山东理工大学
SHANDONGUNIVERSITY OF TECHNOLOGY

毕 业 论 文

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

学生姓名：赵齐贤

学　　号：14121203044

指导教师：郭晶

2018 年 6 月

摘要

中国电影市场正处于快速发展的阶段，2012 年中国超越日本，成为仅次于美国的全球第二大电影市场。截止 2017 年 12 月 20 日，中国内地银幕数超过北美（美国和加拿大），成为全球银幕数最多的国家。虽然中国电影总票房逐年递增，2017 年，中国电影总票房高达 559.11 亿，相比 2016 年增长了 13.45%，但在这繁荣数字背后却是超过 80%国产影片的亏损现象。基于电影具有高投入高风险的特性,研究电影票房的影响因素及其影响程度是确保电影盈利、控制风险的必要手段,对于当下国产电影的投资与决策具有重要参考价值。目前我国对电影票房影响因素分析多侧重于定性分析，实证研究和样本数据较少，得出的结论带有主观色彩。

本文在吸取国内外专家学者研究成果的基础上,根据中国电影市场的实际情况，通过定性和定量分析选取和定义了 8 个类别 23 个指标构建了电影票房影响因素的指标体系，这其中包含了技术效果(3D/IMAX)、发行公司影响力、主要发行国家等以往国内研究中没有或较少考虑的指标,全面反映电影的商业性质及票房号召力。

本文通过 Python 软件对 CBO 中国票房数据进行抓取、清洗、分析、挖掘、建模及预测，以 2008-2017 年在国内公开上映的 240 部电影数据为依据，运用多元线性回归的方法建立模型，对电影票房影响因素进行了实证分析，研究我国电影票房的影响因素及其影响程度。实证结果表明，最终模型的拟合程优度接近 80%。多元线性回归模型估计出了有显著影响的自变量的影响系数，其数值和正负符号与预期基本相符，可以很好的解释因变量和自变量间的关系，对于现实生活中的电影票房预测问题具有一定的参考价值。

本文根据回归结果得出了盗版和剧情类电影对于电影票房有负面影响,而银幕数、导演影响力、网络口碑(关注人数)、演员影响力、发行公司、故事熟悉程度(改编)和技术效果(IMAX)对于电影票房有促进作用的结论,并根据这些结论针对我国电影行业提出了合理的建议，为电影投资决策提供参考价值。

关键词：电影票房；影响因素；多元回归；Python

Abstract

Chinese movie market has been in rapid development. In 2012, China surpassed Japan and became the second largest movie market in the world, just after the United States. And in December 20, 2017, the number of screens in China surpassed North America Area (US and Canada), becoming the country with the largest number of screens in the world. The total box office in China has increased year by year, however, behind this prosperity figure is the losses of more than 80% of domestic movies. Based on the features of high investment and high risk, the research on the influencing factors and impact degree of movie box office is a necessary means to ensure movie profitability and control risk, which has important reference value for the investment and decision-making of current domestic movies.

At present, the analysis of the influence factors on movie box office in China focuses on qualitative analysis. There are few empirical studies and sample data, and the conclusions obtained are subjective. Based on the research results of experts and scholars at home and abroad, this article selects and defines 23 invitations in 8 categories according to the actual conditions of films shown in Chinese cinemas, and constructs an index system for the influence factors of movie box office. This includes Technical indicators (3D/IMAX), distribution company influence, major issue countries, and other indicators that were not considered or considered in previous domestic studies fully reflect the commercial nature of the film and the box office appeal.

This article uses Python software to crawl, cleanse, analyze, excavate, model and predict CBO's China box office data. Based on the 240 movie data released in China in 2008-2017, the model is built with multiple linear regression methods. An empirical analysis was made of the influencing factors at the box office of the movie to study the influencing factors and influence degree of the movie box office in China. The empirical results show that the goodness of fitting of the final selection model approaches 80%. The multivariate linear regression model estimates the coefficient of influence of the independent variables that have a significant effect. The numerical value and sign of the positive and negative signs are basically consistent with the expectations. It can explain the relationship between the dependent variable and the

independent variable well and predict the movie box office in real life. The problem has a certain reference value.

Based on the results of the stepwise regression, the paper concludes that the IMAX and 3D movies have improved the audience's visual experience and at the same time promoted movie box office. Comedy and horror movies have a positive impact on the box office, and the United States and South Korea have won high box office as the main release country's movies. The probability of the increase in the conclusions, and based on these conclusions for our country's film industry made reasonable recommendations for the film investment decision-making to provide a reference value.

Keywords: Movie Box Office; Influencing Factors; Multiple Regression; Python

目 录

第一章 绪 论

1.1 研究背景和问题提出

1.1.1 研究背景

近十年来，在中国经济迅速发展的大背景下，除去 2016 年电影产业因产业整合、升级而出现的短暂低迷，整个电影产业出现了高速增长的局面。电影作为人们日常生活中重要的休闲娱乐、精神享受的重要文化途径，电影产业的稳定发展对更好推动人的全面发展、社会全面进步具有重要意义。

在图 1-1 至图 1-4 中，从电影总票房、银幕块数、观影人次和电影院数量的角度，可以看出 2008-2017 年我国电影行业总体呈现上升发展的趋势。值得注意的是，尽管中国电影总票房、银幕块数、观影人次和电影院数量逐年递增，观影人次增长率自 2014-2017 年却是逐年递减，2017 年的增长率仅为 0.95%。从观影人次的角度，中国电影市场有接近饱和的趋势。但是从银幕块数和电影院数量的角度，中国电影市场仍处于高速发展的阶段，未来电影院和银幕数量会继续增加，这与近年来实际观影人次所反映出的电影市场饱和相矛盾，从整体结构和配置上出现了差异化的空间，有资源配置浪费的倾向。

