UNIONE EUROPEA
Fondo sociale europeo

FSE 2007-2013
POR SARDEGNA
obiettivo competitività regionale e occupazione

Università degli Studi di Sassari

Essays on Empirical Cultural Economics

Direttore della scuola di Dottorato:

Prof. Michele M. Comenale Pinto

Tutor:

Prof. Gerardo Marletto

Tesi di Dottorato di Ricerca di:

Gianpiero Meloni

La presente tesi è stata prodotta durante la frequenza del corso di dottorato in Diritto ed Economia dei sistemi produttivi dell'Università degli Studi di Sassari, a.a. 2011/2012 - XXVII ciclo, con il supporto di una borsa di studio finanziata con le risorse del P.O.R. SARDEGNA F.S.E. 2007-2013 - Obiettivo competitività regionale e occupazione, Asse IV Capitale umano, Linea di Attività I.3.1.

I've seen things you people
wouldn't believe

Blade Runner

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Movie manufacturing received a lot of attention from academic research in recent years. The role of the market is prominent in the global media industry and its size is increasing: according to PwC[1], the worldwide revenue will grow from 38 billion U.S. dollars in 2014 to nearly 46 billion in 2018.
For a practitioner this industry present different aspects of great interest: DeVany (2004) found that the relationship between a motion picture's cost and revenue is wildly unpredictable compared to other investments due to the heterogeneity in movie performance with box-office revenues exhibiting heavy right tails. In the words of the author:

> The movie industry is a profoundly uncertain business. The probability distributions of movie box-office revenues and profits are characterized by heavy tails and infinite variance! It is hard to imagine making choices in more difficult circumstances. Past success does not predict future success. Forecasts of expected revenues are meaningless because the possibilities do not converge on a mean; they diverge over the entire outcome space with an infinite variance. This explains precisely why "nobody knows anything" in the movie business.

A broader motivation for studying motion pictures is that the vast majority of empirical work on trade is for manufacturing, with relatively little work on services. Exports of services such as motion pictures are distinct from exports of manufactures in that variable production costs (e.g., exhibiting movies to consumers) are incurred in the country of consumption, rather than the country of production, and the physical cost of transporting goods abroad (e.g., shipping master film prints) is close to nil. For motion pictures and other cultural goods, cross-country differences in language, social mores, or religion may be the significant barriers to trade (Rauch and Trindade, 2009)[2].

[1]Source: http://www.pwc.com/gx/en/global-entertainment-media-outlook/segment-insights/filmed-entertainment.jhtml

[2]Excerpt from Hanson and Xiang (2011), "Trade barriers and trade flows with product heterogeneity: An application to US motion picture exports". *Journal of International Economics* 83, pag 15.

Aim of the first part of my dissertation, is to highlight the impact of cultural differences among importing countries of American movies and how they influence box-office revenues. Previous literature observed the arrival of movies in a country as given (see Lee, 2009) or focused on the determinants of revenues implementing complex econometric procedures like gravity models (Hanson and Xiang 2011, Marvasti and Canterbery 2005) using cultural differences as explanatory variables.
My contribute to the literature comes from the application of the Heckman's (1979) two-step methodology to infer the probability of arrival of American movies in foreign countries and then to evaluate the box-office revenues for different clusters of nations built around Hofstede's (2001) index of cultural distance from the United States and Human Development data (HDI). To do so, I built a data set of 1341 US movies exhibited in 50 countries over the 2002-2013 period: for each of them I collected information on box-office, production budget and idiosyncratic characteristics like genre, sequel, source, Academy Awards nominations and MPAA rating[3].
To sum up the findings discussed in length in Chapter 2, estimation results suggest different strategies to sell Hollywood movies around the world. In general, countries with relatively high HDI and that are close to the American culture tend to be less affected by measures of quality of a movie and show special preferences for action titles. Although the estimation is subject to an identification problem due to the negative correlation between HDI and cultural differences, it suggests that is mostly cultural distance, and not the Human Development level of a country, that leads the consumption in this group of nations. Further, the estimation indicates that once a movie is introduced in a country with low HDI or high cultural distance they are relatively more faithful to the following sequels of these movies.
While Hollywood is the dominant agent in the worldwide movie industry, other markets have a great importance for their size (like India and China in recent years) or their history. The second part of my thesis is devoted to the

[3]Movie ratings from Motion Picture Association of America provide parents with advance information about the content of movies to help them determine what's appropriate for their children. I use this information as a proxy of the content of a movie in terms of violence, sex etc.

analysis of the Italian movie production market which is one of the oldest in the industry and renowned worldwide for its quality. Academic research on Italian domestic market is, to my knowledge, limited to the contribute of Bagella and Becchetti (1999). With a descriptive and econometric analysis on box office performances of movies produced in Italy between 1985 and 1996, they focused on the relationship between popularity of human inputs and the relative impact of state subsidization on box-office revenues.

Using a GMM-HAC [4] approach they find that the ex-ante popularity of human inputs (directors and actors) affects in a nonlinear way box-office performance and the interaction between the two factors' popularity has a positive impact on total admissions. Moreover, authors find that the subsidized films do not have a significantly lower performance in the econometric analysis of total admissions and the net effect of subsidies on the mean of the dependent variable is irrelevant.

In chapter 3, *La Grande Bellezza*, I focus on the impact of state subsidization and try to confirm the results from Bagella and Becchetti looking not only at box-office performances of subsidized movies, but also on their quality. In particular, I collected data for 754 Italian movies exhibited during the 2002-2011 period gathering information on amount of subsidization, genres and festivals presence and related awards granted. Results of box-office revenues estimation using subsidization as a single variable are coherent with the findings of Bagella and Becchetti, showing an overall slightly negative impact and its net effect is negligible. However, when introducing interaction between subsidies and genres the sign of coefficients turn to be positive and suggest that, while weak, the net impact on public financing on the Italian movie industry is affecting positively box-office revenues.

I find these results to be consistent with the econometric analysis of awards won at festivals where, using a standard Poisson model, iteration variables between genres and subsidies have a slightly positive incidence on the ratio of awards granted. Overall, could be said that public financing is, at least, not hurting the sector and as a policy indication my argue to a better performance of public expenditure is to shift subsidization on dramas and thrillers

[4] Generalized Method of Moments Heteroskedasticity and Autocorrelation Consistent.

movies leaving comedies outside of intervention[5].

[5]As shown in Bagella and Becchetti (1999), Italian moviegoers have a strong preference for comedy movies and the impact of the genre on box-office revenues reflect this bias.

Chapter 2

American Beauty

2.1 Introduction

The field of motion picture industry has received great attention by academic research in the recent period (see McKenzie (2012) for a detailed review); in particular, big focus is given to the transnational flow of motion pictures and the changes in the worldwide supply of movies. In 2007, the New York Times noted that 'American movies (and music) have done very well in some countries like Sweden and less in others like India'. Today, economic growth is booming in countries where American popular culture does not dominate, namely India, China and Russia. Moreover, population growth is strong in many Islamic countries, which typically prefer local culture. Nevertheless, some countries that seem little permeable to foreign cultures, are now experiencing an aperture to international movies. After less than three years, The Wall Street Journal (2010) puts in evidence the significant rise of the international box-office and find that this turnabout depends on the fact that 'one of most American of products is now being retooled to suit foreign tastes'. For The Economist (2011) this growth is partially a result of the dollar's weakness, but it also depends from three crucial aspects: a boom in the demand of movies in the emerging world, a concerted effort by the major studios to produce movies that might play well abroad and a global marketing push to ensure this goal.
One of the first economics studies in the field comes from Prag and Casavant (1994) that present an empirical study of the determinants of a motion pictures financial success using a dataset of 652 over a large time period, where a subset of these (195 movies) also have data on advertising expenditures. Among the many factors which are included in this study, results that quality and marketing expenditures are important determinants. Film ratings, production cost, and the presence of star performers are only important determinants when marketing is not included. Marketing expenditures are positively related to production costs, winning Academy awards and the presence of major stars. Looking for simple correlations, the paper states that there is no evidence of a positive relationship between cost of production and film quality. Also, the only genre dummy which is significant is

that for dramas and indicates that being a drama is a negative factor for film revenues. Contrary to popular wisdom, PG13 and R rated films do not perform better at the box office.

De Vany and Walls (1999) tried to consider the mathematical properties of box office revenue and estimate profit data of 2.015 movies released between 1985 and 1996 for the United States and Canada and found that is impossible to attribute the success of a movie to individual casual factors. They evaluate the impact of budget, actors and director power, sequels, genre, rating and release year on "hit" probability, where a hit is defined as a movie grossing over US dollars 50 million. They show that the audience reception (captured by a dummy variable for films lasting more than ten weeks) is the most important variable in determining box office revenues. Consequently, they reject forecasting models of box-office revenues.

Marvasti and Canterbery (2005) observe that the American industry motivation for seeking foreign markets is found in his domestic box-office, average costs and industry structure: with remarkably rapid production costs increase, differentiation trough export can ensure economies of scale. This can be explained with the fact that most of the marketing costs are incurred for exports that potentially add much more to revenues than to costs. The authors state that because of the apparent dependency of domestic box office on a high level of circulating capital is easy to understand the predominance of United States in the movie industry given the difference in gross GDP compared to competitors. This predominance is narrowing due to the changes in GDP observed in countries like China and India. Using an annual pooled cross-section dataset of 33 countries over the period 1991-1995 and developing a complex iceberg-gravity model, the authors study the impact of cultural and trade barriers to US movies export. They find that despite substantial barriers to film imports, including the low percentual of English-speaking population in the sampled countries, other large economies apparently have been unable to internationally extend their domestic markets. However, data from more recent years, show that this equilibrium is changing and new movie industries are emerging (like China and India). Actually the greatest US barrier to foreign competition appears to be the gi-

gantic production and marketing costs required to produce the kind of films now demanded around the world.
This chapter aims to study the production function of American movies that sell in foreign markets focusing on the cultural differences among importing countries and the United States. Trade patterns between two countries are usually justified by aspects like national income, which has a positive impact, and distance between the two, which impacts negatively[1]. However, in recent years, also cultural proximity has been considered like a potential key on international flow between countries. The literature has used different variables to proxy cultural ties, such as common language (Melitz, 2008) or religion (Frankel, 1997). Guiso, Sapienza and Zingales (2009) suggest that 'perceptions rooted in culture are important (and generally omitted) determinants of economic exchange'. In their paper, they show that cultural biases affect economic exchange between countries. This trust is affected not only by the characteristics of the country being trusted, but also by cultural aspects of the match between trusting country and trusted country. The authors highlight the effect of trust on bilateral trade in goods, financial assets, and direct foreign investment. Felbermayr and Toubal (2010), similarly, find that cultural proximity is an important determinant of bilateral trade volumes of European countries, where cultural distance is measured by bilateral score data from Eurovision Song Contest. Grinblatt and Keloharju (2001) documents that investors are more likely to hold, buy, and sell the stocks of (Finnish) firms that are located close to the investor, that communicate in the investor's native tongue, and that have chief executives of the same cultural background. As stated by McKenzie and Walls (2012):

> A substantial body of work that tests the cultural discount hypothesis in the context of the motion-picture industry has evolved over the past decade. Many studies, rely on aggregate or macro-level data to test the cultural discount model. For example, Fu and Sim (2010) and Oh (2001) examined international trade in films and find support for the cultural discount hypothesis.

[1]See the gravitity model starting form the seminal work of Tinbergen (1962)

> Jayakar and Waterman (2000) examined U.S. film exports, and S. W. Lee (2002) examined competitive balance of film trade between the United States and Japan, both studies finding support for the cultural discount hypothesis of media market dominance. Most of the recent studies leverage highly disaggregated film-level data through the application of modern econometric analysis. Fu and Lee (2008) examined the market for films in Singapore, F. L. F. Lee (2006) examined the market for films in Hong Kong, and F. L. F. Lee (2008, 2009) examined the cultural discount hypothesis in a number of East Asian countries. The film-level research uniformly finds evidence of cultural discount in the particular East Asian motion-picture markets under study.

For the purpose of my analysis, the paper from Lee (2009) is particularly relevant as reported in McKenzie (2012): "the author takes an international/cultural perspective on the role of Academy Awards on motion picture demand. Using nominations and awards as indicators of cinematic achievement he investigates the relationship between such achievement and a sample of US films' box office revenues in nine East Asian countries. In the analysis he makes a distinction between 'drama' awards (e.g. best director, best leading and supporting actor/actress, best screenplay and best film editing) and 'non-drama' awards (all other awards) to investigate how films defined in these respects may have cross-cultural appeal. Using data on the top 100 US movies from 2002 to 2007, the results show that non-drama awards relate positively to box office revenues, but drama awards show negative correlations. The interpretation of such results is that films with culturally specific (American) storylines do not translate for East Asian audiences as well as films which, for example, might contain relatively more special effects. Further, he finds that the negative relationship of drama awards and East Asian box office appears more pronounced in countries less culturally similar to the USA in terms of a culture similarity index." In his analysis, Lee implemented an approach coming from international management and international trade literature, the Hofstede's (1980, 2001) cultural dimensions theory, which is a framework for cross-cultural communication that describes the effects of a society's culture on the values of its members, and how these values relate to behavior. As we will see later, I'll use this framework to group countries in two clusters based on the cultural distance from the United States

2.2 Data

I collected information about different features including production budget (adjusted for inflation) of 1431 American movies over the period 2002-2013 as well as information on box-office performance for each of them in 50 countries. I focused on this time span, due mainly to the precision of data on foreign markets revenues available on boxofficemojo.com. To measure performance at box-office for each movie I look at its features splitting in two type of variables: quality variables like budget and Academy Awards nominations in one or more of the main categories[2] and variables of the idiosyncratic features of each title, in particular genres, MPAA ratings[3] , sequels and sources. I gathered these data from opusdata.com and imdb.com, while data for the production budget comes from thenumbers.com[4].Table 1 shows the descriptive statistics for the variables under analysis. Note that excluding production budget, all others are dummy variables.

