Stockholms universitet

Prediction of a film's profit and its IMDB score

Qi Lin

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:22
Matematisk statistik
September 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2018:22**
http://www.math.su.se

Prediction of a film's profit and its IMDB score

Qi Lin*

September 2018

Abstract

This thesis aims to find out which factors are associated with the profit of a film as well as its IMDB-score. We use 600 films which have been collected randomly from the website imdb.com. The films was released between 1990 and 2017. We can use Binary logistic regression to study if a film is profitable or not. The first binary logistic regression model is obtained by comparing stepwise procedure with the purposeful selection. The model contains three interactions and five main effects and has a acceptable predictive capacity. The second proportional odds model of IMDB score is obtained by purposeful selection which only includes six main effects. The results show that different predictors variables are correlated with a profitable film and it's IMDB score. These two fitted models have the acceptable predictive capacity with the data set.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: levxy@hotmail.com. Supervisor: Kristoffer Lindensjö and Felix Wahl.

Acknowledgment

This thesis is a Bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University. I take this opportunity to thank my two supervisors Kristofer Lindensjö and Felix Wahl for their support and advice. I am also grateful to my friends and my family, thank for the suggestions and unceasing encouragement.

Contents

1 Introduction

According to Statistics Portal[1], global box office revenue is expected to increase by approximately 12 billion U.S dollars from 2016 to 2020. The global film industry is flourishing. However, not all of the films make profits. Some movie generate massive profit while other films do not make a profit at all. One factor can be that the market provides various types of films and these films which are directed by different directors. Other factors are correlated with the performances of eminent performers, coupled with the production and distribution of films by different scales of these companies. The large portal site, IMDB updates constantly lots of movie information and people can grade all these movies on the website. Hence, it is interesting to find out which factors correlate with the profit of a film and to predict whether a film is profitable. Whether these factors will affect the audience's rating on the portal is also interesting to study. Furthermore, it can also be helpful to understand which factors have correlations with the IMDB score by fitting a model.

1.1 Aim

The aim of this thesis is to find out which factors, such as the director, film genre or runtime are associated with the profit of a film as well as the IMDB-scores. This thesis will try to answer the following questions:

- Which factors are related to the profit of a movie?
- Which factors are associated with a film's score on the website imdb.com?
- How to develop a model that can predict whether a film is profitable and how to build a model that can predict its IMDB score?

The thesis adopts different modeling selections and comparison methods to select an appropriate model to predict the probability of a film being profitable and the film's IMDB score.

1.2 Disposition of the paper

Section 1 presents the aim and the introduction. In section 2, each variable and data collection are outlined. The statistical methods that are used in the thesis are introduced in section 3. In section 4, we identify which variables are significant and apply different diagnostics to build a suitable model. Results of ultimate models and prediction are explained in section 5. Lastly, the discussion is presented in section 6.

2 Data

The dataset which was downloaded from https://www.imdb.com/interfaces/, contains 8 million different movies and television products, such as movie, tv-episode, short, tv series and video game, and it includes pieces of information about a film's title, genres, runtimeMinutes, and start year. 89677 movies were scraped by using R program. We selected randomly 600 films whose budget is higher than one million dollars and were released between 1990 and 2017. The others seven variables, such as film's budget, certificate, director, actress/actor, production company, issue country, and language, are handled from each film's page with the aid of R.

Table 1: Response variables and explanatory variables

Variables	Description
Response variable:	
Profit	Indicates if a film is profitable
IMDB score	The rating of a film at IMDB.com
Explanatory variable:	
Genre	The 600 films in the dataset have 21 different genres
BigBudget	Indicates if the budget of the film is bigger than a hundred million dollars
Certificate	Indicates if the certificate of a film is G, PG, PG.13 or R
RealeaseDate	Indicates if the release date is SummerHoliday, Workday, Otherday or Christmas
FamousDirector	Indicates if director win Academy awards/ Golden Globes/BAFTA Awards before film issues
Star	Indicates if actress/actor win the Academy awards/ Primetime Emmy Awards/AACTA before film issues
BigCo	Indicates if the company is one of the top 500 production companies
IssueCountry	Indicates if the issuing countries is U.S, English-speaking countries and non-English speakling countries
RuntimeMinutes	Indicates how much time the film play
Language	Indicates if the language of the film is English or multilingual

2.1 Response variables

There are two response variables in this thesis. One is a binary response variable that describes whether a film is profitable. The other one is an ordinal response variable which is to describe the IMDB-score of a film. The binary variable *Profit* is created by whether the income of a film is

higher than the cost. Among the 600 films in the dataset, 408 films (68 percent) are profitable.

Table 2: Summary of response variable **Profit**

Profit	Frequency	Percent
0	192	32%
1	408	68%

The other response variable *IMDB score*, audiences' rating of a movie, has four categories that are the extent of a rating scale: less than 5 points, between 5 and 6.5 points, between 6.5 and 8 points, and more than 8 points. Thus, the response variable *IMDB score* can be treated as ordinal.

Table 3: summary of response variable **IMDB score**

IMDB score	Frequency	Percent
≤ 5	56	9.33%
$(5, 6.5]$	237	39.50%
$(6.5, 8.0]$	267	44.50%
$(8.0, 10.0]$	40	6.67%
Total	600	100%

2.2 Explanatory variables

As mentioned before, the data was manually inserted. Film genres, film budgets, film certificates, film release dates, directors, actress/actors, film production companies, countries and film languages are chosen as our explanatory variables. There are additional factors that may affect our response variables, but we focus on what we have obtained. Now we give a summary of each explanatory variable.

- **FamousDirector**

We define the variable FamousDirector according to whether a director had won the Academy Awards/Golden Globes/British Academy Film Awards before the film was published. In our dataset 510 films is directed by famous directors. Films directed by famous directors are more likely to make a profit, as shown in Table 4 below. IMDB scores of films which are headed by famous directors or not renowned directors, as shown in Table 5.

Table 4: Table of Profit with FamousDirector

Profit	FamousDirector	Not famousDirector
No	149 (29.2%)	43 (47.8%)
Yes	361 (70.8%)	47 (52.2%)
Total	510 (100%)	90 (100%)

Table 5: Table of IMDB score with FamousDirector

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
FamousDirector	38(67.86%)	199 (83.97%)	236(88.39%)	37 (92.50%)
Not famousDirector	18 (32.14%)	38 (16.03%)	31(11.61%)	3 (7.50%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

• **Star**

Variable Star indicates whether an actor/actress is a star. A star is determined by whether the actor/actress had received academy awards/Golden Globes/British Academy Film Awards before the film was issued. The dataset has 479 stars, and 121 actors/actress are not stars. Table 6 shows that films with stars are more likely to make profit. Table 7 displays the IMDB score of a film with stars in it. It seems that a film with a star may obtain possibly more chance to get an IMDB score between 5 and 7.

Table 6: Table of Profit with Star

Profit	Star	Not Star
No	145 (30.2%)	47 (38.8%)
Yes	334 (69.8%)	74 (61.2%)
Total	479 (100%)	121 (100%)

Table 7: Table of IMDB score with Star

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
Star	35 (62.5%)	185 (78.06%)	227(85.02%)	32 (80.00%)
Not Star	21 (37.5%)	52 (21.94%)	40(14.98%)	8 (20.00%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

• **BigCo**

The variable BigCo describes whether a film production company is one of the top 500 production companies. The rating is based on the worldwide financial income of the production companies. In our dataset, we have 389 big production companies, while the rest 211 ones are not. Table 9 shows that comparing with a film produced by small company, the one produced by big companies has higher proportions.

Table 8: Table of Profit with BigCo

Profit	Big Company	Not Big Company
No	89 (22.9%)	103 (48.8%)
Yes	300 (77.1%)	108 (51.2%)
Total	389 (100%)	211 (100%)

Table 9: Table of IMDB score with BigCO

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
Big company	27 (48.21%)	157 (66.24%)	178(66.67%)	27 (67.50%)
Not Big company	29 (51.79%)	80 (33.76%)	89(33.33%)	13 (32.50%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

• **Budget**

As an example, the production cost of a typical horror film does not need to be too high to make a good profit[2]. To avoid reducing the accuracy of the model, we set a minimum budget for each film to 1 million. Besides, we place a film which had a budget of more than 100 million as large budget. In the dataset, 97 films have a budget of more than 100 million.

Table 10: Table of Profit with Budget

Profit	Big Budget	Not Big Budget
No	18 (18.6%)	174 (34.6%)
Yes	79 (81.4%)	329 (65.4%)
Total	97 (100%)	503 (100%)

Table 11: Table of IMDB score with Budget

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
Big Budget	4 (7.14%)	27 (11.39%)	53(19.85%)	13 (32.50%)
Not Big Budget	52 (92.86%)	210 (88.61%)	214(80.15%)	27 (67.50%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

• **RuntimeMinutes**

Variable *RuntimeMinutes* is a continuous variable. Based on the quantiles, we created four partitions to see the differences among different runtime of films. It seems that films are more likely to get a profit when their length is between 137 and 189 minutes, as shown in Table 12.

Table 12: Table of Profit with runtimeMinutes

Proft	[73min, 96min]	(96min, 115min]	(115min, 136min]	(136min, 189min]
No	63(42.9%)	51(34.93%)	44(29.33%)	34(21.66%)
Yes	84(57.1%)	95(65.07%)	106(70.67%)	123(78.34%)
Total	147 (100%)	146 (100%)	150 (100%)	157 (100%)

• **Certificate**

Movie certificate is assessed by MPAA, Motion Picture Association of American which was founded in 1922. In our dataset, we have four certificates. G means that it is suitable for general audiences. PG indicates that it is recommended to use parental guidance because some material is not ideal for children. PG-13 means that parents should be strongly cautioned because some parts of the film may be unsuitable for children under the age of 13. R means restrict. In other words, audiences under 17 years old need to be with parents or under the guardian of an adult. Most of the movies have a certificate R, but we can see that movies with PG certificates are more likely to make a profit, see Table 13 below.

Table 13: Table of Profit with Certificate

Proft	PG	PG.13	R	G
No	24(22%)	67(29.3%)	82(35.2%)	6(54.5%)
Yes	85(78%)	161(70.3%)	151(64.8%)	5(45.5%)
Total	109 (100%)	144 (100%)	155 (100%)	11 (100%)

From Table 14, films with PG certificate have the highest proportion of IMDB-score between $[0, 5]$. However, films with PG-13 have the highest ratio of IMDB-score between $(5, 6.5]$. Films with R certificate have the highest proportion of IMDB-score between $(6.5, 10]$.

Table 14: Table of IMDB score with Certificate

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
PG	19 (33.93%)	42 (17.72%)	44(16.48%)	4 (10.00%)
R	15 (26.79%)	87 (36.71%)	113(42.32%)	18 (45.00%)
PG.13	18 (32.14%)	96 (40.51%)	103(38.58%)	11 (27.50%)
G	4 (7.14%)	12 (6.06%)	7(2.62%)	7 (17.50%)
Total	63(100%)	237(100%)	267(100%)	40(100%)

• **ReleaseDate**

This variable indicates when a movie is released. According to Public holidays in the United States, we divide the years into four periods. Considering that school summer vacation has a sufficient impact on the box office that we set up **SummerHoliday** between 6/1 and 8/31. **Christmas** is set to between 12/24 and 12/25, including the New Year (1/1). **Workday** is set from Monday to Thursday and **Otherday** including the day before weekend, weekend and others holidays in the United States, like Independence Day(7/4), Labor Day(first Monday in September), Thanksgiving Day(Fourth Thursday in November), Halloween(10/31), Easter(3/22-4/25), Saint Patrick's Day (3/17), Mother's Day(second Sunday in May) and Father's Day(third Sunday in June). Some Christmas days and some

SummerHoliday days were on Friday, which means that some films can be both categorize in Christmas and Otherday, as well as SummerHoliday and Otherday. In our dataset, there are 92 films were released on this case.