图 1-1 2008-2017 年中国电影总票房

Figure 1-1 Revenues of movie tickets in China from 2008-2017

数据来源：中国统计年鉴（2008-2017）和国家新闻出版广电总局电影局

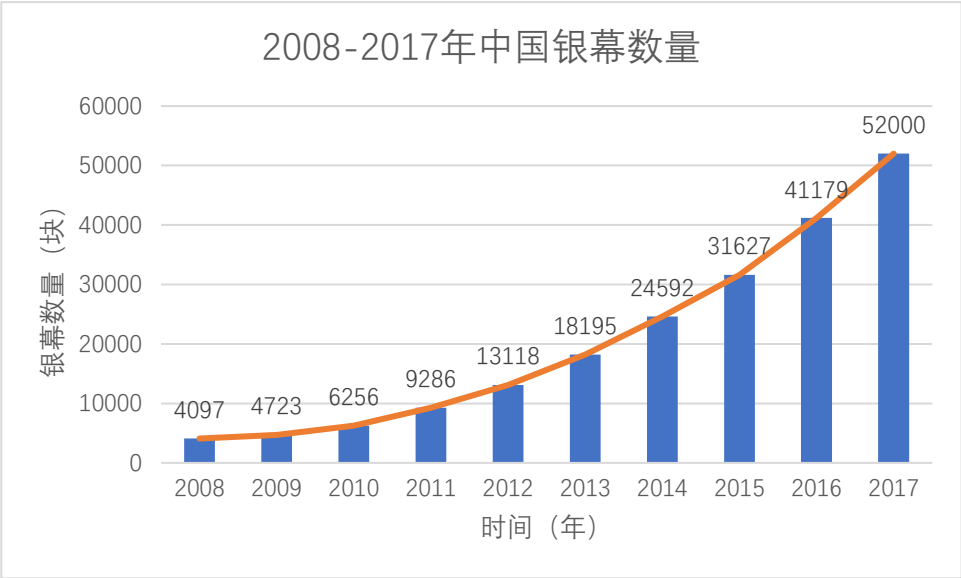

图 1-2 2008-2017 年中国银幕数量

Figure 1-2 Number of screens in China from 2008-2017

数据来源：公开资料整理

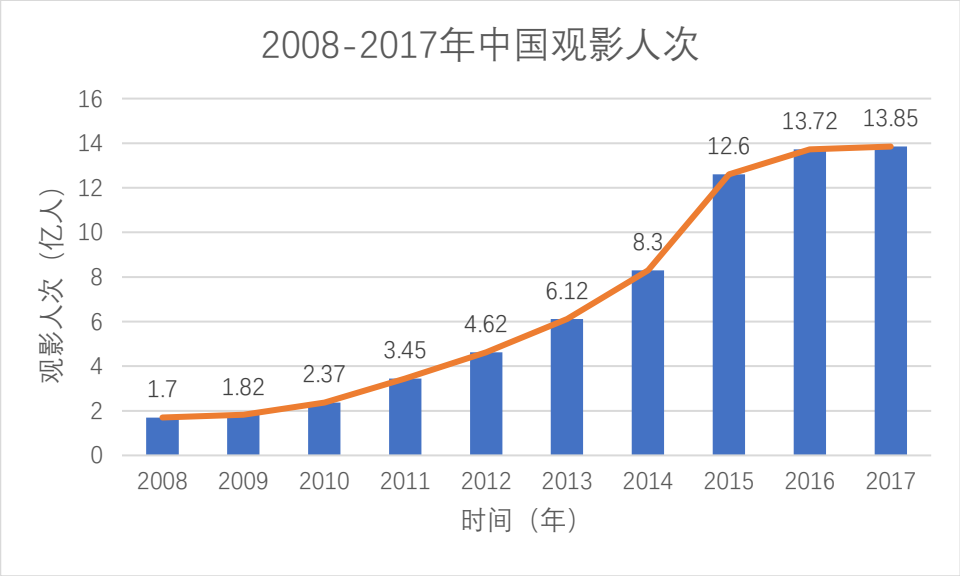

图 1-3 2008-2017 年中国电影观影人次

Figure 1-3 Person-time in China from 2008-2017

数据来源：公开资料整理

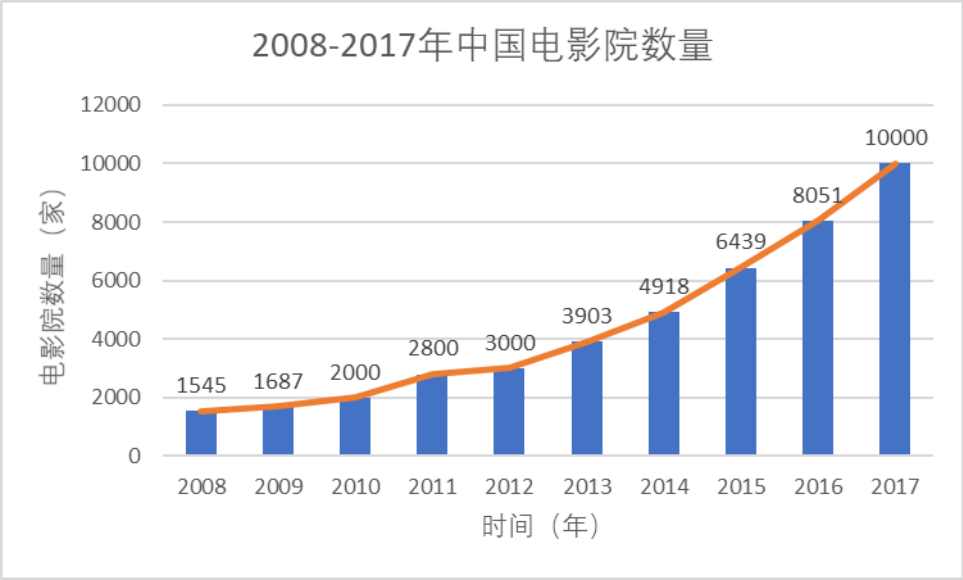

图 1-4 2008-2017 年中国电影院数量

Figure 1-4 Number of cinemas in China from 2008-2017

数据来源：公开资料整理

1.1.2 问题提出

我国电影产业目前仍处于飞速发展的阶段，但是 2015-2016 年的票房增长的速度仅为 3.7%，远低于它 2008-2017 年间的 30%-40%，甚至接近 50%的年平均增长率，创下十年最低增速。除此之外，再考虑中国趋于饱和的观影人次，中国电影市场能否继续高速发展，或者能否像中国经济由飞速发展转变为平稳发展，值得关注。

此外，通过比较研究中国和美国电影产业发展历程，可以发现一些相似之处，中国电影产业可以从美国电影产业发展历程中学习经验。美国电影产业经过1893-1926的无声电影和早期有声电影