The lowest production budget in the sample belongs to the movie *Paranormal Activity* (16.3 thousands US dollars) which is the most profitable movie ever made in terms of return of investment thanks to a box-office revenue of nearly 200 millions dollars. The biggest budget, 337 millions dollars, corresponds to the third installment in the *Pirates of the Caribbean* franchise. Table 1 also shows that the most representative genre is comedy (32% of movies in the sample) which is used as reference category in the estimations.

[2]Best movie, best director, best actor or actress in a leading or supporting role or best animation movie.

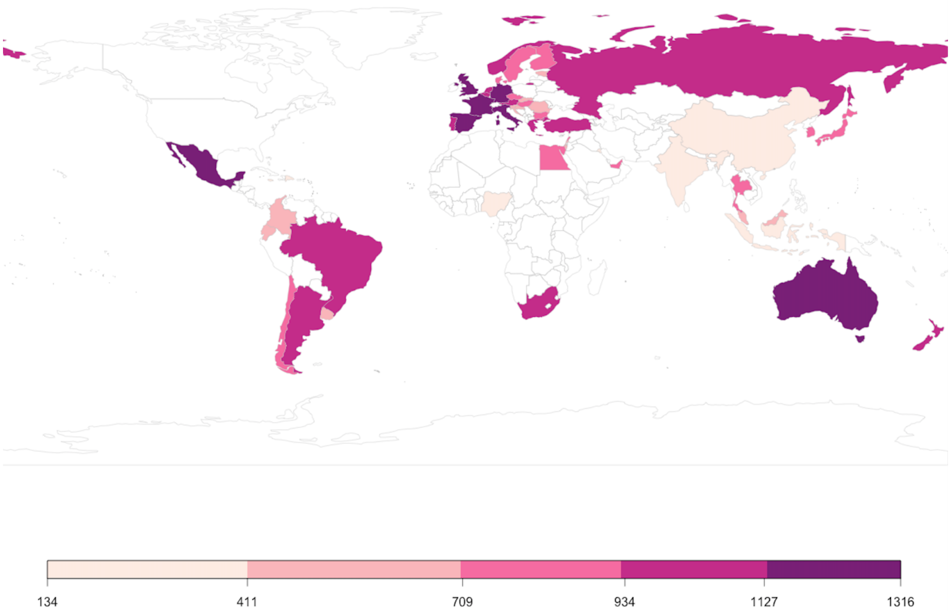
[3]Movie ratings provide parents with advance information about the content of movies to help them determine what's appropriate for their children and it is used in the sample as a proxy of the content of a movie in terms of violence, sex etc. *G* stands for General Audiences; *PG* stands for Parental Guidance Suggested; *PG-13* stands for Parents Strongly Cautioned; *R* stands for Restricted, Under 17 requires accompanying parent or adult guardian"

[4]See http://www.the-numbers.com/movie/budgets/all

Table 2.1: Movies Descriptive Statistics

variable	**mean**	**std. deviation**	**min**	**max**
budget (adjusted)	57077316	53545640	16290	336900000
genres				
drama	0.230	.4213692	0	1
comedy	0.319	.4661191	0	1
action	0.277	.447538	0	1
thriller	0.174	.3792456	0	1
ratings				
G	0.023	.1501491	0	1
PG	0.166	.3718677	0	1
PG-13	0.421	.4938416	0	1
R	0.390	.4880635	0	1
sequel	0.116	.3203402	0	1
sequel2	0.062	.2415925	0	1
nomination	0.121	.3261186	0	1
original screenplay	0.497	.5001649	0	1
Observations	1431			

Figure 2.1: Movies exhibited per country



Besides, sequels and franchises account for 18% of the titles in the dataset and an overall 12% received an Academy Awards nomination.
In the second row of Table 2 we can see the number of movies exhibited in each of the 50 countries of our sample, where not surprisingly the nation with more imported film is United Kingdom, followed by Spain, Australia and Germany. The same information is displayed graphically in figure 1. The fourth column shows the mean value over the sample period of the freedom of trade as measured by the Index of Economic Freedom from the Heritage Foundation which constitutes the instrumental variable I use in the probit model to estimate the probability of arrival of a movie i in a country j.
In the next section we will see the estimation results for several clusters of countries around two dimensions: Human Development Index (HDI) and cultural distance (CD). In particular, for the sake of symmetry the estimation sample is split in two groups depending on the position of each country with respect to the median of the average value of the HDI during the period of analysis and cultural distance, which is fixed through time, which drives to 25 countries for estimation in each of the clusters. The mean values of these

variables and the relative clusters are shown in columns 5-8 of Table 2. The HDI is an index created by the United Nations and summarizes measure of average achievement in key dimensions of human development: standards of leaving, education and life's expectation and quality. According to the Human Development Index technical notes in our sample there is one low developed country (Nigeria, $< .500$), ten medium developed nations ($>= .500$ and $< .700$), thirteen high developed ($>= .700$ and $< .800$) and twenty-six very highly developed countries ($>= .800$). To capture information about the cultural distance between U.S.and each of the countries in the dataset, following Lee (2009), I implement a value-based index developed by Hofstede (1980) built around four dimensions: 1) *power distance*, that expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally; 2) *uncertainty avoidance* expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity; 3) *individualism versus collectivism* and 4) *masculinity versus femininity*[5]. I gathered data for each country and values from Hofstede (2001) and then each country's cultural distance from the United States is computed using Kogut and Singh's (1988) formula:

$$CD_j = \sum_{I=1} \left\{ (I_{ij} - I_{iu})^2 / V_i \right\} / 4$$

Where CD_j is the cultural distance of country j from the United States, I_{ij} is the value for country j on the ith cultural dimension (I_{iu} for the U.S.) and V_i is the variance of the ith cultural dimension. A simple correlation analysis shows that there is a positive correlation between HDI and the number of movies exhibited in a country that is also negatively correlated with the cultural distance from US. Besides, there is a negative correlation between HDI and CD, in fact in a regression between these variables, the coefficient associated to cultural distance is -0.02 with a p-value equal to 0.048. Figure 1 shows the values of these two variables for the 50 countries.

[5]See section I of the Appendix for a briefer description of this index and all other variables included in the dataset.

Figure 2.2: Human Development Index versus Cultural Distance

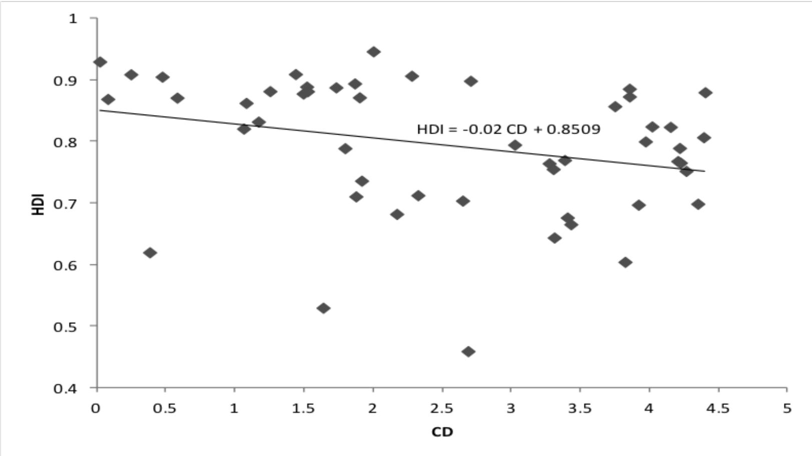


Table 2.2: Countries Descriptive Statistics: General

country	movies	%	free trade	HDI	cl	CD	cl
Argentina	1033	72.2	61.4	0.784	1	1.795	1
Australia	1262	88.2	77	0.927	2	0.02	1
Austria	1059	74	87.5	0.873	2	1.493	1
Belgium	1044	73	81.4	0.889	2	1.518	1
Brazil	985	68.8	69.2	0.71	1	2.323	1
Bulgaria	856	59.8	85.8	0.758	1	3.273	2
Chile	934	65.3	79.2	0.793	1	3.97	2
China	182	12.7	50.6	0.665	1	3.43	2
Colombia	709	49.5	72.2	0.698	1	3.919	2
Croatia	692	48.4	87.5	0.797	1	3.024	2
Czech Republic	897	62.7	87.5	0.857	2	1.08	1
Denmark	831	58.1	79.8	0.893	2	1.866	1
Dominican Republic	134	9.4	80	0.678	1	2.17	1
Ecuador	639	44.7	62.8	0.703	1	4.349	2
Egypt	795	55.6	73.9	0.638	1	3.31	2
Estonia	487	34	82.4	0.833	2	1.169	1
Finland	918	64.2	87.1	0.876	2	1.252	1

Table 2.2: Countries Descriptive Statistics: General

country	movies	%	free trade	HDI	cl	CD	cl
France	1214	84.8	81	0.877	2	1.524	1
Germany	1236	86.4	86	0.899	2	0.472	1
Greece	956	66.8	80.2	0.85	2	3.749	2
Hungary	853	59.6	87.1	0.815	2	1.062	1
India	374	26.1	24.2	0.526	1	1.637	1
Indonesia	404	28.2	74.6	0.598	1	3.822	2
Israel	543	37.9	77.1	0.887	2	1.73	1
Italy	1215	84.9	86.8	0.865	2	0.58	1
Jamaica	151	10.6	70.4	0.705	1	1.874	1
Japan	761	53.2	80.6	0.895	2	2.704	2
Kuwait	151	10.6	77.8	0.786	1	4.217	2
Lebanon	877	61.3	80.5	0.738	1	1.916	1
Malaysia	661	46.2	73.4	0.754	1	4.264	2
Mexico	1230	86	57.6	0.749	1	3.303	2
Netherlands	1085	75.8	87.5	0.907	2	1.437	1
New Zealand	1127	78.8	84.6	0.905	2	0.246	1
Nigeria	411	28.7	61.6	0.465	1	2.686	2
Norway	965	67.4	89.2	0.943	2	2	1
Portugal	1006	70.3	79.8	0.806	1	4.391	2
Romania	617	43.1	86	0.767	1	4.204	2
Russia	1049	73.3	62.6	0.76	1	4.223	2
Singapore	875	61.1	85	0.876	2	3.854	2
Slovakia	635	44.4	87.1	0.825	2	4.151	2
Slovenia	654	45.7	86.5	0.88	2	4.402	2
South Africa	1095	76.5	76.3	0.622	1	0.382	1
South Korea	825	57.7	73.6	0.882	2	3.854	2
Spain	1302	91	79.8	0.867	2	1.9	1
Sweden	883	61.7	87.1	0.906	2	2.278	1
Thailand	792	55.3	75.9	0.676	1	3.405	2

Table 2.2: Countries Descriptive Statistics: General

country	movies	%	free trade	HDI	cl	CD	cl
Turkey	965	67.4	73.7	0.697	1	2.647	2
United Arab Emirates	915	63.9	75	0.824	2	4.017	2
United Kingdom	1316	92	87.6	0.865	2	0.079	1
Uruguay	674	47.1	83	0.776	1	3.385	2
Observations	41274						

Table 2.3: Countries Descriptive Statistics: Boxoffice

country	mean	sd deviation	min	max
Argentina	1251744.193	2185664.781	1464	23556116
Australia	6519565.82	8429739.898	1605	114876536
Austria	1295698.231	1710308.112	2330	15339482
Belgium	1400454.826	1911476.309	1729	14744213
Brazil	4038623.611	6616694.647	1374	64863380
Bulgaria	158687.731	228226.804	1343	3699024
Chile	653480.889	1163902.986	2808	11394242
China	16771464.805	26707876.414	8599	197911296
Colombia	1063486.091	1670273.202	2067	14791967
Croatia	157600.695	217043.118	2346	2740992
Czech Republic	439665.291	815728.457	1635	13541869
Denmark	1281129.148	2207207.449	1338	24857536
Dominican Republic	120134.643	165901.79	2269	1420529
Ecuador	393628.831	582329.57	1329	5491452
Egypt	128567.282	175055.301	1526	2482628
Estonia	97227.909	131733.654	1775	1313026
Finland	646796.701	1006001.645	2438	9016192
France	7835170.936	12445908.565	1350	171871504
Germany	7706033.923	13200710.352	2332	171115344

Table 2.3: Countries Descriptive Statistics: Boxoffice

country	mean	sd deviation	min	max
Greece	997674.874	1290114.475	2808	12775577
Hungary	500500.336	661597.267	2161	7928378
India	1349806.032	2778435.522	2665	26299510
Indonesia	1093129.397	1641125.635	30073	15308150
Israel	898788.255	1083924.519	11016	14224082
Italy	4790454.54	7019764.23	1337	90679032
Jamaica	75117.659	101000.584	1446	671902
Japan	13064651.172	24489015.708	9447	243522416
Kuwait	281045.795	256964.726	9094	1947920
Lebanon	129239.599	165178.266	1562	1834933
Malaysia	1113718.557	1536392.376	2557	13909003
Mexico	4885422.433	7023475.08	11565	63420020
Netherlands	1839929.09	2768811.902	1150	25723286
New Zealand	980112.781	1264528.786	2382	13669993
Nigeria	73771.746	232878.517	1720	3930188
Norway	1212640.313	1883928.63	2096	15774829
Portugal	807240.662	973256.034	1844	10094128
Romania	198639.641	344005.18	1585	6034680
Russia	5372056.795	8556759.968	5895	127174120
Singapore	935230.295	1238111.158	5242	11339837
Slovakia	154889.989	265665.508	1172	3167044
Slovenia	149104.451	176354.553	2430	1981953
South Africa	738156.474	851264.411	2042	8449588
South Korea	5270132.872	9576162.332	1306	114557272
Spain	5601730.244	7409987.907	1859	119452120
Sweden	1612282.091	2754706.609	1372	30540692
Thailand	852686.848	1239324.887	1139	9138544
Turkey	976549.552	1354475.709	4523	19075024
Utd Arab Emirates	628826.708	814060.978	1453	7939070

Table 2.3: Countries Descriptive Statistics: Boxoffice

country	mean	sd deviation	min	max
United Kingdom	12945515.414	19378329.138	3899	163594016
Uruguay	111337.734	151339.844	813	1298218

2.3 Methodology

The non-random nature of the sample population put some challenges for the empirical analysis. The principal issue is that not every movie is shown in each country and as a consequence the panel is not balanced. This could lead to endogenous sample selection resulting in inconsistent estimates of the coefficient if, within a model of the revenue performance, the variables that affect the probability of arrival of a movie in a certain country are also highly correlated with the revenue of that movie. The best practice in this kind of scenario is to implement the Heckman's (1979) two-step estimation procedure. This section provides the theoretical background of the sample selection problem starting from its two principal components, truncated and censored distributions[6].