Table 15: Table of Profit with ReleaseDate

Proft	SummerHoliday	Workday	Otherday	Christmas
No	37(27.8%)	21(29.2%)	124(33.3%)	10(43.5%)
Yes	96(62.2%)	51(70.8%)	248(66.7%)	13(56.5%)
Total	133 (100%)	72 (100%)	372 (100%)	23 (100%)

Table 16: Table of IMDB score with ReleaseDate

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
SummerHoliday	10 (17.86%)	58 (11.81%)	56(20.97%)	9 (22.5%)
Workday	9 (16.07%)	24 (10.13%)	35(13.11%)	4 (10.00%)
Otherday	41 (73.21%)	193 (81.43%)	198(74.16%)	32 (80.00%)
Christmas	3 (5.36%)	7 (2.95%)	12(4.49%)	1 (2.50%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

- **Language**

A film can be released in English, or in a non-English language, or in a few languages. Variable **Language** is divided into two categories, **English** and **OtherLanguages**. It seems that films released in other languages are more likely to be profitable than those films in English, see Table 17.

Table 17: Table of Profit with Language

Profit	English	OtherLanguage
No	184 (31.6%)	63 (26%)
Yes	399 (68.4%)	179 (74%)
Total	583 (100%)	242 (100%)

Table 18: Table of IMDB score with Language

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
English	54 (96.43%)	234 (98.73%)	256(95.88%)	39 (69.64%)
OtherLanguage	14 (25.00%)	86 (36.29%)	119(44.57%)	23 (41.07%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

- **IssueCountry**

A movie can be released in several different countries at the same time. It is reasonable to set the variable **issueCountry** into three different categories according to areas. **U.S.A** means that a film is issued in the U.S.A. **EnglishCountry** represents that a movie is distributed in a country not belonging to United States but whose official language is English. Non-English means that a film is distributed in countries whose official language is not English. Table 19 shows that the profit of films from different countries is alike. However, it is obvious that most movies are released in the United States.

Table 19: Table of Profit with issueCountry

Proft	U.S.A	EnglishCountry	NonEng
No	165(43.8%)	55(36.4%)	68(38.2%)
Yes	377(56.2%)	96(63.6%)	110(61.8%)
Total	542 (100%)	151 (100%)	178 (100%)

Table 20: Table of IMDB score with issueCountry

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
U.S.A	48 (85.71%)	222 (93.67%)	235(88.01%)	37 (92.5%)
EnglishCountry	11 (19.64%)	53 (22.36%)	73(27.34%)	14 (35.00%)
NonEng	13 (23.21%)	67 (28.27%)	87(32.58%)	11 (27.50%)
Total	56(100%)	237(100%)	267(100%)	40(100%)

• **Genre**

A film can have one or more different genres. In our dataset, we have 21 different film genres, including action, drama, thriller, animation, adventure, biography, documentary, comedy, music, musical, history, sport, crime, Sci-Fi, family, mystery, fantasy, horror, romance, western and war. We set these 21 types of films as 21 categorical variables, for example, *Action* describes whether a film is an action film or not. Top five high-frequency movie genres are shown in the following tables.

Table 21: Table of Profit with Genre

Proft	Drama	Comedy	Action	Adventure	Crime
No	118(40.6%)	70(34.5%)	41(22.4%)	37(22.2%)	30(26.8%)
Yes	172(59.4%)	133(65.5%)	142(77.6%)	130(77.8%)	82(74.2%)
Total	290 (100%)	203 (100%)	183 (100%)	167(100%)	112(100%)

Table 22: Table of IMDB score with Genre

	$[0, 5]$	$(5, 6.5]$	$(6.5, 8]$	$(8, 10]$
Drama	17	93	154	26
Adventure	16	62	75	14
Action	15	81	71	8
Comedy	31	94	71	7
Crime	9	45	48	10

2.3 Missing data

Missing data is likely to impact the result of an analysis. For handling missing data, there have three different assumptions. The most stringent assumption is Missing Completely at Random (MCAR), which means that the probability of a missing data is independent of other variables and itself. The assumption is not realistic in the most cases. A less stringent assumption is Missing at Random (MAR), which means that the probability of a missing data is independent of itself but depends on other variables. The last assumption, Not Missing at Random (NMAR), which means that the probability of a missing data depends on itself, but independent of other variables. (Agresti 2013, pp.395 - 396).

Missing data is more likely to be MAR when missing data is an answer to a questionnaire. There is a risk that heavy bias is increased by delete missing data unless missing data are MCAR. It is most challenging to build an imputation model based on unobserved data with NMAR. Imputation method is one solution to deal with the missing data. The advantage of this method is that it does not need to remove any useful data except replacing the missing information by estimated values.

There are many different types of imputation methods. In this thesis, we use the k-nearest neighbour method, which was introduced by Jönsson & Wohlin (2006). The k-nearest neighbour method is to input missing data with a similar value from an observation which has complete data. In our data set, there are 13 films do not have certificates. So, by using the nearest neighbour method, certificates of 13 films are replaced by certificates of same type films. We do not present the details about the k-nearest neighbour method in this thesis due to time limitation. Readers are encouraged to read more information in the book(Jönsson & Wohlin 2006, pp.12-14).

3 Theory

Statistical theories and methods are introduced in this section, such as logistic regression, ordinal logistic regression, variable selection methods and model diagnostics. These theories are cited in *Categorical Data Anal-*

ysis[3], *Analysis of Ordinal Categorical Data*[4], *Applied Logistic Regression*[5], *Lineära Statistiska Modeller*[6] and *Applied Statistical Inference*[8].

3.1 Odds ratio

Let probability of a profitable film denotes $P(Y_i = 1) = \pi_i(\mathbf{X})$, with predictor vector $\mathbf{X} = (x_{i1}, x_{i2}, \cdots, x_{ij})$, where x_{ij} denotes the observation number i with the variable j. The odds means that the proportion of probability of a profitable film with the probability of an unprofitable film, defined as:

$$\Omega = \frac{\pi_i(x)}{1 - \pi_i(x)}$$

The Odds is always non-negative when $\pi_i(x)$ between 0 and 1. If $\Omega > 1$, it means a film is more likely to earn a profit. Conversely, When $\Omega < 1$ it says that a film is more likely to be unprofitable.

In logistic regression, we usually use the odds ratio to interpret the difference of probability of a profitable film between two predictors. The odds ratio is the proportion of two odds. Let Ω_{FD} indicate the odds for a profitable film with a famous director, Ω_{NFD} denotes the odds for a profitable film with a not famous director. The odds ratio is defined as:

$$\theta = \frac{\Omega_{FD}}{\Omega_{NFD}} = \frac{\frac{\pi_{FD}}{1-\pi_{FD}}}{\frac{\pi_{NFD}}{1-\pi_{NFD}}}$$

If $\Omega_{FD} = \Omega_{NFD}, \theta = 1$ means that the response variable is independent of explanatory variable *FamousDirector*. $\theta > 1$ indicates that the probability π_{FD} is higher than π_{NFD}. If $\theta < 1$, it means that the probability π_{NFD} is higher than π_{FD}. We should note that the odds ratio is always positive.

3.2 Logistic regression

Logistic regression is a model of categorical response data and used to determine whether a binary response variable correlates with one or more independent explanatory variables. In our data set, the film which is profitable can be considered as a binary response variable. Profitable film Y_i denotes 1 with probability $\pi_i(x) = P(Y_i = 1|X = x_i)$, unprofitable film denotes 0 with probability $P(Y_i = 0|X = x_i) = 1 - \pi_i(x) = 1 - P(Y_i = 1|X = x_i)$. If a film is profitable associating with one predict variable, we can say it is the simple logistic model(Agresti 2013, p.163):

$$\pi_i(x) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

which has the linear relationship :

$$\text{logit}([\pi_i(x)]) = \log\frac{\pi_i(x)}{1 - \pi_i(x)} = \alpha + \beta x_i$$

The profitability whether a film is profitable is associated with multiple predictors variables $\mathbf{X} = (x_{i1}, \cdots, x_{ij})$, where x_{ij} denotes the observation number i with the variable j. The formula of Multiple logistic regression is(Agresti 2013, p.182):

$$\text{logit}[\pi_i(x)] = \log\frac{\pi_i(x)}{1-\pi_i(x)} = \alpha + \sum \beta_i x_{ij} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_j x_{ij}$$

After solving the above formula, we get $\pi_i(x)$:

$$\pi_i(x) = \frac{e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}}}{1+e^{\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}}}$$

The parameter β_j is related to the difference between the log of the odds when $x_{ij} = 1$ and the log of the odds when $x_{ij} = 0$. For instance, β_{FD} shows the difference between a film directed by a famous director and non-famous director in log odds of $P(Y_i = 1)$. However, it is difficult to interpret the difference between one unit increase/decrease of the log-odds. It is easier to explain the change of the odds. The odds is increased multiplied by e^{β_j} when there has a one-unit increase in the explanatory variable x_j. For instance, the odds of a profitable film is $e^{\beta_{FD}}$ times as higher for the film is directed by famous director than the film is directed by the non-famous director.

3.3 Ordinal Logistic regression

For binary response variables, logistic regression is the most common way of analyzing the effects of explanatory variables, even for ordinal response variables.

When the response variable is ordinal, it makes sense to form logits that take the category order into account. The logits can be formed by grouping categories that are continuous on the ordinal scale. Agresti(2010, pp.44-46) introduces three types logits, which are cumulative logits, adjacent-categories logits and continuation-ratio logits. Based on our the other response variable *IMDB score* which we mentioned in section 2.1 has four categories, cumulative logits will be used in this thesis. The interested readers can read more about the other two types loigts in chapter four of Hosmer (2010). Cumulative logits can characterize how the ordinal IMDB score correlates with one or more explanatory variables.

For an ordinal response variable Z with probabilities $\pi_1, \cdots, \pi_c$, the cummulative logits are defined as:

$$\text{logit}[P(Z \leq k)] = \log\frac{P(Z \leq k)}{1-P(Z \leq k)} = \log\frac{\pi_1 + \cdots + \pi_k}{\pi_{k+1} + \cdots + \pi_c}$$

for $k = 1, \cdots, c-1$.

For instance, the ordinal IMDB-score variable has three cumulative logits, and each cumulative logit is:

$$\text{logit}[P(Z \leq 1)] = \log \frac{\pi_1}{\pi_2 + \pi_3 + \pi_4}$$

$$\text{logit}[P(Z \leq 2)] = \log \frac{\pi_1 + \pi_2}{\pi_3 + \pi_4}$$

$$\text{logit}[P(Z \leq 3)] = \log \frac{\pi_1 + \pi_2 + \pi_3}{\pi_4}.$$

For cumulative logit, the model of IMDB score with $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$ is the ordinal response variable and a prediction vector $\mathbf{X} = (x_{i1}, \cdots, x_{ij})$, x_{ij} denotes the j variable with i observation. The model is formalized as:

$$\text{logit}[P(Z \leq k)] = \alpha_k + \boldsymbol{\beta'}\mathbf{X} = \alpha_k + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}. \quad k = 1, ..., c-1$$

The parameter vector $\boldsymbol{\beta}$ describes the effect of multiple predictor variables. We should know each cumulative probability has its intercept α_k but the effects β remains the same for each cumulative logit. Intercept α_k increases means the cumulative probability increases. We can even compare the probability of a smaller IMDB score level with the probability of a more substantial IMDB score level.

