主成分分析得到的成分对原模型进行回归。结果如表 17：

表 17 模型总结

变量名	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.36E-02	1.93E-02	3.327	0.0014**
RC1	2.37E-02	1.78E-02	1.33	0.184
RC2	1.36E-01	1.87E-02	7.279	2.08e-12 ***
RC3	3.60E-01	2.03E-02	17.773	< 2e-16 ***
RC4	1.41E-01	2.40E-02	5.871	9.71e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.623 on 366 degrees of freedom

Multiple R-squared: 0.7161, Adjusted R-squared: 0.7119

F-statistic: 146.9 on 4 and 366 DF, p-value: < 2.2e-16

结果如表所示，可以看到所有成分对对数票房有一定正向影响，其中 RC3 的影响最显著，R 方在 0.7 以上，拟合效果较好。

（四）岭回归

对于 AIC 模型中的多重共线性问题，我们还考虑了使用岭回归方法进行研究。关于岭回归中的 λ 值，我们使用两种方法来确定，即可视化方法和交叉验证法。

我们首先用可视化方法寻找 λ 的大致范围，分别计算不同参数 λ 所对应的每一个回归系数，我们得到如图 22 所示的岭迹图，可以看到回归系数随着 λ 值的增加而趋近于稳定的点。当 Log lambda 在-2 附近时，绝大多数变量的回归系数趋于稳定，故认为 λ 值可以选择在 0.1 附近。

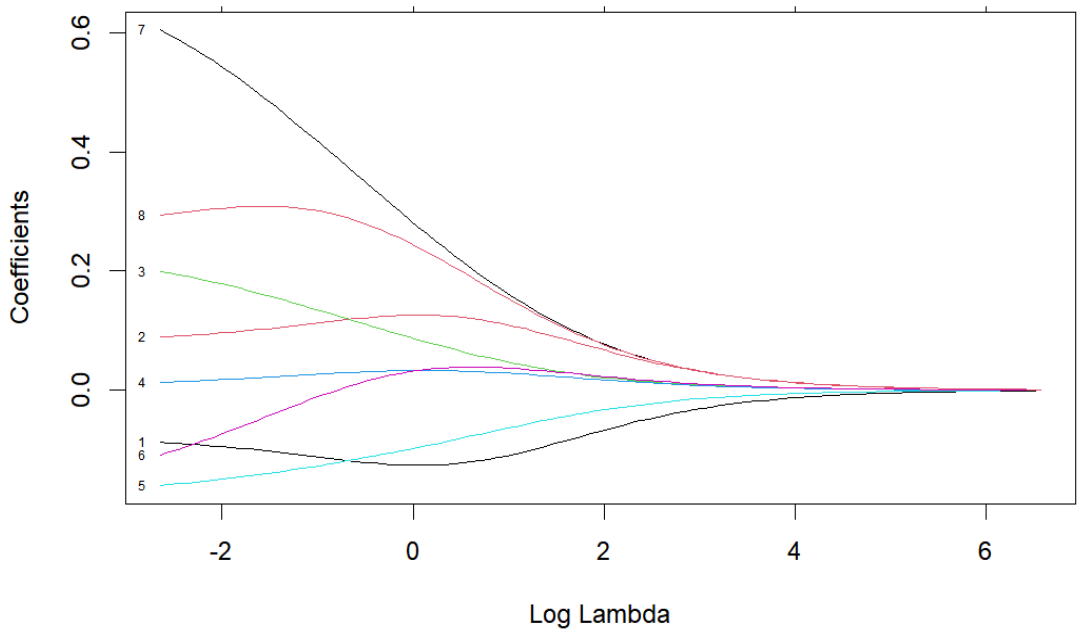

图 22 岭迹图

可视化方法可以帮助我们确定 λ 值的大概范围，为了能够更精确地找到最佳的 λ 值，我们使用 k 重交叉验证的方法（k-fold Cross-validation）。k 重交叉验证法是先将数据集 D 随机划分为 k 个大小相同的互斥子集，每次随机的选择 k-1 份作为训练集，剩下的 1 份做测试集。当这一轮完成后，重新随机选择 k 份来训练数据。若干轮（小于 k ）之后，选择损失函数评估最优的模型和参数（用于数据量不是特别充分时）。利用 R 中的 glmnet 包，我们得到如图 23 所示，Log lambda 的范围在（-3，-1）。对于每一个 λ 值计算平均均方误差（MSE），从中挑选出最小的平均均方误差，并将对应的 λ 值挑选出来，作为最佳的惩罚项系数 λ 的值，进一步计算可以确定最佳的 λ 值为：0.07095129。

