Accounting for latent classes in movie box office modeling

Antipov, Evgeny and Pokryshevskaya, Elena
Higher School of Economics, Higher School of Economics

10. December 2010

Accounting for latent classes in movie box office modeling

Evgeny Antipov, The State University – Higher School of Economics

Elena Pokryshevskaya, The State University – Higher School of Economics

Abstract

This paper addresses the issue of unobserved heterogeneity in film characteristics influence on box-office. We argue that the analysis of pooled samples, most common among researchers, does not shed light on underlying segmentations and leads to significantly different estimates obtained by researchers running similar regressions for movie success modeling. For instance, it may be expected that a restrictive MPAA rating is a box office poison for a family comedy, while it insignificantly influences an action movie's revenues. Using a finite mixture model we extract two latent groups, the differences between which can be explained in part by the movie genre, the source, the creative type and the production method. Based on this result, the authors recommend developing separate movie success models for different segments, rather than adopting an approach, that was commonly used in previous research, when one explanatory or predictive model is developed for the whole sample of movies.

Keywords: finite mixture model, box office, latent class, movie success, quantile regression

1 Introduction

Although present day movies are distributed not only through movie theaters, but also on DVD, in the form of digital content for mobile devices, etc., theatrical success still remains very important, because it accounts for a significant share of total revenues and influences the popularity of the film in secondary distribution channels (Elberse and Eliashberg, 2002)[1]. That is why box office modeling is still interesting for academics and practitioners.

Knowledge about the key factors of the financial success of a movie is of great use in making investment and production related decisions. However, researchers often achieve very controversial results about the influence of different film characteristics on its box office, which makes it difficult to use such results to optimize movie production and distribution. For example, estimates of budget elasticity of the box office revenue vary from 0.55 (De Vany and Walls, 2004)[2] to 1.44 (Ravid and Basuroy, 2004)[3]. While some authors reported on the negative influence of the MPAA R-rating on box-office (e.g. Litman, 1983[4]; Sawhney and Eliashberg, 1996[5]), others have found the interrelation to be insignificant (e.g., Prag and Cassavant, 1994)[6]. Litman (1983)[4] concluded that winter releases are the most successful, while Sochay (1994)[7] found summer months to be more beneficial. We believe the reasons for such controversies are not only different estimation procedures, regressors, time periods and sample sizes, but also the existence of latent classes of movies, across which parameters of regression coefficients differ significantly. Thus researchers tend to obtain results which reveal the dependencies typical of the latent class, which dominates within their sample. Even if a researcher has a representative dataset, conclusions based on a single model for a heterogeneous sample are likely to be misleading, since, for instance, R-rating may be bad for a comedy, but not for an action movie, budget elasticity of box office may be higher for low-budget films than for high-budget movies, etc. Including various interactions between variables in order to test such hypothesis would result in too many parameters and therefore is not feasible since the typical sample of movies available to researchers consists of 300-1500 movies and cannot be much larger since it should not cover too many years.

In this paper we start from running a few pooled regressions and after that use a finite mixture regression in order to reveal the latent classes of movies, to detect the differences between the segments and to describe them. This model allows for different parameter values for movies in different latent classes, thus providing a rich and flexible functional form. We show that pooled regressions can be very misleading because of significant differences between latent segments of motion pictures.

2 Data

We used the data from *Nash Information Services, LLC,* which are publicly available from www.the-numbers.com. Our analysis is based on a 2001-2009 sample (1276 observations with known US Gross box office revenues, among which 1271 - with known first weekend box office revenues). For each movie in the dataset we have the following characteristics available:

1. Production budget (inflation-adjusted to 2009 prices).
2. Genre (adventure, action, comedy, drama, horror, thriller/suspense, romantic comedy).
3. Distributor. We have chosen 13 distributors that have the greatest number of movies released in 2001-2009 plus “another distributor” category. The names of the companies are not disclosed in the statistical analysis output, because the aim is not to interpret the individual effects of each distributor.
4. MPAA rating (G, PG, PG-13, R/NC-17).
5. Source (Original screenplay, Based on a book/short story, Sequel, Remake, Based on real life events, Based on TV, Based on a comic/graphic novel, Based on a play, Other).
6. Creative type (Contemporary fiction, Dramatization, Factual, Fantasy, Historical fiction, Kids fiction, Science fiction, Super hero).
7. Production method (Live action, Animation/live action, Digital/stop-motion animation, Hand animation/rotoscoping).
8. High stars presence. In our study high stars are actors who have the cumulative US gross box office of over 1 billion dollars for all movies they have acted in before the release of the movie currently in consideration. Inflation is not taken into account, which makes recent roles much more significant than those which were played decades ago. Unfortunately, the number of high stars is known only for movies from “Top 1000 Movies with highest combined star gross” list and we assume that other movies have no stars at all. Despite this rough assumption we expect the participation of high stars to be an important predictor,

since both mean and median US Gross and First Weekend Box Office sales are more than 3 times higher for movies with high stars than for those without.

9. The week of the release since the beginning of the year (from 1 to 53). For the purposes of regression analysis we have recoded this variable into 6 dummy variables: weeks 1-9, weeks 10-18, weeks 19-27, weeks 28-36, weeks 37-45, weeks 46-53. Such intervals not only help to detect seasonality, but are also similar to the intervals chosen by IBM SPSS Statistics CHAID routine.
10. The number of competitors (number of other movies from the sample, which were released within 2 weeks before or after the movie's release).

For better interpretation of regression output the reference categories we used for each qualitative variable are listed in Table 1.

Table 1. Reference categories for qualitative explanatory variables

Variable	Reference category
Genre	action
Distributor	another distributor
MPAA rating	G
Source	original screenplay
Creative Type	contemporary fiction
Production method	live action
The week of the release	weeks 1-9

The dependent variable is the US first weekend box office receipts (inflation-adjusted).

The features of our dataset that distinguish it from those used in other research are:

1. Reasonably large sample, including most movies released in 2001-2009, which allowed us to include a large number of regressors and make sure it is well-balanced, if not representative.
2. Good typology of movies not only by genre, but also by their source, creative type and production method. This allowed us to describe the creative characteristics of each movie in detail.

3 Box-office modeling

3.1 Pooled models of first weekend box office receipts

We use the first weekend box office revenue as a dependent variable, since, on one hand, it is correlated to the total box office revenue and, on the other hand, it is less dependent on uncontrollable characteristics such as the length of theatrical release or marketing efforts after the first weekend. Although some researchers use linear (in variables) regression specification, we do not think that characteristics like R-rating or winter release really add some fixed amount of money to box office – they are more likely to be associated with a constant percentage increase in the generated revenue. That is why we take the logarithm of box office sales as the dependent variable.

We run three types of regression: ordinary least squares (OLS) with heteroscedasticity robust standard errors, quantile regression and robust regression (see Table 2). Robust regression first performs an initial screening based on Cook's distance to eliminate gross outliers before calculating starting values and then performs Huber iterations (Huber, 1964)[8] followed by biweight iterations, as suggested by Li (1985)[9]. Quantile regression allows estimating the median of the dependent variable, conditional on the values of the independent variables. An excellent introduction to quantile regression can be found in Hao and Naiman (2007)[10]. The results produced by quantile regression and robust regression are similar, but the standard errors are different. Robust regression has smaller standard errors because it is not as sensitive to the exact placement of observations near the median. In order to make prediction quality comparable among the three model we use squared correlations between actual and fitted values of the dependent variable (pseudo R^2).