The cumulative probability (Agresti 2010, p.47) is defined as following:

$$P(Z \leq k) = \frac{e^{\alpha_k + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_k + \boldsymbol{\beta'}\boldsymbol{X}}}. \quad k = 1, ..., c-1$$

For exemple, our ordinal response variable, each cell probability is:

$$P(\text{IMDB score} \leq 5.0) = P(Z \leq 1) = \frac{e^{\alpha_1 + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_1 + \boldsymbol{\beta'}\boldsymbol{X}}}$$

$$\begin{aligned} P(5.0 < \text{IMDB score} \leq 6.5) &= P(Z \leq 2) - P(Z \leq 1) \\ &= \frac{e^{\alpha_2 + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_2 + \boldsymbol{\beta'}\boldsymbol{X}}} - \frac{e^{\alpha_1 + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_1 + \boldsymbol{\beta'}\boldsymbol{X}}} \end{aligned}$$

$$\begin{aligned} P(6.5 < \text{IMDB score} \leq 8.0) &= P(Z \leq 3) - P(Z \leq 2) \\ &= \frac{e^{\alpha_3 + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_3 + \boldsymbol{\beta'}\boldsymbol{X}}} - \frac{e^{\alpha_2 + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_2 + \boldsymbol{\beta'}\boldsymbol{X}}} \end{aligned}$$

$$P(\text{IMDB score} > 8.0) = 1 - P(Z \leq 3)$$

$$= 1 - \frac{e^{\alpha_3 + \boldsymbol{\beta'}\boldsymbol{X}}}{1 + e^{\alpha_3 + \boldsymbol{\beta'}\boldsymbol{X}}}$$

The model of cumulative logits is also called the proportional odds model which is used in the following section.

3.4 Likelihood function and maximum likelihood estimate

A predictor vektor $\mathbf{X} = (x_{i1}, x_{i2}, \cdots, x_{ij})$ has a joint density function $f(X|\theta)$ with a fixed parameter θ. Given observed datum $X_1 = x_{i1}, X_2 = x_{i2}, \cdots, X_j = x_{ij}$, the likelihood function is defined as a function of θ(Held&Bové 2014, p.14):

$$L(\theta|x_{i1}, x_{i2}, \cdots, x_{ij}) = f(x_{i1}, x_{i2}, \cdots, x_{ij}|\theta)$$

For estimating parameter we use maximum likelihood estimate. The maximum likelihood estimate $\hat{\theta}_{ML}$ is the solution of the partial derivatives $\frac{\partial l(\theta)}{\partial(\theta)} = 0$, where $l(\theta)$ is the logarithm of the likelihood function. Agresti(2013, p.9) point out that maximum likelihood estimate $\hat{\theta}_{ML}$ has the highest probability of which one case happens with observed data.

3.5 Variable Selection

In this thesis, we will compare two different variable selection methods to fit a model. One is *Purposeful Selection*, the other is *Stepwise Procedures*(Sundberg 2016, pp.71-72). Stepwise procedures are the most widely used, which is embedded in the R program. Hosmer(2013) points out that we have no chance to examine the model before stepwise procedures obtain the final model. Compared to stepwise procedures, the Purposeful selection which is given by Hosmer, Lemeshow and Sturdivant (2013, pp.89-93) gives analytical control over every step among the selection process and confounding can be considered.

3.5.1 Stepwise Procedures

Stepwise Procedures usually has three types methods, which are *Forward selection*, *Backward Elimination* and *Stepwise Selection*.

Forward selection *The forward selection* adds terms sequentially. It begins with the intercept in the model. All explanatory variables which are not in the model can be added back to the model if their p-value is lower than 0.05. The process stops until there are no new variables which can be added back into the model.

Backward elimination *The Backward elimination* begins with the saturated model and removed terms sequentially. At each step, the variable which is not significant at level 0.05 is excluded. The process stops when each variable in the model is significant.

Stepwise selection *Stepwise selection* is based on the *Forward selection*, at the same time, it combines *Backward elimination*. Similar to the forward selection, it only starts with intercept. In the process, variables that are not significant at the 0.05 level are eliminated from the model. Meanwhile, variables which are significant at level 0.05 will be added to the model.

The process is repeated until there are no more additions or eliminations which do not significantly improve the model fit.

3.5.2 Purposeful Selection

The other variable selection method is the purposeful selection which has seven steps(Hosmer et al. 2013, pp.89-93).

Step 1: At the beginning of the purposeful selection, each independent variable should be analyzed carefully. Those explanatory variables whose p-value is less than 0.25 should be contained in our first multivariable model. Hosmer points out the reason why we here do not use traditional level (such as 0.05) is because that it often fails to identify variables seemed to be important.

Step 2: Now we have all the predictors with p-value less than 0.25 in a multivariable logistic regression model. These variables with a p-value larger than 0.05 were excluded by using the p-value of their Wald statistic(in section 3.5.2). By using a likelihood ratio test, the smaller, new model is compared with the larger, previous model.

Step 3: Compare the estimated coefficients in the smaller, new model with these in the larger, old model. We should pay attention to those whose coefficients are estimated to change too much. Using an indicator $\Delta\hat{\beta}_i = |(\hat{\theta}_i - \hat{\beta}_i)/\hat{\beta}_i| > 0.2$ to check whether an estimated coefficient changes more than 20%, in which $\hat{\beta}_i$ denotes the estimated parameter of variable i in the larger, previous model and $\hat{\theta}_i$ represents the estimated parameter of variable i in the smaller, new model. If $\Delta\hat{\beta}_i > 0.2$, we should consider to add the excluded variables back to the model until all important variables are included in the model. Meanwhile, the likelihood ratio test is needed when each excluded variable is added back into the model.

Step 4: Add each variable that is insignificant in Step 1 one by one to the smaller model is obtained in Step 3 and check its significance. At this step, we aim to find variables which are insignificant have an impact at the variables in the present model.

Step 5: Variable in the model should be checked closely in this step. Each category for categorical variables should be reasonable, and each continuous variable should have a linear relationship with the logit. The model at the end of Step 5 is referred to as the *main effects model*.

Step 6: Now we check the interactions between explanatory variables in the *main effects model*. Interaction two predictors means that the effect of a

predictor on the response variable is inconstant over the levels of the other predictor. Furthermore, statistical reason and realistic situations should be considered when an interaction is included in the model. We add possible interactions to the *main effects model* one by one, and check the statistical significance of the interaction by using a likelihood ratio test(see in section 3.5.1) with a low p-value(level 5% or even 1%). At this step, no main effects are removed from the model and repeated from step two. We refer to the model at the end of Step 6 is called the *preliminary final model*.

Step 7: We should examine how well that the *preliminary final model* fits the data before it becomes our final model. Goodness-of-fit methods are presented in the following section.

3.6 Model fit and diagnostics

The purpose of the model building is to find a good statistical model that can fit the data well while it is as simple as possible. We will use different goodness-of-fit methods to test the accuracy of the model that approximates the observed data and its predictive capacity. Furthermore, by using these statistic test, two models are compared and an explanatory variable is determined whether to keep in the model. In this section, we will present statistic tests which are used in this thesis.

3.6.1 Testing the linearity of continuous explanatory variable

Continuous variable should be controlled if it has either increased or decreased linearly relationship with the logit. Hosmer et al. (2013, pp.94-107) introduce four methods for dealing with the continuous variable. In this thesis, we use the second method which is easily performed in all statistical packages for testing the linearity of the continuous variable(Hosmer et al. 2013, pp.95-96). Based on the three cutpoints of quartiles of the runtime of films we create a categorical variable with four levels. Then replacing the continuous variable by the categorical variable to fit the multivariable model. The lowest category is referred to as reference group whose coefficient is equal to zero. To check if it looks linearly, one can plot those estimated coefficients against the midpoint of the upper three quantiles. For those who are interested, the other three methods can be found in Chapter 4.2.1.

3.6.2 Likelihood-ratio test

Likelihood-ratio test(Agresti 2002, p.11) is used for comparing two logistic models, a smaller model M_0 and a lager model M_1. Null hypothesis test and the alternative hypothesis between M_0 and M_1 are formulated:

$$H_0 : M_0 \text{ holds}$$

$$H_1 : M_1 \text{ holds but not } M_0.$$

The likelihood ratio test between M_0 and M_1 is formed:

$$-2 \cdot log\left(\frac{L_0}{L_1}\right) = -2(l_0 - l_1)$$

where L_0 is the maximized likelihood function under the null hypothesis and L_1 is the maximized likelihood function under the alternative hypothesis. l_0 and l_1 are the corresponding maximized log-likelihood functions.

The likelihood ratio test statistic is asymptotically distributed as a *chi-squared* under the null hypothesis,

$$-2(l_0 - l_1) \overset{H_0}{\approx} \chi^2_{df}$$

where degrees of freedom(df) is equal to the difference between the number of parameters in the two models.

3.6.3 Wald statistic test

By using the Wald statistic test, an explanatory variable is determined whether it is significant or not. We have a null hypothesis:

$$\mathrm{H}_0 : \theta = \theta_0$$

and alternative hypothesis:

$$\mathrm{H}_1 : \theta \neq \theta_0$$

The Wald statistic test is defined as:

$$W = \frac{\hat{\theta}_{ML} - \theta_0}{\mathrm{se}(\hat{\theta}_{\mathrm{ML}})}$$

where $\hat{\theta}_{ML}$ is the maximum likelihood estimated of θ_0, and $\mathrm{se}(\hat{\theta}_{\mathrm{ML}})$ is the standard error of the maximum likelihood estimation of $\hat{\theta}_{\mathrm{ML}}$. Wald statistic test can approximate an asymptotically standard normal distribution under H_0 (Held&Bové 2014, p.99).

$$W = \frac{\hat{\theta}_{ML} - \theta_0}{\mathrm{se}(\hat{\theta}_{\mathrm{ML}})} \approx N(0, 1)$$

Moreover, the Wald confidence interval can be calculated under the null hypothesis by using the Wald statistic test. In the thesis, we use a significant level 0.05 to determine whether an explanatory variable is significant or not. The Wald confidence interval is defined as:

$$(\hat{\theta}_{ML} - z_{\frac{1+\gamma}{2}} \cdot SE_{\hat{\theta}_{ML}}, \hat{\theta}_{ML} + z_{\frac{1+\gamma}{2}} \cdot SE_{\hat{\theta}_{ML}})$$

$$=$$

$$(\hat{\theta}_{ML} - 1.96 \cdot SE_{\hat{\theta}_{ML}}, \hat{\theta}_{ML} + 1.96 \cdot SE_{\hat{\theta}_{ML}})$$

, where $\gamma = 0.95$ and $z_{\frac{1+\gamma}{2}} = 1.96$ which is the value of the 97.5 percentile point of standard normal distribution.

3.6.4 The Hosmer-Lemeshow test

Traditional goodness-of-fit tests, such as Pearson Chi-Square statistic, Deviance, do not have asymptotic χ^2-distributions when the number of parameters is almost equal to the number of total observations and there exist one or more continuous explanatory variables(Hosmer et al. 2013, pp.155-157). So, the Hosmer-Lemeshow test is presented as a more suitable goodness-of-fit test for these problems (Hosmer et al. 2013, pp.157-158).

If the number of parameters is equal to the sample size, we assume that n estimated probabilities are corresponding to n columns. The first column is grouped into the smallest value, and the n:th column are grouped into maximum values. Under the grouping, we have two strategies. One is group by percentage of the estimated probabilities. The other is group by fixed values of the estimated probability.