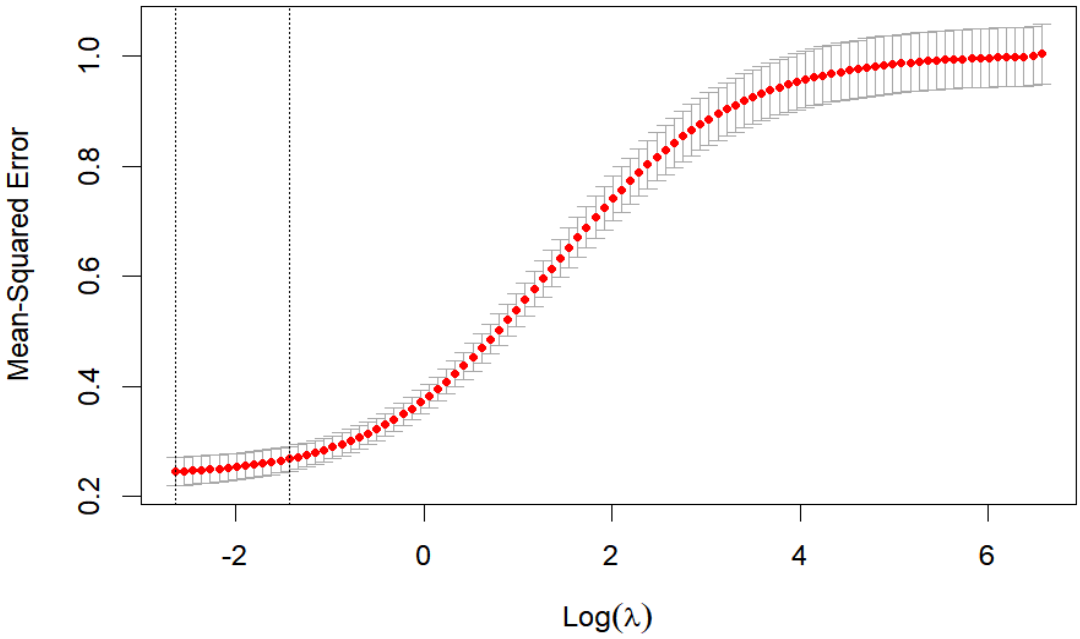

图 23 Log(λ)取值图

根据最佳的 λ 值，我们构建岭回归模型，模型的相关信息如表 18 所示。

表 18 模型展示

变量名	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)
Symbol_0	-3.74E-02	-7.20E-01	2.48E-01	2.901	0.00373 **
Symbol_1	3.74E-02	7.20E-01	2.48E-01	2.901	0.00373 **
PG	1.99E-01	3.82E+00	3.88E-01	9.841	< 2e-16 ***
PG-13	1.20E-02	2.32E-01	2.95E-01	0.784	0.43322
R	-1.61E-01	-3.09E+00	3.05E-01	10.138	< 2e-16 ***
score	-1.09E-01	-2.09E+00	5.32E-01	3.932	8.42e-05 ***
logvotes	6.05E-01	1.16E+01	6.13E-01	18.961	< 2e-16 ***
logbudget	2.93E-01	5.64E+00	5.56E-01	10.148	< 2e-16 ***

可以看到，电影级别为 PG、对数评分人数、对数成本对对数票房有较为显著的正向影响；电影级别为 R、评分对对数票房有负向影响。

（五）lasso 回归

岭回归不管怎么缩减，都会始终保留建模时的所有变量，无法降低模型的复杂度，为了克服这个缺点，我们考虑运用 LASSO 回归进行处理。与岭回归模型类似，LASSO 回归同样属于缩减性估计，而且在回归系数的缩减过程中，可以将一些不重要的回归系数直接缩减为 0，即达到变量筛选的功能。LASSO 回归将在岭回归模型中的惩罚项由平方和改成了绝对值，即惩罚项为 L1 正则式。