Table 2. Parameter estimates of pooled first weekend box office models

	(1) OLS, White std. errors		(2) Quantile Regression		(3) Robust Regression	
lnBudget	0.764^{***}	(0.062)	0.727^{***}	(0.042)	0.778^{***}	(0.040)

HighstarsPresence	-1.495**	(0.486)	-0.576	(0.417)	-1.033*	(0.401)
HighstarsPresence*lnBudget	0.418***	(0.111)	0.225*	(0.099)	0.311**	(0.095)
number of competitors	-0.010	(0.009)	-0.015	(0.008)	-0.012	(0.008)
pg	-0.287	(0.173)	-0.044	(0.237)	-0.135	(0.233)
pg-13	-0.304	(0.171)	0.033	(0.236)	-0.053	(0.232)
r/nc-17	-0.803***	(0.196)	-0.322	(0.242)	-0.404	(0.238)
Distributor1	1.449***	(0.167)	1.465***	(0.152)	1.299***	(0.147)
Distributor2	1.180***	(0.194)	1.352***	(0.168)	1.186***	(0.162)
Distributor3	1.276***	(0.218)	1.323***	(0.240)	1.068***	(0.232)
Distributor4	0.277	(0.281)	-0.265	(0.216)	-0.049	(0.209)
Distributor5	1.258***	(0.255)	1.440***	(0.180)	1.419***	(0.173)
Distributor6	0.697**	(0.246)	1.069***	(0.183)	0.834***	(0.177)
Distributor7	0.131	(0.243)	-0.028	(0.190)	0.002	(0.183)
Distributor8	1.126***	(0.220)	1.307***	(0.179)	1.090***	(0.172)
Distributor9	1.389***	(0.175)	1.346***	(0.154)	1.205***	(0.148)
Distributor10	1.203***	(0.184)	1.346***	(0.140)	1.224***	(0.136)
Distributor11	-1.534***	(0.268)	-1.585***	(0.232)	-1.807***	(0.229)
Distributor12	1.457***	(0.179)	1.348***	(0.154)	1.281***	(0.149)
Distributor13	1.226***	(0.171)	1.295***	(0.140)	1.114***	(0.135)
week 10-18	-0.229	(0.139)	-0.181	(0.126)	-0.161	(0.123)
week 19-27	-0.428**	(0.142)	-0.245	(0.131)	-0.277*	(0.127)
week 28-36	-0.171	(0.129)	-0.168	(0.125)	-0.238	(0.122)
week 37-45	-0.230	(0.147)	-0.157	(0.129)	-0.174	(0.125)
week 46-53	-0.971***	(0.165)	-0.607***	(0.131)	-0.708***	(0.127)
adventure	-0.278	(0.161)	-0.072	(0.153)	-0.098	(0.149)
comedy	-0.121	(0.145)	-0.068	(0.126)	-0.063	(0.123)
drama	-0.908***	(0.159)	-0.738***	(0.128)	-0.858***	(0.124)
horror	1.048***	(0.172)	0.745***	(0.166)	0.912***	(0.160)
romantic comedy	-0.130	(0.175)	-0.048	(0.173)	-0.084	(0.168)
thriller/suspense	0.055	(0.153)	0.002	(0.149)	-0.037	(0.144)
Constant	-0.624	(0.348)	-0.847**	(0.328)	-0.799*	(0.320)
Observations	1271		1271		1271	
Pseudo R^2	0.62		0.61		0.61	

Standard errors in parentheses
* *p* < 0.05, ** *p* < 0.01, *** *p* < 0.001

The following empirical facts can be inferred from the robust regression analysis of the whole sample of movies (we deliberately avoid any speculation on why such patterns occur, because in the next section of the paper we argue that such aggregate estimates are rather misleading and have little value):

1. A percentage increase in a movie's budget leads to 0.778% increase in the movie's first weekend box office revenue. This value falls within the range of estimates from Walls (2005)[11].

2. Ceteris paribus, the presence of high stars increases budget elasticity of demand by about 0.311.
3. Number of competitors has no statistically significant impact on the first weekend box office revenues.
4. Neither PG nor PG-13 ratings lead to a statistically significant decrease in the first weekend box office, while the R-rating's influence is still controversial (but more likely to be negative).
5. The individual effects of most of the important distributors are large and significant with only Distributor 11 negatively impacting the box office.
6. Unlike the beginning of the year weeks 46-53 are bad for a movie release. The second worst period is weeks 19-27 (i.e. May-June).
7. Horror films have a huge advantage compared to action movies (at least on the first weekend), while dramas are likely to provide a far more modest box office result.