之后，1929-1946年为高速发展阶段，但是1947-1970年陷入衰落，其中除去政治因素外，电影市场本身关于反垄断的运动和之后的派拉蒙判决亦是主要原因。而造成2016年的十年最低增速的主要原因，亦是因为中国电影产业暴露出的一些问题，比如2016年以《叶问3》为首的一系列电影的虚假票房事件。此外还有以万达院线、光线传媒、华谊兄弟等中国影视公司巨头由于自身运营、资本出现的问题，市值跌落，为中国电影产业蒙上了一层阴影。从图1-5中美电影总票房比较可见，中国电影产业如何能从高速发展顺利转为稳定发展，类似于当下美国电影产业的发展，是当下中国电影产业值得探讨研究的重要内容。

图1-4 2008-2017年中美电影总票房比较

Figure 1-1 Revenues of movie tickets comparison between China and the USA

数据来源：中国统计年鉴 和 Box Office Mojo

观影人次趋于饱和的信号，对电影的票房号召力提出了更高的要求 。如何提高电影的票房号召力，对稳定当下的电影市场以及电影产业的平稳增长具有重要意义。电影的票房收入是电影的票房号召力的重要体现，哪些因素影响电影的票房收入，影响程度正是本文所要研究的问题。

1.2 研究意义和目的

1.2.1 研究意义

电影的票房收入是电影的票房号召力的重要体现，本文的研究正是基于中国电影市场的实际情况，分析并量化中国电影票房的影响因素及其影响程度，最后进行实证分析给出合理的结论和建议，为中国电影投资决策提供参考意见，本文将从提高电影的票房号召力的角度，促使中国电影产业继续高速发展或者由高速

发展顺利转变为稳定发展，从而避免电影产业步入萧条期或者衰落期，以及由此而造成从事电影及其相关行业的工作人员的失业甚至文化产业的衰退。

1.2.2 研究目的

本文的研究目的主要有三个：

（1）立足于中国电影市场的实际情况，从理论上分析中国电影票房的影响因素，进而构建中国电影票房影响因素的指标体系。

（2）搜集数据，运用统计方法和 python 软件对数据进行量化处理，尝试建立票房影响因素的模型，实证分析中国电影票房的影响因素及其影响程度。

（3）根据研究结果，从高票房号召力的角度，对电影投资提出合理建议。

1.3 研究现状

经过查阅各种文献资料发现，国内外的专家学者们很早就对电影票房的影响因素进行过相关研究。从上世纪八十年代起，西方便开始有学者研究影响电影票房收入的相关因素以及它们对票房收入的影响程度。国内的研究起步较晚，但是也得出了许多电影票房收入的相关因素。

1.3.1 国外研究现状

当代西方电影票房研究开始于 20 世纪 80 年代，中国研究都是以北美地区（美国、加拿大）发行的电影为例，研究电影的票房及其导致其成功的因素进行实证分析，建立票房影响因素模型，并进行预测，观察模型预测值于实际值的差距。

美国电影经济学家巴瑞·李特曼[1]于 1989 年发表的论文《电影成功预测:基于八十年代人的经验》奠定了电影票房研究的基本方法和模型。巴瑞将电影票房的影响因素分为创意、发行及上映、电影营销三个层次，在每个层次中分别选择了一些变量建立多元线性回归模型。

上世纪 90 年代，斯格特·苏凯[2]在研究票房的影响因素时，考虑了更多的变量，比如电影放映期间的上映总天数。之所以将这一变量用来衡量电影票房是否获得成功，是因为上映总天数课以反映影片的生存周期，对电影票房进行动态研究。

近几年最为轰动的电影票房预测模型及影响因素分析来自于谷歌公司。众所周知，谷歌是全球最大的搜索引擎公司，依托于互联网搜索、云计算、广告技术等业务，建立的票房预测模型与实际票房非常接近，达到94%[3]。相比较其他模型，谷歌公司将电影及主演的点击率、搜索量、电影的片花播放量等其他通过谷歌公司所独有资源所获取的变量考虑进模型中，获得了相当高的准确度。

1.3.2 国内研究现状

国内电影票房的相关研究整体来说还是有很大的提升空间，相关的研究主要还是停留在案例分析和定性分析上，定量分析研究较少。主要原因是中国电影相较于美国等其他国家起步较晚，电影市场不够规范，相关部门的监管力度较小，可获得的数据较少或者同类数据差异较大。

基于以上前提，国内的电影票房影响因素研究多考虑几个容易获得影响因素。而由于数据本身的精确性、不稳定性，不同的作者针对相同的影响因素，有时可能会得出完全不同的结论。

史伟、王洪伟和何绍义[4]从新浪微博入手，挖掘微博中的情感信息，得出网络口碑，并以此分析电影票房收入，并做出预测。罗敏[5]则是考虑中国电影市场受同档期国外影片尤其是好莱坞商业大片的影响，将国外影片对票房的影响考虑进来。

聂鸿迪[6]从相关的中国电影网站上尽可能多的收集相关数据，综合考虑多个影响因素。最终选取10类22个变量，包括电影的类型，档期是否包含假期、豆瓣评分、想看指数等多个不同方面的影响因素。利用多元回归模型解释各个影响因素对票房的影响。

1.4 研究方法

本文在国内外学者对电影票房影响因素研究的基础上，使用多元线性回归模型，并选取逐步回归算法进行预测。多元线性回归模型对于变量的解释性较较好，因此可以较好地解释电影票房的影响因素及其影响程度。

1.5 论文结构

本文的主要研究内容有四个部分：

（1） 针对 2008-2017 年中国电影票房的相关数据，确定票房的影响因素，并通过爬虫技术和 Python 软件实现电影特征的相关数据的获取。

（2） 对原始数据进行预处理操作以及提取相关特征。

（3） 根据属性特征，分别建立基于多元回归分析和神经网络的电影票房预测模型。

（4） 对不同的模型建立预测分析，并比较结果。

本文主体共有六个章节，论文结构如下：

第一章，阐述电影票房的影响因素及票房预测的研究现状、选题背景和研究意义，并阐述研究目的及章节安排。

第二章，分析中国电影票房的影响因素有哪些，同时指明本文电影票房数据地获取方法是利用 Python 软件中爬虫技术的相关软件包，从相关网站上获取。在分析影响因素的同时，对各变量进行数据预处理并进行解释说明。

第三章，阐述本文将会用到的多元回归模型的理论基础，并建立多元回归模型。在信息量准则和提高拟合优度的基础上，选出 AIC、BIC 和拟合优度最好的模型。在解释变量集最优的基础上，选出逐步回归模型。

