

本科毕业论文

题 目 中国电影票房影响因素分析

|  |  |
| --- | --- |
| 学 院 | 数学与统计学院 |
| 专 业 | 统计学 |
| 年 级 | 2013级 |
| 学 号 | 222013314032023 |
| 姓 名 | 林淑颖 |
| 指 导 教 师 | 徐文昕 |
| 成 绩 |  |

2017年4月19日

目录

[摘要 1](#_Toc480393899)

[Abstract 1](#_Toc480393900)

[1. 研究目的 1](#_Toc480393901)

[2. 背景介绍 1](#_Toc480393902)

[3. 指标设计 3](#_Toc480393903)

[4. 描述分析 6](#_Toc480393904)

[4.1 简单描述分析 6](#_Toc480393905)

[4.2 聚类分析 9](#_Toc480393906)

[4.3 聚类分析结果评估 14](#_Toc480393907)

[5. 模型分析 14](#_Toc480393908)

[5.1 方差分析 14](#_Toc480393909)

[5.2 逻辑回归模型 17](#_Toc480393910)

[5.3 逻辑回归模型评估 17](#_Toc480393911)

[6. 总结讨论 18](#_Toc480393912)

[参考文献 18](#_Toc480393913)

[致谢 18](#_Toc480393914)

**中国电影票房影响因素分析**

林淑颖

西南大学数学与统计学院，重庆 400715

**摘要：**本文通过对2010-2013年度中国国产电影票房历史数据的分析，试图找出可能影响票房的各个重要因素，并通过建立方差分析和逻辑回归模型，进一步探究这些因素与电影票房的相关性。除此之外，我们还进行了聚类分析，研究不同票房水平的影片的特征，深入了解当前国产电影的特点，总结其优势及不足，从而提出提高国产影片票房的有针对性的建议。

**关键词**：电影票房；上映月份；导演；主演；方差分析；逻辑回归；聚类分析

**Analysis of influential factors of Chinese film box office**

Lin Shuying

School of Mathematics and Statistics, Southwest University, Chongqing 400715, China

**Abstract:** This paper tries to find out the important factors that may affect the box office by analyzing the historical data of Chinese domestic film box office 2010-2013, and explores the relativity between these factors and the box office by establishing the analysis of variance and logistic regression model. In addition, we also carry out cluster analysis to study the characteristics of different box-office levels of the films, in-depth understanding of the current characteristics of domestic films, and sum up its advantages and shortcomings, thus make targeted recommendations to improve the box office of domestic films.

**Key words:** Movie Box Office; Release Month; Director; Starring; Analysis of Variance; Logical Regression; Cluster Analysis

# 研究目的

通过对2010年至2013年间中国电影发行放映协会统计的过千万票房的国产电影的相关数据的研究，更深入的了解中国电影行业的现状，同时探究什么样的国产电影将会更有市场。

# 背景介绍

随着经济的发展，人民收入水平的提升，人民对精神食粮的渴求以及其他娱乐方式的匮乏直接导致看电影的人数和票房收入节节攀升，自2003年电影产业化改革启动以来的十余年间，中国电影的市场规模便一直保持着每年30%左右的增长。2005年，中国电影在度过百年诞辰后，迎来了蓬勃发展的“新元年”，电影产业在国内经济高速发展、综合国力不断增强、电影产业化改革和市场需求的拉动下，焕发出旺盛的生命力和可持续发展的巨大潜力。中国的电影市场进入了前所未有的繁荣发展时期。

2010年以后，我国电影市场总体保持在较高的水平，中国电影产业的发展具备了扎实稳健的内在功力。国家新闻出版广电总局电影局公布的数据显示，2013年我国城市影院观影人次总和已经超过6亿，城市影院观影人次已经连续四年同比增长超过30%，中国电影已经取得了不错的成绩。

下面我们先来看一下中国电影2010-2013年的票房数据：

表1 中国电影2010-2013年的票房数据

|  |  |  |
| --- | --- | --- |
| 年份 | 总票房（亿元人民币） | 同比增长率 |
| 2010 | 101.72 | 63.9% |
| 2011 | 131.15 | 28.9% |
| 2012 | 170.73 | 30.2% |
| 2013 | 217.69 | 27.5% |

2010年中国电影票房迅速超过百亿，增长至101.72亿人民币，进入全球排名前五，仅次于美国、日本、英国、法国等国。2013年中国电影票房再次翻番达到217.69亿人民币，超过日本成为全球第二大电影票房市场，位列美国之后，电影票房份额达到全球的10%。这对国外影片也产生了直接的影响，他们开始增加影片中的“中国元素”。

在中国电影产业蒸蒸日上的发展状况下，我们再来看一下国产影片的实际发展情况。2010年度，国产影片的票房为57.34亿，占总票房的56.37%，2011年度，国产电影的票房为70.31亿，占总票房的53.61%，2012年度，国产影片的票房为82.73亿，占总票房的48.46%，2013年度，国产电影的票房为127.67亿，占总票房的58.65%。从上述数据中可以看出，虽然国产影片的票房在持续增长，可是其相对总票房的占比并不高。