Usually, we use the first method to set $g = 10$ groups, where $1/10$ smallest estimated probabilities in the first group and $1/10$ most substantial estimated probabilities in the last groups. Asymptotically, the Pearson statistic, $\hat{C}$, approximates by the χ^2-distribution with $(g-2)$ degrees of freedom when the logistic regression model is the correct model. Interested readers can read more details in the Chapter 5.2.2 of Hosmer and Lemeshow (2013).

3.6.5 Lipsitz test

Unlike binary logistic regression model, traditional goodness-of-fit such as Pearson Chi-Square statistic and the Deviance are not suitable for a proportional odds model which has one or more continuous explanatory variables. Hosmer et al. (2013, pp.303) introduce Lipsitz test which is a goodness of fit test for the proportional odds model.

The observed data is grouped into percentiles. The test requires an ordinal response variable which is divided into g equal-sized groups. We define the relationship between predictors and categorize outcome as $I_{ij} = 1$ if the IMDB score is in the group j and $I_{ij} = 0$ if the IMDB score is in other categories. The value of Lipsitz test is obtained by the sum of the predictive probability of each ordinal scale with the integral of product of each I_{ij} and each grouped sample.

We should note that the analysis approximates by a χ^2-distribution with $(g-1)$ degrees of freedom. For having enough groups and numbers of subjects to show the difference from model assumptions, groups g are suggested to be $6 \leq g \leq i/5k$, where i is the number of observations, k is the ordinal scales. Usually, the numbers of groups g is equal to 10(Matthew 2017, p.3). More details regarding Lipsitz test can be find in Chapter 8.2.1 of Hosmer and Lemeshow.

3.6.6 AIC

The Akaike information criterion (AIC) is commonly used to compare models with different numbers of parameters from the same dataset(Agresti 2013, p.212). AIC is defined as:

$$AIC = -2(\text{maximized log likelihood} - \text{number of parameters in the model})$$

$$= -2(\log(L) - j)$$

where L is the maximized likelihood function, and j is the number of parameters in the model. Among the given models, the model with the smallest AIC value is preferred.

3.7 Prediction capacity

After building a logistic regression model and the proportional odds model, the predictive capacity of both models should be tested. Receiver operating characteristic curve(ROC), the area under ROC-curve(AUC), and McFadden's pseudo-R squared is a way to test the predictive power of a model.

3.7.1 ROC and AUC

To evaluate the predictive capacity of a model, we usually create a receiver operating characteristic curve(ROC-curve). Let, predictive value $\hat{y}$ is 1 when $\hat{\pi}_i > \pi_0$ and predictive value $\hat{y}$ is 0 when $\hat{\pi}_i < \pi_0$, for some cutoff π_0. ROC-curve plots the probability of sensitivity and 1 - specificity for an entire range of possible cutoff π_0, where sensitivity = 1 - specificity. Agresti(2013, p.223) gives the definitions of specificity and sensitivity for all posibile values of cutoff π_0:

$$sensitivity = P(\hat{y}_i = 1|y = 1) \quad \text{and} \quad specificity = P(\hat{y}_i = 0|y = 0).$$

So, a ROC curve usually has a concave curve which connects the points (0,0) and (1,1). The area under ROC can describe the predictive accuracy of a fitted model. The more significant area under a ROC curve, the better the

predictive power of the fit model. As the same, the higher the sensitivity, the better predictive power.

For an ordinal response vairable Z in the propotional odds model, let $x = 1$ denote that a case happens and $x = 0$ denote that a case does not happen. A positive response with an ordinal scale $Z \leq k$, the sensitivity and specificity is defined as(Agresti 2010, p.133):

$$sensitivity = P(Z \leq k|x = 1) \quad \text{and} \quad specificity = P(Z > k|x = 0).$$

The ROC curve is plotted with points for $k = 0, 1, 2, \cdots, c$. When $k = c$ is the cutoff point, the ROC curve connects the point $(0, 0)$; when $k = 0$, the ROC curve connects the point $(1, 1)$. the ROC curve usually has a concave curve above the straight line connecting the points $(0, 0)$ and $(1, 1)$ for $k = 1, ..., \cdots, c - 1$. For instance, in our proportional odds model, three ROC curves are plotted for $k = 1, 2, 3$.

In addition, Hosmer and Lemeshow (2013, p.177) give guidelines on how to interpret the area under ROC-curve($\boldsymbol{AUC}$):

$$\text{if} = \begin{cases} 0.5 < AUC < 0.7 & Poor \\ 0.7 < AUC < 0.8 & Acceptable \\ 0.8 < AUC < 0.9 & Excellent \\ 0.9 < AUC & Outstanding \end{cases}$$

The area under the ROC is 0.5 which means it under a straight line connecting the points $(0, 0)$ and $(1, 1)$ in XY-plane. Meanwhile, the model's predictive accuracy is no better than random guessing (Agresti 2013, p.224).

3.7.2 McFadden's pseudo R squared

Relating to the fact that the model contains too many paramters while AUC is the only measure to test the predictive power. In addition to AUC, R squared is presented. R^2 measure describes the accuracy of a model how it approximates the observed data. McFadden(1973, p.123) presents McFadden's pseudo-R squared which is defined as:

$$R^2_{McFadden} = 1 - \frac{\log(L_{\hat{\theta}_{ML}})}{\log(L_{\theta_0})} = 1 - \frac{l_{\hat{\theta}_{ML}}}{l_{\hat{\theta}_0}}$$

where $l_{\hat{\theta}_{ML}}$ stands for the maximized log likelihood value of the present fitted model, and $l_{\hat{\theta}_0}$ denotes the value of the maximized log likelihood function for the model with the only intercept.

The binary logistic regression has outcome either 0 or 1. If the outcome is close to 1, the log likelihood is close to zero, a loglikelihood contribution is largely negative. If the model has no predictive ability, loglikelihood of

the current fitted model is larger than loglikelihood of null model, so $\frac{l_{\hat{\theta}_{ML}}}{l_{\hat{\theta}_0}} \approx 1 \Rightarrow R^2_{McFadden} \approx 0$. Similarly, if the model has good predictive ability and outcome is 1, so log likelihood is near zero, $\frac{l_{\hat{\theta}_{ML}}}{l_{\hat{\theta}_0}} \approx 0 \Rightarrow R^2_{McFadden} \approx 1$. McFadden(1979, p.307) point out that the value of $R^2_{McFadden}$ is usually lower than traditional R^2, that means $0.2 \leq R^2_{McFadden} \leq 0.4$ which means fitted model is excellent.

4 Model Building

In section 4.1, the response variable *Profit* is analyzed with the binary logistic regression, and the other ordinal response variable *IMDB-score* is studied with ordinal logistic regression in section 4.2. For fitting the first binary logistic model, we use two variable selection methods, respectively Stepwise selection and purposeful selection, to fit the model which can show a correlation between profit and predictors. For fitting the proportional odds model, we use purposeful selection. The proportional odds model shows which predictors are associated with IMDB score of a film. Our data set is divided into two parts. For model building, we use films which were released between 1990 and 2009. Films which were published between 2010 and 2017 is used to test the predictive capacity of the final selected model in section 5.

4.1 Fitting the model of a profitable film

In this section, we want to find out which predictor correlates with the binary response variable *Profit*.

4.1.1 Purposeful selection of predictors

Step 1: At the beginning, we examine each independent explanatory variable. Here we use the limit 0.25 p-value of a Wald statistic test to find out which explanatory variables are significant. The results of the first step are shown in the Appendix(Table 37).

Step 2: Now we have our first multivariable logistic regression model M_0 contains all variables in Step 1 which the p-value is less than 0.25.

Table 23: Results of the first multivariable logistic regression model M_0 from *Step 1*.

Parameter	Coeff.	SE	P-value
RuntimeMinutes	0.0268	0.0083	0.0012
FamousDirector	0.3907	0.3280	0.2336
BigBudget	-0.2826	0.4889	0.5632
BigCo	0.8899	0.2518	0.0004
PG	0.5506	0.3558	0.1218
U.S.A	0.0961	0.4650	0.8363
EnglishCountry	-0.2641	0.3074	0.3902
NonEng	-0.7987	0.2864	0.0053
OtherLanguage	0.4711	0.2657	0.0762
Action	0.3293	0.3147	0.2955
Drama	-0.7323	0.2881	0.0110
Crime	0.3399	0.3289	0.3014
Fantasy	0.1207	0.4737	0.7989
Mystery	-0.3419	0.3913	0.3822
Adventure	-0.1908	0.3441	0.5792

We check the p-value of the Wald statistic for each variable in the model M_0. Variables with a p-value above 0.05 are eliminated from the model M_0. The new smaller model is presented in the following table.

Table 24: Results of the multivariable logistic regression model M_1 after excluding variables in *Step 2*.

Parameter	Coeff.	SE	P-value
RuntimeMinutes	0.0287	0.0075	0.0002
BigCo	0.9441	0.2357	$6.21 \cdot 10^{-5}$
NonEng	-0.6926	0.2589	0.0075
Drama	-0.8112	0.2524	0.0013

By using likelihood ratio statistic, we compare the larger model M_0 with the smaller model M_1.

Hypothesis test between M_0 and M_1 can be performed as follows:

$$H_0 : M_0 \text{ holds}$$

$$H_1 : M_1 \text{ holds but not } M_0.$$

$$G^2(M_0|M_1) = -2(l_0 - l_1) = -2(416.00 - 428.29) = 24.58 > \chi^2_{0.05}(11) = 19.675$$

The result shows that the null hypothesis is rejected and M_1 holds.

Step 3: We now compare the estimated coefficients of each explanatory variable in the smaller model M_1 with the ones in the larger model M_0.

Table 25: Comparison of estimated coefficients $\hat{\theta}_i$ in model M_1 and estimated coefficients $\hat{\beta}_i$ in model M_0

Parameter	$\hat{\theta}_i$	$\hat{\beta}_i$	$\Delta\hat{\beta}_i$
RuntimeMinutes	0.0287	0.0268	0.0522
BigCo	0.9441	0.8927	0.0544
NonEng	-0.6926	-0.8024	0.1584
Drama	-0.8112	-0.7058	0.1298

Table 25 shows that the difference in estimated coefficients between the two models is less than the indicator value 0.20. Those Excluded variable are not added back into the model M_1.

Step 4: Now we check the variables which are not significant in *step 1* by adding them back to model M_1 one by one. We found that the estimated parameters remained almost the same. Model M_1 is considered our *preliminary main effects model*.

Step 5: Now we check closely about variables in the model M_1. M_1 contains three categorical variables, they are *BigCo*, *Drama*, and *NonEng* should be check reasonble . The estimated coefficient of *BigCo* is 0.9441. It means that the probability of getting profit increases when a big film production company produces the film. $\beta_{Drama} = -0.8112$ means that a drama film has a negative association with the profit of a film. In our dataset, it seems that a film is a drama film, the probability of a profitable film is decreased. Similarly, the negative coefficient of variable *NonEng* means that a film released by a non-English speaking country will decrease the probability of a film being profitable.

To exam, the variable *RuntimeMinutes* is a continuous variable which has a linear relationship with the logit. We use the method which mentioned in section 3.6.1. Categorical variable *RuntimeCat* with four levels is created and replacing the variable *RuntimeMinutes* to fit a model M_2. To check if it looks linearly, the estimated coefficients of the categorical variable *RuntimeCat* is plotted in Figure 1. Estimated coefficients of explanatory variables in the model M_2 are shown in Table 26.