第二章 电影票房分析及其影响因素

本文的因变量是电影票房（Box Office）。选取的影响因素，即自变量包括电影类型、电影档期、电影制式、演员、导演及发行公司。自变量既包括连续型变量，也包括离散型变量。

2.1 电影票房

电影票房（Box Office）。电影票房是一部电影在放映期间累积的门票收入，是衡量电影销售情况的重要指标。票房可以用放映场次，观影平均人数以及门票价格来计算。在当下的电影产业，票房已成为衡量一部商业电影是否成功的重要指标。

2.2 影响因素的选择

电影票房受许多因素的影响，因此需要综合考虑非常多的因素，然后才能对电影票房进行预测。但是由于电影票房的很多数据，比如实际投资、宣传投资、发行公司的实际盈利等数据并不公开，从而无法获得。只能对可获得的数据，利用爬虫等其他网站数据抓取技术，获取电影的相关数据，尝试探索不同影响因素对电影票房的影响程度。

本文在国内外学者对电影票房及其影响因素研究的基础上，选择了在电影网站上可以获取到的影响因素数据。本文根据中国实际的电影市场，选取了八类变量，分别是电影的平均票价、平均观影人数、电影上映首日是否为假期、电影类型（剧情、动作、喜剧、惊悚、科幻、动画、战争、灾难、爱情、奇幻）、主演影响力、导演影响力、发行公司影响力、电影制式（3D，IMAX）、主要发行国家（中国、美国、韩国、日本）。

2.3 数据获取及来源

本文的数据全部来源于 CBO 中国票房数据（http://www.cbooo.cn/）。通过 Python 软件中 Beautiful Soup 软件包，对 2008-2017 年的年度电影票房数据页面进行抓取，以 2008 年电影《赤壁(上)》数据为例，具体过程如下：

年度票房　年度首周票房　　标红为最新上榜影片，最新数据仍在统计中

2008年

排名：影片名	类型	总票房（万）	平均票价	场均人次	国家及地区	上映日期
1.赤壁(上)	战争	27490	33	41	中国/中国香港/韩国	2008-07-10
2.画皮	爱情	20453	30	41	中国/中国香港/新加坡	2008-09-26
3.非诚勿扰	爱情	17641	34	62	中国/中国香港	2008-12-18
4.功夫熊猫	动画	15150	27	36	美国	2008-06-20

```
<img alt="赤壁(上)" onerror="this.src='../../Content/images/nopic.jpg'" src="http://www.cbooo.cn/moviepic/24709.jpg"> == $0
▶<p>…</p>
</a>
</td>
<td>战争</td>
<td>27490</td>
<td>33</td>
<td>41</td>
<td>中国/中国香港/韩国</td>
<td>2008-07-10</td>
```

图 2-1 数据抓取示例 1

	Year	Top	Title	Boxoffice	Genre	Price	Person	Country	Date	Actor1	...	Actor_1	Actor_2	Dire	com	3D	IMAX	中国	美国	韩国	日本
0	2008	1	赤壁(上)	27490	战争	33	41	中国/中国香港/韩国	2008-07-10	金城武	...	0.625000	0.625000	0.833333	0.416667	0	0	1	0	1	0
1	2008	2	画皮	20453	爱情	30	41	中国/中国香港/新加坡	2008-09-26	陈坤	...	1.041667	1.041667	1.250000	44.166667	0	0	1	0	0	0
2	2008	3	非诚勿扰	17641	爱情	34	62	中国/中国香港	2008-12-18	葛优	...	1.250000	1.250000	3.333333	5.833333	0	0	1	0	0	0
3	2008	4	功夫熊猫	15150	动画	27	36	美国	2008-06-20	杰克·布莱克	...	0.416667	0.208333	0.416667	0.833333	0	0	0	1	0	0
4	2008	6	功夫之王	14560	动作	32	31	美国/中国	2008-04-24	成龙	...	1.666667	0.833333	0.416667	0.833333	0	0	1	1	0	0

图 2-2 数据截图

2.4 样本选取

本文选取的样本为2008-2017年在中国公开上映每年票房排名前24位的影片，共240部电影。这里的电影既包含主要发行国家为中国的电影，也包含美国、英国、韩国、日本等其它国家发行的电影。电影的类型包括剧情、动作、喜剧、惊悚、科幻、动画、战争、灾难、爱情和奇幻。

选择这些样本，主要基于以下几点原因：

（1）如前文所述，自1978年改革开放，中国电影产业迅速发展。2008年作为改革开放30年的时间结点，电影数量和票房质量较前期相比都有质的飞跃，因此选取2008-2017年这一时间段的电影具有很强的时效性。

（2）鉴于电影数据获取难度大，部分排名靠后电影的相关数据缺失，将每年上映的全部电影作为样本研究可操作性差，且并不现实，而每年排名在前电影数据相对容易获得且信息相对完整，关注度高且信息相对准确。

（3）由于电影票房具有偏态性，每年排名靠前的电影的票房之和占全年总票房的绝大多数，比如 2008 年《赤壁（上）》、《画皮》、《非诚勿扰》等排名前 8 的电影占当年总票房的 28%，前 24 部电影占 55%，2018 年仅《战狼 2》一部电影占当年总票房的 10%，前 24 部电影占 57.7%。2008-2017 十年间，每年票房前 24 名电影的总票房都至少占全年总票房一半以上,具有很强的票房影响力和号召力，能够体现出当年电影市场的特点。

（4）挖掘国产电影票房竞争力是本文的主要研究目的之一，选取中国以外的其它主要发行国家的电影，是希望立足于优秀的外国电影，国产电影能借鉴其优秀特质，提高自身竞争力。

（5）选取的电影类型有十类，基本包括所有的电影类型，便于分析和总结电影类型对电影票房的影响力。

2.5 数据预处理

根据中国实际的电影市场，选取了八类变量，分别是电影的平均票价、平均观影人数、电影上映首日是否为假期、电影类型（剧情、动作、喜剧、惊悚、科幻、动画、战争、灾难、爱情、奇幻）、主演影响力、导演影响力、发行公司影响力、电影制式（3D，IMAX）、主要发行国家（中国、美国、韩国、日本）。除平均票价、平均观影人数、演影响力、导演影响力、发行公司影响力是连续型变量，其它均为离散型二分类变量，是为 1，否为 0。比如，电影首日上映为国家法定节假日，则该变量的值为 1，电影首日上映不是国家法定节假日，则该变量的值为 0。