同时根据调研结果，19～35岁的年轻观众已经成为中国影院观众的主要人群，占所有影院观众的85.9%。并且有消费能力的年轻人逐渐成为电影消费的主力军，成为票房的重要保证。那么，该怎样迎合这一部分观众的口味就成为了我们关注的问题。同时为了提高国产影片的质量以及提升其社会效果，赢得更多的国内观众的信任，拓展海外市场，研究影响国产影片票房的因素就是十分必要的了。

# 指标设计

首先我们应该考虑一个问题：究竟什么样的电影算是成功的电影呢？衡量一部电影是否成功有许多指标，但有一项是毋庸置疑的，那就是票房。所以我们选择票房数据作为因变量，但是票房的数值大小差别十分大。其实我们真正关心的并不是这些数值，而是票房是处于较高水平还是较低水平。曾经电影行业十分流行“亿元导演俱乐部”的说法，对票房过亿的影片的推崇足以说明过亿的票房是相对较高的，既然我们只关心票房是否属于高票房，那么我们的票房数据就只有两个取值0,1,0=票房未过亿，即低票房，1=票房过亿，即高票房。但是，票房是多种因素共同作用的结果，因此，仅选用票房指标来衡量电影是否成功是不够的，所以我们将同时选取另外两个指标作为因变量进行分析，即观影人数以及电影放映场数，单位分别为万人、千场。

为了能够更好地探究什么样的电影会在中国卖得好，从而建立更加符合实际的票房预测模型，我们将考虑以下几个指标，并对选择其作为解释性变量的原因进行详细的讨论。

●影片类别

首先考虑的变量是影片类别，影片类别实际上反应的是电影的题材，所以该变量为分类变量，共有15个类型，15个类型分别是：喜剧、魔幻、动作、爱情、灾难、历史剧情、励志、主旋律、警匪、悬疑、动画、儿童、惊悚、家庭伦理。电影是艺术，但同样，也是商品，在消费文化的长时间影响下，电影正在形成各种拥有相对稳定欣赏群体的类别，所以为了获得大量的观众群体，大多数影片都兼具了许多种影片类型，尽管各种电影的彼此交叉给科学分类带来了困难，但是探讨其中规律却是对掌握市场走向和运作方式有着特殊的价值，因此我们仅选取了该片最主要的类型。

●上映年份

不同的上映年份可能会对电影票房产生不同的影响，如果某年经济形势严峻，那么势必电影票房也会降低，所以我们选择这一变量作为解释性变量，该变量取值为2010-2013，即数据来源于这四年之间。

●上映月份

该变量的取值为1-12，电影票房很容易受到假期的影响，假期多的月份观影人数也一定会随之增长，比如暑期档、贺岁档的电影就可能会吸引更多人观影。

●上映周数

该变量的取值为1-53，从当年1月1号起记为第一周，选择上映周数作为解释性变量与上映月份的理由相似，均是为了分析上映时间对票房的影响。

●电影宣发公司

电影宣发公司在某种程度上能够反映电影的宣传力度，较大型的公司的宣传可能更到位，并且影响力、票房号召力更强，该变量有四个取值：C（国有宣发公司，例如：中国光影集团）,S（明星私企，例如：光线传媒）,L（小私营公司）,G（联合发行）。

●时长

即电影的分钟数，过长或过短的电影都会对观众的观影体验产生影响，找到合理的时长，从而能够有效的控制电影的剪辑，迎合观众的喜好，这是十分重要的。

●作品是否改编

该指标是电影产品的核心元素，观众对电影的审美能力日益成熟，对于国产影片，观众越来越看重故事的精彩程度。该指标主要包括两个水平，一个是畅销小说改成剧本，另一个则是非改编剧本，也就是原创剧本，我们用N表示改编的剧本，S表示原创剧本。

●作品是否真人真事

艺术源于生活，来自真人真事的作品往往可以引起更广泛的关注，从而引起更多的共鸣，因此我们选择这一指标作为解释变量，该指标主要包括两个水平0,1，1=真实，0=虚构。

●作品是否翻拍

翻拍作品大多数都是曾经的成功作品，已经拥有了一部分观众，但同时也存在另一个问题，那些忠于原作的观众是否能够接受这类作品？该指标同样包括两个水平0,1，1=翻拍，0=非翻拍