2.3.1 Truncation

A truncated distribution is the part of an untruncated distribution that is above or below some specified value. For instance, in the sample under analysis, I could subset the population cutting off those movies with a production budget below one million dollars. In order to derive the first and second moments of the truncated distribution we must introduce first the density of a truncated random variable.

Theorem 1 (Density of a Truncated Random Variable)**.** *If a continuous random variable x has a probability density function $f(x)$ and a is a constant,*

[6]Notation and order of the arguments follow Greene (2003) *Econometric Analysis - Sixth Edition*, Pearson Education.

then

$$f(x|x > a) = \frac{f(x)}{Prob(x > a)}$$

If x has a normal distribution with mean μ and standard deviation σ, then

$$Prob(x > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

where $\alpha = (a - \mu)/\sigma$ and $\phi(\cdot)$ is the standard normal cumulative density function. The density of the truncated normal distribution is then

$$f(x|x > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/(2\sigma^2)}}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{a-\mu}{\sigma}\right)}{1 - \Phi(\alpha)}$$

where $\phi(\cdot)$ is the standard normal pdf.
We can now derive the moments of a truncated normal distribution as follows:

Theorem 2 (Moments of the Truncated Normal Distribution)**.** *If* $x \sim N[\mu, \sigma^2]$ *and* a *is a constant, then*

$$\begin{aligned} E[x|truncation] &= \mu + \sigma\lambda(\alpha), \\ Var[x|truncation] &= \sigma^2[1 - \delta(\alpha)], \end{aligned}$$

where $\alpha = (a - \mu)/\sigma$*,* $\phi(\alpha)$ *is the standard normal density and*

$$\lambda(\alpha) = \frac{\phi}{1 - \Phi(\alpha)} \text{ if truncation is } x > a,$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha] \text{ with } \delta \in (0, 1).$$

The function $\lambda(\alpha)$ called the inverse Mills Ratio[7], named after John P. Mills, as we can see from the formula is the ratio of the probability density function to the cumulative distribution function of a distribution. Its role is crucial in the following regression analysis to take in account of possible

[7]it is also called the hazard function for the standard normal distribution.

selection bias, as proposed by Heckman (1979).

2.3.2 Censored data

Censoring is a condition in which the value of an observation is only partially known. When the dependent variable is censored, values in a certain range are all transformed to a single value. This point will be useful in the next subsection when, modeling the probability of exhibition of a movie in a certain country, we define the selection variable z^*. Meanwhile, let derive the censored normal distribution as we did for truncation. Define a new random variable y as a transformation of the original one y^* by

$$\begin{aligned} y &= 0 \; if \; y^* \leq 0 \\ y &= y^* \; if \; y^* > 0 \end{aligned}$$

The distribution that applies if $y^* \sim N[\mu, \sigma^2]$ is

$$Prob(y = 0) = Prob(y^* \leq 0) = \Phi\left(-\frac{\mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu}{\sigma}\right),$$

and, if $y^* > 0$, then y has the density of y^*.

Theorem 3 (Moments of the Censored Normal Variable)**.** *If $y^* \sim N[\mu, \sigma^2]$ and $y = a$ if $y^* \leq a$ or else $y = y^*$, then*

$$\begin{aligned} E[y] &= \Phi a + (1 - \Phi)(\mu + \sigma\lambda) \\ Var[y] &= \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi] \end{aligned}$$

where

$$\Phi[\frac{(a - \mu)}{\sigma}] = \Phi(\alpha) = Prob(y^* \leq a) = \Phi, \; \lambda = \frac{\phi}{(1 - \Phi)} \; and \; \delta = \lambda^2 - \lambda\alpha.$$

2.3.3 Sample selection (incidental truncation)

Many samples are truncated on the basis of a variable that is correlated with the dependent variable. For example, let assume that the international distribution of a movie is set by the production companies only that choose to export a film in a country if the expected revenue exceeds their reservation revenue and choose to stay out of that country otherwise[8]. If the dependent variable (box-office revenues in my sample) is correlated with the difference between reservation and expected revenues, least squares yields inconsistent estimates. In this case, the sample is said to have been selected on the basis of this difference.
Suppose that y and z have a bivariate distribution with correlation ρ. With respect to the previous example, we are interested in the distribution of y given that z exceeds a particular value. As before, we are interested in the form of the incidentally truncated distribution and the mean and variance of the incidentally truncated random variable.
The truncated joint density of y and z is

$$f(y, z|z > a) = \frac{f(y, z)}{Prob(z > a)}$$

To obtain the incidentally truncated marginal density for y, we would then integrate z out of this expression. The moments of the incidentally truncated normal distribution are given in the following theorem.

Theorem 4 (Moments of the Incidentally Truncated Bivariate Normal Distribution)**.** *If y and z have a bivariate normal distribution with means μ_y and μ_z, standard deviations σ_y and σ_z, and correlation ρ, then*

$$\begin{aligned} E[y|z > a] &= \mu_y + \rho\sigma_y\lambda(\alpha_z), \\ Var[y|z > a] &= \sigma_y^2[1 - \rho^2\delta(\alpha_z) \end{aligned}$$

[8]Note that this kind of assumption is not realistic as it neglect the importing decision for each country.

where

$$\alpha_z = \frac{(a-\mu_z)}{\sigma_z},\ \lambda(\alpha_z) = \frac{\phi(\alpha_z)}{[1-\Phi(\alpha_z)]}\ and\ \delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z)-\alpha_z]$$

We are now able to derive a general framework, let

$$z_i^* = \boldsymbol{w'}_i\boldsymbol{\gamma} + u_i$$

be the equation that determines the sample selection and

$$y_i^* = \boldsymbol{x'}_i\boldsymbol{\beta} + \varepsilon_i$$

the equation of primary interest. ε_i and u_i have bivariate normal distribution with zero means and correlation ρ. The sample rule is that y_i is observed only when $z_i^* > 0$. Applying Theorem 4 we obtain the model

$$\begin{aligned}
E[y_i|y_i \text{ is observed}] &= E[y_i|z_i^* > 0] \\
&= E[y_i|u_i > -\boldsymbol{w'}_i\boldsymbol{\gamma}] \\
&= E[\boldsymbol{x'}_i\boldsymbol{\beta} + \varepsilon_i|u_i > -\boldsymbol{w'}_i\boldsymbol{\gamma}] \\
&= \boldsymbol{x'}_i\boldsymbol{\beta} + E[\varepsilon_i|u_i > -\boldsymbol{w'}_i\boldsymbol{\gamma}] \\
&= \boldsymbol{x'}_i\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda_i(\alpha_u) \\
&= \boldsymbol{x'}_i\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u)
\end{aligned}$$

where

$$\alpha_u = \frac{-\boldsymbol{w'}_i\boldsymbol{\gamma}}{\sigma_u} \text{ and } \lambda(\alpha_u) = \frac{\phi(\boldsymbol{w'}_i\boldsymbol{\gamma})/\sigma_u}{\Phi(\boldsymbol{w'}_i\boldsymbol{\gamma})/\sigma_u}$$

then

$$\begin{aligned}
y_i|z_i &= E[y_i|z_i > 0] + \upsilon_i \\
&= \boldsymbol{x'}_i\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u) + \upsilon_i
\end{aligned}$$

If λ is omitted the specification error of an omitted variable is committed and the OLS regression produces inconsistent estimations of β.
It is important to note that the selection variable z^* is not observed, rather we

observe only its sign, but not its magnitude. The absence of information on the scale of z^* implies that the disturbance variance in the selection equation cannot be estimated. Thus, a reformulation of the model is mandatory. Let assume that z_i and $\boldsymbol{w}_i$ are observed for a random sample of observations but y_i is observed only when $z_i = 1$, then selection mechanism becomes:

$$z_i^* = \boldsymbol{w'}_i\boldsymbol{\gamma} + u_i \text{ with } \begin{cases} z_i = 1\ if\ z_i^* > 0 \\ z_i = 0 \text{ otherwise} \end{cases}$$

$$Prob(z_i = 1|\boldsymbol{w}) = \Phi(\boldsymbol{w'}_i\boldsymbol{\gamma})$$

$$Prob(z_i = 0|\boldsymbol{w}) = 1 - \Phi(\boldsymbol{w'}_i\boldsymbol{\gamma})$$

Then the regression model is

$$y_i^* = \boldsymbol{x'}_i\boldsymbol{\beta} + \varepsilon_i \text{ observed only if } z = 1$$

$$(u_i, z_i) \sim \text{bivariate normal}[0, 0, 1, \sigma_\varepsilon, \rho]$$

with

$$E[y_i|z_i, \boldsymbol{x}_i, \boldsymbol{w}_i] = \boldsymbol{x'}_i\boldsymbol{\beta} + \rho\sigma_\varepsilon\lambda(\boldsymbol{w'}_i\boldsymbol{\gamma})$$

The parameters of the sample selection model are usually estimated using Heckman's (1979) two-step procedure[9] that works as follows:

1. Estimate the probit equation by maximum likelihood to obtain estimates of $\boldsymbol{\gamma}$. For each observation in the sample compute the inverse Mills ratio $\hat{\lambda}_i = \phi(\boldsymbol{w'}_i\hat{\boldsymbol{\gamma}})/\Phi(\boldsymbol{w'}_i\hat{\boldsymbol{\gamma}})$

2. Estimate $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma_\varepsilon$ by least squares regression of y on $\boldsymbol{x}$ and $\hat{\lambda}$.

The probit model belongs to the family of binary response models that are used when the dependent variable is dichotomic, so that can take only two values and is usually coded as 0 and 1. Typical economic examples are the participation to the labor force, in which an agent chooses between two alternatives or differences in wage related to gender, in which being a female or a male can drive to different salaries. If we consider the sample of American

[9]The model could be also estimated by maximum likelihood.

movies under analysis, the dependent variable will take value 1 if a movie is exhibited in a certain country and 0 otherwise. Let P_i be the probability that $y_i = 1$ conditional on the information set Ω_i, which is characterized by exogenous variables. The aim is to model the conditional probability and since the values of the dependent variables are 0 and 1, P_i is also the expectation of y_i conditional on Ω_i[10]

$$P_i \equiv (y_t = 1 \mid \Omega_i) = E(y_t = 1 \mid \Omega_i)$$

It is important to understand why the implementation of a regression model is note feasible when we face this kind of dependent variable: suppose that $\mathbf{X}_i \subset \Omega_i$ is a row vector of length k in which the first term is a constant. Then a linear regression model would specify $E(y_i = 1 \mid \Omega_i)$ as $\mathbf{X}_i\beta$, failing to impose the condition that $0 \leq E(y_i = 1 \mid \Omega_i) \leq 1$, which must holds because $E(y_i = 1 \mid \Omega_i)$ is a probability. And since it makes no sense to estimate negative probabilities or greater than 1, regressing y_i on $\mathbf{X}_i$ is not a feasible approach to model the conditional expectation of a binary variable.
To ensure that $P_i \in [0, 1]$, a model must specify that:

$$P_i \equiv E(y_i = 1 \mid \Omega_i) = F(\mathbf{X}_i\beta).$$

Where $\mathbf{X}_i\beta$ is an index function which maps from the vector of explanatory variables $\mathbf{X}_i$ and the vector β of parameters to a scalar index, and $F(x)$ is a transformation function with the following properties:

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad f(x) \equiv \frac{dF(x)}{dx} > 0$$

Which are the properties of the cumulative distribution function of a probability distribution, and ensure that $F(\mathbf{X}_i\beta) \in [0, 1]$ while allowing the index function $\mathbf{X}_i\beta$ to take any value on the real line.

[10]Thus a binary response model can also be thought of as modeling the conditional expectation.

A binary response model is a probit when $F(\mathbf{X}_i\beta) = \Phi(\mathbf{X}_i\beta)$ where

$$\Phi \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp\left(-\frac{1}{2}X^2\right) dX$$

and its first derivative is the standard normal density function $\phi(x)$. An attractive feature of the probit model is that it can be derived from a model involving a latent, unobserved, variable z_i^*.As seen before, let be

$$z_i^* = \mathbf{X}_i\beta + u_i, \quad u_i \sim NID(0,1).$$

We observe only the sign of z_i^*, which determines the value of the observed binary variable y_i with the relationship

$$y_i = 1 \quad if \quad z_i^* > 0; \quad y_i = 1 \quad if \quad z_i^* \leq 0$$

The two previous equations define a latent variable model, the intuition is that z_i^* is an index of the net utility associated with some action; only if its value is positive then the action is undertaken. Then it is now possible to compute P_i, the probability that $y_i = 1$ as

$$\begin{aligned} P(y_i = 1) &= P(z_i^* > 0) = P(\mathbf{X}_i\beta + u_i > 0) \\ &= P(u_i \leq \mathbf{X}_i\beta) = \Phi(\mathbf{X}_i\beta). \end{aligned}$$

2.4 Estimation Results

A general approach in economic literature explains film success as a function of production budget, awarded prizes and features of the movie like genre, rating, being a sequel and so on. This approach is particularly useful when it is applied to countries abroad the United States given that, although the different explanatory variables may fail to be exogenous in the local market as the can be deemed to be affected by the expected revenue, in general they can be considered as exogenous with respect to the revenue in each single country.