关于 LASSO 回归中 λ 值，我们同样使用两种方法来确定，即可视化方法和交叉验证法。

首先用可视化方法寻找 λ 的大致范围，我们得到图 24。可以看到，与岭回归模型绘制的折线图类似，图中出现了喇叭形折线，说明该变量存在多重共线性。回归系数随着 λ 值的增加而趋近于稳定的一点，当 Log lambda 在-3 附近时，绝大多数变量的回归系数趋于稳定。

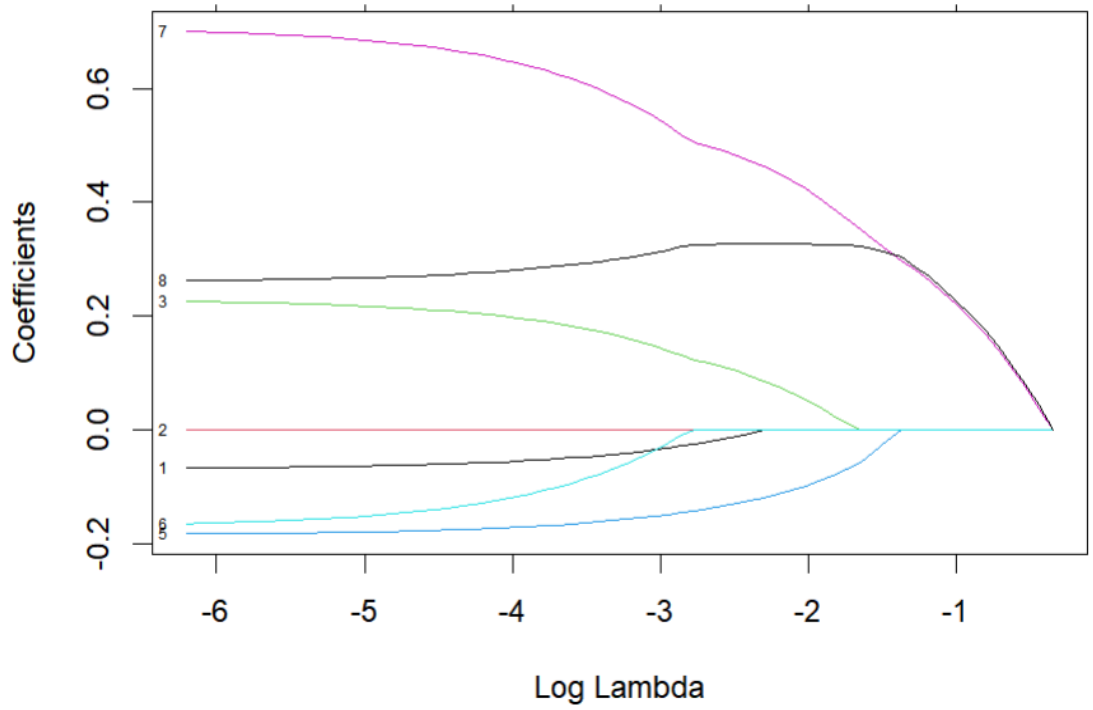

图 24 Log(λ)折线图

之后，我们使用 k 重交叉验证的方法，得到如图 25 所示，Log lambda 的范围在（-4.5，-2.5）。对于每一个 λ 值计算平均均方误差（MSE），从中挑选出最小的平均均方误差，并将对应的 λ 值挑选出来，作为最佳的惩罚项系数 λ 的值，进一步计算可以确定最佳的 λ 值为：0.01078397。