●作品是否续集

续集作品有其自身独特的优势，一是好拍，二是票房好。那么在在国产影片市场中，这两个优势能否得到发挥，而观众又是否买账？这是值得我们探究的。

●导演姓名

不同导演的风格不同，并且因其已有作品的口碑不同，势必会对电影的票房产生影响。

●导演年代

导演年代在某种程度上会影响导演的风格，同时也能展现出导演的资历，甚至传达出在那个年代所形成的价值观念，因此导演年代也是我们纳入考虑的解释性变量。

●第几部作品

该变量刻画了影片是该导演的第几部作品。票房最高的电影《人再囧途之泰囧》是导演徐峥的处女作，但是拿到了12亿的票房，作为导演，经验能在多大程度上影响票房？我们是否能从导演的作品数判断其作品将会取得的成就？这是我们选择这一变量进行探究的目的。

●导演得奖情况

导演得奖情况能够在某种程度上反应导演的才华，并且是对他曾经的作品的一种肯定，那么是否得奖较多的导演就意味着较高的票房呢？这仍需我们进一步探究，该指标包括的两个水平为：1=得奖，0=没得奖。

●导演是否转型

随着电影产业的发展，越来越多的人进入这一行业，导演的数量也是与日俱增，而导演又是一门能够自学成才的工种，在中国电影市场中，演员出身，作家出身的导演并不少见，我们希望看到已经在自己的领域取得成就的那些名人，随着工种的转变是否会对票房有影响。

●主演1

中国的粉丝效应是不容忽视的，因此，不同演员的演技甚至其粉丝基数的大小都可能会对电影票房产生影响。

●主演1 的百度指数

影片的主演在分析中是一个字符变量，为了更好的使用这一指标，同时优化工作效率，我们决定在相同的时间段，选择所有演员百度指数的平均值作为衡量主演的标准。因为大多数电影不会仅有一个主演，所以我们选取了两个主演。百度指数越高，代表该演员的知名度越高。

●主演2

一部成功的电影不可能只有一个主演，与选择主演1作为解释性变量的原因类似，我们选择主演2作为解释性变量。

●主演2 的百度指数

同样用百度指数衡量主演2.

以上就是本研究所选取的全部变量。

针对以上变量，我们将通过聚类分析来进行分组，然后对所得的分类变量进行描述性分析，而在模型分析部分，我们将探究影响影片的观影人数、放映场数的因素，此时我们仍使用原始数据进行方差分析。

# 描述分析

## 4.1 简单描述分析

在正式的模型分析之前，我们首先对因变量及解释变量做必要的描述分析，以便更好的了解数据。主要结果如表所示。

表2 因变量及部分解释性变量的描述分析

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变量名称 | 样本量 | 均值 | 标准差 | 最小值 | 中位数 | 最大值 |
| 放映场数 | 266 | 125.24 | 123.451 | 9.69 | 77.94 | 833.71 |
| 人数 | 266 | 336.84 | 478.978 | 31.91 | 159.71 | 3920.55 |
| 票房 | 266 | 0.338 | 0.474 | 0 | 0 | 1 |
| 上映月份 | 266 | 6.556 | 1.115 | 1 | 7 | 12 |
| 上映周数 | 266 | 27.33 | 3.362 | 1 | 28 | 53 |
| 时长 | 266 | 104.9 | 13.810 | 77 | 103 | 156 |
| 作品是否改编 | 266 | 0.906 | 0.292 | 0 | 1 | 1 |
| 作品是否真人真事 | 266 | 0.090 | 0.287 | 0 | 0 | 1 |
| 作品是否翻拍 | 266 | 0.098 | 0.298 | 0 | 0 | 1 |
| 作品是否续集 | 266 | 0.162 | 0.369 | 0 | 0 | 1 |
| 导演年代 | 266 | 60.94 | 8.916 | 30 | 60 | 80 |
| 第几部作品 | 266 | 13.55 | 18.482 | 1 | 7 | 110 |
| 导演得奖情况 | 266 | 0.5865 | 0.493 | 0 | 1 | 1 |
| 导演是否转型 | 266 | 0.139 | 0.347 | 0 | 0 | 1 |
| 主演1百度指数 | 266 | 9536 | 12852.52 | 0 | 5534 | 84936 |
| 主演2 百度指数 | 266 | 7651 | 11944.19 | 0 | 3716 | 84936 |

从上述表格可以看出，剔除含有缺失值的样本后，我们一共选取了266个样本。

●关于放映场数，其中位数为77.94，反应出了放映场数的平均水平，而其最小值为9.69，最大值为833.71，相差十分悬殊。

●关于观影人数，中位数为159.71，最小值为31.91，最大值为3920.55，同样是相差十分悬殊，说明数据的变异性大。

●关于票房，其均值为0.338，说明有33.8%的影片的票房是过亿的，即我们认为有33.8%的影片是属于高票房的。

●关于上映月份，其均值为6.556，说明多数影片会选择在暑期档的时候上映。

●关于上映周数，其均值为27.33，同样是暑期档的时间。

●关于时长，中位数为103，说明这一时长的电影更易被观众所接受，既不至于因为太短而不尽兴，又不会因为太长而失去耐心。

●关于作品是否改编，其均值为0.906，说明90.6%作品都是改编而来的。

●关于作品是否真人真事，其均值为0.09，说明只有9%的作品是来源于真人真事。

●关于作品是否翻拍，其均值为0.098，说明只有9.8%的作品是翻拍的，我们推测翻拍作品在中国并不具备市场，或者从侧面反映，先入为主的观念使得中国观众更接受第一部作品。