Therefore the baseline specification explains the revenue of a movie i (in logs and adjusted for inflation) in a country j as a function of two main groups of variables: indicators of the quality of the film, $budget_i$ and $nomination_i$ and variables related to the different features of a movie in order to check how these characteristics have an impact on its box-office performance. The following model is considered:

$$\begin{aligned} ln\, revenue_{ij} = \beta_0 + \beta_1 ln\, budget_i + \beta_2 nomination_i + \beta_3 sequel_i + \beta_4 franchise_i \\ + \beta_5 \mathbf{ratings}_i + \beta_6 \mathbf{genres}_i + \beta_7 original_i + year_i + country_j + \epsilon_i \end{aligned}$$

where $budget_i$ is the log of the production budget for the ith film expressed in American dollars and adjusted for inflation; $nomination_i$ is a dummy variable that takes value 1 if the ith movie received an Academy Award Nomination in one or more of the main categories: best movie, best director, best actor or actress in a leading or supporting role or best animation movie; $sequel_i$ and $franchise_i$ denote if the ith is a sequel or a subsequent title in a serie; $\mathbf{ratings}_i$ is a vector of dummy variables that includes G_i, PG_i and R_i, that according to the MPAA Film Rating System stands for general audience, parental guidance suggested and parents strongly cautioned respectively; $\mathbf{genres}_i$ is a vector that includes $drama_i$, $action_i$, $thriller_i$; $original_i$ is a factor variable which takes value 1 if the screenplay is an original subject and 0 otherwise; $country_j$ and $year_i$ are included to control for different unobserved factors that could explain heterogeneity of movies revenues in different nations and time periods; the terms β_r for $r = [1, 7]$ are parameters of the model and $\epsilon_{i,j}$ is an error term.[11]

The sample amounts to 1431 US movies observed in a total of 50 different countries, where US/Canada is not included[12], from 2002 and 2013. However, the panel is not balanced because not every movie is shown in each of the countries. The possible presence of endogenous sample selection may then result in inconsistent estimates of the coefficients in a model that ac-

[11] Variables $PG-13_i$ and $commedy_i$ are leaved outside age rating and genre groups respectively to avoid perfect multicollinearity.

[12] MPAA, www.boxofficemojo.com and other data providers aggregate United States and Canada as an unique market, often called *domestic*.

counts for film revenues if the shock that affect the probability that a given film is exhibited in a certain country are highly correlated with the shocks that determines its revenue. Based on this premise I run an estimation pool model for film revenues in all countries employing Heckman's (1979) two-step methodology. In the first step, I estimate a probit model for the probability that a movie is exhibited in a country; this allows to obtain the Mills ratios needed to correct the OLS estimates of the primary equation in the second stage. In order to identify the model it is necessary to choose at least one instrumental variable to be included only in the probit correlated with the probability of exhibition but uncorrelated with the unobservable error term. A first best solution is to use information about entry barriers for foreign movies in each country: an important advantage of this type of variable is that given it has been defined at national level it is plausible to assume that it is exogenous to the expected revenue of each individual movie. However, note that protection laws in favor of local movie industries could come in a variety of different forms such as definitions of quotas for foreign movies like in France. Also, some of these laws are quite old (in Italy one should trace back to 1936) and, while are still in place, are not actively applied. Although it is impossible to get enough information about all the possible types of restrictions in the film industry for each country in the sample, this information can be successfully proxied by the trade of freedom index included in the Economic Freedom report of the Heritage Foundation. Beside this variable I also considered the inclusion of two alternative instruments defined at movie level: 1) opening week revenue in the domestic market[13] ; and 2) an indicator of whether the nearest neighbor movie released in a particular country two years before.[14]. However, over-identifying tests clearly indicates that these are not valid instruments as they are both significant at the conventional levels in the primary equation.

Table 1.4 shows the results of equation (1) for both the Heckman model and a typical OLS estimation of a pool regression for the 50 countries in the sample.

[13]see McKenzie and Walls (2012).

[14]The definition of this variable is based on the minimization of the canonical distance of each film with all other movies exhibited two years earlier using the same variables defined in equation (1) and it was implemented by R package FNN.

Results from the probit model estimation are reported in table 1.5 and can

Table 2.4: Determinants of film revenues

	Heckit model	OLS regression
budget	0.933(***) (62.67)	0.640(***) (55.58)
nomination	0.973(***) (38.62)	0.637(***) (35.52)
drama	-0.290(***) (-13.17)	-0.258(***) (-11.74)
action	0.331(***) (17.63)	0.236(***) (13.43)
thriller	0.479(***) (20.87)	0.267(***) (14.01)
G	-0.0886(*) (-2.29)	0.0392 (1.03)
PG	-0.00379 (-0.21)	0.0460(**) (2.60)
R	-0.0658(***) (-4.49)	-0.0489(***) (-3.31)
sequel	0.582(***) (28.01)	0.469(***) (23.80)
sequel2	0.530(***) (22.38)	0.427(***) (18.60)
Inverse Mills Ratio	1.409(***) (18.53)	
N	41274	41274
Adjusted R^2	0.615	0.610

Omitted: comedy, PG-13; t statistics in parentheses.
(*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$

be seen that the estimated coefficient associated to the instrumental variable *trade freedom* is positive and significant at the conventional values and the qualitative impact of all other variables is similar to their estimated effect in the revenue equation. I find evidence of endogenous samples selection as the inverse mills ratio is highly significant in the primary equation and has a positive sign which is consistent with the logical argument that films that are expected to have higher revenue in a given country are also more likely to be exhibited there.

However, in spite of these findings, the comparison with the OLS estimation indicates that most of the qualitative results are unaffected by the Heckman correction. To control for potential correlation of the error term in the primary and selection equations I also tested the Mundlak-Chamberlain approach as proposed by Wooldridge (2010), with no qualitative change in

Table 2.5: Probit results

trade freedom	0.015(*) (2.23)
budget	0.933(***) (62.67)
nomination	0.973(***) (38.62)
drama	-0.290(***) (-13.17)
action	0.331(***) (17.63)
thriller	0.479(***) (20.87)
G	-0.0886(*) (-2.29)
PG	-0.00379 (-0.21)
R	-0.0658(***) (-4.49)
sequel	0.582(***) (28.01)
sequel2	0.530(***) (22.38)
original screenplay	0.0462(***) (3.63)
N	41274

Omitted: comedy, PG-13; t statistics in parentheses.
(*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$

the estimated results or in the subsequent ranking of countries presented next.

It can be observed that variables related to the quality of a movie such as budget and nomination have a positive impact both in the revenue and the probability of exhibition. Moreover, mild movies for the general audience are less successful in terms of revenue than stronger movies for which parental guidance is suggested which seems to indicate that, at least outside the domestic market, Hollywood movies are not mainly produced for family consumption as in principle one could suggest. The fact that sequels and subsequent movies in a serie have a positive impact on revenue outlines that it is profitable in many circumstances to take advantage of an existing product instead of introducing a completely new movie in the market (and with that new characters and plots). Should be noted that drama movies are less profitable than other genres such as thriller and action movies.

Although I control for country individual effects in the above estimation, there is a risk of aggregation bias given that the model imposes that explanatory variables have the same impact on revenues regardless the type of country under analysis. In a first approximation to circumvent this problem I estimate equation (1) for two groups of countries depending on their human development level and cultural distance to the United States. In particular, for the sake of symmetry the estimation sample is split in two groups depending on the position of each country with respect to the median of the average value of the HDI during the period of analysis and cultural distance, which is fixed through time, which drives to 25 countries for estimation in each of the clusters. The estimation output is shown in Table 1.6. Note that estimated parameters for countries with high HDI are very similar to countries that are close to the American culture. In fact, there is a high correlation between these two groups of countries that can be observed in the fact that for the 25 countries with HDI over the median, 18 have a CD below the median.

In general, countries with high HDI (or with low CD) see to be less affected by nomination and budget that which points to the fact they are more prone to consume all type of American movies instead of only high quality ones. Another potential explanation for this result is that more developed countries could be interested in less commercial movies. Besides, high HDI and low CD countries dislike relatively more mild films that are authorized for all the public which is probably an indication that society is less restrictive in these countries; surprisingly they are less faithful to sequels, suggesting that they have more alternative leisure options to previous successful movies. Regarding their taste about genres, countries in the high HDI and low CD cluster show relatively more preferences for action films and less preferences for drama (with respect to comedies).

One possible way to circumvent the identification problem about the role played by HDI and CD to explain tastes for US movies is to estimate the previous equation only for countries with HDI above the median but that are not close to the American culture (Greece, Japan, Singapore, Slovakia, Slovenia, South Korea and United Arab Emirates) and for countries with HDI below the median and low CD (Argentina, Brazil, Dominican Republic, India, Jamaica, Lebanon and South Africa). Results of this estimation are shown in Table 1.6. Note that estimation results for countries with a cultural proximity to the United States but with low HDI are more similar to the average estimation for high and low CD shown in columns 1 and 4 of table 1.6. On the contrary, nations with high HDI and high CD show higher preference for budget and nomination as they are relatively more attracted by only high quality movies, less preference for mild films and more preference for drama, action and thriller movies with respect to the reference genre, comedy. In fact, comedies have typically very idiosyncratic values that makes them difficult to export to countries with different cultural values.

Table 2.6: movie revenues for different clusters of countries

	low HDI	high HDI	high CD	low CD	high HDI/CD	low HDI/CD
budget	0.896*** (39.71)	0.973*** (49.81)	0.954*** (46.77)	0.972*** (41.78)	1.095*** (19.83)	0.802*** (16.33)
nomination	0.689*** (19.85)	1.220*** (33.37)	1.228*** (33.86)	0.742*** (19.94)	0.946*** (10.33)	0.729*** (10.44)
drama	-0.372*** (-12.16)	-0.239*** (-7.88)	-0.264*** (-8.61)	-0.345*** (-11.32)	-0.192*** (-3.33)	-0.307*** (-5.33)
action	0.356*** (12.88)	0.301*** (11.91)	0.230*** (9.33)	0.446*** (15.25)	0.534*** (8.08)	0.190*** (3.55)
thriller	0.525*** (14.99)	0.437*** (14.44)	0.345*** (11.59)	0.669*** (18.09)	0.785*** (9.13)	0.236*** (3.37)
G	-0.314*** (-5.60)	0.0791 (1.53)	0.0863 (1.67)	-0.331*** (-5.88)	-0.406*** (-3.54)	-0.296* (-2.47)
PG	0.0269 (1.17)	-0.0393 (-1.53)	-0.0300 (-1.15)	0.00960 (0.42)	-0.0902 (-1.94)	-0.0129 (-0.28)
R	-0.185*** (-8.65)	0.0298 (1.47)	0.00945 (0.46)	-0.160*** (-7.67)	-0.132*** (-3.33)	-0.197*** (-4.15)
sequel	0.522*** (18.33)	0.638*** (21.85)	0.666*** (22.34)	0.511*** (17.79)	0.528*** (8.97)	0.541*** (8.91)
sequel2	0.444*** (12.97)	0.603*** (18.60)	0.596*** (18.05)	0.468*** (13.78)	0.516*** (7.86)	0.439*** (6.26)
original screenplay	0.0903*** (5.19)	0.00848 (0.47)	0.0139 (0.76)	0.0831*** (4.79)	0.0559 (1.65)	0.0790* (2.28)
Inverse Mills Ratio	1.233*** (10.84)	1.609*** (15.40)	1.631*** (13.95)	1.427*** (12.82)	1.741*** (6.46)	0.836** (2.71)
_cons	-4.272*** (-8.64)	-3.469*** (-8.70)	-5.313*** (-12.29)	-8.026*** (-15.08)	-8.284*** (-6.55)	-2.101 (-1.95)
Observations	17416	23858	22886	18388	5621	4649
Adjusted R^2	0.658	0.559	0.562	0.659	0.625	0.619

Omitted: comedy, PG-13; t statistics in parentheses.
(*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$

Chapter 3

La Grande Bellezza

3.1 Introduction

The Italian film industry is pretty active in terms of movies produced per year. The Anica report for 2011[1] shows that domestic theaters exhibited 241 titles entirely produced in Italy plus 55 co-produced with other countries over a total of 901 exhibited movies[2]. However, to my knowledge, academic research on the field didn't pay much attention to this market.

The only notable exception comes from Bagella and Becchetti's (1999) paper that studies some of the critical issues in the Italian movies market with a descriptive and econometric analysis on box office performances of movies produced in Italy between 1985 and 1996. In particular they focused on the relationship between popularity of human inputs (director and cast of actors) and box office revenues and the specialization in comedy production.

They use a database which gathers information on all movies produced in Italy from 1985 to 1996 with a sample of 977 films. For each movie they consider producing or co-producing companies, total admissions, distribution companies, box office revenues and programming days, ex ante popularity of actors and directors and availability of state subsidies. The total number of admissions is used as a dependent variable to measure box office performance. The descriptive analysis presented documents a reduction in the number of movies financed in the relevant sample period, and a sharp reduction in per screen daily admissions and revenues, paralleled by a positive trend in absolute admissions. In the econometric analysis the focus is on three crucial issues: the first is relative to how popularity of human inputs affects motion picture performance. The second is on the relative impact of state subsidization, while the third regards the relative impact of additional factors such as organizational and marketing capacity of production houses and the Italian viewers taste specificity affecting the relate success of specialization genres.

Using a GMM-HAC (Generalized Method of Moments Heteroskedasticity and Autocorrelation Consistent) approach they find that the ex-ante popularity of human inputs affects in a nonlinear way box-office performance and

[1] www.anica.it/online/allegati/dati/19042012_DATI2011.pdf

[2] 313 movies came from United States, 226 from other European countries and 66 from the rest of the world.

the interaction between the two factors' popularity has a positive impact on total admissions. With regard to the second question, a quite surprising result shows that the subsidized films do not have a significantly lower performance in the econometric analysis of total admissions, daily revenues and per screen daily admissions despite the far lower ex ante popularity of cast and directors of subsidized films as compared to non subsidized films. As for the third question, when looking at the results on net impact of different Italian production houses on the dependent variable, they find that only one (*Filmauro*) has a significant positive effect on total admissions. The positive and significant effect of the comic genre on total admission shows that the choice of producing these types of films has an independent positive effect on box office revenues net of ex ante cast and director popularity.
In this chapter my aim is to focus on the impact of subsidization on box-office revenues (the quantitative side of the market) with a new set of data, and then to expand the study from Bagella and Becchetti controlling also for possible impact on the quality of financed movies[3].