本文获取的电影票房的单位是万元。为减少不同数据间的差距过大，消除回归模型中的异方差现象，将原始票房数据取对数。对电影的平均票价、平均观影人数同样采取对数处理。

电影类型分为十类，分别是剧情、动作、喜剧、惊悚、科幻、动画、战争、灾难、爱情、奇幻。在本文中，将电影类型设置为虚拟变量，将对应该类型电影的值设定为 1，否则为 0。一部电影可以对应多个类型，比如灾难战争片。

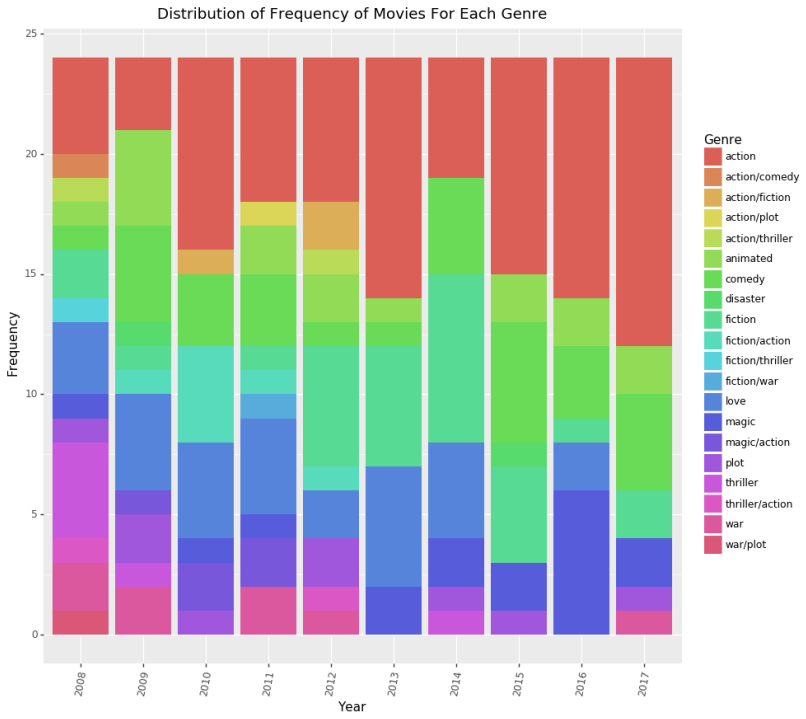

图 2-3 2008-2017 年选取电影的类型分布

主演影响力、导演影响力、发行公司影响力所对应的原始变量都是文本，不能直接用来建模，因此分别转换为影响力的 0 到 100 的数值来衡量。影响力的数值为 2008-2017 年电影数据中出现的频数占总次数的百分比。比如说，计算导演冯小刚的影响力，在 2008-2017 年每年前 24 部电影中，他指导的电影一共有 9 部，那它的影响力为 9/240*100 = 3.333333。

	Count	Dir	Percent
0	8	冯小刚 Xiaogang Feng	3.333333
1	5	李仁港 Daniel Lee	2.083333
2	5	徐克 Hark Tsui	2.083333
3	4	叶伟信 Wilson Yip	1.666667
4	4	张艺谋 Yimou Zhang	1.666667
5	4	王晶 Jing Wong	1.666667
6	4	迈克尔·贝 Michael Bay	1.666667
7	3	麦兆辉 Alan Mak	1.250000
8	3	张一白 Yibai Zhang	1.250000
9	3	乔恩·费儒 Jon Favreau	1.250000
10	3	彼得·杰克逊 Peter Jackson	1.250000
11	3	大卫·叶茨 David Yates	1.250000
12	3	克里斯托弗·诺兰 Christopher Nolan	1.250000
13	3	扎克·施奈德 Zack Snyder	1.250000
14	3	郭敬明 Jingming Guo	1.250000

2-4 2008-2017 年选取电影的导演影响力（前 15）

电影上映的时间，也是电影能否在票房上取得成功的重要因素之一。本文并没有对贺岁档、五一档、国庆档进行分类，而是统一处理所有国家法定节假日。如果电影是在国家法定节假日内上映，则是否假期上映这一变量为 1，否则为 0。

电影的主要发行国家在一定程度上也会影响我国上映的电影票房。众所周知，最近几年中，部分好莱坞商业大片在北美地区遭遇滑铁卢，却在中国市场获得惊人票房，因此本文将美国考虑为主要发行国家的变量之一。中国电影今年在本土的票房影响力越来越大，因此本文同样将中国考虑为主要发行国家之一。日韩文化及日韩明星在近年来一直在中国有一定的影响力，因此本文同样将日本、韩国考虑为主要发行国家之一。将电影发行国家设置为虚拟变量，本文共考虑中国、美国、韩国、日本四个主要的电影发行国家。将对应电影发行国家的值设定为 1，否则为 0。一部电影可以对应多个不同的发行国家，比如《功夫熊猫 3》的主要发行国家包括美国、中国。

近些年，随着影片制作技术的不断提高与创新，3D 和 IMAX 技术的普及，为观众带来更好的视觉体验。本文主要考虑 3D 和 IMAX 两种电影制式，如果电影是 3D 或 IMAX 的，则其对应的 3D 和 IMAX 的变量值为 1，否则为 0。