●关于作品是否续集，均值为0.162，说明16.2%的作品是续集。

●关于导演年代，均值为60.94，说明作为一个导演应该是拥有一定阅历的，所以年纪可能会偏大。

●关于第几部作品，中位数为7，说明多数导演都不只拥有一部作品。

●关于导演得奖情况，均值为0.5865，说明58.65%的导演都曾得过奖。

●关于导演是否转型，均值为0.139，说明13.9%的导演是转型的，其比例并不大。

●关于主演1 和主演2的百度指数，明显主演1的百度指数较高。

●关于影片类别，如下图：

图1 影片类别的描述分析

从影片类别来看，爱情片占据了中国电影业的主要市场。

●关于导演，如下图：

图2 导演的描述分析

从上述数据，看出知名导演的作品数量也会比较多。

●关于主演1，如下图：

图3 主演1的描述分析

作为主演，多数为当红一线明星。

●关于主演2，如下图：

图4 主演2的描述分析

同样为一线明星较多。

●关于宣发公司，如下图：

图5 宣发公司的描述分析

宣发公司的样本数量相差不大。

## 4.2 聚类分析

为了更深入了解变量特征，我们选择用k-均值聚类法对影片放映场数、观影人数以及票房进行聚类，将数据聚为三类，聚类结果如下：

表3 聚类分析结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类别 | 样本数 | 放映场数 | 观影人数 | 票房 |
| 1 | 15 | 492.17000 | 1917.4187 | 70261.775 |
| 2 | 184 | 63.76092 | 121.9814 | 3825.346 |
| 3 | 67 | 211.92075 | 573.0243 | 19171.219 |

从上述结果可以得到聚类的三个类别的均值，根据其均值，可以将第一类定义为高票房电影，而第二类为低票房影片，第三类为票房处于两者之间的，同时我们得到三类影片的样本点分别为：《狄仁杰之神都龙王》、《魁拔之大战元泱界》、《消失的子弹》，究竟以这三部影片为代表样本的电影分别有怎样的特点，我们来进行具体的分析。

表4 三组中影片类别结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 影片类别 | 1 | 2 | 3 | 合计 |
| 爱情 | 4(1.5) | 46(17.3) | 13(4.9) | 63(23.7) |
| 动画 | 0(0) | 19(7.1) | 4(1.5) | 23(8.6) |
| 动作 | 4(1.5) | 21(7.9) | 18(6.8) | 43(16.2) |
| 儿童 | 0(0) | 4(1.5) | 0(0) | 4(1.5) |
| 家庭伦理 | 0(0) | 6(2.3) | 0(0) | 6(2.3) |
| 惊悚 | 0(0) | 27(10.2) | 0(0) | 27(10.2) |
| 警匪 | 0(0) | 11(4.1) | 7(2.6) | 18(6.8) |
| 历史剧情 | 1(0.4) | 6(2.3) | 5(1.9) | 12(4.5) |
| 励志 | 1(0.4) | 3(1.1) | 0(0) | 4(1.5) |
| 魔幻 | 2(0.8) | 3(1.1) | 6(2.3) | 11(4.1) |
| 喜剧 | 2(0.8) | 27(10.2) | 7(2.6) | 36(13.5) |
| 悬疑 | 0(0) | 1(0.4) | 5(1.9) | 6(2.3) |
| 灾难 | 1(0.4) | 1(0.4) | 1(0.4) | 3(1.1) |
| 主旋律 | 0(0) | 9(3.4) | 1(0.4) | 10(3.8) |
| 合计 | 15(5.6) | 184(69.2) | 67(25.2) | 266(100) |

注：括号内的数据为百分比。

上述结果给出了266部国产电影的类型，从数据看，高票房组排名靠前的两种类型为爱情和动作，低票房组票房排名靠前的类型为爱情、惊悚、喜剧，爱情片的数量远远超过了另外两种类型，票房处于中间位置的一组中，排名靠前的影片类型为动作和爱情，由此可见，爱情是中国影片中永恒不变的题材，可是爱情片虽然有更广泛的观众群体，可是，在爱情片泛滥的今天，取得成功的难度也极大的增加了，同样我们可以看出，动作片也越来越受欢迎。而喜剧片和惊悚片大多处于低票房组中，说明两种类型片在中国的市场并不好，喜剧片和惊悚片的意义并不仅仅是为了让观众大笑或是感到刺激，而是要让所有观看过电影的人都能够产生新的思考和感悟，所以制作这类作品的难度也更高些。

