Empir Econ (2009) 37:447–461
DOI 10.1007/s00181-008-0240-z

ORIGINAL PAPER

Screen wars, star wars, and sequels

Nonparametric reanalysis of movie profitability

W. D. Walls

Received: 15 November 2006 / Accepted: 15 July 2008 / Published online: 23 September 2008

Abstract In this paper we use nonparametric statistical tools to quantify motion-picture profit. We quantify the unconditional distribution of profit, the distribution of profit conditional on stars and sequels, and we also model the conditional expectation of movie profits using a nonparametric data-driven regression model. The flexibility of the nonparametric approach accommodates the full range of possible relationships among the variables without prior specification of a functional form, thereby capturing nonlinearities and interactions without introducing possible specification bias. We find that marginal returns to budgets and opening screens vary over the domain of these variables. We also find that the conditional distribution of movie profit and the expected level of profit are related to the use of movie stars and sequels.

Keywords Motion-picture profit · Nonparametric regression · Hollywood economics

JEL Classification L82 · C14

1 Introduction

The motion-picture industry is routinely criticized for producing films that have ever bigger budgets, that showcase brand name movie stars, that are released on an increasingly large number of cinema screens, and that are often sequels. Competition for the audience has intensified in the past decade and film budgets have escalated rapidly as films are increasingly being promoted and advertised nationally. The strategy of the

W. D. Walls (✉)
Department of Economics, University of Calgary,
2500 University Drive NW, Calgary, AB T2N 1N4, Canada
e-mail: wdwalls@ucalgary.ca

movie industry is to use big-budget widely-released star-filled films to generate high profits. In this paper, we use nonparametric statistical methods to analyze empirically how motion-picture profits are related to big budgets, wide releases, movie stars, and other film attributes.

Several empirical studies have examined the role of budgets, screens, stars, and sequels on box-office revenue.[1] De Vany and Walls (2004) provide an analysis of motion-picture profit focusing on the stable Paretian distribution; in their paper they estimate the density of profits for a sample of motion pictures with and without stars. The empirical analysis in this paper extends the earlier work of De Vany and Walls (2004) on modeling the distribution of motion-picture profit in two ways. First we take a data-driven nonparametric approach to modeling the distribution of motion-picture profit. Second, the completely nonparametric approach is applied in a regression framework permitting the estimation of the marginal effects of budgets, screens, sequels, and stars on motion-picture profit without imposing any assumptions of distribution or functional form on the statistical analysis.

There are several reasons to take a nonparametric approach to modeling motion-picture profit. As shown by De Vany and Walls (1999, 2002, 2004), motion-picture revenues and profits are characterized by skewness and heavy tails, and standard statistical models such as the Gaussian and log Gaussian can be rejected.[2] Their approach is to model the unconditional distribution of profit using the skew-stable model, a model which fits the skewness and tails of the distribution well but which does not lend itself to quantification of how various explanatory variables affect the parameters of the distribution or to estimation of the marginal effects of the attributes of a film and its theatrical release. As an alternative to the parametric stable Paretian approach, we investigate the use of recently-developed nonparametric statistical tools to quantify the distribution of movie profit and how various explanatory variables are related to movie profit. Thus, the analysis in this paper is more general than what has come before in that it is nonparametric, and in that it explicitly accounts for the attributes of a movie and its theatrical release in a model that can flexibly accommodate the skewness and heavy tails of motion-picture profit.[3]

Using nonparametric methods we quantify the unconditional distribution of profit, and the distribution of profit conditional on stars and sequels. We also model the conditional expectation of movie profits using a completely nonparametric data-driven regression model based on the multivariate product kernel with mixed data types (continuous and discrete) developed by Li and Racine (2004) and Racine and Li (2004).

[1] See, for example, Albert (1998, 1999); De Vany and Walls (1996, 1997, 1999, 2002, 2004); Litman (1983); Litman and Ahn (1998); Litman and Kohl (1989); Nelson et al. (2001); Prag and Cassavant (1994); Ravid (1999); Sedgwick and Pokorny (1999); Smith and Smith (1986), and Wallace et al. (1993).

[2] As shown in Fig. 1 below, the distribution of motion-picture profit has too much mass in the center of the distribution, as well as too much mass in the upper tail, to be consistent with a normal distribution. The nonparametric model does not impose a priori restrictions on the shape of the distribution, though heavy tails can pose estimation issues; this point is discussed further below.

[3] Although the nonparametric model has the flexibility to accommodate skewness and heavy tails in the conditional distribution of movie profit, the estimator—in particular the fixed bandwidth estimator—may be affected by the presence of heavy tails that contain few observations that are distant from the sample mean.

The nonparametric approach captures the full range of possible relations among the variables without prior specification of a functional form or making distributional assumptions. In our empirical application to a sample of nearly two thousand films, we find increasing marginal returns to budgets and opening screens over the domain of these variables, a striking result that helps to explain the industry's move toward widely released big-budget movies. Our results confirm the 'curse of the superstar' result of De Vany and Walls (2004), but we find no such 'curse' for sequels.