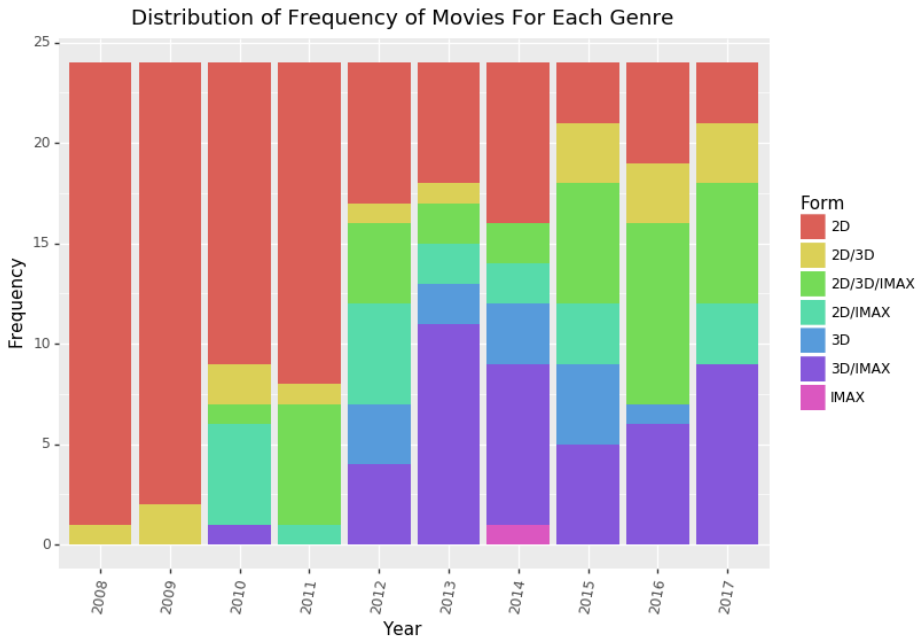

图 2-5 2008-2017 年选取电影的不同制式分布

Year	Top	Title	Boxoffice	Genre	Price	Person	Country	Date	Actor1	...	Actor_1	Actor_2	Dire	Company	3D	IMAX	China	America	Korea	Japan
2008	1	赤壁(上)	27490	war	33	41	中国/中国香港/韩国	2008-07-10	金城武	...	0.625000	0.625000	0.833333	0.416667	0	0	1	0	1	0
2008	2	画皮	20453	love	30	41	中国/中国香港/新加坡	2008-09-26	陈坤	...	1.041667	1.041667	1.250000	44.166667	0	0	1	0	0	0
2008	3	非诚勿扰	17641	love	34	62	中国/中国香港	2008-12-18	葛优	...	1.250000	1.250000	3.333333	5.833333	0	0	1	0	0	0
2008	4	功夫熊猫	15150	animated	27	36	美国	2008-06-20	杰克·布莱克	...	0.416667	0.208333	0.416667	0.833333	0	0	0	1	0	0
2008	6	功夫之王	14560	action	32	31	美国/中国	2008-04-24	成龙	...	1.666667	0.833333	0.416667	0.833333	0	0	1	1	0	0
2008	7	007：大破量子	12046	action/thriller	31	29	英国/美国	2008-11-05	丹尼尔·克雷格	...	0.625000	0.208333	0.416667	0.416667	0	0	0	1	0	0

图 2-6 数据预处理后截图

第三章 基于回归的电影票房预测模型

3.1 多元线性回归模型的基本理论

多元线性回归模型[7]指的是有多个解释变量的线性回归模型。一个变量往往受多个变量的影响，因此多元线性回归模型用于表示被解释变量与多个解释变量之间的线性关系。多元线性回归的数学模型为

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{1-1}$$

ε 为随机误差，满足：

$$E(\varepsilon) = 0, \mathrm{var}(\varepsilon) = \delta^2 \tag{1-2}$$

利用最小二乘法估计每一个回归系数。最终建立多元线性回归模型。

对式(1-1)两边求期望，则有

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1-3}$$

式(1-3)就是多元线性回归方程，我们需要通过样本数据来估算方程中的未知参数 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$，于是有

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{1-4}$$

式(1-4)为多元线性经验回归方程，$\hat{\beta}_i$ 表示当其他解释变量不变时，x_i 每变动一个单位所引起的被解释变量 y 的平均变动数量。

(1)模型确定之后，并不能马上用于预测使用，模型必须要通过一些模型检验，确定模型的正确性。为判断自变量是否从整体上对因变量 y 是否有显著性影响，对回归方程进行显著性检验，构造 F 统计量，如式 (1-5)

$$\mathrm{F} = \frac{\mathrm{SSR} / \mathrm{p}}{\mathrm{SSE} / \mathrm{n-p-1}} \tag{1-5}$$

其中 SSR 为回归平方和，SSE 为误差平方和，p 为自变量的个数，n 为样本的数量。计算相对应的 P-值，若 P-值<显著性水平α，则可以拒绝回归方程不具有显著性的原假设，认为该回归方程具有显著性。

(2)在检验了回归方程整体的显著性后，再对每个解释变量的回归系数进行显著性检验，单独检验每个解释变量是否对被解释变量有显著性的影响。如果没有显著性影响，则该多元回归方程是没有意义的，需要将其剔除。通过 T 统计量，如式(1-6)

$$t = \widehat{\beta_j}/\sigma \quad (1\text{-}6)$$

计算相对应的 P-值，若 P-值<显著性水平α，则可以拒绝回归系数不具有显著性的原假设，认为该回归系数具有显著性。

(3)变量选择是模型中很重要的一环，如果某些预测变量（因变量）与响应变量（自变量）不相关，可能会造成模型过拟合。利用逐步回归，针对每一次建立的模型，通过赤池信息量准则 AIC，贝叶斯信息量准则 BIC 和调整R^2等统计量指标来判断模型的质量。根据赤池信息量准则 AIC 和贝叶斯信息量准则 BIC 越小模型越准确、调整R^2越大模型能够用自变量解释的部分越大的准则，选出最优的模型。

3.2 多元线性回归研究过程

3.2.1 模型构建

在上一节多元回归模型的理论基础上，结合本文对中国电影票房影响因素研究的具体设定，建立以下回归模型：

$$Y = \beta_0 + \beta_1 * 平均票价 + \beta_2 * 平均观影人数 + \beta_3 * 是否假期上映 + \beta_4 * 剧情 + \beta_5 * 动作 + \beta_6 * 喜剧 + \beta_7 * 惊悚 + \beta_8 * 科幻 + \beta_9 * 动画 + \beta_{10} * 战争 + \beta_{11} * 灾难 + \beta_{12} * 爱情 + \beta_{13} * 奇幻 + \beta_{14} * 主演\ 1\ 影响力 + \beta_{15} * 主演\ 2\ 影响力 + \beta_{16} * 导演影响力 + \beta_{17} * 发行公司影响力 + \beta_{18} * 3D + \beta_{19} * IMAX + \beta_{20} * 中国 + \beta_{21} * 美国 + \beta_{22} * 韩国 + \beta_{23} * 日本 + \varepsilon$$

其中，Y 表示电影票房，β_0，$\beta_1 \ldots \beta_{23}$是待估计的回归系数，ε是随机误差项。

表 3-1 变量符号及含意

序号	变量符号	变量含义	序号	变量符号	变量含义
1	X_1	平均票价	13	X_{13}	奇幻
2	X_2	平均观影人数	14	X_{14}	主演 1 影响力
3	X_3	假期上映	15	X_{15}	主演 2 影响力
4	X_4	剧情	16	X_{16}	导演影响力
5	X_5	动作	17	X_{17}	发行公司影响力
6	X_6	喜剧	18	X_{18}	3D