表5 三组中上映年份结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 上映年份 | 1 | 2 | 3 | 合计 |
| 2010 | 3 | 42 | 13 | 58 |
| 2011 | 2 | 47 | 14 | 63 |
| 2012 | 3 | 51 | 16 | 70 |
| 2013 | 7 | 44 | 24 | 75 |
| 合计 | 15 | 184 | 67 | 266 |

从结果可以看出，国产影片的数量在四年间是逐渐上升的，虽然增幅并不大，票房较高的两组中，2013年的影片数量相比之前有很大的提升，而低票房组中，反而是11年、12年的影片数量较多，我们可以推测造成这种结果的原因可能是随着时间变化，国产电影的质量得到了提升，但也可以猜测为随着经济的发展，愿意到影院看电影的观众数目增多，但是我们不能得出一个确定的结论，因为影响影片票房的因素有很多，政治、经济、文化政策都可以对其产生很大的影响。

表6 三组中上映月份结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 上映月份 | 1 | 2 | 3 | 合计 |
| 1 | 0 | 11 | 11 | 22 |
| 2 | 1 | 15 | 6 | 22 |
| 3 | 1 | 19 | 2 | 22 |
| 4 | 1 | 12 | 7 | 20 |
| 5 | 1 | 7 | 0 | 8 |
| 6 | 2 | 22 | 3 | 27 |
| 7 | 1 | 24 | 6 | 31 |
| 8 | 0 | 22 | 7 | 29 |
| 9 | 1 | 19 | 10 | 30 |
| 10 | 0 | 14 | 2 | 16 |
| 11 | 0 | 11 | 6 | 17 |
| 12 | 7 | 8 | 7 | 22 |
| 合计 | 15 | 184 | 67 | 266 |

从上述结果可看出，影片上映时间多集中于7,8,9三个月以及12月、1月、2月，即为暑期档和贺岁档的时间，选择这两个时间段上映的影片的票房较高的可能性更大些。

表7 三组中上映周数结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 上映周数 | 1 | 2 | 3 | 合计 |
| 1~13 | 2 | 41 | 19 | 62 |
| 14~26 | 2 | 42 | 10 | 54 |
| 27~39 | 4 | 62 | 19 | 85 |
| 40~53 | 7 | 39 | 19 | 65 |
| 合计 | 15 | 184 | 67 | 266 |

将影片上映周数分为了四个阶段，从结果看，第三季度和第四季度的影片较多，相应较高票房的影片集中在三、四两季度，同样对应于暑期档，以及假期较多的时间。

表8 三组中宣发公司结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 宣发公司 | 1 | 2 | 3 | 合计 |
| C | 4 | 42 | 11 | 57 |
| G | 2 | 44 | 13 | 59 |
| L | 0 | 58 | 10 | 68 |
| S | 9 | 40 | 33 | 82 |
| 合计 | 15 | 184 | 67 | 266 |

从宣发公司看，S公司的电影具有更高的票房，S代表的是明星私企，说明现在某些私企的影片质量反而超过了大型国企，这是值得某些国企借鉴的。

表9 三组中影片时长结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 影片时长 | 1 | 2 | 3 | 合计 |
| 76~95 | 0 | 71 | 7 | 78 |
| 96~115 | 3 | 91 | 34 | 128 |
| 116~135 | 10 | 20 | 24 | 54 |
| 136~156 | 2 | 2 | 2 | 6 |
| 合计 | 15 | 184 | 67 | 266 |

从时长的结果看，多数影片的时长集中在96~115之间，从高票房数据组看，时长为116~135的影片居多，可见这一时长将会给观众带来更好的观影体验，所以影片剪辑是应注意将时长控制在这一时间段内。

表10 三组中作品是否改编结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 作品是否改编 | 1 | 2 | 3 | 合计 |
| N | 6 | 9 | 10 | 25 |
| S | 9 | 175 | 57 | 241 |
| 合计 | 15 | 184 | 67 | 266 |

从结果看，尽管改编作品中只有36%处于低票房组中，但是改编作品的数量较少，但是由于数据量的限制，我们没办法确定得出改编作品就一定可以带来更高的票房的结论。

表11 三组中作品是否真人真事结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 真人真事 | 1 | 2 | 3 | 合计 |
| 0 | 13 | 169 | 60 | 242 |
| 1 | 2 | 15 | 7 | 24 |
| 合计 | 15 | 184 | 67 | 266 |

从结果看，仍然是只有很少一部分影片是真人真事的，同样，虽然在真人真事的影片中，有62.5%的影片处于低票房组中，我们仍然不能够确定真人真事的影片并不受欢迎。

表12 三组中作品是否翻拍结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 翻拍 | 1 | 2 | 3 | 合计 |
| 0 | 14 | 172 | 54 | 240 |
| 1 | 1 | 12 | 13 | 26 |
| 合计 | 15 | 184 | 67 | 266 |