3.2 Finite mixture model

The results of the quantile regression estimation describe the pooled sample of movies. Since our dataset covers a reasonably large sample (in terms of sample to population ratio) and we have used methods, which are robust to outliers and the violation of the normality assumption, it may be expected that rather reliable estimates, which reflect the market situation well, have been obtained.

At the same time, we hypothesize that feature importance is variable across different segments of movies. Such idea follows from the distributions of the logarithm of first weekend box office (Fig. 1) and the logarithms of US Gross box office (Fig. 2), as well as from common sense. For instance, restrictive MPAA rating may be critical for family-oriented movies, but not very important or even positive for action movies; the participation of high stars may be important for dramas, but not very important for horror films, etc.

Fig. 1. Kernel density (logarithm of the first weekend box office revenue)

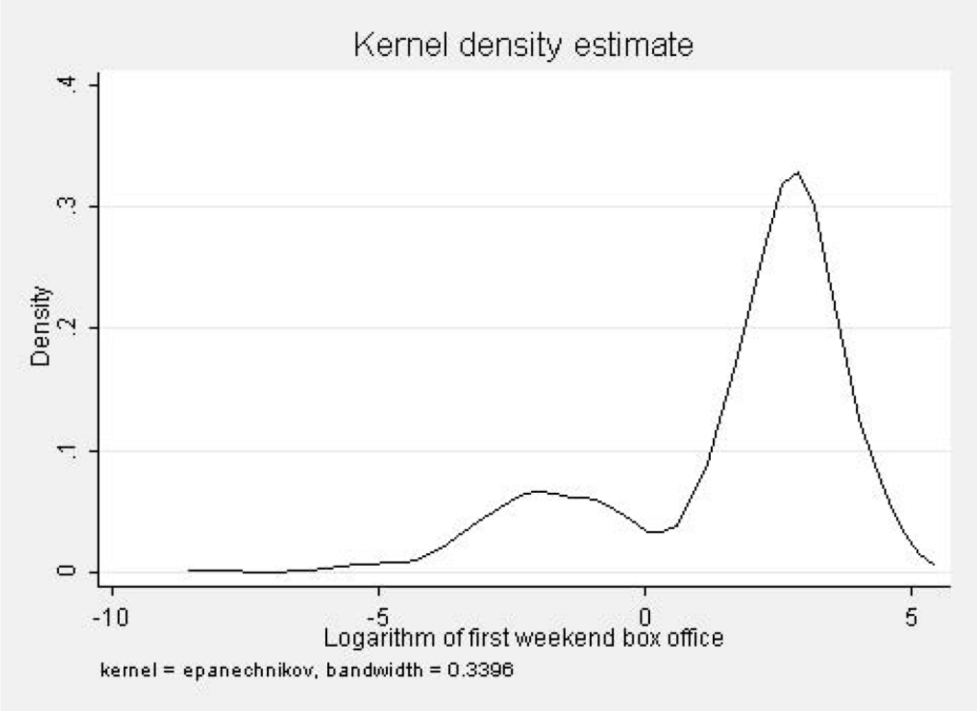
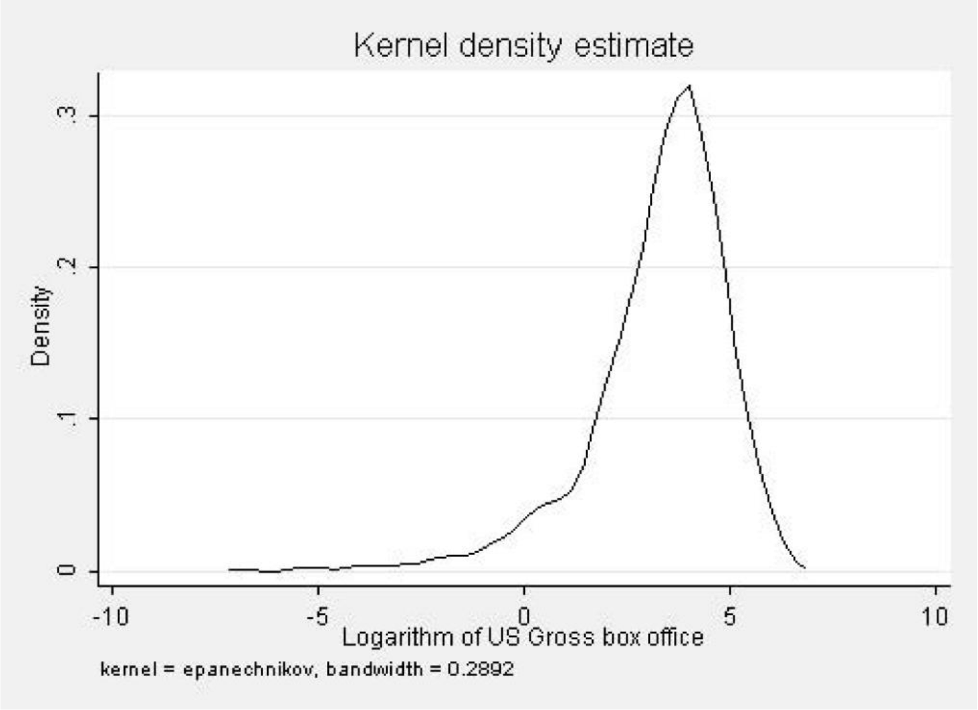