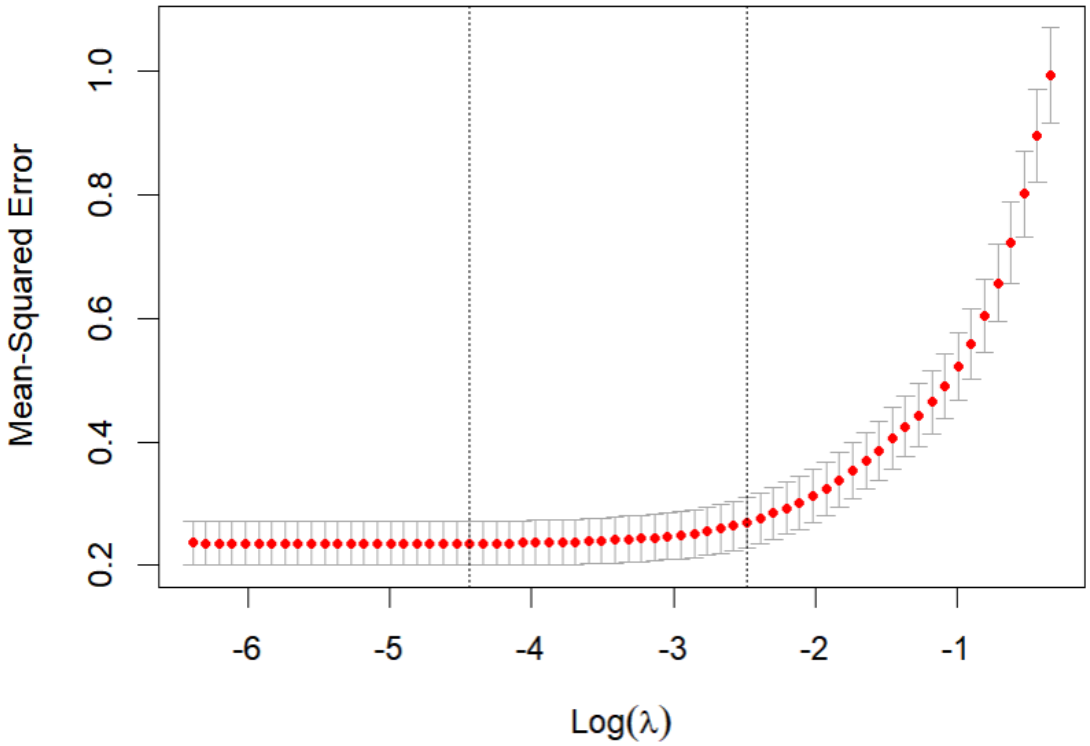

图 25 Log(λ)取值图

根据最佳 λ 值，我们构建 Lasso 回归模型，模型的相关信息如表 19 所示：

表 19 模型展示

变量名	Estimate	Scaled estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)
Symbol_0	-3.38E-02	-6.51E-01	2.56E-01	2.541	0.011 *
Symbol_1	3.38E-02	6.51E-01	2.56E-01	2.541	0.011 *
PG	2.32E-01	4.46E+00	4.27E-01	10.429	< 2e-16 ***
PG-13	4.41E-03	8.49E-02	3.07E-01	0.276	0.782
R	-1.78E-01	-3.43E+00	3.26E-01	10.498	< 2e-16 ***
score	-1.67E-01	-3.22E+00	6.21E-01	5.183	2.18e-07 ***
logvotes	7.04E-01	1.35E+01	7.67E-01	17.636	< 2e-16 ***
logbudget	2.62E-01	5.04E+00	6.59E-01	7.642	2.13e-14 ***

可以看到，Lasso 回归模型得出的结论与岭回归模型得出的结论相似：电影级别为 PG、对数评分人数、对数成本对对数票房有较为显著的正向影响；电影级别为 R、评分对对数票房有负向影响。

（六）分位数回归

由于电影票房受偶然因素影响较大，均值易受到极大或极小值点影响，因此为了增强预测的稳健性，我们进行了分位数回归。分位数回归是估计一组解释变量与被解释变量的分位数之间线性关系的统计建模方法，重点关注被解释变量的条件分位数。相比于线性回归关注的均值，分位数对离群值具有更强的稳健性。并且，选择不同的分位数进行回归，我们可以得到更多与数据分布有关的信息，诸如高票房电影与低票房电影受各因素影响程度的差距大小等。