Figure 1: Plotting estimated coefficients of RuntimeCat with first interval [73, 95], second interval (95, 105], third interval(105, 119], fourth interval(119, 189] in Model M_2

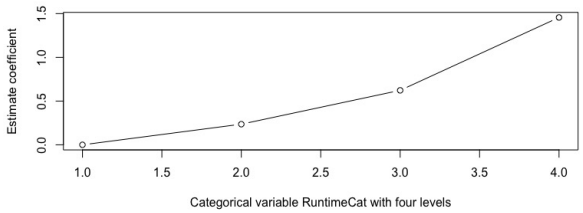


Table 26: Estimated parameter with categorical *RuntimeCat*

Parameter	Coeff.	SE	P-value
RuntimeCat2	0.2358	0.3291	0.4736
RuntimeCat3	0.6226	0.3438	0.0694
RuntimeCat4	1.45589	0.3790	0.0001
BigCo	0.9562	0.2372	$5.53 \cdot 10^{-5}$
NonEng	-0.6926	0.2617	0.0081
Dram	-0.8341	0.2604	0.0014

In Table 26, the categorical variable *RuntimeCat* has a positive estimated coefficient, but *RuntimeCat2* and *RuntimeCat3* are not significant at level 0.05. It seems that *RuntimeCat* follows a linear trend from Figure 1.

In addition to examining the categorical variable *RuntimeCat*, likelihood ratio test can be performed. We compare the following two models. One is ours preliminary main effect model M_1. The other is the model with the categorical variable *Runtime* of M_2.

Hypothesis test between M_1 and M_2 is performed as follows:

$$H_0 : M_1 \text{ holds}$$

$$H_1 : M_2 \text{ holds but not } M_1.$$

$$G^2(M_1|M_2) = -2(l_0 - l_1) = 2(428.29 - 425.80) = 4.98 < \chi^2_{0.05}(3) = 7.815$$

The value of likelihood-ratio test 4.98 does not exceed the $\chi^2_{0.05}(3)$. Moreover we use AIC to compare model M_1 with model M_2. The AIC of model M_1 (438.29) is also a bit smaller than AIC of model M_2 (439.8). Summarizing all of the above statistical results, we decide to treat *RuntimeMinutes*

as a continuous variable in the model. Model M_1 is considered as our *main effects model*.

Step 6: Now we check the interactions between explanatory variables in the model M_1. Only two-way interactions are considered, as it is complicated to interpret models with three-way interactions or multi-directional interactions in realistic terms. We check six two-way interactions between four main effects in M_1. Six two-way interactions are added back into the model once a time. We find that none of the two-way interactions are significant at level 0.05. Predictors remain the same in the model M_1 which denotes as M_p.

Step 7: After the previous 6 step, we get the model M_p as our *preliminary final model*. We will compare M_p with other models which obtained by using *Stepwise Procedures* in the next section.

4.1.2 Stepwise Procedures of predictors

Another model selection method, such as stepwise produces, is an alternative way to fit a binary logistic model if a film is profitable.

• **Stepwise selection**

In each explanatory variable, we set a starting model with intercept and a saturated model which has two-way interaction. By applying *Stepwise selection* in the R, we get a model *ModelStepboth1* (see Table 38 in Appendix) which has a value of AIC 431.1 and some explanatory variables are not significant at level 0.05. These insignificant explanatory variables are removed once a time from the model. In the second model M_{both}, we find one explanatory variable *OtherLanguage* which is not significant at the significant level 0.05. However, it is still retained in the model, because the interaction *Action: OtherLanguage* between explanatory variable *OtherLanguage* and variable *Action* is significant at level 0.05. The value of AIC is raised to 437.67 compare with the one in the model *ModelStepboth1*. The process stops here because no additional effects met the 0.05 significant level for moving from the model anymore. The model within *Stepwise selection* is:

Table 27: Results of Fitting the Model M_{both} with the *Stepwise selection*

Parameter	Coeff.	SE	P-value
BigCo	1.0387	0.2398	$1.48 \cdot 10^{-5}$
RuntimeMinutes	0.0365	0.0095	0.0001
OtherLanguage	-0.0264	0.2887	0.9273
Drama	-1.0831	0.3161	0.0006
Action	3.6354	1.7760	0.0406
Action:RuntimeMinutes	-0.0403	0.0167	0.0161
Action:OtherLanguage	1.2448	0.5845	0.0332
Action:Drama	1.3059	0.6391	0.0410

• **Forward selection**

We apply the *Forward selection* by starting to test the model only with the intercept. The saturated model contains all predictors and two-way interactions. Implementing **forward selection** in the R program. Implementing *Forward selection* in the R program, we get the first model *ModelForward*(see Table 39 in the Appendix) which has a value of AIC 431.45 and of which some explanatory variables are not significant at level 0.05. These insignificant explanatory variables are eliminated step by step from the model *ModelForward*. In the smaller, new model M_{For}, each explanatory variable is significant at level 0.05, hence the AIC value is raised to 443.45. The process stops here because no additional effects to satisfy the 0.05 significant level for moving from the model anymore. The model within *Forward selection* is:

Table 28: Results of Fitting the Model M_{for}

Parameter	Coeff.	SE	P-value
RuntimeMinutes	0.0257	0.0073	0.0004
BigCo	0.9940	0.2328	$1.96 \cdot 10^{-5}$
Dram	-0.8073	0.2499	0.0012

In contrast to the forward selection, *Backward elimination* begins with a saturated model which has two-ways interactions between each explanatory variable. The function stoped with a high AIC value 2000, so we did not consider the model obtained by *Backward Elimination*.

4.1.3 Comparison of above models

In this subsection, we compare models obtained by different selection methods from the previous section. M_p obtained through the purposeful selection. M_{both} obtained by stepwise selection, and M_{for} from forward selection. We apply some common statistical model selection criteria, such as the *Akaike information criterion*, *McFadden R square*, *AUC and ROC* and *Hosmer Lemeshow test* to find an adequate model. The following table shows the results.

Table 29: Comparison between M_p, M_{both}, and M_{For}

Model	AIC	Hosmer Lemeshow test	$R^2_{McFadden}$	AUC
M_p	438.29	$\chi^2_8 = 10.977$ with p-value=0.2018	0.1004	0.732
M_{both}	437.67	$\chi^2_8 = 4.7587$ with p-value=0.7830	0.1143	0.746
M_{For}	443.45	$\chi^2_8 = 6.2193$ with p-value=0.6227	0.0853	0.712

Akaike information criterion: The above table shows that the AIC of the model M_{both} has the lowest value. However, M_{both} is more complicated than the other models, which have five explanatory variables and three two-way interactions as explanatory variables.

Hosmer Lemeshow test: All three models are not significant. But model M_{both} has the highest p-value and the lowest $\chi^2(8) = 4.7587$ compared with the other two models.

McFadden R square: A higher R^2 indicates that a model has a higher predictive capacity. In comparison with model M_p and M_{for}, model M_{both} get the highest $R^2_{McFadden}$.

Area under Receiver Operating Characteristic curve: The bigger area under ROC-curve indicates the model has higher predict power(see section 3.6.1). Model M_{both} has the highest AUC value of 0.746, compare it with AUC value 0.732 of M_p and AUC value 0.712 of M_{for}.

An excellent statistical model can fit the data well while it is as simple as possible. Hence, the model M_{both} is the most complicated model compare to the other two models. The model M_{for} is the simplest which has three predictors, and the model M_p has four predictors.

This thesis aims to build a model which can predict a film's profit. So we focus on the predictive power of the model. To summarize the above statistical analysis, model M_{both} has the highest AUC and $R^2_{McFadden}$ value means the predictive capacity of the model is best. Moreover, it has the highest p-value from the Hosmer Lemeshow test and the lowest AIC value. So we choose model M_{both} as our final model. The predictive capacity of the model M_{both} will be discussed in section **results**.

4.2 Fitting the model of IMDB score

In the previous section, we focused on fitting a binary logistic model if the film is profitable. Next, we will determine predictors that are correlated with a film's IMDB score. It would be interesting to find out whether the same predictors are related to the film's profitability, and it's IMDB score. IMDB score is an ordinal response variable, which can be analyzed with the ordinal logistic regression. Hosmer et al.(2013) point out that purposeful selection can be used to building a proportional odds model. A more complicated statistical analysis of the IMDB score will be studied in this section by using purposeful selection, and we also try to fit a model that can estimate the probability of a film's IMDB score.

4.2.1 Purposeful selection

Step 1: For the proportional odds model, we begin to control the p-value of each independent explanatory variable that is not exceed 0.25. The results of the first step are shown in the Appendix(Table 40).

Step 2: The first multivariable proportional odds model M_{olm} contains variables that p-value is less than 0.25 in step 1.

Table 30: Results of the first multivariable proportional odds model M_{olm}

Parameter	Coeff.	SE	P-value
RuntimeMinutes	0.0364	0.0072	< 0.0001
Star	0.0102	0.2806	0.9710
FamousDirector	0.2661	0.2976	0.3712
BigBudget	0.5464	0.4071	0.1795
PG	-0.0082	0.3397	0.9807
R	0.4569	0.2369	0.0538
EnglishCountry	0.3487	0.2639	0.1864
English	-0.9347	0.6823	0.1707
Christmas	-0.1928	0.5672	0.7339
OtherLanguage	0.3370	0.2270	0.1377
OtherDay	-0.0574	0.3099	0.8532
Action	-1.0129	0.2814	0.0003
Documentary	2.4725	1.2402	0.0462
Comedy	-0.7490	0.2734	0.0061
Musicial	0.3432	1.1698	0.7692
Drama	-0.1818	0.2746	0.5081
Family	-1.0791	0.4099	0.0085
Horror	-0.9531	0.3905	0.0147
SciFi	-0.6356	0.4441	0.1524
Biography	0.5993	0.5190	0.2482

We check the p-value of Wald statistic for each variable in the model M_{olm} which with an AIC value 756.90. Variables with p-values above 0.05 are excluded from the model M_{olm}. The new smaller model M_{op} is presented in the following Table 31.

Table 31: Results of the multivariable proportional odds model M_{op}

Parameter	Coeff.	SE	P-value
RuntimeMinutes	0.0415	0.0063	< 0.0001
Action	-0.9451	0.2417	< 0.0001
Documentary	2.5942	1.1857	0.0287
Comedy	-0.7782	0.2457	0.0015
Family	-1.0123	0.3574	0.0046
Horror	-0.8459	0.3480	0.0151

Step 3: We now compare the estimated coefficient of each explanatory variable in the smaller model M_{op} with the ones in the larger model M_{olm}.

Table 32: Results of changed estimated cofficients in the model M_{op} and model M_{olm}

Parameter	$\hat{\theta}_i$	$\hat{\beta}_i$	$\Delta\hat{\beta}_i$
RuntimeMinutes	0.0415	0.0364	0.1228
Action	-0.9451	-1.0129	0.0717
Documentary	2.5942	2.4725	0.0469
Comedy	-0.7782	-0.7490	0.0375
Family	-1.0123	-1.0791	0.0659
Horror	-0.8459	-0.9531	0.1267

Table 32 shows that the difference of the estimated coefficients between the two models is less than the indicator value 0.20. So, those excluded variable are not added back into the model M_{02}.