7	X_7	惊悚	19	X_{19}	IMAX
8	X_8	科幻	20	X_{20}	中国发行
9	X_9	动画	21	X_{21}	美国发行
10	X_{10}	战争	22	X_{22}	韩国发行
11	X_{11}	灾难	23	X_{23}	日本发行
12	X_{12}	爱情			

上表中，除了X_1，X_2，X_{14}，X_{15}，X_{16}，X_{17}是连续型变量，其他变量均为离散型二分类变量。X_3表示电影首日上映是否是在国家法定节假日期间，$X_4 - X_{13}$表示电影主要类型的哑变量，X_{18}和X_{19}表示电影放映技术的哑变量，$X_{20} - X_{23}$表示电影主要发行国家的哑变量。

基于上述对模型的假定，在 Python 中，利用最小二乘法进行回归，得到了各个回归系数及对回归方程和回归系数作了显著性检验，其中回归方程的判定系数R^2等于 0.938，调整后的R^2等于 0.934，结果如表 3-2 所示：

表 3-2 多元回归结果 1

OLS Regression Results

Dep. Variable:	票房	R-squared:	0.996
Model:	OLS	Adj. R-squared:	0.995
Method:	Least Squares	F-statistic:	2178.
Date:	Tue, 19 Jun 2018	Prob (F-statistic):	7.71e-243
Time:	18:43:48	Log-Likelihood:	-253.26
No. Observations:	240	AIC:	552.5
Df Residuals:	217	BIC:	632.6
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
平均票价	2.8672	0.203	14.109	0.000	2.467	3.268
平均观影人数	-0.0445	0.192	-0.231	0.817	-0.424	0.335
假期上映	0.2992	0.149	2.004	0.046	0.005	0.593
剧情	0.0371	0.298	0.125	0.901	-0.550	0.625
动作	0.0116	0.199	0.058	0.954	-0.380	0.403
喜剧	0.4907	0.257	1.907	0.058	-0.017	0.998
惊悚	-0.2975	0.285	-1.044	0.298	-0.859	0.264
科幻	-0.2898	0.225	-1.287	0.200	-0.734	0.154
动画	-0.0391	0.312	-0.125	0.900	-0.654	0.576
战争	-0.0620	0.344	-0.180	0.857	-0.739	0.615
灾难	0.9408	0.591	1.591	0.113	-0.224	2.106
爱情	0.1483	0.271	0.546	0.585	-0.387	0.683
奇幻	-0.2602	0.243	-1.073	0.285	-0.738	0.218
主演1影响力	-0.1794	0.114	-1.572	0.117	-0.404	0.046
主演2影响力	-0.1151	0.129	-0.891	0.374	-0.370	0.140
导演影响力	-0.0632	0.087	-0.728	0.468	-0.234	0.108
发行公司影响力	0.0028	0.003	0.980	0.328	-0.003	0.008
3D	0.2525	0.145	1.746	0.082	-0.033	0.538
IMAX	0.7790	0.142	5.504	0.000	0.500	1.058
中国	0.3974	0.230	1.728	0.085	-0.056	0.851
美国	-0.0232	0.238	-0.098	0.922	-0.493	0.446
韩国	-0.6283	0.430	-1.461	0.145	-1.476	0.219
日本	0.9210	0.543	1.696	0.091	-0.150	1.991

Omnibus:	0.028	Durbin-Watson:	1.166
Prob(Omnibus):	0.986	Jarque-Bera (JB):	0.003
Skew:	0.003	Prob(JB):	0.998
Kurtosis:	2.983	Cond. No.	455.

通过上表中数据可以得到，在该模型中，虽然判定系数R^2等于 0.996，拟合程度很好，但是有很大一部分的自变量没有通过显著性检验。

3.2.2 模型调整

基于上述全模型的讨论，再次利用最小二乘法进行回归，得到了各个回归系数及对回归方程和回归系数作了显著性检验，其中回归方程的判定系数R^2等于 0.938，调整后的R^2等于 0.934，结果如表 3-3 所示：

表 3-3 多元回归结果 2

OLS Regression Results

Dep. Variable:	票房	R-squared:	0.938
Model:	OLS	Adj. R-squared:	0.934
Method:	Least Squares	F-statistic:	245.2
Date:	Tue, 19 Jun 2018	Prob (F-statistic):	5.45e-128
Time:	18:48:50	Log-Likelihood:	-572.67
No. Observations:	240	AIC:	1173.
Df Residuals:	226	BIC:	1222.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
IMAX	2.6041	0.492	5.288	0.000	1.634	3.575
3D	1.6632	0.463	3.590	0.000	0.750	2.576
喜剧	2.9989	0.583	5.148	0.000	1.851	4.147
惊悚	4.5772	0.841	5.444	0.000	2.920	6.234
灾难	5.8888	1.952	3.016	0.003	2.042	9.736
爱情	2.5553	0.626	4.080	0.000	1.321	3.789
发行公司影响力	0.0574	0.009	6.206	0.000	0.039	0.076
中国	2.8293	0.478	5.922	0.000	1.888	3.771
主演1影响力	2.0271	0.369	5.495	0.000	1.300	2.754
主演2影响力	1.1946	0.467	2.559	0.011	0.275	2.114
导演影响力	1.3694	0.287	4.773	0.000	0.804	1.935
科幻	3.1646	0.523	6.057	0.000	2.135	4.194
动画	5.3357	0.743	7.178	0.000	3.871	6.800
奇幻	1.8698	0.665	2.812	0.005	0.560	3.180

Omnibus:	0.168	Durbin-Watson:	1.717
Prob(Omnibus):	0.919	Jarque-Bera (JB):	0.102
Skew:	-0.050	Prob(JB):	0.950
Kurtosis:	3.008	Cond. No.	329.