1.四分位数回归

为了研究低票房电影受哪些因变量影响，我们首先进行四分位数回归，得到结果如下：

表 20 四分位数回归

变量名	0.25 分位数回归系数估计值	0.75 分位数回归系数估计值	0.25 分位数回归 p 值	0.75 分位数回归 p 值
(Intercept)	-8965.26579	-7560.94811	0***	0***
len	29.51563	-45.28185	0.16494	0.06616.
symbol1	103.31011	134.71787	0.2178	0.1314
ratingPG	1008.65356	1086.22396	0***	0***
ratingPG-13	468.29158	571.77475	0***	0***
genreAdventure	187.66702	113.27887	0.59356	0.62036
genreAnimation	297.24417	558.89682	0.16506	0.00977*
genreBiography	-388.65229	-391.78385	0.0026**	0.0051*
genreComedy	-180.16085	13.65025	0.15734	0.89806
genreCrime	-649.13936	28.948	0***	0.94895
genreDrama	-192.66878	-8.30927	0.14645	0.93572

genreHorror	594.57726	827.9779	0.00192**	0.00018***
year2018	119.32552	120.09514	0.13001	0.11424
year2019	189.17786	349.47146	0.01765*	0.00007***
monthApr	-121.89402	90.09059	0.61732	0.48937
monthAug	-248.20186	-17.22214	0.25263	0.88427
monthDec	-157.3658	179.92319	0.52897	0.05507.
monthFeb	-299.7057	39.97253	0.23789	0.78161
monthJul	5.85324	117.47558	0.97844	0.5545
monthJun	-141.49408	-63.51798	0.54348	0.52837
monthMar	-191.22783	-150.42024	0.42499	0.46682
monthMay	-193.88277	74.90217	0.48288	0.70806
monthNov	-267.69257	36.45379	0.22257	0.81928
monthOct	-240.5103	340.94843	0.27758	0.17291
monthSep	-123.4365	-89.69876	0.62889	0.46507
score	-68.58474	-151.50808	0.31727	0.01761*
logvotes	758.21858	688.73732	0***	0***
region1 Not.NorthAmerica	-100.44253	-59.68519	0.37635	0.59259
logbudget	333.67454	261.85462	0***	0***
runtime1	-288.50524	595.25941	0.09932	0.00767*

检验上下四分位数回归系数是否存在显著差异，结果如下：

表 21 回归系数

Df	Resid Df	F value	Pr（>F）
29	713	1.723	0.011*

p 值较小，因此拒绝原假设，认为上下四分位数回归的系数具有显著差异。

2.中位数回归

表 22 模型显著性检验

变量名	p 值	变量	p 值
(Intercept)	0 ***	monthAug	0.29131
len	0.65648	monthDec	0.85054
symbol1	0.00836 *	monthFeb	0.20455
ratingPG	0 ***	monthJul	0.78029
ratingPG-13	0 ***	monthJun	0.39924
genreAdventure	0.00396 *	monthMar	0.00461 *
genreAnimation	0.00141 **	monthMay	0.78685
genreBiography	0.07189	monthNov	0.58227
genreComedy	0.53611	monthOct	0.28118
genreCrime	0.92044	monthSep	0.48692
genreDrama	0.3963	score	0.00002 ***
genreHorror	0 ***	logvotes	0 ***
year2018	0.10048	region1Not.NorthAmerica	0.00713 *
year2019	0.06063	logbudget	0 ***
monthApr	0.54856	runtime1	0.29825

运用逐步回归，并根据 BIC 准则对模型进行选择，得到结果如表 23：

表 23 模型选择

变量名	回归系数估计值	p 值
(Intercept)	-8573.70362	0 ***
symbol1	125.25067	0.03247 .
ratingPG	916.23239	0 ***
ratingPG-13	580.57517	0 ***
genreAdventure	262.61481	0.03815 .
genreAnimation	567.93251	0.00861 *
genreBiography	-301.69753	0.00347 *

genreComedy	67.07104	0.48683
genreCrime	-250.40882	0.16545
genreDrama	-75.09226	0.48936
genreHorror	695.72245	0 ***
score	-158.80178	0.00029 ***
logvotes	745.25463	0 ***
region1Not.NorthAmerica	-151.57621	0.02135 .
logbudget	307.76888	0 ***
runtime1	241.09999	0.0998

3.模型对比

(1)四分位数回归对比

对比上下四分位数回归，发现对于高票房电影，动画题材、十二月上映、评分和电影时长对票房有显著影响，而低票房电影中，犯罪题材电影对票房有显著的负面影响。结合上下四分位数回归的系数，我们得出结论如下：