Step 4: Now we check the variables which are not significant in *step* 1 by adding them back to the model M_{op} once a time. We find that estimated parameters remain almost unchanged. Model M_{op} is considered as ours *preliminary main effects model*

Step 5: Model M_{op} contains six explanatory variable *RuntimeMinutes* which we need to check if it has a linear relationship with the logit, and five categorical variables *Action*, *Documentary*, *Comedy*, *Family* and *Horror*. The coefficient of variable *Documentary* is 2.5942. It means that a documentary film has a higher probability to get a high IMDB score. Estimated coefficients β_{Action} is negative, which means it has a negative association with action film and IMDB score. Similarly, the negative coefficients of variable *Comedy* mean that a comedy film has decreased probability to get high IMDB score. β_{Family} is -1.0123 which means a family film has a lower

probability of getting high IMDB score. Likewise, negatively estimated coefficient β_{Horror} means that a horror film has a low probability to get high IMDB score.

As we did in the model of film's profitability, by using the method which mentioned in section 3.6.1. We plot the estimated coefficients of categorical variable *Runtime* with four levels in the model M_{opt}, as shown in Figure 2. In Table 33, we can see the estimated coefficients of explanatory variable in model M_{opt}.

Figure 2: Plotting estimated coefficients of RuntimeCat with first interval[73, 95], second interval(95, 105], third interval(105, 119], fourth interval(119, 189] in Model M_{opt}

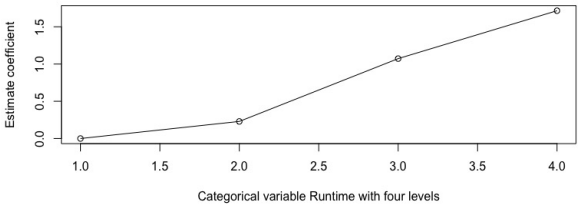


Table 33: Estimated parameters of the Model M_{opt} with the categorical variable *Runtime*

Parameter	Coeff.	SE	P-value
Runtime2	0.2294	0.2943	0.4357
Runtime3	1.0727	0.3104	0.0005
Runtime4	1.7160	0.3217	< 0.0001
Action	-1.0392	0.2454	< 0.0001
Documentary	2.0468	1.2013	0.0884
Comedy	-0.8670	0.2448	0.0004
Family	-0.9396	0.3583	0.0087
Horror	-0.9402	0.3456	0.0065

From above Table 33, we can see that categorical variables *Runtime* has 3 positive estimated parameters, and it appears to follow a linear trend in Figure 2. In addition to examining whether variable *RuntimeMinutes* should be treated as categorical, AIC value is performed. AIC value is raised from 745.13 to 758.23 by comparing model M_{op} with model M_{opt}. Summarized all the statistical results, we decide to treat *RuntimeMinutes* as a continuous variable in the model. Model M_{op} is considered our *main effects model*.

Step 6: Now we check the interactions between variables in the model M_{op}. We control 12 two-way interactions between 6 main effects in the

model M_{op}. Interactions are added back into model M_{op} once a time, but none of the two-way interactions is significant at level 0.05. So model M_{op} is our *preliminary final model*.

Step 7: After the previous step 6, we use the model M_{op} as our *preliminary final model*, which predictive capacity will be tested in section 5.2.2.

4.2.2 Model Diagnostic

The ordinal logistic regression model M_{op} contains a continuous variable *RuntimeMinutes*, so traditional goodness-of-fit such as Pearson Chi-Square statistic and the Deviance cannot be used(Agresti 2013, pp.155-157). The Lipsitz test described in section 3.5.4 was proposed to evaluate the model. For M_{op}, the value of Lipsitz test with $g = 10$ groups is equal to 3.4398, with df = 9 and the p-value is 0.9443. The difference between the observed counts and fitted values is quite small. In other words, the model M_{op} fits the data quite well.

5 Result

In this section, we discuss the predictive power of the binary logistic model of film profitability and the proportional odds model of IMDB score. These 268 films which have been released between 2010 and 2017 are used as prediction data.

5.1 The Model of a profitable film

Determined by using binary logistic regression predictors which are related to whether a film is profitable. The model can be used to predict if a movie is profitable or not. Purposeful selection and stepwise producers obtain different models. Through model comparison in the previous section, model M_{both} which has better goodness-of-fit is considered as our final model.

The selection model M_{both} contains predictors *BigCo*, *RuntimeMinutes*, *OtherLanguage*, *Drama*, *Action* and interaction terms *Action*RuntimeMinutes*, *Action*OtherLanguage*, *Action*Drama*.

Our binary logistic model:

$$\text{logit}[\pi_i(x)] = P(Y_{Profitable} \mid x_{BigCo}, ..., x_{Action})$$

$$= \alpha + \sum_{i=1}^{8} \beta_i x_{ij}$$

$$= \alpha + \beta_{BigCo} \cdot x_{BigCo} + \beta_{RuntimeMin} \cdot x_{RuntimeMin} + \beta_{OtherLanguage} \cdot x_{OtherLanguage} +$$

$$\beta_{Drama} \cdot x_{Drama} + \beta_{Action} \cdot x_{Action} + \beta_{Action*Drama} \cdot x_{Action} x_{Drama} +$$

$$\beta_{Action*RuntimeMinutes} \cdot x_{Action} x_{RuntimeMinutes} + \beta_{Action*OtherLanguage} \cdot x_{Action} x_{OtherLanguage}$$

5.1.1 Interpretation of Odds Ratio

In Table 34, the odds ratios are significantly different from 1 for model M_{both}. Also, 95% Wald confidence intervals are presented.

Table 34: Odds Ratio and corresponding 95% confidence intervals of M_{both}

	Odds Ratio estimate	95% Wald Confidence interval	
BigCo	2.8254	1.7659	4.5207
RuntimeMinutes	1.0371	1.0180	1.0566
OtherLanguage	1.0267	0.5536	1.7137
Drama	0.3385	0.1822	0.6290
Action	20.8092	1.1681	1230.76
Action:RuntimeMinutes	0.9605	0.9453	0.9760
Action:OtherLanguage	3.4722	1.07782	10.9179
Action:Drama	3.6909	1.0547	12.9165

BigCo: The estimated odds ratio of *BigCo*:

$$\theta_{\text{BigCo}} = \frac{\text{Odd}_{\text{BigCo}}}{\text{Odd}_{\text{NotBigCo}}} = e^{\beta_{BigCo}} = e^{1.0387} = 2.8254$$

which means the odds of a profitable film is estimated 2.8254 times as higher for the film is manufactured by a large production company than for the film is made by a small production company. In producing a film, a big production company is more likely to get a profit than a small production company.

RuntimeMinutes: Variable *RuntimeMinutes* was treated as continuous (see in section 4.1.1). Film duration increases by one minute, the odds of a film has profit multiply by 1.0371. Hence, the estimated odds ratios differ not so much from 1, which means that the film durations are not much different to determine that the film is profitable.

OtherLanguage: The predictor *OtherLanguage* that describes whether a film language is non-English or not. The odds ratio is 1.0267. It does not differ so much from 1 which mean that no big difference with film language to make a profit.

Drama: The odds ratio of *Drama* shows that the odds of a film's profitability is estimated 0.3385 times as lower for a drama film than for the other type of movie. A movie with different genres has a higher probability of making a profit than a drama film.

Action: The odds ratio of *Action* shows that the odds of a film which is profitable is estimated 20.8092 times higher than other types of movies. An action film has a higher probability of getting profit than other types of film. We should pay attention that the 95% confidence interval of the odds ratio is wide. There have three two-way interactions between variable *Action* and variable *RuntimeMinutes*, *OtherLanguage* and *Drama*. The odds ratio

of *Action*RuntimeMinutes* differs not so much from 1. It means that the duration of action film increases by one minute or decrease one minute has almost the same probability of making a profit. The odds ratio of *Action* OtherLanguage* estimates 3.4722, which means that action movies made in other languages have a higher probability of making profits than Enligh-speaking action movies.

The odds ratio of *Action* Drama* estimates 3.6909. It means that an action drama film has a higher probability of getting profit than a film that has other genre combined with action.

We can see from the Table 34 that if a movie is profitable, the variable *Action* is the most effective, interaction *Action*Drama*, interaction *Action* OtherLanguage* and *BigCo* also have a significant effect on a film's profitability. *RuntimMinutes*, *OtherLanguage* and interaction *Action*RuntimeMinutes* have not so much effect on a film's profitablility. Surprisingly, *Drama* hurts a film's profit.

5.1.2 Prediction

The predictive power of the binary logistic model M_{both} is discussed in this section. We try to predict the probabilities of a film to be profitable or not. We use film *Machine Gun Preacher* to illustrate model M_{both} in our forecast data set. We estimate the probability of this film which is an action, biography, crime film and produced by a big film production company, meanwhile, it displays by other languages in 129 minutes.

Let $\mathbf{X}_{Machine} = (BigCo_{Machine}, RuntimeMinutes_{Machine}, Action_{Machine}, Action{*}OtherLanguage_{Machine}, Action{*}Drama_{Machine}, Drama_{Machine}, Action{*}RuntimeMinutes_{Machine}, OtherLanguage_{Machine})$ be the values of the moive *Machine Gun Preacher*.

$$\begin{aligned}
&\text{logit}[P(\text{The film is profitable} \mid \mathbf{X}_{Machine})] \\
&= \alpha + \beta_{RuntimeMinutes} \cdot RuntimeMinutes_{Machine} + \beta_{OtherLanguage} \cdot OtherLanguage_{Machine} \\
&+ \beta_{Drama} \cdot Drama_{Machine} + \beta_{Action*RuntimeMinutes} \cdot (Action{*}RuntimeMinutes)_{Machine} \\
&+ \beta_{Action} \cdot Action_{Machine} + \beta_{Action*Drama} \cdot (Action * Drama)_{Machine} + \\
&\beta_{Action*OtherLanguage} \cdot (Action{*}OtherLanguage)_{Machine} + \beta_{BigCo} \cdot BigCo_{Machine} \\
&= 5.4023
\end{aligned}$$

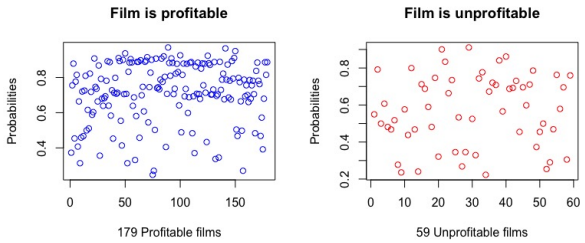
The estimated probability of film profit is calculated as follows:

$$P(\text{The film is profitable} \mid \mathbf{X}_{Machine}) = \frac{e^{5.4023}}{1 + e^{5.4023}} = 99.51\%$$

Film *Machine Gun Preacher* has 99.51% probability to have a profit which matches the film's realistic result. Furthermore, to display model

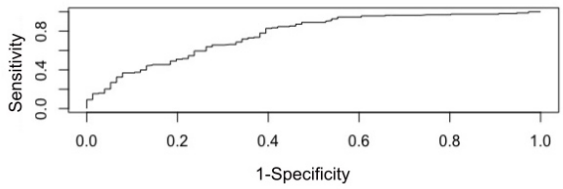
prediction capabilities, we use 268 prediction data to predict probabilities of cinema which being profitable/unprofitable. Figure 3, the predicted probabilities of films being profitable/unprofitable are plotted. We see that the probability of a film is profitable, generally higher than the film is unprofitable, but there are still profitable films with low probabilities, as well as unprofitable films with high probabilities.

Figure 3: Predict probabilities for profitable/unprofitable film



ROC-curve which is mentioned in section 3.5.6 is a way to determine predictive capacity. The corresponding ROC-curve for model M_{both} is presented in Figure 4.

Figure 4: ROC curve for model M_{both}



A receiver operating curve(ROC) which describes the accuracy of the model M_{both} is displayed above. The model M_{both} has an AUC value of 0.71. From the previous section, we know that if the AUC of a model is higher than 0.7, it indicates that our model has an *acceptable* predictive capacity (see section 3.5.6). For prediction, the model M_{both} has acceptable predictive accuracy.