Fig. 2. Kernel density (logarithm of the US gross box office revenue)



The finite mixture model provides a natural and intuitively attractive representation of heterogeneity in a small number of latent classes, each of which may be regarded as a 'type', or a 'segment'. In such a model, the population may be thought to consist of a mixture of K subpopulations each with its own set of regression coefficients. Below we assume K=2, since a mixture of three groups seems to provide a very small increase in the likelihood measure of quality compared to the 2-component solution. Each subpopulation is described by a standard normal linear model. A central feature of the model is that the composition of subpopulations cannot be explicitly identified a priori (Morduch and Stern, 1997)[12].

The results of some studies (Heckman and Singer, 1984)[13] suggest that such finite mixture models provide good approximations even if the underlying mixing distribution is continuous. In addition, the finite mixture approach is semiparametric - it does not require any distributional assumptions for the mixing variable. The finite mixture models are estimated using maximum likelihood and cluster-corrected robust standard errors are used throughout for inference purposes. These are implemented using the Stata package *fmm*[1], which fits a finite mixture regression model of a dependent variable on independent variables using maximum likelihood estimation. The model is a K-component finite mixture of densities, with the density within a class k allowed to vary in location and scale. Optionally, the mixing probabilities may be specified with covariates. For a rigorous introduction to finite mixture models please refer to McLachlan and Peel (2000)[14].

We build a 2-component regression model of the first weekend box office (see Table 3) – the indicator of movie success for which the presence of a mixture of two normal distributions is most likely the case according to the shape of its distribution (Fig. 1). Log pseudolikelihood measure for the 2-component model is -1979, that is significantly larger than the same indicator for a single component model (-2278).

Table 3. Parameter estimates of 2-component FMM

[1] *fmm* package was developed by Partha Deb (Hunter College and The Graduate Center, City University of New York, USA) and can be accessed through net resources search available in Stata