在该模型中，虽然判定系数R^2等于 0.938，低于全模型的判定系数R^2，但是通过回归方程和所有系数的显著性检验。F 统计量的 P-值都接近于 0，说明所选择的回归模型作用显著。t 统计量的 P-值都接近于 0，表明回归系数显著。所有的自变量系数均大于 0，表明这些因素与票房收益正相关。比如 IMAX 的回归系数为 2.6041，表明电影制式为 IMAX 的电影获得高票房的概率大于电影制式不是 IMAX 的电影。Durbin-Waston 检验的值为 1.717，表明电影票房数据不存在序列相关性。该模型亦通过残差检验，残差服从均值为 0 且等方差的正态分布，残差序列独立。

从解释变量集最优且模型不受多重共线性干扰的角度，运用逐步回归的思想，每引入一个解释变量后都要进行 F 检验，并对已经选入的解释变量逐个进行 t 检验，当原来引入的解释变量由于后面解释变量的引入变得不再显著时，则将其删除。将选入和剔除变量的显著性α都控制为 0.05，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止。此时选入的电影票房影响因素包括 6 个变量，依次是是 IMAX,3D,美国，喜剧，韩国，惊悚。具体结果如表 3-3 所示：

表 3-4 多元回归结果 3

OLS Regression Results

Dep. Variable:	票房	**R-squared:**	0.730
Model:	OLS	**Adj. R-squared:**	0.723
Method:	Least Squares	**F-statistic:**	105.6
Date:	Wed, 20 Jun 2018	**Prob (F-statistic):**	9.93e-64
Time:	14:55:06	**Log-Likelihood:**	-749.55
No. Observations:	240	**AIC:**	1511.
Df Residuals:	234	**BIC:**	1532.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
IMAX	3.2756	0.931	3.518	0.001	1.441	5.110
3D	4.8948	0.849	5.765	0.000	3.222	6.568
美国	4.8345	0.785	6.157	0.000	3.288	6.381
喜剧	9.1697	1.021	8.979	0.000	7.158	11.182
韩国	9.2332	2.784	3.317	0.001	3.749	14.717
惊悚	5.3978	1.767	3.054	0.003	1.916	8.880

Omnibus:	82.448	**Durbin-Watson:**	1.336
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	20.150
Skew:	0.439	**Prob(JB):**	4.21e-05
Kurtosis:	1.884	**Cond. No.**	8.16

该模型亦通过残差检验，残差服从均值为 0 且等方差的正态分布，残差序列独立。

3.2.3 结果分析

通过上述建模，解释变量集最优模型为:$Y = 3.2756 * IMAX + 4.8948 * 3D + 4.8345 * 美国 + 9.1697 * 喜剧 + 9.2332 * 韩国 + 5.3978 * 惊悚$

从模型中可以看到，IMAX 和 3D 制式的电影在提升观众视觉体验的同时提升了电影票房，喜剧和惊悚类型的电影对票房产生正面影响，美国和韩国作为主要发行国家的电影取得高票房的概率增大。

结论及建议

本文利用 2008-2017 年中国电影票房及其相关数据行了建模研究，最后确立的逐步回归模型包括的主要影响因素为电影的制式是否为 3D 或者 IMAX，发行的国家是否为美国或韩国，电影的类型是否为喜剧片或惊悚片。本文在此基础上得到的结论是，制式为 3D 或者 IMAX，发行国家为美国或韩国，类型为喜剧片或惊悚片的电影获得高票房的可能性更大。最后确立的拟合度较高的选模型包括 IMAX、3D、喜剧、惊悚、发行公司影响力、主演影响力及其他 5 个变量，得到的结论除了包括逐步回归模型得到的结论，还包括灾难、惊悚、动画类型电影相较于其他类型的电影更可能收获高票房，相对于发行公司的影响力，主演和导演的影响力对电影票房的提高力度更大。

基于以上结论，我们从电影票房盈利的角度，对电影的出品方提出的建议如下：

1.注重电影作品的质量，提升拍摄技术，提高观众的视觉体验。

2.在导演和主要演员的选择上，可以多选取票房纪录较好，累计票房高的导演和主演阵容。

3.我国的发行公司可以积极地向美国、韩国的发行公司学习电影的营销手段，提高电影的热度，引起观众的注意。

参考文献

[1] Barry R.Litman and Linda S.Kohl, Predicting financial success of motion pictures: The ‘80s experience[J]. Journal of Medical Economics. 1989，2(02): 35-50.

[2] Sochay S. Predicting the performance of motion pictures[J]. Journal of Medical Economics, 1994,7(4): 1-20.

[3] Panaligan R, Chen A. Quantifying movie magic with google search[J]. Google Whitepaper-Industry Perspectives + User Insights, 2013.

[4] 史伟，王洪伟，何绍义. 基于微博情感分析的电影票房预测研究[J]. 华中师范大学学报（自然科学版）.2015,49 (1) :66-71.

[5] 罗捷. 基于电影评价的进口影片票房预测研究[D]. 重庆：重庆大学, 2015,4.

[6] 聂鸿迪. 中国电影票房的影响因素及其证实研究[D]. 北京：北京交通大学, 2015.

[7] R Muche. Logistic Regression: a useful in rehabilitation research [J]. Rehabilitation, 2008, 47(01): 56-62.

[8] 王雪娟.电影票房预测研究发展史简论[D].重庆：重庆大学, 2015.

[9] 任丹.基于多元线性回归模型的电影票房预测系统设计与实现[D].广东：中山大学，2015.

[10] Simonoff, J. S. and Sparrow, I. R. Predicting movie grosses: Winners and losers, blockbusters and sleepers. In Chance, 2000.

[11] Joshi, M., Das, D., Gimpel, K., and Smith, N. A. Movie Reviews and Revenues: An Experiment in Text Regression[C]. In Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference, 2010.

[12] Sharda, R. and Delen, D. Predicting box-office success of motion pictures with neural networks. In Expert Systems with Applications, 2006.

[13] 何双男. 中国大陆地区电影票房影响因素实证研究[J]. 电影文学, 2017，(22):4-8.

[14] 崔凝凝, 唐嘉庚. 基于回归分析的中国电影票房影响因素研究[J]. 江苏商论, 2012, (08):35-39.

[15] 张雪. 基于深度学习卷积神经网络的电影票房预测[D]. 首都经济贸易大学, 2017.

[16] 陈文俊. 2014-2015 内地电影市场网络口碑对影片票房的影响研究[D].广西大学,2016.

致　谢

衷心感谢我的指导老师郭晶老师。从论文选题到搜集资料，从完成初稿到反复修改，直到定稿，郭晶老师始终认真负责，不厌其烦地给予我深刻而细致的指导，帮助我开拓研究思路，精心点拨。

感谢统计系的全体老师和同学多年来的关心和支持！感谢所有关心和帮助过我的人们！