5.2 IMDB score

By using purposeful selection, we determine predictors which are correlated to ordinal response variable IMDB score in the proportional odds model. We obtain the model M_{op} containing six main effects variables. Estimated parameters as shown in the following table:

Table 35: Estimated parameters in the proportional odds Model M_{op}

Parameter	Coeff.	SE	P-value
intercept > 2	-1.3232	0.7270	0.0487
intercept > 3	-3.8584	0.7393	< 0.0001
intercept > 4	-7.1173	0.8367	< 0.0001
runtimeMinutes	0.0415	0.0063	< 0.0001
Action	-0.9451	0.2417	< 0.0001
Documentary	2.5942	1.1857	0.0287
Comedy	-0.7782	0.2457	0.0015
Family	-1.0123	0.3574	0.0046
Horror	-0.8459	0.3480	0.0151

5.2.1 Interpretation of Odds Ratio

Similarly to odds ratios in the binary logistic regression, we use the same method to interpret the proportional odds ratio in the ordinal regression model.

Table 36: Table of the proportional odds ratios and corresponding 95% confidence intervals

	Odds Ratio estimate	95% Wald Confidence interval	
RuntimeMinutes	1.0423	1.0296	1.0553
Action	0.3886	0.2419	0.6241
Documentary	13.3859	1.3102	136.7562
Comedy	0.4592	0.2837	0.7433
Family	0.3633	0.1804	0.7321
Horror	0.4292	0.2169	0.8482

RuntimeMinutes: RuntimeMinutes is treated as continuous which was discussed in section 4.2.1. For one-minute increases in a film duration, the odds of obtaining an IMDB score of more than 5 in contrast to an IMDB score of less than five is 1.04 times higher. Similarly, the odds of obtaining an IMDB score of more than 6.5 in contrast to an IMDB score less than 6.5 is 1.04 times higher for one-minute increases in a film duration. The odds of achieving an IMDB score of more than 8 in comparison to an IMDB score of less than eight is 1.04 times higher for one-minute increases in a film

duration. That is, a film with a longer duration has a higher probability of getting higher IMDB score.

Action: The odds of having an IMDB score of more than 5 is 61% lower for an action film compared to another type of movie. The odds of having an IMDB score of more than 6.5 is 61% lower for an action film compared to another kind of film. The odds of having an IMDB score of more than 8.0 is 61% lower for an action film compared to another type of movie. An action film is most likely to get lower IMDB score than another kind of film.

Comedy, **Family**, and **Horror**: Odds ratio of these three predictors do not differ too much. The odds of getting a higher IMDB score is 0.45 times lower for a comedy film compared to another type of film. A family film is most likely to get lower IMDB score to compare to other types of film, such as a horror film.

Documentary: In Table 36, we can see obviously that the odds ratio of a documentary is highest. That is, A documentary film has the highest probability to get high IMDB score. The odds of getting an IMDB score of more than 5 in contrast to an IMDB score of less than 5 is 13.38 times higher. The same is between IMDB scores more than 6.5 and IMDB scores less than 6.5, such as more than 8 IMDB score and less than 8 IMDB score.

5.2.2 Prediction

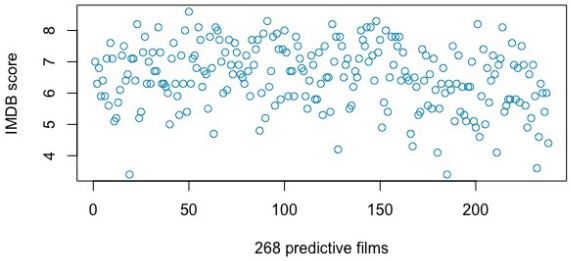
We predict probabilities by IMDB score to demonstrate the model's predictive capacity as well. We also use film *Machine Gun Preacher* which in our prediction data set to illustrate model M_{op}.

Let $\mathbf{X}_{Machine} = (RuntimeMinutes_{Machine}, Action_{Machine}, Family_{Machine}, Comedy_{Machine}, Documentary_{Machine}, Horror_{Machine})$ be the values of the moive Machine Gun Preacher. The estimated cumulative logit probability of the film *Machine Gun Preacher*(calculation see in Appendix B):

$$P(IMDBscore \leq 5 \mid \mathbf{X}_{Machine}) = 4.38\%$$
$$P(5 < IMDBscore \leq 6.5 \mid \mathbf{X}_{Machine}) = 32.21\%$$
$$P(6.5 < IMDBscore \leq 8 \mid \mathbf{X}_{Machine}) = 57.17\%$$
$$P(8 < IMDBscore \leq 10 \mid \mathbf{X}_{Machine}) = 6.24\%$$

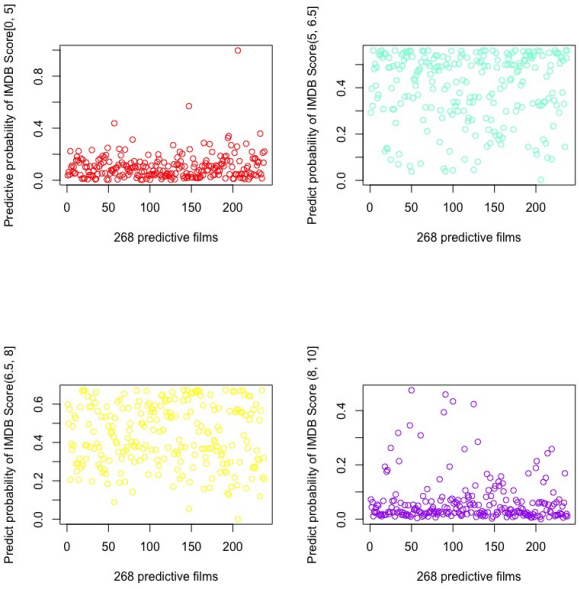
Film *Machine Gun Preacher* has the highest probability to get IMDB scores between 6.5 and 8, and lowest probability to get IMDB scores less than five which match film IMDB scores 6.8. Furthermore, we will use prediction data to predict probabilities of film IMDB score. In figure 5, we plot the IMDB score of 268 predictive films. Most of the IMDB scores are distributed between 5 and 8.

Figure 5: Plot of Total IMDB score



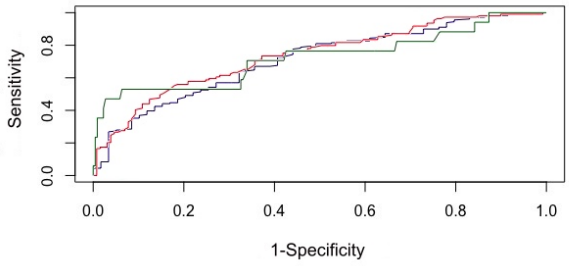
In the following Figure 6, predictive probabilities of four different categories of the IMDB score are plotted. Most films get higher probabilities to get IMDB scores in interval $(5, 6.5]$ and interval $(6.5, 8]$ than those in the interval $[0, 5]$ and interval $(8, 10]$.

Figure 6: Predictive probabilities of films' IMDB score



The area under ROC-curve describe the predictive accuracy of the model M_{op}. The proportional odds model has an ordinal response variable, so ROC-curve with three cumulative logit is plotted, as shown in Figure 7.

Figure 7: ROC curve for the model M_{op}



Based on the model M_{op}, the blue ROC curve that is plotted when category $k = 1$ is the cutoff point for a positive outcome [1-specifity, sensiivity] = $[P(Y \geq 1|x = 0), P(Y \geq 1|x = 1)]$. Likewise, the red ROC curve and green ROC curve, category $k = 2$ and category $k = 3$ respectively.

The corresponding AUC value of model M_{op} is 0.768 which exceed the acceptable criterion 0.70, so the model M_{op} has an acceptable predictive capacity.

6 Discussion

One of the purposes of this thesis is to identify the factors which are related to film profits and to build a model that can predict a film's profitability. Purposeful selection and stepwise procedure obtained three binary logistic regression models. The model obtained by stepwise selection has the highest ability to describe data and has an acceptable predictive capacity. This model contains the following variables: film production company, film duration, film language, whether a film is an action film, drama film, interaction between that whether a movie is action film and film duration, interaction between that whether a movie is action film and non-English films, interaction between that whether a movie is action film and that whether a movie is drama film. Interestingly, action films have the most significant impact on their profits among all explanatory variables. The interval of the odds ratio of variable action is broad, between 1.17 and 1230.76, which means that the benefit of different action films varies greatly. An action film with a non-English language also has a substantial effect on film profit. A drama film has a lower probability of making a profit than other types of film, but an action combined with drama film has a high probability of a film's profit. It is possible that an audience loses interest when watching a drama film, but an action combined with drama film can attract more audience's attention. That whether a production company is big also has also a substantial effect. We know that a big film production company can decide film certificate, film genre, film duration, how many stars will act in a film, if the famous/unfamous director will direct a film, so it is no surprise that a big film production company has a high probability of affecting a film's profitability. Film duration does not have any direct effect on the film profit. I find it surprising that whether a film has star(s) acting in it or not have no association with the profitability.

Does this model have a high predictive capacity to predict which film will rake in returns? We divided 268 predictive films into two parts, one part with profitable films, the other part with unprofitable films. We saw that profitable films generally have a high probability, but there are still some films have a low probability, as well as unprofitable films with extreme probabilities. To evaluate the predictive power, the ROC-curve is plotted and the value of AUC is 0.71, which means that the model has acceptable predictive ability. Summarize the above predictor, will a non-English drama action film released by a major film production company be profitable? There is a possibility that other factors are associated with a film's profitability, such as politic factor, a problem with a production company, which are not considered in the thesis.

The other aim of this thesis is to feature a model of IMDB score, which has an ordinal response variable. The same predictors are also concerned to be associated with IMDB score. Model Mop is obtained by purposeful

selection. This model has 6 predictors, which are *Action film*, *film duration*, *documentary film*, *comedy film*, *family film* and *horror film*. It is interesting that the model of the IMDB score is different compared to the model of film profit. Neither the size of film production company nor film language has an association with IMDB score. An action film has a lower probability of having a high IMDB score, although the film has high profitability. Does an audience think that is pleasing to watch an action film, but it is not rated with a high IMDB score? Similarly, a comedy film also has a low probability of having a high IMDB score, the same for a family film, a horror film. On the contrary, a documentary film has a high probability of getting a high IMDB score, but it does not associate with a film's profit. IMDB score does not differ much in film duration. To test the model's predictive capacity, we compared the realistic IMDB score with the predicted IMDB score. We found that high probabilities of IMDB scores are distributed in the interval (5, 6.5] and (6.5, 8], which is the same as the distribution of the IMDB score. The AUC value of the model is 0.768 which indicates the model has a acceptable predictive power.

The limitation of this study is that the data are just collected from the IMDB website. These movies were collected on the site, information about some films was incomplete. Therefore, the difficulty of randomly selecting data has increased, which may have an impact on when we build a model for film profit and IMDB score. We should know that IMDB scores are obtained in February 2018, which are not obtained in the same year of the movie release. Furthermore, both the 95% confidence interval for the odds ratios of Action in the logistic regression model and the confidence interval for the odds ratio of Documentary in the proportional odds model are wide. To Improve the predictive power and accuracy of the model whether a movie is profitable and its rating, we need much more films from different countries. If we continued with the study, I would suggest using an extensive data set to increase the model's predictive capacity. Meanwhile, ordinal logistic regression is also be widely used. The response variable *profit* can set to be an ordinal scale, then build a proportional odds model of a film's profit. In this way, we can get the highest probability of which level a film can make a profit.