3.2 Data

In order to conduct my analysis, I collected data for 754 Italian movies exhibited during the 2002-2011 period. Similarly to what I have done for the American data set, I gathered information about box-office revenue (expressed in euros and adjusted for inflation, base year 2011) of each movie and its genre[4]. With respect to the data set on Hollywood movies three main differences arise: 1) production budget is missing due to the absence of publicly available information ; 2) I collected data for all the public subsidization from MiBACT (Ministero dei Beni e delle Attività Culturali e del

[3]Should be noted that until now only other two papers in the literature of movie economics analyzed in some way the role of public subsidization: Jansen (2005) for the German market and McKenzie and Walls (2012b) for the Australian industry.

[4]From the first database I keep the same distinction of drama, comedy, documentary and thriller movies. The only difference is that in the Italian production market action titles are absent: this kind of movies almost always relay on high production budgets and heavy use of special effects, both aspects difficult to attain for the Italian industry given its relative size.

Turismo) and 3) I also gathered information on appearances and prizes won at film festivals.
Table 3.1 shows descriptive statistics of the sample. Of the total 754 movies, a sub-sample of 311 was granted public subsidizes from MiBACT to promote relevant cultural aspects of a movie or the work of new directors[5]. Over the period under analysis the average public financing per movie was 636 thousands of euros with a maximum at 4.2M.
The whole sample shows a strong predominance of dramas and comedies against thrillers and documentaries, with the first accounting for 45% of the sample and the latter for 43%. I want to highlight how the shares of comedies and dramas change when we consider the sub-sample of financed movies: drama quota rise to 53%, while comedies drop to 33%. This shift can be explained by multiple concurrent factors regarding comedies: 1) this kind of movies are less likely to contain cultural aspects of public interest; 2) as shown in Bagella and Becchetti (1999) Italian moviegoers exhibit a strong preference for this genre, with box-offices revenues over the mean so that production companies are less prone to seek for public financing and 3) the presence of young directors that make comedy movies and are granted subsidies ensure that the drop is not even steeper.

For a sub-sample of 529 motion pictures I got information on the participation (or not) to film festivals and prizes won at them. 461 movies was exhibited at festivals and were eligible for awards[6] and 279 of them were public financed movies, which account for 90% of the subsidized movies sample. Tables 3.2 and 3.3 highlight some interesting facts in the distribution of these variables. On average each movie in the subset competed in 26 festivals, winning 5.67 prizes. These values slightly rise for financed movies becoming respectively 28.64 participation and 6.21 prizes. For both groups can be seen that there is a predominance of zero awards associated with a pretty low median value (2 for the whole subset and 3 for subsidized movies).

[5]Other than financing movies of particular cultural interest, MiBACT grants financial aid to the first or second movie of new directors. The degree of discretion on the allocation of resources from MiBACT to the domestic industry and what can be defined ”cultural interest” is over the scope of this dissertation.

[6]Out of competition appearances are not recorded.

Table 3.1: Movies Descriptive Statistics

variable	**mean**	**std. deviation**	**min**	**max**
		Whole Sample		
subsidies (adjusted)	636898	1011733	0	4200919
genres				
drama	0.448		0	1
comedy	0.435		0	1
documentary	0.059		0	1
thriller	0.058		0	1
Observations	754			
		Subsidized Movies		
genres				
drama	0.534		0	1
comedy	0.334		0	1
documentary	0.061		0	1
thriller	0.071		0	1
festivals	25.698	27.961	0	139
prizes	5.575	9.158	0	51
Observations	311			
		Data on festivals		
festivals	22.686	25.559	0	139
prizes	4.941	8.161	0	51
Observations	529			

The analysis of percentiles shows that the distribution of prizes is heavily shifted to the right, meaning that a small amount of movies conquer the majority of awards. The third column of Tables 3.2 and 3.3 show the ratio between prizes and festivals appearances. While a simple correlation analysis of the two variables indicates a strong reciprocity (∼0.8), it is of particular interest the fact that the mean and median values are attested around 16-19%, from which we can conclude that an heavy participation commitment to festivals doesn't automatically drives to more awards.

Table 3.2: Festivals and Prizes

	festivals		**prizes**		**win ratio**	
		smallest		smallest		smallest
1%	1	1	0	0	0	0
5%	2	1	0	0	0	0
10%	3	1	0	0	0	0
25%	8	1	1	0	0.04	0
50%	18		2		0.16	
		largest		largest		largest
75%	35	125	7	44	0.27	1
90%	61	128	15	44	0.42	1
95%	81	130	22	50	0.50	1
99%	129	139	40	51	1	1
mean		26.03		5.67		0.19
std deviation		25.73		8.50		0.22
Observations	461					

Table 3.3: Festivals and Prizes – subsidized movies

	festivals		**prizes**		**win ratio**	
		smallest		smallest		smallest
1%	1	1	0	0	0	0
5%	2	1	0	0	0	0
10%	4	1	0	0	0	0
25%	9	1	1	0	0.05	0
50%	20		3		0.15	
		largest		largest		largest
75%	39	125	7	44	0.25	1
90%	72	128	16	44	0.40	1
95%	92	130	33	50	0.47	1
99%	128	139	44	51	1	1
mean		28.64		6.21		0.19
std deviation		28.05		9.46		0.23
Observations	279					

3.3 Methodology

Econometric procedures involved in my analysis of the Italian movie industry are pretty straightforward and include Ordinary Least Squares for the study of box-office revenues and the implementation of a count data model when looking at the quality aspect of prizes won by movies. The next section covers in details the theory behind the most used approach when dealing with count variables, the Poisson regression model.

3.3.1 The Poisson regression model

Many economic studies relay on variables that are non-negative integers. Examples includes data on patents granted to firms, number of contacts for call centers or visits to the hospital by an individual. On the same fashion, to analyze the impact of public subsidization on the quality of a film, I will study the prizes won by a movie at film festivals. Data of this type are called count data and the empirical analysis of them is based on models of events. In principle, the study could be done implementing multiple linear regression; however, the preponderance of zeros and small values in the dependent variable and its discrete nature, suggest that it is possible to improve using a different methodology. Another option that naturally comes to mind is to implement an ordered discrete choice model, like ordered probit. However, this is not usually feasible, because this kind of model requires the number of possible outcomes to be fixed and known.
Concerning the dataset under analysis, from tables 2.1 and 2.2 we can recall that the dependent variable on prizes assume values from 0 to 51 with a strong presence of zeros. To deal with these characteristics we need a model for which any non-negative integer value is a valid, although possibly very unlikely, value. We then first turn to a distribution which has this propriety: the Poisson distribution. Named after French mathematician Siméon Denis Poisson, is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and or space if these events occur with a known average rate and independently of the time since the last event.

If a discrete random variable Y follows the Poisson distribution, then

$$Prob(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \qquad y = 0, 1, 2, ...$$

Similarly the poisson regression model specifies that each y_i is drawn from a Poisson distribution with parameter λ_i, which is related to the regressors $\mathbf{x}_i$. The primary equation of the model then is

$$Prob(Y = y|\mathbf{x}_i) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \qquad y = 0, 1, 2, ...$$

The model, like the distribution, is characterize by a single parameter, λ. The most common formulation for λ_i is the loglinear model,

$$ln\lambda_i = \mathbf{x'}_i\beta.$$

The expected number of events per period (per festival, in our sample) is given by

$$E[y|\mathbf{x}_i] = Var[y|\mathbf{x}_i] = \lambda_i = e^{\mathbf{x'}_i\beta}$$

then

$$\frac{\partial E[y|\mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i\beta.$$

The easiest way to estimate the parameters of the model is with maximum likelihood techniques. The log-likelihood function will be

$$ln\ L = \sum_{i=1}^{n}[-\lambda_i + y_i\mathbf{x'}_i\beta - ln\ y_i!].$$

The likelihood equation is

$$\frac{\partial ln\ L}{\partial \beta} = \sum_{i=1}^{n}(y_i - \lambda_i)\mathbf{x}_i = \mathbf{0}$$

and the Hessian matrix is

$$\frac{\partial^2 ln\ L}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x'}_i.$$

Since the Hessian is negative definite for all $\mathbf{x}$ and β, optimization techniques based on Newton's Method generally work very well and converge rapidly. Given the estimates, the prediction for observation i is $\hat{\lambda}_i = exp(\mathbf{x'}_i\beta)$.
Because the conditional mean function is nonlinear and the regression is heteroschedastic, the Poisson model doesn't produce a counterpart to the R^2 typical of linear regression models. An alternative based on the standardized residuals that compares the fit of the model with a restricted version with only the constant term is given by the so called *pseudo* R^2:

$$R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left[\frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}\right]}{\sum_{i=1}^{n}\left[\frac{y_i - \overline{y}}{\sqrt{\overline{y}}}\right]}.$$

Davidson and MacKinnon (2005) points out that "Although its simplicity makes it attractive, the Poisson regression model is rarely entirely satisfactory. In practice, even though it may predict the mean event count accurately, it frequently tends to underpredict the frequency of zeros and large counts, because the variance of the actual data is larger than the variance predicted by the Poisson model. This failure of the model is called overdispersion".
As we will see in the next section, I perform a robustness check treating available data with an alternative model: negative binomial. The result of Likelihood ratio test suggests that the model of choice should be the Poisson.

3.4 Estimation Results

As stated in the previous sections, my study of the Italian movie industry resolve around two dimensions: quantity (box-offices revenues) and quality (prizes won at film festivals). For the analysis of box-office performance the baseline specification explain the revenue of a movie i (in logs and adjusted for inflation as usual) as a function of subsidization in log form if any and genres: comedy, drama and thriller, with documentary omitted and taken as reference category. The following model is considered:

$$\begin{aligned} ln\ revenue_i = &\beta_0 + \beta_1 ln\ subsidization_i + \beta_2 comedy_i \\ &+ \beta_3 drama_i + \beta_4 thriller_i + \varepsilon_i \end{aligned}$$

where β_r for $r = [1, 4]$ are parameters of the model and ϵ_i is an error term. The sample amount to 754 Italian movies exhibited in the domestic market during the 2002-2011 period. I evaluate data with the typical OLS approach, grouping observations by year and comparing results for random and fixed effects models. The random effects assumption is that the individual specific effects are uncorrelated with the independent variables. The fixed effect assumption is that the individual specific effect is correlated with the independent variables. If the random effects assumption holds, the random effects model is more efficient than the fixed effects model. To establish which model fits the data I then perform the Hausman's (1978)[7] test, which evaluates the consistency of an estimator when compared to an alternative, less efficient, estimator which is already known to be consistent. in other words, helps to evaluate if a statistical model corresponds to the data. In the contest of Panel data, the Hausman test can be used to differentiate between fixed effects and random effects models. Running the test on the data under analysis, we can conclude that the random effect model is to be preferred under the null hypothesis due to higher efficiency (should be noted that both specifications are consistent).

Table 3.4 shows estimations results for the given specification. These first results are consistent with the findings of Bagella and Becchetti (1999) in the sense that highlight the crucial role of comedy genre in driving the box-office performance of Italian movies and the irrelevance and slightly negative effect of subsidization. The following step is to evaluate the impact of public

[7]Also called the Wu–Hausman test, Hausman specification test, and Durbin–Wu–Hausman test.

Table 3.4: Italian movies revenues - base specification

	Fixed Effects	Random Effects
ln subsidization	-0.0352** (-2.95)	-0.0676*** (-4.96)
drama	1.145** (3.22)	1.146** (3.24)
comedy	2.490*** (6.99)	2.484*** (7.05)
thriller	1.361** (2.88)	1.319** (2.81)
documentary	*(omitted)*	*(omitted)*
R^2		
within	0.149	0.143
between	0.308	0.303
overall	0.119	0.125
N	754	754

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

financing for different kind of movies, that is look at the iteration between genres and subsidization. To do so, I consider the following specification:

$$\begin{aligned} ln\ revenue_i =& \beta_0 + \beta_1 l\mathbf{subsizied\ genres}_i \\ &+ \beta_2' \mathbf{nonsubsidized\ genres'}_i + \varepsilon_i \end{aligned}$$

where *subsidized genres* is a vector of iteration variables between the four genres and the log of subsidization (so can take values from 0 to n) and *nonsubsidized genres* is a vector of dummy variables which take value 1 if a movie belong to a given genre but hasn't received public funding, β_1 and β_2 are parameters of the model and ϵ_i is an error term.
As for the base specification, I consider random and fixed effects models using ordinary least squares and, again, the Hausman test suggest that the preferred model should be the random one. Results are shown in Table 3.5. The main difference with the previous specification is that we now find evidence of a positive impact of subsidization for three genres out of four. Beside, should be noted that the magnitude of coefficients of not financed movies is higher than in the previous specification, highlighting a bigger impact on revenues. One could conclude that, given the coefficients, financing comedies

guarantee the best resources allocation, however if we take in account the aforementioned bias of Italian consumers for this genre of movies the policy implication should be to shift public expenditure over dramas and thrillers.

To conclude the analysis of the impact of public subsidization on the Italian

Table 3.5: Italian movies revenues - budget iteration with genres

	Fixed Effects	Random Effects
subs comedy	0.439** (2.63)	0.442* (2.57)
nosubs comedy	7.540** (3.29)	7.822** (3.30)
subs drama	0.388* (2.41)	0.392* (2.35)
nosubs drama	5.699* (2.48)	5.991* (2.53)
subs thriller	0.413* (2.48)	0.412* (2.39)
nosubs thriller	5.837* (2.51)	6.105* (2.55)
subs documentary	0.305 (1.73)	0.308 (1.69)
nosubs documentary	4.891* (2.11)	5.157* (2.16)
R^2		
within	0.173	0.172
between	0.210	0.187
overall	0.144	0.148
N	754	754

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

motion picture industry, let now turn our attention to the quality of produced movies. As said in the data section of this chapter, I collected data for film festivals participation and prizes won for a subset of 461 movies: the idea is to use information on prizes to evaluate the implicit quality of motion pictures with a special focus for those who received public funding. From table 3.3 we recall that 279 of 311 financed movies participated as competitors in at least one festival.