2 Nonparametric Kernel regression

The statistical technique used in this paper is based on Fan and Gijbels's (1992, 1996) local linear nonparametric regression model. The nonparametric regression model has been operationalized by data-driven cross-validated bandwidth selection and recent advances in generalized kernel estimation due to Li and Racine (2004) and Racine and Li (2004). In the standard nonparametric regression model a dependent variable y is related to a vector of independent variables x through some unknown and unspecified function $m(\cdot)$

$$y_i = m(x_i) + \epsilon_i \tag{1}$$

where i indexes observations, $m(x_i)$ is an unknown smooth function with argument $x_i = (x_i^c, x_i^u, x_i^o)$, where x_i^c is a vector of continuous regressors such as budget, x_i^u is a vector of regressors that assume unordered discrete values such as genres, x_i^o is a vector of regressors that assume ordered discrete values such as time effects, and ϵ_i is an additive stochastic disturbance.

Expanding a first-order Taylor series about the regression equation (1) at x_j yields

$$y_i \approx m(x_j) + (x_i^c - x_j^c)\beta(x_j) + \epsilon_i \tag{2}$$

where $\beta(x_j)$ is the partial derivative of $m(x_j)$ with respect to x^c, also called the marginal effect and similar in interpretation to a regression coefficient in a linear regression model. There is, however, an important difference between the marginal effect in the nonparametric model and in a parametric model such as a linear regression: the marginal effects in the nonparametric regression model are not restricted to be constant over the domain of the independent variable; in a linear regression model the regression coefficients are constants.

The estimator of the vector of unknowns $\delta(x_j) \equiv (m(x_j), \beta(x_j))'$ is given by

$$\hat{\delta}_{x_j} = \begin{pmatrix} m(x_j) \\ \beta(x_j) \end{pmatrix} = \left[\sum_i K_{\hat{h}} \begin{pmatrix} 1 & (x_i^c - x_j^c) \\ (x_i^c - x_j^c) & (x_i^c - x_j^c)(x_i^c - x_j^c)' \end{pmatrix} \right]^{-1} \times \left[\sum_i K_{\hat{h}} \begin{pmatrix} 1 \\ (x_i^c - x_j^c) \end{pmatrix} y_i \right] \tag{3}$$

where

$$K_{\hat{h}} = \prod_{s=1}^{q} \hat{h}_s^{-1} l^c(x^c(si - x_{sj}^c)/\hat{\lambda}_s^c) \prod_{s=1}^{r} l^u(x_{si}^u, x_{sj}^u, \hat{\lambda}_s^u) \prod_{s=1}^{p} l^o(x_{si}^o, x_{sj}^o, \hat{\lambda}_s^o). \quad (4)$$

K_h is the product kernel, which is the product of the individual kernels for the continuous and discrete variables: l^c is the standard normal kernel function with window width h_s associated with the sth component of x^c; l^u is a variation of Aitchison and Aitken's (1976) kernel function which equals one if $x_{si}^u = x_{sj}^u$ and λ_s^u otherwise; and l^o is the Wang and Van Ryzin (1981) kernel function which equals one if $x_{si}^o = x_{sj}^o$ and $(\lambda_s^0)^{|x_{si}^o - x_{sj}^o|}$ otherwise.[4]

Implementing the nonparametric kernel regression involves two choices: the kernel and the window width (or bandwidth). The difference between the optimal kernel and most kernels used in practice is small, so the choice of kernel is not problematic.[5] But the choice of bandwidths $(h, \lambda^u, \lambda^o)$ can be a sticky issue with a well-known tradeoff: A small bandwidth means that there may not be enough data points resulting in an undersmoothed estimate having low bias and high variance, but choosing a large bandwidth including many data points may result in an oversmoothed estimate having high bias and low variance. There are several alternative data-driven selection methods for the bandwidth. Among the many alternatives, we employ Hurvich et al. (1998) Expected Kullback Leibler (AICc) criteria. This method chooses smoothing parameters using an improved version of a criterion based on the Akaike Information Criteria and has been shown to perform well in small samples.[6] In our empirical application the bandwidths are chosen specifically to minimize

$$\text{AICc}(h, \lambda^u, \lambda^o) = \log(\hat{\sigma}^2) + (1 + tr(H)/N)/[1 - (tr(H) + 2)/N] \quad (5)$$

where

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{N} (y_j - \hat{m}_{-j}(x_j))^2 = \frac{1}{N} y'(I - H)'(I - H)y \quad (6)$$

and $\hat{m}_{-j}(x_j) = Hy_j$ is the leave-one-out estimator of $m(x_j)$. In the application to film earnings set out below, we use the fully nonparametric local linear estimator described above with the bandwidths determined using the AICc criterion.