- 对于低票房电影，犯罪题材将对票房产生极大负面影响。
- 对于高票房电影，选择十二月圣诞节档期上映有利于票房上升，而低票房电影选择圣诞节档期上映将会被高票房电影压缩市场，造成低票房。
- 电影时长较长的高票房电影故事情节更完整、引人入胜，会取得更高票房。
- 不论票房如何，动画题材均有利于票房上升。
- 评分越高的电影反而票房越低。
- 电影时长对高票房电影和低票房电影的影响有显著不同，如图 26 所示。

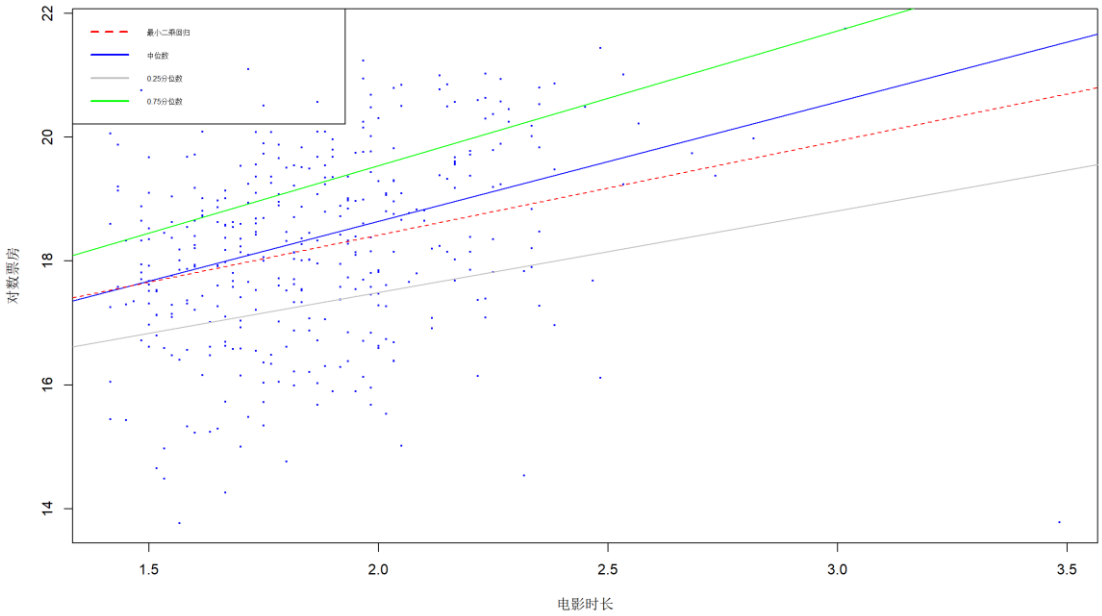

图 26 电影时长分位数回归图

其中第五条结论看似违反常理，其实不然。追求票房的一般为商业片，商业片的评分会低于文艺片，但大部分观众对文艺片的兴趣不如商业片。此外，越高的票房代表着越多的观众，观众上升的同时，不同的审美、争议也会造成评分的下降，票房口碑双丰收的影片在电影史中也只是少数。因此，票房越高的电影评分反而越低。

(2)中位数回归和最小二乘回归对比

表 24 回归对比

变量名	中位数回归系数估计值	最小二乘法系数估计值	中位数回归 p 值	最小二乘法回归 p 值
(Intercept)	-8573.7036	3.34992	0 ***	< 2e-16 ***
symboll	125.25067	208.13	0.03247 .	0.012 *
ratingPG	916.23239	1289.43	0 ***	< 2e-16 ***
ratingPG-13	580.57517	481.75	0 ***	4.64e-10 ***
genreAdventure	262.61481	null	0.03815 .	null
genreAnimation	567.93251	null	0.00861 *	null
genreBiography	-301.69753	null	0.00347 *	null
genreComedy	67.07104	null	0.48683	null
genreCrime	-250.40882	null	0.16545	null

genreDrama	-75.09226	null	0.48936	null
genreHorror	695.72245	null	0 ***	null
score	-158.80178	-250.99	0.00029 ***	3.09e-07 ***
logvotes	745.25463	803.67	0 ***	< 2e-16 ***
region1 Not.NorthAmerica	-151.57621	null	0.02135 .	null
logbudget	307.76888	282.52	0 ***	3.72e-13 ***
runtime1	241.09999	null	0.0998	null