The dependent variable *prizes* is a so called count variable and need to be treated with a Poisson model (see methodology section of this chapter for a detailed explanation). As a robustness check I tested the data with negative binomial model; however, likelihood ratio test results suggest that the model

of choice should be the Poisson[8].
The following model is considered:

$$prizes_i = \beta_0 + \beta_1 festivals_i + \beta_2 ln\ subsidization_i + \beta_3 comedy_i + \beta_4 drama_i + \beta_5 thriller_i + \varepsilon_i$$

Where prizes for the ith movie are a function of festivals, subsidization if any and genres. β_r for $r = [1, 5]$ are parameters of the model and ϵ_i is an error term. Estimation results are shown in table 3.6 and are presented both in terms of coefficients and of incidence rate ratios, which is a relative difference measure used to compare the incidence rates of events occurring at any given point in time or space.
At a first glance we can see that incidence ratio for *festivals* confirms what we discovered with the descriptive analysis: participation to festivals doesn't automatically drives to more awards. Moreover, subsidization behaves as previously seen in the base specification of the revenue equation, his rate ratio for prizes would be expected to decrease by a factor of 0.98, while holding all other variables in the model constant. Magnitude and sign of the coefficient are very similar with the findings in the revenue equation of my analysis and in what stated by Bagella and Becchetti (1999). Beside, not surprisingly the genre with a greatest impact in the occurring event of winning a prize turns out to be drama, which is coherent with the idea that quality is more perceived in movies with a "serious" plot and dramatic characterization.
One final test I want to perform, like the second specification of the revenue equation, is to allow the iteration between subsidization (in log form) and genres. To do so, I model as follows:

$$prizes_i = \beta_0 + \beta_1 festivals_i + \beta'_2 \mathbf{subsidized\ genres}_i + \beta'_3 \mathbf{nonsubsidized\ genres}_i + \varepsilon_i$$

where *subsdized genres* and *non subsidized genres* are vector of variables that behave as in the second revenue equation, β_r for $r = [1, 3]$ are parameters of the model and ϵ_i is an error term. We can see estimation results in table 3.7.

[8]Estimation results for this model can be found in Appendix III.

Table 3.6: Poisson model for prizes

	coefficients	incidence ratio
festivals	0.0283*** (55.48)	1.02
ln subsidization	-0.0152*** (-4.84)	0.98
comedy	0.656*** (6.30)	1.93
drama	0.868*** (8.57)	2.38
thriller	0.731*** (5.23)	2.08
documentary	*(omitted)*	
N		461
Pseudo R^2		0.524

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Analysis of coefficients and incidence rate ratios suggest that, while positive in sign, the impact of subsidization for each genre is negligible.

The overall picture of the impact of public funding on the Italian motion

Table 3.7: Poisson model for prizes with iteration between budget and genres

	coefficients	incidence ratio
festivals	0.0284*** (53.31)	1.03
nosubs comedy	0.898* (1.72)	2.45
subs comedy	0.0645* (1.69)	1.06
nosubs drama	1.316** (2.54)	3.72
subs drama	0.0684* (1.89)	1.07
nosubs thriller	0.598 (1.06)	1.81
subs thriller	0.0758** (1.99)	1.08
nosubs documentary	0.547 (1.02)	1.73
subs documentary	0.000470 (0.01)	1.00
N		461
Pseudo R^2		0.530

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

picture sector tells us that its performance is slightly positive in terms of expected revenues and almost negligible on the quality of a movie. Moreover, regression analysis suggest to steer public expenditure over dramas and thrillers leaving comedies, Italian consumer's preferred genre, outside of intervention.

Chapter 4

Final Remarks

First part of this thesis studied foreign performance of American movies in 50 countries over the 2002-2013 period. The focus was on the impact of cultural differences among importing countries of American movies and how these differences can explain box-office revenues. Previous literature in movie economics observed the arrival of motion pictures in a country as given or focused on the determinants of revenues using cultural differences as explanatory variables into complex econometric modelling like gravity models.
My contribute to this literature comes from the application of the Heckman's (1979) two-step methodology to infer the probability of arrival of American movies in foreign countries and then to evaluate the box-office revenues for different clusters of nations built around Hofstede's (2001) index of cultural distance from the United States and Human Development data.
Overall, the analysis I conducted suggests that countries with relatively high HDI and that are close to the American culture tend to be less affected by measures of quality of a movie and show special preferences for action titles. Although the estimation is subject to an identification problem due to the negative correlation between HDI and cultural differences, it suggests that is mostly cultural distance, and not the Human Development level of a country, that leads the consumption in this group of nations. Further, the estimation indicates that once a movie is introduced in a country with low HDI or high cultural distance they are relatively more faithful to the following sequels of these movies.
The second part of my thesis is devoted to the analysis of the Italian movie production market which is one of the oldest in the industry and renowned worldwide for its quality. Previously, the only academic research on Italian domestic market is, to my knowledge, limited to the contribute of Bagella and Becchetti (1999). With a descriptive and econometric analysis on box office performances of movies produced in Italy between 1985 and 1996, they focused on the relationship between popularity of human inputs and the relative impact of state subsidization on box-office revenues.
Starting from their contribute, I focus on the impact of state subsidization looking not only at box-office performances of subsidized movies, but also on their quality. Results of box-office revenues estimation using subsidization

as a single variable are coherent with the findings of Bagella and Becchetti, showing an overall slightly negative impact and its net effect negligible. However, when introducing interaction between subsidies and genres the sign of coefficients turn to be positive and suggest that, while weak, the net impact on public financing on the Italian movie industry is affecting positively box-office revenues.
Same results arise in the econometric analysis of awards won at festivals where iteration variables between genres and subsidies have a slightly positive incidence on the ratio of awards granted.

Chapter 5

Appendix I: Data Description

This first appendix describes the two datasets that constitute the core of this thesis, one for the global market of American movies and one for the Italian movie industry. For each of them I present a list of data providers, how the base has been built and a brief description of each of the variables included.

5.1 Data description

5.1.1 dataset I: American Beauty

data providers:

- `www.boxofficemojo.com` (movie characteristics)
- `www.imdb.com` (movie characteristics)
- `www.thenumbers.com` (movie characteristics)
- `www.mpaa.org` (movie ratings)
- `http://www.heritage.org/index/trade-freedom` (freedom of trade)
- `hdr.undp.org` (Human Development Index)
- `www.gert-hofstede.com` (Cultural Distance Index)

The dataset consist of 1431 movies exhibited during the period 2002-2013 in 50 foreign countries. For each movie the following variables are available:
boxoffice: amount of money earned in a single country, expressed in american dollars and adjusted for inflation.
budget: estimated production budget, expressed in american dollars and adjusted for inflation. In some cases production companies reveal the exact production budget for a movie, in other cases it is estimated by imdb.com or thenumbers.com given the characteristics of a film, like cast, locations, use of special effects and so on.
sequel: factor variable which takes value 1 if a movie continues the narrative of a preexisting one (e.g. Kill Bill Vol.2).

franchise: factor variable which takes value 1 if a movie is the third of subsequent movie in a serie (e.g. the Harry Potter saga).
nomination: factor variable which takes value 1 if a movie received an Academy Award Nomination in one of the main categories: best movie, best director, best actor or actress in a leading or supporting role, best animation movie. This variable works as an ex-post proxy of the quality of a movie as stated by Academy members. In those cases in which a movie is released in a country after have received a nomination, it also works as an ex-ante signal that could possibly attract consumers to theaters.
action: factor variable which takes value 1 if a movie belongs to action, western or sci-fi genres and 0 otherwise.
comedy: factor variable which takes value 1 if a movie belongs to comedy, romantic comedy, family movies genres or if it is an animation movie (and 0 otherwise).
drama: factor variable which takes value 1 if a movie is of dramatic genre and 0 otherwise.
documentary: factor variable which takes value 1 if a movie is a documentary and 0 otherwise.
thriller: factor variable which takes value 1 if a movie belongs to thriller or horror genres and 0 otherwise.
original screenplay: factor variable which takes value 1 if a movie is based on an original screenplay and 0 if it is adapted. An adapted screenplay is a movie based on previously known sources, like novels, tv-shows, comics or real life events.
G: stands for General Audiences. It is a factor variable which takes value 1 if a movie is rated G by the MPAA. The MPAA Film Rating System describes this category "All Ages Admitted - Nothing that would offend parents for viewing by children".
PG: stands for Parental Guidance Suggested. It is a factor variable which takes value 1 if, regarding the content of a movie, "Some material may be not be suitable for children". The Film Rating Systems states that "Parents urged to give parental advice. May contain some material parents might not like for their young children".

PG-13: stands for Parents Strongly Cautioned. It is a factor variable which takes value 1 if "Some material may be inappropriate for children under 13". The MPAA rating system suggests that "Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers".
R: stands for Restricted. Factor variable which takes value 1 if a movie "Contains adult material. Parents are urged to learn more about the film before taking their children with them". The MPAA rules out that "Under 17 requires accompanying parent or adult guardian".
release date: release date of a movie in a single country.
Other than quantitative and qualitative variables of each movie, the dataset includes country specific variables needed to evaluate differences in the box-office performance.
trade freedom: According to the 2014 Index of Economic Freedom from the Heritage foundation[1] the Trade freedom is a composite measure of the absence of tariff and non-tariff barriers that affect imports and exports of goods and services. Its score is based on two inputs:

- The trade-weighted average tariff rate and
- Non-tariff barriers (NTBs).

Different imports entering a country can, and often do, face different tariffs. The weighted average tariff uses weights for each tariff based on the share of imports for each good. Weighted average tariffs are a purely quantitative measure and account for the basic calculation of the score using the following equation:

$$Trade\ Freedom_i = \left(\frac{Tariff_{max} - Tariff_i}{Tariff_{max} - Tariff_{min}} * 100 \right) - NTB_i$$

where Trade Freedom$_i$ represents the trade freedom in country i; Tariff$_{max}$ and Tariff$_{min}$ represent the upper and lower bounds for tariff rates (%); and Tariff$_i$ represents the weighted average tariff rate (%) in country i. The minimum tariff is naturally zero percent, and the upper bound was set as

[1]see: http://www.heritage.org/index/trade-freedom

50 percent. An NTB penalty is then subtracted from the base score. The penalty of 5, 10, 15, or 20 points is assigned according to the following scale:

- 20 NTBs are used extensively across many goods and services and/or act to effectively impede a significant amount of international trade.
- 15 NTBs are widespread across many goods and services and/or act to impede a majority of potential international trade.
- 10 NTBs are used to protect certain goods and services and impede some international trade.
- 5 NTBs are uncommon, protecting few goods and services, and/or have very limited impact on international trade.
- 0 NTBs are not used to limit international trade.

Rhe extent of NTBs in a country's trade policy regime is determined using both qualitative and quantitative information. Restrictive rules that hinder trade vary widely, and their overlapping and shifting nature makes their complexity difficult to gauge. The categories of NTBs considered as penalties include:

- Quantity restrictions - import quotas; export limitations; voluntary export restraints; import–export embargoes and bans; countertrade, etc.
- Price restrictions - antidumping duties; countervailing duties; border tax adjustments; variable levies/tariff rate quotas.
- Regulatory restrictions - licensing; domestic content and mixing requirements; sanitary and phytosanitary standards (SPSs); safety and industrial standards regulations; packaging, labeling, and trademark regulations; advertising and media regulations.
- Investment restrictions - exchange and other financial controls.

- Customs restrictions - advance deposit requirements; customs valuation procedures; customs classification procedures; customs clearance procedures.

- Direct government intervention - subsidies and other aid; government industrial policy and regional development measures; government financed research and other technology policies; national taxes and social insurance; competition policies; immigration policies; government procurement policies; state trading, government monopolies, and exclusive franchises.

Human Development Index (HDI): is an index built by United Nations for assessing the development of a country outside of economic growth. In the official description of the Human Development Index can be read that: The HDI is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions. The health dimension is assessed by life expectancy at birth component of the HDI is calculated using a minimum value of 20 years and maximum value of 85 years. The education component of the HDI is measured by mean of years of schooling for adults aged 25 years and expected years of schooling for children of school entering age. Mean years of schooling is estimated by UNESCO Institute for Statistics based on educational attainment data from censuses and surveys available in its database. Expected years of schooling estimates are based on enrollment by age at all levels of education. This indicator is produced by UNESCO Institute for Statistics. Expected years of schooling is capped at 18 years. The indicators are normalized using a minimum value of zero and maximum aspirational values of 15 and 18 years respectively. The two indices are combined into an education index using arithmetic mean. The standard of living dimension is measured by gross national income per capita. The goalpost for minimum income is $100 (PPP) and the maximum is $75,000 (PPP). The minimum value for GNI per capita, set at $100, is justified by the considerable amount of unmeasured subsistence and non-market production

in economies close to the minimum that is not captured in the official data. The HDI uses the logarithm of income, to reflect the diminishing importance of income with increasing GNI.
cultural distance (CD): to capture information about the cultural distance between U.S.and each of the countries in the dataset, following Lee (2009), I implement a value-based index developed by Hofstede (1980) built around four dimensions:

- power distance[2] - This dimension expresses the degree to which the less powerful members of a society accept and expect that power is distributed unequally.

- uncertainty avoidance - The uncertainty avoidance dimension expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity. The fundamental issue here is how a society deals with the fact that the future can never be known.

- individualism versus collectivism - The high side of this dimension, called individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families. Its opposite, collectivism, represents a preference for a tightly-knit framework in society in which individuals can expect their relatives or members of a particular in-group to look after them in exchange for unquestioning loyalty.