	Component1		Component2	
lnBudget	0.377***	(0.040)	0.771***	(0.075)
HighstarsPresence	-1.357***	(0.346)	-1.484*	(0.714)
HighstarsPresence*lnBudget	0.436***	(0.082)	0.385*	(0.179)
number of competitors	-0.015*	(0.006)	-0.018	(0.017)
pg	0.052	(0.164)	-0.913	(0.574)
pg-13	0.164	(0.158)	-0.924	(0.562)
r/nc-17	0.115	(0.164)	-1.542**	(0.563)
Distributor1	0.288*	(0.113)	2.226***	(0.360)
Distributor2	0.327**	(0.124)	1.521***	(0.376)
Distributor3	0.252	(0.190)	2.096***	(0.520)
Distributor4	0.174	(0.198)	0.496	(0.343)
Distributor5	0.404**	(0.141)	1.005*	(0.473)
Distributor6	0.033	(0.125)	0.306	(0.485)
Distributor7	-0.360*	(0.183)	0.291	(0.335)
Distributor8	0.311*	(0.152)	1.403***	(0.378)
Distributor9	0.082	(0.124)	2.306***	(0.351)
Distributor10	0.273**	(0.103)	1.565***	(0.340)
Distributor11	-4.640***	(0.258)	-0.099	(0.498)
Distributor12	0.203	(0.117)	2.264***	(0.348)
Distributor13	0.171	(0.107)	1.748***	(0.322)
week 10-18	-0.107	(0.092)	-0.200	(0.292)
week 19-27	-0.106	(0.097)	-0.344	(0.296)
week 28-36	-0.140	(0.096)	0.040	(0.286)
week 37-45	-0.072	(0.099)	-0.407	(0.283)
week 46-53	-0.249*	(0.102)	-1.490***	(0.288)
Constant	1.374***	(0.244)	-1.211	(0.668)
Proportion (%)	57.5		42.5	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The differences between the two groups in both the significance and the magnitude of impact are impressive. This evidence suggests that the movie market is composed of two unlike groups. About 57.5% of movies belong to the first component. They appeared to be irresponsive to restrictive ratings, have lower budget elasticity, are much less influenced by the distributor, but very unlikely to succeed if distributed by company 11. Finally this segment is much less influenced by the time of release than segment 2. The second latent class is very sensitive to restrictive ratings, especially R/NC-17. Distributors 1, 3, 9 and 12 are strongly preferred for movies from this segment, while no major distributor is associated with a significant decrease in first weekend box office. If project managers target collecting large first weekend revenues, weeks 46-53 are practically prohibited.

Therefore judgments based on single-equation models are generally misleading and are more or less valid only in the case when the analysis is based on some homogeneous (in terms of box-office response to characteristics) sample of movies. However, it is worth mentioning, that the differences between the two latent classes appeared to be somewhat less impressive when the US Gross box office revenue was used as a dependent variable. This can possibly be explained by the actions taken by distributors to change their marketing efforts, while their film is showing in cinemas, which helps to fix some mistakes made at the earlier stages.

We expect creative characteristics and movie budget to do a good job in classifying movies into two classes. The distribution of the posterior probability of belonging to class 1 is far from normal, that is why we again use quantile regression (see Table 4) to see how well the genre, the creative type, the production method and the script source explain the probability of belonging to segment 1, which is the one much less sensitive to movie characteristics.

Table 4. Parameter estimates of quantile regression (dependent variable: posterior probability of belonging to class 1 (%))

	Quantile regression	
adventure	-1.568	(2.811)
comedy	-1.764	(2.293)
drama	-59.627^{***}	(2.422)
horror	12.760^{***}	(3.034)
romantic comedy	0.815	(3.166)
thriller/suspense	-1.297	(2.740)
dramatization	2.245	(3.584)
factual	-25.775^{***}	(2.413)
fantasy	-1.456	(2.385)
historical fiction	-5.617^{*}	(2.204)
kids fiction	3.006	(3.182)
science fiction	-3.637	(2.621)
super hero	-4.686	(4.603)
animation/live action	1.656	(3.879)
digital/stop-motion animation animation	-4.323	(3.862)
hand animation/rotoscoping	-17.027^{**}	(5.308)
based on book/short story	0.771	(1.732)
sequel	1.395	(2.256)
remake	3.600	(2.905)
based on real life events	-11.943^{**}	(4.181)
based on tv	-2.362	(3.781)

based on comic/graphic novel	2.036	(4.114)
based on play	-11.468*	(5.196)
other	-7.658*	(3.600)
budget	0.028	(0.016)
Constant	75.990***	(2.285)
Observations	1271	
Pseudo R^2	0.2	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Although the precision of this regression model leaves much to be desired (however, R^2 is not that low for a model, where most regressors are dummy variables), every group of dummy regressors has at least one variable that has a significant influence on the posterior probability of belonging to class 1, which supports our hypothesis. Since the constant term is about 76, the estimated probability of belonging to class 1 is 76% for a contemporary fiction low-budget action movie, based on original screenplay, produced using the live action method. Horror films are especially likely to belong to latent class 1. The following characteristics are indicative of type 2 movies: genre – drama, creative type – factual, production method – hand animation/rotoscoping, source – based on real life events or based on a play or some source not included in our analysis.