将经过 BIC 选择的中位数回归与未中心化标准化的最小二乘法回归对比，对于中位数回归中显著而最小二乘回归中不显著的变量，我们认为主要是由于部分数据偏离样本总体过大，导致均值被拉高或压低，因此与其他组的区别在最小二乘回归中不明显。因此中位数回归对异常值更加稳健，其变量的显著性更能说明该组数据与样本总体间存在区别。以电影类型中是否含有符号为例，说明其受离群值影响较大，故其在中位数回归中显著，而在最小二乘法回归中不显著。

（七）模型预测

我们分别用 BIC 准则选择出的模型、主成分分析后的回归模型、岭回归得到的模型、lasso 回归得到的模型和分位数回归得到的模型对 2017-2019 年上映的其他 10 部电影的票房进行预测，并将预测值与实际值进行对比。10 部电影的名称、票房数据和自变量数据如图 27 所示：

name	len	symbol	rating	genre	year	released.month	score	votes	region	budget	gross	runtime
The Lego Ninjago Movie	4	0	PG	Animation	2017	Sep	6	24000	NorthAmericas	70000000	1.23E+08	101
Thor: Ragnarok	2	1	PG-13	Action	2017	Nov	7.9	628000	NorthAmericas	1.80E+08	8.54E+08	130
The Nutcracker and the Four Realms	6	0	PG	Adventure	2018	Nov	5.6	30000	NorthAmericas	1.20E+08	1.74E+08	99
The Predator	2	0	R	Action	2018	Sep	5.3	120000	NorthAmericas	88000000	1.61E+08	107
Terminator: Dark Fate	3	1	R	Action	2019	Nov	6.2	159000	NorthAmericas	1.85E+08	2.61E+08	128
Dumbo	1	0	PG	Adventure	2019	Mar	6.3	69000	NorthAmericas	1.70E+08	3.53E+08	112
Alita: Battle Angel	3	1	PG-13	Action	2019	Feb	7.3	240000	NorthAmericas	1.70E+08	4.05E+08	122
Ralph Breaks the Internet	4	0	PG	Animation	2018	Nov	7	140000	NorthAmericas	1.75E+08	5.29E+08	112
Transformers: The Last Knight	4	1	PG-13	Action	2017	Jun	5.2	140000	NorthAmericas	2.17E+08	6.05E+08	154
Guardians of the Galaxy Vol. 2	6	1	PG-13	Action	2017	May	7.6	596000	NorthAmericas	2.00E+08	8.64E+08	136

图 27 数据展示

采用各个模型对该 10 部电影对数票房的预测结果及均方误差如下图所示：

模型名称	RMSE	预测1	预测2	预测3	预测4	预测5	预测6	预测7	预测8	预测9	预测10
BIC模型	0.10485	18.72078	20.47112	19.17218	18.97919	19.40339	19.75321	19.89541	20.12525	20.04593	20.52593
主成分分析模型	0.52497	18.48783	20.71241	18.57687	17.80354	19.03484	19.28048	20.20034	19.79693	19.29698	20.62199
岭回归模型	0.11347	18.69839	20.4431	19.10716	18.87886	19.40534	19.67189	19.91784	20.03054	19.95955	20.48311
lasso回归模型	0.10517	18.71577	20.46853	19.16348	18.97236	19.40008	19.74551	19.89468	20.11844	20.03287	20.52167
1/4分位数回归模型	0.33317	18.43452	20.21041	18.97434	18.61022	19.09839	19.54319	19.77302	20.08711	19.39188	20.30034
中位数回归模型	0.15447	18.94695	20.60918	19.03255	18.86396	19.38049	19.69681	20.01282	20.3669	20.13655	20.665
3/4分位数回归模型	0.42701	19.2455	20.7745	19.23915	19.18795	19.84151	20.16031	20.41189	20.67468	20.4935	20.75395
loggross	\	18.62836	20.56542	18.97434	18.89407	19.38049	19.68278	19.81935	20.08711	20.22144	20.5798