- masculinity versus femininity - The masculinity side of this dimension represents a preference in society for achievement, heroism, assertiveness and material rewards for success. Society at large is more competitive. Its opposite, femininity, stands for a preference for cooperation, modesty, caring for the weak and quality of life.

I collected data for each country and values from Hofstede (2001) and then each country's cultural distance from the United States is computed using

[2] description of each dimension is taken from http://geert-hofstede.com/dimensions.html

Kogut and Singh's (1988) formula:

$$CD_j = \sum_{I=1} \left\{ (I_{ij} - I_{iu})^2 / V_i \right\} /4$$

Where CD_j is the cultural distance of country j from the United States, I_{ij} is the value for country j on the ith cultural dimension (I_{iu} for the U.S.) and V_i is the variance of the ith cultural dimension.
This index is widely used and accepted in the international business or cross-cultural communications literature.

5.1.2 dataset II: La Grande Bellezza

data providers:

- `www.cinemaitaliano.info` (movie characteristics and revenues)
- `www.comingsoon.it` (movie characteristics and revenues)
- `http://www.cinema.beniculturali.it/` (public subsidization data)

The dataset consist of 754 movies produced in Italy and exhibited during the period 2002-2011. For each movie the following variables are available:
boxoffice: amount of money earned by each movie, expressed in euros and adjusted for inflation.
subsidization: amount of public subsidization granted from MiBACT (Ministero dei Beni e delle Attività Culturali e del Turismo), expressed in euros and adjusted for inflation.
festivals: variable that accounts participation at film festivals when a movie is eligible for awards. Out of competition appearances are not recorded.
prizes: prizes won at film festivals.
comedy: factor variable which takes value 1 if a movie belongs to comedy, romantic comedy, family movies genres or if it is an animation movie (and 0 otherwise).
drama: factor variable which takes value 1 if a movie is of dramatic genre and 0 otherwise.

documentary: factor variable which takes value 1 if a movie is a documentary and 0 otherwise.
thriller: factor variable which takes value 1 if a movie belongs to thriller or horror genres and 0 otherwise.

Chapter 6

Appendix II: Script codes

This Appendix shows the script code written to implement the empirical analysis of my dissertation. Each code is presented twice: the first one with extended comments for each command, while the second is the raw script, more convenient to copy and paste into STATA. The first section provides a detailed explanation of what done in the study of Hollywood movies in the global market, while the second is devoted to the inference on Italian movie industry data. At the end of the appendix, links to online manuals for each command are provided.

6.1 Script I: American Beauty

6.1.1 Commented code

```
fillin movie country
```

`fillin` adds observations with missing data so that all interactions between movies and countries exist, thus making a complete rectangularization of the data, needed to implement the probit model.

```
generate spot = 0
replace spot = 1 if log_boxoffice!=.
```

the variable *spot* is the dependent variable for the probit and takes value 1 if a movie is exhibited in a country and 0 otherwise.

```
encode country, gen(nation)
```

the `encode` command generates a new variable (*nation*) based on a string variable (*country*). In this case it assigns a number from 1 to 50 for each country in the dataset.
After these preliminary steps the code moves to the first regressions, namely a standard OLS of the whole sample. The goal of this regressions is to offer a benchmark for the subsequents estimations.

```
regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
```

```
original_screenplay i.year i.nation, vce(robust)
est store ols_pool
```

`regress` fits a linear model on the dependent variable *log_budget* and the indipendent variables. The `i.` prefix in `i.year` and `i.nation` instructs STATA to treat these variables as implicit dummies so is not needed to generate new ones; in this case the regression will take in account the sample period 2002-2013 and the 50 countries of the dataset. According to the software manual, the `vce(robust)` uses the robust or sandwich estimator of variance. This estimator is robust to some types of misspecification so long as the observations are independent[1]. With `est store` the estimation results are stored for further analysis and output possibilities.
The next section of the code is dedicated to Chamberlain procedure. It starts with the analysis of the whole sample and it is divided in three steps: first, an ordinary probit model is computed for each year in the sample to obtain consistent estimates of the parameters of the selection equation and find, in the second step, the selection hazard. The third step is the devoted to the OLS where the regression is run incorporating information from the inverse Mills ratios.

```
bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation, vce(robust)
```

To compute a probit regression for each year of the sample the composite command `bysort` is implemented: most Stata commands allow the `by` prefix, which repeats the command for each group of observations for which the values of the variables in a list are the same. Here is used in conjunction with `sort` that, as the name suggests, sort the data for a specific dimension (*year* in our case). The `eststo` command stores the estimation results in progression (i.e est1, est2, etc.) allowing to recall them in a successive moment to calculate the inverse Mills ratios. Next, `probit` fits maximum likelihood

[1]see Huber (1967) and White (1980, 1982).

models with dichotomous dependent variables coded as 0/1, or to be more precise, coded as 0 and not 0. As stated earlier, the dependent variable here is *spot*.

```
local j=1
gen counter=1
while counter <= 12 {
    est restore est`j'
    predict p1, xb
    replace p1 = -p1
    gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
    gen capphi = normal(p1)
    gen invmills_`j' = phi/(1-capphi)
    drop p1 phi capphi
    local j = `j'+1
    replace counter = counter+1
}
drop _est_est* counter
```

Given the software doesn't include an automatic way to obtain the selection hazards, some calculation is needed. The first line introduces a macro, a tool used in programming STATA. Formally, a macro is a string of characters, called the *macroname*, that stands for another string of characters, called the macro contents.. They can be either global or local. While the first is public and influences the behavior of the entire software, `local`, as the name suggests, exist solely within the program or do-file in which is defined; if that program or do-file calls another program or do-file, the local macros previously defined temporarily cease to exist. We will see shortly after why we need it and can't count solely on the other operator introduced (`counter`). The Inverse Mills ratios calculation is done inside the subsequent `while` loop. A while loop works evaluating an expression (`counter <= 10`, since there are ten valid years in the dataset) and executes the commands enclosed in the braces until it is true. To better understand the code inside the brackets let's first recall the inverse Mills ratio formula from Appendix I on methodology:

$$E(v_t|z_t = 1, \mathbf{W}_t) = E(v_t|v_t > \mathbf{W}_{t\gamma}, \mathbf{W}_t) = \frac{\phi \mathbf{W}_t \gamma}{\Phi \mathbf{W}_t \gamma}$$

Where $\mathbf{W}_t$ is the set of explanatory variables.

Now let's look inside the loop. Since we need separate selection hazards for each year of the sample, the first line in the code `est restore est‘j’` load into software memory the regression results j specific to the previous run of probit models. Here comes in role the use of the local macro: `‘j’` changes value in each iteration of the loop until the expression `counter <= 10` holds true, so at every iteration new regression results are loaded. Then `predict` calculates predictions, residuals, influence statistics, and the like after estimation; with the `xb` option calculates the linear prediction from the fitted model. The successive lines

```
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_‘j’ = phi/(1-capphi)
```

generate the ϕ and Φ as seen in Appendix I and then calculate the inverse Mills ratio for a given probit regression. Either inside and outside the loop a `drop` call delete those variables and estimations results that are not needed anymore. After this computation it is possible to run the OLS part of the procedure, including the correction given by the inverse Mills ratios as follows:

```
regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_1-invmills_12, vce(robust)
est store pool
```

Where `invmills_1-invmills_12` serve to include the ten ratios just calculated.

The next section of the code applies the same Chamberlain procedure, but taking into account countries differences in terms of Human Development (see

Appendix II). The idea is to group countries into two groups using quantiles and then run the regressions for each group.

```
bysort country: egen hdi_mean= mean(hdi)
xtile hdi_quant= hdi_mean
```

To do that, the first step is to calculate the mean value of the human development index for each country during the sample period. Sorting by country, `egen` creates a new variable equal to some function, in this case with the `mean` argument. Then `xtile` creates a categorical variable that contains the corresponding two quantiles of *hdi_mean*.
The next steps are the same as the procedure for the whole sample. The only difference is that probit and OLS are executed for the two groups:

```
bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if hdi_quant==1, vce(robust)

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if hdi_quant==2, vce(robust)

local j = 13
gen counter = 13
while counter <= 36 {
   est restore est‘j’
   predict p1, xb
   replace p1 = -p1
   gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
   gen capphi = normal(p1)
   gen invmills_‘j’ = phi/(1-capphi)
   drop p1 phi capphi
   local j = ‘j’+1
```

```
   replace counter = counter+1
}
drop _est_est* counter

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_13-invmills_24 if hdi_quant==1, vce(robust)
est store hdi1

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_25-invmills_36 if hdi_quant==2, vce(robust)
est store hdi2
```

The same idea and procedure are applied to take in account differences in cultural distance between United States, which is the producer of movies, and each country. Again, `xtile` is used to generate a categorical variable that group countries in two quantiles.

```
xtile cdist_quant= c_distance

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if cdist_quant==1, vce(robust)

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if cdist_quant==2, vce(robust)

local j=37
```

```
gen counter=37
while counter<=60 {
   est restore est‘j’
   predict p1, xb
   replace p1=-p1
   gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
   gen capphi = normal(p1)
   gen invmills_‘j’ = phi/(1-capphi)
   drop p1 phi capphi
   local j= ‘j’+1
   replace counter=counter+1
}
drop _est_est* counter

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_37-invmills_48 if cdist_quant==1, vce(robust)
est store cdist1

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_49-invmills_60 if cdist_quant==2, vce(robust)
est store cdist2
```

The same regressions (whole sample, Human Development groups and cultural distance groups) are then performed using the classical Heckman two step approach. The code involved is similar of what seen previously, the main difference is that a single selection hazard is needed for every regression, so instead of multiple probit procedures we need only one and the loop sequence is not necessary anymore. Then, for the pool regression the code is as follows:

```
probit spot trade_freedom log_budget nomination G
```

```
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation, vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_heck_pool = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_heck_pool, vce(robust)
est store heck_pool
```

Then come regressions for Human Development groups:

```
probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if hdi_quant==1,
vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_heck_hdiq1 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_heck_hdiq1
if hdi_quant==1, vce(robust)
```

```
est store heck_hdi1

probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if hdi_quant==2,
vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_heck_hdiq2 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_heck_hdiq2
if hdi_quant==2, vce(robust)
est store heck_hdi2
```

And finally the same approach is applied grouping for cultural distance quantiles:

```
probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if cdist_quant==1,
vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_cdist1 = phi/(1-capphi)
drop p1 phi capphi
```

```
regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_cdist1
if cdist_quant==1, vce(robust)
est store heck_cdist1

probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if cdist_quant==2,
vce(robust)

predict p1, xb
replace p1=-p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_cdist2 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_cdist2
if cdist_quant==2, vce(robust)
est store heck_cdist2
```

6.1.2 Naked code

```
fillin movie country
generate spot = 0
replace spot = 1 if log_boxoffice!=.
encode country, gen(nation)

regress log_boxoffice log_budget nomination G PG R
```

```
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation, vce(robust)
est store ols_pool

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation, vce(robust)

local j=1
gen counter=1
while counter <= 12 {
    est restore est`j'
    predict p1, xb
    replace p1 = -p1
    gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
    gen capphi = normal(p1)
    gen invmills_`j' = phi/(1-capphi)
    drop p1 phi capphi
    local j = `j'+1
    replace counter = counter+1
}
drop _est_est* counter

          gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
          gen capphi = normal(p1)
          gen invmills_`j' = phi/(1-capphi)

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_1-invmills_12, vce(robust)
est store pool
```

```
bysort country: egen hdi_mean= mean(hdi)
xtile hdi_quant= hdi_mean

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if hdi_quant==1, vce(robust)

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if hdi_quant==2, vce(robust)

local j = 13
gen counter = 13
while counter <= 36 {
   est restore est`j'
   predict p1, xb
   replace p1 = -p1
   gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
   gen capphi = normal(p1)
   gen invmills_`j' = phi/(1-capphi)
   drop p1 phi capphi
   local j = `j'+1
   replace counter = counter+1
}
drop _est_est* counter

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_13-invmills_24 if hdi_quant==1, vce(robust)
```

```
est store hdi1

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_25-invmills_36 if hdi_quant==2, vce(robust)
est store hdi2

xtile cdist_quant= c_distance

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if cdist_quant==1, vce(robust)

bysort year: eststo: probit spot trade_freedom
log_budget nomination G PG R sequel sequel_plus
actionall drama thrillerall original_screenplay
i.year i.nation if cdist_quant==2, vce(robust)

local j=37
gen counter=37
while counter<=60 {
   est restore est‘j’
   predict p1, xb
   replace p1=-p1
   gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
   gen capphi = normal(p1)
   gen invmills_‘j’ = phi/(1-capphi)
   drop p1 phi capphi
   local j= ‘j’+1
   replace counter=counter+1
}
```

```
drop _est_est* counter

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_37-invmills_48 if cdist_quant==1, vce(robust)
est store cdist1

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_49-invmills_60 if cdist_quant==2, vce(robust)
est store cdist2

probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation, vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_heck_pool = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation
invmills_heck_pool, vce(robust)
est store heck_pool

probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
```

```
original_screenplay i.year i.nation if hdi_quant==1,
vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_heck_hdiq1 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_heck_hdiq1
if hdi_quant==1, vce(robust)
est store heck_hdi1

probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if hdi_quant==2,
vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_heck_hdiq2 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_heck_hdiq2
if hdi_quant==2, vce(robust)
est store heck_hdi2
```

```
probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if cdist_quant==1,
vce(robust)

predict p1, xb
replace p1 = -p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_cdist1 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation invmills_cdist1
if cdist_quant==1, vce(robust)
est store heck_cdist1

probit spot trade_freedom log_budget nomination G
PG R sequel sequel_plus actionall drama thrillerall
original_screenplay i.year i.nation if cdist_quant==2,
vce(robust)

predict p1, xb
replace p1=-p1
gen phi = (1/sqrt(2*_pi))*exp(-(p1^2/2))
gen capphi = normal(p1)
gen invmills_cdist2 = phi/(1-capphi)
drop p1 phi capphi

regress log_boxoffice log_budget nomination G PG R
sequel sequel_plus actionall drama thrillerall
```

```
original_screenplay i.year i.nation invmills_cdist2
if cdist_quant==2, vce(robust)
est store heck_cdist2
```