从结果看，翻拍的作品并不多，其中高票房的影片只有一部，其实结合我们个人的观影体验来说，大多数时候翻拍作品都不会太受欢迎，因为翻拍的作品大多是经典，而经典之所以为经典，就是它不可超越，但这些也只是我们的推测，同样由于数据量的问题，我们不能确切得出结论。

表13 三组中作品是否续集结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 续集 | 1 | 2 | 3 | 合计 |
| 0 | 10 | 158 | 55 | 223 |
| 1 | 5 | 26 | 12 | 43 |
| 合计 | 15 | 184 | 67 | 266 |

续集作品的样本量相对有所增加，但是多数续集作品仍集中在低票房组中，可见，续集作品也并不一定是票房的保证。

我们只关注了高票房组的导演，情况如下：冯小刚3部，徐克2部，姜文、郭敬明、徐峥、薛晓璐、张艺谋、赵薇、周星驰、陈可辛、成龙、乌尔善各一部，几乎均为我们所熟知的导演，而在低票房组中，几乎没有这些导演的作品。

表14 三组中导演年代结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 导演年代 | 1 | 2 | 3 | 合计 |
| 30 | 0 | 1 | 0 | 1 |
| 40 | 0 | 3 | 0 | 3 |
| 50 | 7 | 39 | 21 | 67 |
| 60 | 3 | 72 | 31 | 106 |
| 70 | 4 | 61 | 12 | 77 |
| 80 | 1 | 8 | 3 | 12 |
| 合计 | 15 | 184 | 67 | 266 |

从结果看，大多数导演都是50，60，70年代的，尤其在高票房组这一特征更为明显，我们可以这样理解，这一时期的导演既有足够的经历，同时依旧能够接受新生事物，因此作品的观众群体更为广泛。

表15 三组中第几部作品结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 第几部作品 | 1 | 2 | 3 | 合计 |
| 1~10 | 7 | 120 | 33 | 160 |
| 11~20 | 5 | 36 | 16 | 57 |
| 21~30 | 1 | 13 | 9 | 23 |
| 31~40 | 0 | 2 | 5 | 7 |
| 41~50 | 2 | 2 | 0 | 4 |
| 51~60 | 0 | 3 | 2 | 5 |
| 61~70 | 0 | 1 | 1 | 2 |
| 71~80 | 0 | 3 | 0 | 3 |
| 81~90 | 0 | 0 | 0 | 0 |
| 91~100 | 0 | 3 | 1 | 4 |
| 101~110 | 0 | 1 | 0 | 1 |
| 合计 | 15 | 184 | 67 | 266 |

从结果看，大多数导演的作品都是在1~10组的，但是值得引起我们注意的是，有几位导演的作品数超过了50，甚至达到了110，然而这些导演的作品却几乎都出现在了低票房组和中间票房组，这足以说明导演这一职业不能只是依靠勤奋以及经验，即使已经拍过众多作品，也不能保证影片的票房。

表16 三组中导演得奖情况结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 导演得奖情况 | 1 | 2 | 3 | 合计 |
| 0 | 2 | 93 | 15 | 110 |
| 1 | 13 | 91 | 52 | 156 |
| 合计 | 15 | 184 | 67 | 266 |

从结果看，超过的一半的导演都得过奖，而在高票房组的得奖率高达87%，足以证明导演得奖情况对票房有重要的影响。

表17 三组中导演是否转型结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 是否转型 | 1 | 2 | 3 | 合计 |
| 0 | 7 | 167 | 55 | 229 |
| 1 | 8 | 17 | 12 | 37 |
| 合计 | 15 | 184 | 67 | 266 |

在高票房组转型导演与非转型导演的数量相差不大，而在另外两组中非转型导演数量远远超过了转型导演的数目，我们无法确定这种结果是否是因为高票房组数据量较少，但是值得我们注意的是，转型导演当中有22%的导演处于高票房组，32%的导演处于中间票房组，说明转型导演还是有很大的概率取得成功的。

同样我们关注了高票房组的主演1，葛优两部作品，成龙、冯绍峰、黄晓明、姜文、李连杰、倪妮、汤唯、文章、徐帆、徐峥、杨幂、杨子珊、赵薇各一部作品，几乎均为一线且知名度、活跃度均较高的演员，但同时低票房组也不乏这些演员的作品，并且较多部影片的主演为黄渤、郭富城、李冰冰等影帝影后，虽然演员可以在某种程度上影响票房，但如果影片本身质量不高，再大牌的演员也不能挽救票房。

## 4.3 聚类分析结果评估

聚类分析所得组间平方和占总平方和的83.6%，该数据足够大，表明聚类效果较好。

# 模型分析

## 5.1 方差分析

在描述分析的基础上，我们进行了模型分析。首先建立了线性回归模型，我们利用step函数筛选出变量得到最终的模型后，通过方差分析对各个因素同放映场数之间做了模型分析。