图 28 预测结果

由图 28 可见，各个模型所得预测值的均方误差均较小，说明模型预测效果较好。而在以上 7 个模型中，BIC 模型预测值的均方误差最小，因此将 BIC 模型作为对数电影票房的最优预测模型。BIC 模型预测的置信区间及对数电影票房的真实值如下表所示：

表 25 置信区间

电影序号	lwr	upr	loggross
1	18.52609	18.911	18.6284
2	20.30446	20.635	20.5654
3	18.97056	19.369	18.9743
4	18.79601	19.158	18.8941
5	19.20874	19.594	19.3805
6	19.56998	19.933	19.6828
7	19.73357	20.054	19.8194
8	19.94702	20.3	20.0871
9	19.85947	20.228	20.2214
10	20.36657	20.682	20.5798

由上表可见，10 部电影票房的预测值均落在置信区间之内，说明 BIC 模型的预测效果十分出色。

五、结论及建议

基于已有数据和上述统计模型，我们对可能影响电影票房的因素进行了全面的分析，对电影市场的良性发展和电影吸引力的提升得到以下结论和建议：

首先，标题中存在特殊符号的电影，往往能收获更高的票房。这说明了观

众更在意电影给他们留下的第一印象，而一个引人注目的标题也更容易吸引观众的注意。所以，对于一部优秀的电影来说，想要获得更高的票房，需要在标题上多下功夫，例如通过编写存在特殊符号的标题，可能达到吸引更多观众的效果。

其次，评分更高的电影，却容易收获更低的票房。这种现象是对电影制作方观点的一个验证，说明在泛娱乐化的现代社会，人们更愿意为爆米花式的、注重快感的电影买单，而缺乏去理解、感悟一部内涵丰富、引人深思的电影的耐心。从而，尽管意蕴深刻的电影往往容易拥有更高的评分，但它们却难以得到较高的票房，而这也容易打击导演和编剧的积极性。因此，如果高分电影想要获得更高的票房，一方面，制作方需要在如宣发、标题等其他影响因素上投入更多精力；另一方面，官方机构和在电影界有影响力的组织也需要对这类电影加以支持。

最后，尽管在通常情况下，电影预算和电影票房成正相关关系，但也存在例外的情况。例如，在我们删除的强影响点数据中，其预算高达 1.59 亿美元，但其仅收获了不足 100 万美元的票房。因此，电影制作方不能只是一味地加大电影预算以谋求高收益，而应该进一步考虑电影应当从哪些方面吸引观众，并从标题、类型和主演等方面进行充分的考虑，才能在制作一部好电影的同时，让电影票房与之相称。

六、缺陷与改进

目前已有的模型仍存在一定不足，我们认为可从以下几个角度进行改进：

首先，在数据源方面，由于数据集主要收集的是国外电影的数据，且电影预算、电影分级等数据不易获取，我们很难收集到国产电影相关的、全面的数据，从而难以对国产电影的情况进行完整、准确的分析。如果能在结合现有数据的基础上收集到国产电影的完整数据，可能得到更加全面的分析结果。

其次，在变量选择方面，由于难以量化，我们将电影公司、电影主演等变量忽略了，但实际上由于明星效应和羊群效应，这些变量往往容易对电影票房产生较大影响，由此可能导致模型误差偏大。如果能够通过主演粉丝数量、公司知名度等指标将这些变量考虑在内，可能得到更精确的分析结果。

最后，针对我们研究的问题，可能存在比起线性回归解释性更好的模型。未来可以尝试更多类型的模型对数据进行拟合，对我们的分析结果加以改进。

附录 I 参考文献

[1]王松桂,史建红,尹素菊等.线性模型引论[M].北京:科学出版社,2004.

[2]王汉生.应用商务统计[M].北京:北京大学出版社,2008.

[3]满敬銮,杨薇.基于多重共线性的处理方法[J].数学理论与应用,2010,20(02):105-109

[4]林海明,杜子芳.主成分分析综合评价应该注意的问题[J].统计研究,2013,30(08):25-31.