References

[1] Statistics Portal. (2016). *Film and Movie Industry - Statistics Facts*. Retrieved February 24, 2018, from `https://www.statista.com/topics/964/film/`.

[2] Wojnar, Z. (2016). *13 Low Budget Horror Movies That Hit It Big*. Retrieved February 24, 2018, from `https://screenrant.com/low-budget-horror-movies-hits/`.

[3] Agresti, A. (2013). *Categorical Data Analysis*(3rd ed.). New Jersey: Johan Wiley & Sons.

[4] Agresti, A. (2010). *Analysis of Ordinal Categorical Data*(2nd ed.). New Jersey: Johan Wiley & Sons.

[5] Hosmer, D.W., Lemeshow, s., Sturdivant, R.X. (2013). *Applied Logistic Regression* (3rd ed.). New Jersey: Johan Wiley & Sons.

[6] Sundberg, R. (2016). *Lineära Statistiska Modeller*(3rd ed.). Stockholm: Stockholm Universitet.

[7] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Eds.). *Frontiers in Econometrics*,pp.105-142. New York: Academic.

[8] McFadden, D. (1979). Quantitative methods for analysing travel behavior of individuals: Some recent developments. In D. A. Hensher & P. R. Stopher (Eds.). *Behavioural travel modelling*, pp.279-318. London: Croom Helm.

[9] Held, L., Bové, D.S. (2014). *Applied Statistical Inference*(1st ed.). Springer-Verlag Berlin and Heidelberg GmbH Co. K.

[10] Jönsson, P.,Wohlin C. (2006): Benchmarking k-Nearest Neighbour Imputation with Homogeneous Likert Data *Empirical Software Engineering: An International Journal*, Vol. 11, No. 3, pp. 463-489. Retrieved February 24, 2018, from `http://www.wohlin.eu/emse06.pdf`.

[11] Matthew, J. (2017). *Goodness of Fit Tests for Logistic Regression Models*. Retrieved February 24, 2018, from `https://cran.r-project.org/web/packages/generalhoslem/generalhoslem.pdf`.

7 Appendix

7.1 A Tables

Table 37: Results of fitting a univariable binary logistic model with purposeful selection

Parameter	Estimate Coff.	Std.Error	P-value
RuntimeMinutes	0.0169	0.0063	0.0074
Star	0.1899	0.2632	0.4707
FamousDirector	0.7789	0.2916	0.0076
BigBudget	0.5679	0.4032	0.1590
BigCo	1.0469	0.2259	$3.58 \cdot 10^{-6}$
PG	0.4737	0.3024	0.1170
PG.13	0.1549	0.2310	0.5025
R	-0.1782	0.2197	0.417
U.S.A	0.7477	0.3958	0.0589
EnglishCountry	-0.4442	0.2657	0.0946
NonEng	-0.6785	0.2405	0.0048
English	0.3728	0.6156	0.5450
OtherLanguage	0.3447	0.2270	0.1289
Workday	0.2255	0.4433	0.6110
Otherday	-0.3017	0.3075	0.3265
SummerHoliday	-0.0768	0.2662	0.773
Christmas	0.4283	0.5295	0.418
Action	0.5943	0.2615	0.0230
Documentary	0.1510	1.2296	0.9020
Comedy	0.0770	0.2272	0.7345
Music	0.5725	0.6758	0.3970
Musical	-1.2473	1.2296	0.3100
History	0.1548	0.6223	0.804
Drama	-0.5781	0.2213	0.0089
Sport	0.3122	0.6988	0.6550
Crime	0.3986	0.2866	0.164
SciFi	-0.3007	0.4355	0.4900
Family	0.4104	0.3912	0.294
Mystery	-0.4262	0.3496	0.2230
Thriller	0.2983	0.3009	0.3220
Fantasy	0.5072	0.4281	0.236
Horror	0.2116	0.3695	0.5670
Romance	-0.07266	0.3018	0.8100
Western	14.0315	624.1938	0.9820
War	-0.5596	0.6421	0.3830
Animination	0.4637	0.4917	0.3460
Adventure	0.3103	0.2668	0.2448
Biography	-0.1461	0.4700	0.7560

Table 38: Results of Fitting the Model $M_{ModelStepboth1}$ with *Stepwise selection*

Parameter	Coeff.	SE	P-value
BigCo	1.2455	0.2936 5	$2.21 \cdot 10^{-5}$
RuntimeMinutes	0.0395	0.009855	$6.05 \cdot 10^{-5}$
Drama	-1.0801	0.3218	0.0008
NonEng	-0.4228	0.3998	0.2903
OtherLanguage	0.1079	0.3059	0.7244
Action	3.9418	1.8437	0.0325
BigCo:NonEng	-0.8237	0.5329	0.1222
Action:RuntimeMinutes	-0.0419	0.0173	0.0154
Action:OtherLanguage	1.1825	0.5909	0.0454
Action:Drama	1.3238	0.6594	0.0447

Table 39: Results of Fitting the Model *MoldelForward* with *forward selection*

Parameter	Coeff.	SE	P-value
RuntimeMinutes	0.0399	0.0098	$5.55 \cdot 10^{-5}$
BigCo	1.2498	0.2947	$2.22 \cdot 10^{-5}$
NonEng	-0.4087	0.4003	0.0.7159
OtherLanguage	0.1115	0.3063	0.7159
Drama	-1.1165	0.3239	0.0006
SciFi	-0.6628	0.5132	0.1965
Action	3.8517	1.8656	0.0389
BigCo:NonEng	-0.8537	0.5346	0.1103
RuntimeMinutes:Action	-0.0396	0.0176	0.0241
OtherLanguage:Action	1.1062	0.5952	0.0631
Dram:Action	1.1598	0.6719	0.0843

Table 40: Results of Fitting Univariable **proportional odds model**

Parameter	Estimate Coff.	Std.Error	P-value
RuntimeMinutes	0.0460	0.0059	< 0.0001
Star	0.3950	0.2451	0.1071
FamousDirector	0.5650	0.2776	0.0419
BigBudget	0.7744	0.3382	0.0221
BigCo	-0.0673	0.2024	0.7395
PG	-0.7730	0.2681	0.0039
PG.13	-0.1755	0.2045	0.3910
R	0.5233	0.2003	0.0090
U.S.A	-0.4261	0.3796	0.2616
EnglishCountry	0.3885	0.2507	0.1213
NonEng	0.1733	0.2201	0.4312
English	-0.8152	0.5875	0.1653
OtherLanguage	0.6532	0.2056	0.0015
Workday	-0.1776	0.3911	0.6497
Otherday	-0.3232	0.2725	0.2355
Christmas	0.6095	0.4948	0.2181
SummerHoliday	0.0763	0.2426	0.7530
Action	-0.4511	0.2189	0.0393
Documentary	2.3197	1.0778	0.0314
Comedy	-1.0126	0.2115	< 0.0001
Music	-0.5767	0.5933	0.3311
Musical	1.3812	1.0141	0.1732
History	0.6200	0.5583	0.2668
Drama	0.9454	0.2034	¡0.0001
Sport	0.6339	0.5716	0.2675
Crime	0.0359	0.2482	0.8850
SciFi	-0.5041	0.3966	0.2037
Family	-1.1493	0.3468	0.0009
Mystery	-0.0118	0.3225	0.9709
Thriller	0.0990	0.2622	0.7058
Fantasy	0.0100	0.3600	0.9778
Horror	-0.6700	0.3159	0.0339
Romance	0.2505	0.2636	0.3419
Western	1.4548	1.6759	0.3854
Animination	0.4178	0.4657	0.3697
Adventure	-0.1707	0.2358	0.4691
Biography	1.8172	0.4543	< 0.0001

7.2 B calculation

Calculation of probabilites of IMDB score

Comulativ logits:

$$\text{logit}[\text{P}(Z \leq k) \mid \mathbf{X}] = \alpha_k + \boldsymbol{\beta'}\mathbf{X} = \alpha_k + \beta_1 x_{i1} + \cdots + \beta_j x_{ij}. \quad k = 1, ..., c-1$$

The equivalent model expression for the cumulative probabilities is:

$$\mathbf{P}(Z \leq k \mid \mathbf{X}) = \frac{e^{(\alpha_k + \boldsymbol{\beta'}\mathbf{x})}}{1 + e^{\alpha_k + \boldsymbol{\beta'}\mathbf{x}}}. \quad k = 1, ..., c-1$$

Table 41: estimated parameters in the proportional odds Model M_{op}

Parameter	Coeff.	SE	P-value
intercept > 2	-1.3232	0.7270	0.0487
intercept > 3	-3.8584	0.7393	< 0.0001
intercept > 4	-7.1173	0.8367	< 0.0001
RuntimeMinutes	0.0415	0.0063	< 0.0001
Action	-0.9451	0.2417	< 0.0001
Documentary	2.5942	1.1857	0.0287
Comedy	-0.7782	0.2457	0.0015
Family	-1.0123	0.3574	0.0046
Horror	-0.8459	0.3480	0.0151

Test film *Machine Gun Preacher* which in our prediction data set to illustrate model M_{both}. To estimate the probability of the film *Machine Gun Preacher* which is an action, biography, crime film and is produced by a big film production company, displayed with other languages in 129 minutes. Let $\mathbf{X}_{Machine} = (RuntimeMinutes_{Machine}, Action_{Machine}, Documentary_{Machine}, Comedy_{Machine}, Family_{Machine}, Horror_{Machine})$ be the values of the moive Machine Gun Preacher.

$$\text{logit}[P(\text{IMDB Score} > 8.0 \mid \mathbf{X}_{Machine})] = \alpha_4 + \beta_{RuntimeMinutes} \cdot \text{RuntimeMinutes}_{Machine}$$
$$+\beta_{Action} \cdot \text{Action}_{Machine} = -7.1173 + 0.0415 \cdot 129 - 0.9451 = -2.7089$$

.

$$\text{logit}[P(\text{IMDB Score} > 6.5 \mid \mathbf{X}_{Machine})] = \alpha_3 + \beta_{RuntimeMinutes} \cdot \text{RuntimeMinutes}_{Machine}$$
$$+\beta_{Action} \cdot \text{Action}_{Machine} = -3.8584 + 0.0415 \cdot 129 - 0.9451 = 0.55$$

.

$$\text{logit}[P(\text{IMDB Score} > 5.0 \mid \mathbf{X}_{Machine})] = \alpha_2 + \beta_{RuntimeMinutes} \cdot \text{RuntimeMinutes}_{Machine}$$

$$+\beta_{Action} \cdot \text{Action}_{Machine} = -1.3232 + 0.0415 \cdot 129 - 0.9451 = 3.0852$$

.

Estimated cumulative probabilities of IMDB score is:

$$P(\text{IMDB Score} > 8.0) = \frac{e^{-2.7089}}{1 + e^{-2.7089}} = 6.24\%$$

$$P(\text{IMDB Score} > 6.5) = \frac{e^{0.55}}{1 + e^{0.55}} = 63.42\%$$

$$P(\text{IMDB Score} > 5) = \frac{e^{3.0852}}{1 + e^{3.0852}} = 95.63\%$$

The cell probabilities can also be estimated:

$$P(\text{IMDB Score} > 8.0) = 6.24\%$$

$$P(6.5 < \text{IMDB Score} \leq 8.0) = 63.42\% - 6.24\% = 57.18\%$$

$$P(5.0 < \text{IMDB Score} \leq 6.5) = 95.63\% - 63.42\% = 32.12\%$$

$$P(\text{IMDB Score} \leq 5.0) = 1 - 95.63\% = 4.37\%$$