4 Conclusion and future research

Using three regression techniques (OLS with heteroscedasticity robust standard errors, quantile and robust regressions) we have obtained a few empirical facts about the movie market in general. Using a finite mixture model (FMM) we have shown that it is reasonable to account for latent classes of movies when explaining the first weekend box office results. The finite mixture model is supported both by a priori reasoning and by meaningful a posteriori differences in the behavior of the latent segments. Our results suggest that there are significant differences in the influence of MPAA rating, production budget, high stars participation and distributor on the first weekend box office across two groups. These differences are revealed by the FMM approach but are masked by a single equation method that assumes homogeneity in response. In the case of the US gross theatrical revenue some differences become less significant, but still worth accounting for.

The class to which a movie belongs can be predicted using some creative characteristics, such as the genre, the creative type of a movie, its source and production method. However, our model for detecting to which class a particular movie belongs is far from being perfect. Consulting a movie industry expert enhanced by the insights from our study may help researchers to better reveal segments which should be analyzed separately.

Despite the fact that we used a rich, yet publicly available, data, more refined methodologies for measuring independent variables (especially star power and competitors) are required. We also plan to develop similar predictive models of the relative performance, which, in our view, could be of greater value for investors than box-office revenue models, which are usually far from being accurate.

References

1. Elberse, A. and Eliashberg, J. (2002) Dynamic Behavior of Consumers and Retailers Regarding Sequentially Released Products in International Markets: The Case of Motion Pictures. The Wharton School, University of Pennsylvania. Working paper.
2. De Vany, A.S. and Walls, W.D. (2004) Big budgets, big openings and legs: Analysis of the blockbuster strategy. *Asian Economic Review*, 47(2): 308-328.
3. Ravid, S.A. and Basuroy, S. (2004) Beyond morality and ethics: Executive objective function, the R-rating puzzle, and the production of violent films. *Journal of Business*, 77(2): 155–192.
4. Litman, B.R. (1983) Predicting the success of theatrical movies: An empirical study. *Journal of Popular Culture*, 16: 159-175.
5. Sawhney, M.S. and Eliashberg, J. (1996) A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science*, 15(2): 113– 131.

6. Prag, J. and Cassavant, J. (1994) An empirical study of determinants of revenues and marketing expenditures in the motion picture industry. *Journal of Cultural Economics*, 18(3): 217-35.
7. Sochay, S. (1994) Predicting the performance of motion pictures. *Journal of Media Economics*, 7(4): 1–20.
8. Huber, P.J. (1964) Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35: 73–101.
9. Li, G. (1985) Robust regression. In: D.C. Hoaglin, F. Mosteller, and J.W. Tukey (eds.) *Exploring Data Tables, Trends, and Shapes*. New York: Wiley, pp. 281–340.
10. Hao, L. and Naiman, D.Q. (2007) *Quantile Regression*. Thousand Oaks, CA: Sage.
11. Walls, W.D. (2005) Modelling heavy tails and skewness in film returns. *Applied Financial Economics*, 15(17): 1181-1188.
12. Morduch, J. and Stern, H.S. (1997) Using Mixture Models to Detect Sex Bias in Health Outcomes in Bangladesh. *Journal of Econometrics*, 77: 259-276.
13. Heckman, J. and Singer, B. (1984) A Method of Minimizing the Distributional Impact in Econometric Models for Duration Data. *Econometrica*, 52: 271-320.
14. McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models.* New York: John Wiley.