表18 方差分析检验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名称 | 自由度 | F-统计量 | p-值 |
| 上映年份 | 1 | 32.6713 | <0.05 |
| 上映月份 | 1 | 10.2633 | <0.05 |
| 宣发公司 | 3 | 10.7137 | <0.05 |
| 时长 | 1 | 64.8833 | <0.05 |
| 作品是否改编 | 1 | 7.4326 | <0.05 |
| 作品是否续集 | 1 | 11.6848 | <0.05 |
| 导演得奖情况 | 1 | 4.2113 | <0.05 |
| 导演是否转型 | 1 | 20.6685 | <0.05 |
| 百度指数1 | 1 | 5.5438 | <0.05 |

筛选变量以后，此时方差分析的结果表明，模型整体的F-检验高度显著（p-值<0.0001），并且所有筛选出来的变量都是显著的，即放映场数与上映年份、上映月份、宣发公司、时长、作品是否改编、作品是否续集、导演得奖情况、导演是否转型以及主演1的百度指数等因素有关。判决系数为42.73%，在涉及诸多影响因素的实际问题中，此模型的拟合优度是可以被我们所接受的，相关参数估计及检验结果如下表：

表19 各参数估计检验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 因素名称 | 参数估计 | t-统计量 | p-值 |
| 截距项 | -62610 | -5.804 | <0.05 |
| 上映年份 | 31.03 | 5.789 | <0.05 |
| 上映月份 | 3.704 | 2.051 | <0.05 |
| 宣发公司G | -20.68 | -1.126 | 0.26113 |
| 宣发公司L | -42.6 | -2.442 | <0.05 |
| 宣发公司S | 9.127 | 0.530 | 0.59626 |
| 时长 | 2.744 | 5.598 | <0.05 |
| 作品是否改编 | -42.34 | -1.933 | 0.05431 |
| 作品是否续集 | 54.26 | 3.332 | <0.05 |
| 导演得奖情况 | 32.65 | 2.526 | <0.05 |
| 导演是否转型 | 73.06 | 4.191 | <0.05 |
| 百度指数1 | 0.001133 | 2.355 | <0.05 |

从表中可以得到的结论有：（1）上映年份、作品是否续集、导演得奖情况、导演是否转型的系数均为正，并且较大，说明上映年份越迟，放映场数越多，这可能是由于随着时间增长，影院越来越多，这也从侧面说明了中国电影业的发展，续集的作品放映场数更多，续集的作品本来就是有一定的观众基础的，也许续集可以作为中国电影业发展的另外一个方向，得奖的导演、转型的导演的作品放映场数更多，这反映了一部作品的好坏，导演起着重要的作用。（2）与宣发公司C相比，L与G处于劣势地位，而S则优于C。（3）作品是否改编的系数为负，且较大，说明改编的作品并不受欢迎。（4）上映月份，时长以及主演1 的百度指数均为正，但比较小，也能够说明上映月份越迟，时长越长、主演1 的百度指数越高，电影的放映场数都会越多。

下面是关于影响观影人数的因素的模型分析。

表20 方差分析检验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名称 | 自由度 | F-统计量 | p-值 |
| 上映年份 | 1 | 6.9769 | <0.05 |
| 上映月份 | 1 | 6.5756 | <0.05 |
| 宣发公司 | 3 | 10.5823 | <0.05 |
| 时长 | 1 | 56.6698 | <0.05 |
| 作品是否改编 | 1 | 9.7030 | <0.05 |
| 作品是否续集 | 1 | 13.5087 | <0.05 |
| 导演得奖情况 | 1 | 4.7160 | <0.05 |
| 导演是否转型 | 1 | 24.5103 | <0.05 |

筛选变量以后，此时方差分析的结果表明，模型整体的F-检验高度显著（p-值<0.0001），并且所有筛选出来的变量都是显著的，即观影人数与上映年份、上映月份、宣发公司、时长、作品是否改编、作品是否续集、导演得奖情况、导演是否转型等因素有关。判决系数接近40%，在涉及诸多影响因素的实际问题中，此模型的拟合优度是可以被我们所接受的，相关参数估计及检验结果如下表：

表21 各参数估计检验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 因素名称 | 参数估计 | t-统计量 | p-值 |
| 截距项 | -125700 | -2.891 | <0.05 |
| 上映年份 | 62.13 | 2.875 | <0.05 |
| 上映月份 | 11.05 | 1.518 | 0.130172 |
| 宣发公司G | -38.45 | -0.520 | 0.603790 |
| 宣发公司L | -151.0 | -2.147 | <0.05 |
| 宣发公司S | 64.96 | 0.932 | 0.352202 |
| 时长 | 9.908 | 5.014 | <0.05 |
| 作品是否改编 | -210.7 | -2.387 | <0.05 |
| 作品是否续集 | 232.8 | 3.547 | <0.05 |
| 导演得奖情况 | 138.1 | 2.651 | <0.05 |
| 导演是否转型 | 329.9 | 4.695 | <0.05 |