6.2 Script II: La Grande Bellezza

6.2.1 Commented code

```
iis year
```

The seldom used `iis` command declare the time dimension of the dataset without the need of declaring also the panel variable as in `xtset`.

```
foreach var of varlist comedy-documentary {
qui gen subs_‘var’=‘var’*log_subs
qui replace subs_‘var’=0 if subs_‘var’==.
}
foreach var of varlist comedy-documentary {
qui gen nosubs_‘var’=‘var’
qui replace nosubs_‘var’=0 if subs_‘var’!=0
}
```

The first loop generate iteration variables between genre and subsidization . The command `foreach` call variables from the list comedy, drama, thriller, documentary. The second loop is then used to generate a dummy variable that takes value 1 if a movie belongs to a given genre but didn't received public funding and 0 otherwise.

```
xtreg log_box log_subs drama comedy thriller, fe
est store fe_reg

xtreg log_box log_subs drama comedy thriller, re
est store re_reg
```

`xtreg` command fits regression models to panel data. with the `fe` option, it fits fixed-effects models (by using the within regression estimator); and with the `re` option, it fits random-effects models by using the GLS estimator (producing a matrix-weighted average of the between and within results).

```
hausman fe_reg re_reg
```

To choose between random and fixed effects, the Hausman test is performed.

```
xtreg log_box subs_comedy nosubs_comedy
subs_drama nosubs_drama subs_thriller nosubs_thriller
subs_documentary nosubs_documentary, fe
est store re_iter_reg

xtreg log_box subs_comedy nosubs_comedy
subs_drama nosubs_drama subs_thriller nosubs_thriller
subs_documentary nosubs_documentary, re
est store fe_iter_reg

hausman fe_iter_reg re_iter_reg

poisson prizes festivals log_subs comedy drama thriller
documentary if festivals>0, irr
est store prizes_pois
```

Poisson regression fits models of the number of occurrences (counts) of an event. Here the condition `if festivals>0` limit the estimation to those movies that competed at film festivals. The `irr` option reports estimated coefficients transformed to incidence-rate ratios, that is, e^{β_i} rather than β_i. Standard errors and confidence intervals are similarly transformed.

```
nbreg prizes festivals log_subs comedy drama thriller
documentary if festivals>0, irr
est store prizes_nbreg
```

With the same restriction as above, the model is estimated with a negative binomial. In this model, the count variable is believed to be generated by a Poisson-like process, except that the variation is greater than that of a true Poisson.

```
poisson prizes festivals  nosubs_comedy subs_comedy
nosubs_drama subs_drama nosubs_thriller subs_thriller
```

```
nosubs_documentary subs_documentary if festivals>0, irr
est store prizes_poiss_iter

nbreg prizes festivals subs_comedy subs_drama  subs_thriller
subs_documentary nosubs_comedy nosubs_drama
nosubs_thriller nosubs_documentary if festivals>0, irr
est store prizes_nbreg_iter
```

6.2.2 Naked code

```
iis year

foreach var of varlist comedy-documentary {
qui gen subs_‘var’=‘var’*log_subs
qui replace subs_‘var’=0 if subs_‘var’==.
}

foreach var of varlist comedy-documentary {
qui gen nosubs_‘var’=‘var’
qui replace nosubs_‘var’=0 if subs_‘var’!=0
}

xtreg log_box log_subs drama comedy thriller
documentary, fe
est store fe_reg

xtreg log_box log_subs drama comedy thriller
documentary, re
est store re_reg

hausman fe_reg re_reg

xtreg log_box subs_comedy nosubs_comedy
```

```
subs_drama nosubs_drama subs_thriller nosubs_thriller
subs_documentary nosubs_documentary, fe
est store re_iter_reg

xtreg log_box subs_comedy nosubs_comedy
subs_drama nosubs_drama subs_thriller nosubs_thriller
subs_documentary nosubs_documentary, re
est store fe_iter_reg

hausman fe_iter_reg re_iter_reg

poisson prizes festivals log_subs comedy drama thriller
documentary if festivals>0, irr
est store prizes_pois

nbreg prizes festivals log_subs comedy drama thriller
documentary if festivals>0, irr
est store prizes_nbreg

poisson prizes festivals  nosubs_comedy subs_comedy
nosubs_drama subs_drama nosubs_thriller subs_thriller
nosubs_documentary subs_documentary if festivals>0, irr
est store prizes_poiss_iter

nbreg prizes festivals subs_comedy subs_drama  subs_thriller
subs_documentary nosubs_comedy nosubs_drama
nosubs_thriller nosubs_documentary if festivals>0, irr
est store prizes_nbreg_iter
```

6.3 Commands references

- bysort http://www.stata.com/manuals13/dby.pdf
- drop http://www.stata.com/manuals13/ddrop.pdf
- egen http://www.stata.com/manuals13/degen.pdf
- encode http://www.stata.com/manuals13/dencode.pdf
- fillin http://www.stata.com/manuals13/dfillin.pdf
- foreach http://www.stata.com/manuals13/pforeach.pdf
- hausman http://www.stata.com/manuals13/rhausman.pdf
- iis http://www.stata.com/manuals13/xtxtset.pdf
- local http://www.stata.com/manuals13/pmacro.pdf
- nbreg ttp://www.stata.com/manuals13/rnbreg.pdf
- poisson ttp://www.stata.com/manuals13/rpoisson.pdf
- predict http://www.stata.com/manuals13/p1_predict.pdf
- probit http://www.stata.com/manuals13/rprobit.pdf
- regress http://www.stata.com/manuals13/rregress.pdf
- vce(robust) http://www.stata.com/manuals13/xtvce1_options.pdf
- while http://www.stata.com/manuals13/pwhile.pdf
- xtile http://www.stata.com/manuals13/dpctile.pdf
- xtreg ttp://www.stata.com/manuals13/xtxtreg.pdf

Chapter 7

Appendix III: Regression Tables

Table 7.1: Chamberlain approach - Determinants of film revenues

budget	1.143*** (50.84)
nomination	1.094*** (33.03)
drama	-0.395*** (-9.05)
action	0.173*** (6.00)
thriller	0.296*** (8.08)
G	-0.450*** (-6.92)
PG	-0.112*** (-3.88)
R	-0.0576 (-1.90)
sequel	0.294*** (8.57)
sequel2	0.458*** (11.42)
original screenplay	0.0735** (3.07)
Adjusted R^2	0.594
N	37992

Omitted: comedy, PG-13; t statistics in parentheses.
(*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$

Table 7.2: Chamberlain approach - movies revenues — HDI groups

	low/avg Human Development	high Human Development
budget	1.107*** (27.24)	1.139*** (42.34)
nomination	0.631*** (11.81)	1.208*** (26.30)
drama	-0.596*** (-9.28)	-0.0871 (-1.67)
actionall	0.146*** (3.36)	0.214*** (5.90)
thriller	0.292*** (5.04)	0.379*** (7.73)
G	-0.265* (-2.33)	0 (.)
PG	-0.106* (-2.34)	-0.00344 (-0.09)
R	-0.401*** (-8.39)	0.0858* (2.25)
sequel	0.402*** (7.51)	0.365*** (8.09)
sequel2	0.280*** (4.27)	0.411*** (7.51)
original screenplay	0.223*** (6.56)	0.0781* (2.32)
Adjusted R^2	0.656	0.536
N	15493	21501

Omitted: comedy, PG-13; t statistics in parentheses.
(*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$

Table 7.3: Chamberlain approach - movies revenues — CD groups

	low Cultural Distance	high Cultural Distance
budget	1.441*** (19.57)	1.191*** (35.15)
nomination	1.388*** (17.51)	0.939*** (19.42)
drama	-0.164** (-2.81)	-0.648*** (-10.59)
action	0.153** (3.28)	0.234*** (6.42)
thriller	0.342*** (6.38)	0.379*** (7.06)
G	0 (.)	-0.525*** (-5.11)
PG	-0.150*** (-3.40)	-0.0995* (-2.40)
R	0.0509 (0.90)	-0.245*** (-6.81)
sequel	0.445*** (9.10)	0.444*** (8.54)
sequel2	0.510*** (8.09)	0.306*** (4.61)
original screenplay	-0.0171 (-0.46)	0.197*** (5.80)
Adjusted R^2	0.524	0.658
N	20283	17208

Omitted: comedy, PG-13; t statistics in parentheses.
(*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$

Table 7.4: OLS of prizes

	(1)	(2)
festivals	0.281*** (19.08)	0.283*** (18.86)
subsidization	-0.0350 (-1.09)	
comedy	-0.672 (-0.86)	
drama	-0.560 (-0.69)	
thriller	*(omitted)*	
documentary	-4.095*** (-4.00)	
nosubs comedy		2.955 (0.68)
subs comedy		0.187 (0.60)
nosubs drama		3.686 (0.83)
subs drama		0.170 (0.55)
nosubs thriller		2.702 (0.61)
subs thriller		0.266 (0.83)
nosubs documentary		-2.417 (-0.54)
subs documentary		0.0242 (0.07)
N	461	461
Adjusted R^2	0.720	0.724

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7.5: Negative binomial model for prizes

	coefficients	incidence ratio
festivals	0.0358*** (22.17)	1.03
ln subsidization	-0.0138* (-2.34)	0.98
comedy	0.573** (2.89)	1.77
drama	0.878*** (4.49)	2.40
thriller	0.684** (2.61)	1.98
documentary	*(omitted)*	
N		461
Pseudo R^2		0.165

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7.6: Negative binomial model for prizes with iteration between budget and genres

	coefficients	incidence ratio
festivals	0.0358*** (21.91)	1.04
nosubs comedy	0.898 (0.92)	2.45
subs comedy	0.0569 (0.8)	1.06
nosubs drama	1.316 (1.36)	3.73
subs drama	0.0706 (1.04)	1.07
nosubs thriller	0.531 (0.51)	1.70
subs thriller	0.0744 (1.99)	1.08
nosubs documentary	0.420 (0.42)	1.52
subs documentary	0.00828 (0.11)	1.01
N		461
Pseudo R^2		0.167

t statistics in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Bibliography

Bagella, M. and Becchetti, L. (1999) The determinants of motion picture box office performance: evidence from motion pictures produced in Italy. *Journal of Cultural Economics* 23: 237–256.

Davidson, R. and MacKinnon, J.G. (2005). Econometric Theory and Methods. Oxford University Press.

De Vany, A. (2004) Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry. London: Routledge.

De Vany, A. and Walls, D. (1999) Uncertainty in the movie industry: does star power reduce the terror of the box office? *Journal of Cultural Economics* 23: 285–318.

Frankel, J. A., Stein, E., and Wei, S.-J. (1997). Regional trading blocs in the world economic system. Washington, DC: Institute for International Economics.

Fu, W. W., and Lee, T. (2008). Economic and cultural influences on the theatrical consumption of foreign films in Singapore. *Journal of Media Economics*, 21, 1–27.

Fu, W. W., and Sim, C. (2010). Examining international country-to-country flow of theatrical films. *Journal of Communication*, 60, 120–143.

Greene, W.H. (2003). Econometric Analysis. Pearson Education.

Guiso, Sapienza, Zingales (2003)

Hausman, J. A. (1978). "Specification Tests in Econometrics". *Econometrica* 46 (6): 1251–1271.

Hanson and Xiang (2011), "Trade barriers and trade flows with product heterogeneity: An application to US motion picture exports". *Journal* *of International Economics* 83: 14-26.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.

Hofstede, G. (2001). Culture's Consequences, Comparing Values, Behaviors, Institutions, and Organizations Across Nations. Sage, Thousand Oaks CA.

Jansen, C. (2005) The performance of German motion pictures, profits and subsidies: some empirical evidence. *Journal of Cultural Economics* 29(3): 191–212.

Jayakar, K. P., and Waterman, D. (2000). The economics of American theatrical movie exports: An empirical analysis. *Journal of Media Economics*, 13, 153–169.

Lee, F. L. F. (2006). Cultural discount and cross-culture predictability: Examining US movies' box office in Hong Kong. *Journal of Media Economics*, 19, 259–278.

Lee, F. L. F. (2008). Hollywood movies in East Asia: examining cultural discount and performance predictability at the box office. *Asian Journal of Communication*, 18, 117–136.

Lee, F. L. F. (2009). Cultural discount of cinematic achievement: The academy awards and U.S. movies' East Asian box office. *Journal of Cultural Economics*, 33, 239–263.

Lee, S. W. (2002). An economic analysis of the movie industry in Japan. *Journal of Media Economics*, 15, 125–139.

Marvasti, A. and Canterbery, R. (2005) Cultural and other barriers to motion picture trade. *Economics Inquiry* 43(1): 39–54.

Melitz, M. (2008). International Trade and Heterogeneous Firms, in New Palgrave Dictionary of Economics, 2nd Edition. Palgrave Macmillan

McKenzie, J. (2012) The Economics of movies: a literature survey. *Journal of Economic Surveys*, 26(1): 42-70.

McKenzie, J. and Walls, W. D. (2012). The Changing Role of Hollywood in the Global Movie Market, *Journal of Media Economics* 25(4): 198-219.

McKenzie, J. and Walls, W. D. (2013). Australian films at the Australian box office: performance, distribution, and subsidies, Journal of Cultural Economics, 37(2): 247-269

Prag, J. and Casavant, J. (1994) An empirical study of the determinants of revenue and marketing expenditures in the motion picture industry. *Journal of Cultural Economics* 18(3): 217–235.

Rauch, J. E., Trindade, V. (2009). Neckties in the tropics: a model of international trade and cultural diversity. *Canadian Journal of Economics* 42, 809–843.

Wooldridge J.M (2010). Econometric Analysis of Cross Section and Panel Data. MIT press.