从表中可以看出结论与对各个因素同放映场数之间做了方差分析的结果类似，可以得到：（1）上映月份作为解释性变量并不显著。（2）上映年份、作品是否续集、导演得奖情况、导演是否转型的系数均为正，并且较大，说明上映年份越迟，观影人数越多，随着电影产业的发展，接受看电影这一娱乐方式的人越来越多，续集的作品观影人数更多，同样说明续集的作品在中国电影市场占据重要的位置，同样得奖的导演、转型的导演的作品观影人数更多。（3）与宣发公司C相比，L与G处于劣势地位，而S则优于C。（4）作品是否改编的系数为负，且较大，改编作品的观影人数并不理想。（5）时长的系数为正，但比较小，也能够说明时长越长，电影的观影人数越多。

## 5.2 逻辑回归模型

关于票房，在描述性分析的基础上，通过方差分析对各个因素同票房之间的关系做了逻辑回归模型。我们将数据分为测试集和训练集，随机筛选数据的80%为训练集，剩余20%为测试集，对训练集建立模型的参数估计如下，下表只显示出了显著的四个解释变量，即影片类别为悬疑的、时长、翻拍的作品、以及导演得奖情况，且极大似然估计量均为正，说明时长越长、导演的奖数量越多的作品过亿的可能性越大，且为悬疑类型的影片以及翻拍的作品过亿的可能性更大。

表22 卡方检验结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 参数估计 | Z-统计量 | p-值 |
| as.factor(影片类别)悬疑 | 3.321 | 2.541 | <0.05 |
| 时长 | 0.006302 | 3.278 | <0.05 |
| as.factor(作品是否翻拍)1 | 1.637 | 2.350 | <0.05 |
| 导演得奖情况 | 1.820 | 3.079 | <0.05 |

以上就是模型分析部分的全部内容。

## 5.3 逻辑回归模型评估

基于上述逻辑回归模型，我们将模型应用于测试集，对测试集数据进行预测，预测结果如下：

表23 逻辑回归预测结果

|  |  |  |  |
| --- | --- | --- | --- |
| 预测值 | 0 | 1 | 合计 |
| 0 | 31 | 4 | 35 |
| 1 | 4 | 14 | 18 |
| 合计 | 35 | 18 | 53 |

我们计算了错判率为15%，预测效果较好。

对于预测结果，我们绘制了ROC曲线。

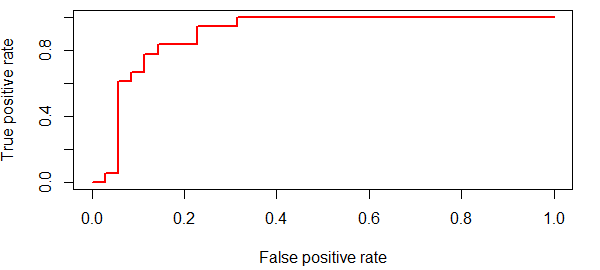


图6 ROC曲线

ROC曲线上升较快，说明模型具有较好的效果。

# 总结讨论

本研究通过对2010-2013年中国电影数据的分析，针对影响票房的各个重要因素做了方差分析，发现影响影片票房的重要因素，即上映年份、上映月份、宣发公司、时长、作品是否改编、作品是否续集、导演得奖情况、导演是否转型等因素。同时我们建立逻辑回归模型预测影片票房能否过亿，通过聚类分析将影片分为高票房组、低票房组、以及票房处于两者之间的组，然后进行深入了解每个组内的影片特征。

但是本研究同样存在不足，因为影响影片票房的因素远不止这些，并且，在聚类分析中，因为某些样本较少，所以并不能得出确定的结论。在接下来的研究中，我们应该进一步扩充数据，考虑更多的因素。

**参考文献：**

[1] 王汉生. 商务数据分析与应用. 中国人民大学出版社，2011.75-136.

[2] 黄文，王正林，等. 数据挖掘：R语言实战. 电子工业出版社，2014.121-150.

[3] 朱旭辉，荆丽娜，逯杉楠. 从电影票房看中国电影产业的现状与发展趋势.2014.22-25.

[4] 蓝乐. 中国电影文化产业的现状及发展. 2014.3.21.1-6.

[5] 黄冬虹. 中国电影产业发展现状研究. 人民网，2014.11.26.

**致谢：**

本论文是在我的导师徐文昕老师的指导下完成的，徐老师不仅给予了学术性的意见，而且耐心指导我修改论文，在此向老师表示衷心的感谢，因为他全程指点和帮助，才有了这篇完整的论文。同时，我也想感谢在我大学四年间所有曾教授过我知识的老师，他们教会我的都是我未来人生中的财富。在论文写作的过程中，我得到了许多人的帮助与支持，包括我的同学以及朋友们，在此，同样向他们表达我最诚挚的谢意！