[4] See Pagan and Ullah (1999) for a general introduction and Hall et al. (2004); Li and Ouyang (2005); Li and Racine (2004, 2006) and Racine and Li (2004) for the technical details of this estimator and its calculation.

[5] See, for example, the discussion in (Pagan and Ullah, 1999, pp. 23–28) and the references therein.

[6] Henderson and Kumbhakar (2006) note that one cause for undersmoothing when using the least-squares cross-validation procedure is the presence of outliers; they also note that AICc was robust to outliers and performed well in comparison to their own robust cross-validation procedure.

Table 1 Film earnings, production budget, and opening screens

	Mean	Median	Std Dev
Profit	−3.35 m	−3.82 m	12 m
Box-office revenue	17.2 m	7.1 m	26.9 m
Production budget	11.9 m	9.3 m	10.3 m
Opening screens	844	858	761
	Number	% of sample	
Movies with stars	326	16.39	
Sequel movies	169	8.50	

'm' indicates millions. Data source is AC Nielson EDI

3 Empirical analysis

3.1 Data description and nonparametric density plots

The data set consists of observations on 1,989 films exhibited in North America beginning in 1985 and continuing through 1996. The data were extracted from the AC Nielson EDI historical database and include cumulative box-office revenue as well as the attributes of a film and its theatrical release. From the EDI database, we selected all films for which data on the variables of interest were available and this resulted in the final sample size of 1,989 films. The EDI data are compiled from the North American distributor-reported box-office figures and are widely regarded as the standard industry source for published information on motion picture theatrical revenues. Table 1 provides summary statistics on the data.[7] Because these are the same data used in a number of other published articles, the empirical results will be directly comparable to those results reported by researchers using different statistical methods. We seek to determine the usefulness of nonparametric regression in analyzing the motion-picture industry, and for this purpose using the same data as prior studies allows us to squarely confront the results obtained in them.

Figure 1 plots the nonparametric kernel density estimate for motion-picture profit against a histogram and an overlaid fitted Gaussian density for the purpose of comparison. Profit clearly does not follow a Gaussian distribution and standard statistical tests reject the null hypothesis that profit is normally distributed.[8]

Figure 2 plots nonparametric kernel density estimates for the profits of movies that feature 'star' talent and those that do not; the figure also plots the upper and lower two-standard-error bounds on the density function estimate.[9] The figure shows that the

[7] A thorough description of the data source, very detailed descriptive statistics and cross-tabulations, and a regression analysis of the EDI data are contained in De Vany and Walls (1999).

[8] For example, the Kolmogorov–Smirnov test, Shapiro–Francia test, Shapiro–Wilk test, and skewness-kurtosis test all result in marginal significance levels of zero.

[9] An actor or director appearing on *Premier*'s annual listing of the hundred most powerful people in Hollywood or on James Ulmer's list of A and A+ actors was considered to be a star in the empirical analysis; this is the same definition of 'star' used in De Vany and Walls (1999). Many thanks to Cassey Lee for compiling the list of stars.

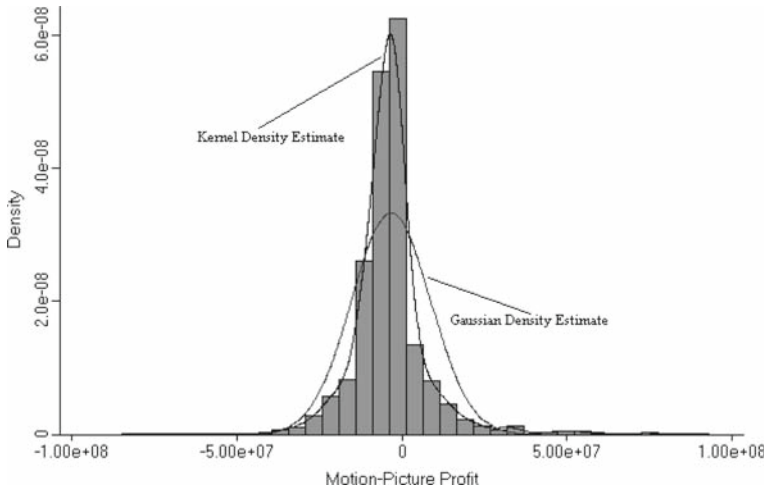

Fig. 1 Unconditional density of motion-picture profit

distribution of profits differ substantially across movies depending on the presence of star talent. The peak of the star profit density occurs at a profit of about −3.8 million, though the peak is much lower than for the non-star density. The peak of the non-star profit density occurs at a profit of −3.3 million, but it has much heavier tails than the density function for star movies.[10] We can formally reject that equality of profit distributions for star and non-star movies and this is consistent with the visual test that the two-standard-error bounds are not overlapping.[11]

Figure 3 plots the nonparametric kernel density estimates for the profits of movies that are sequels (or prequels) and for movies that are non-sequels. It is evident from panel A of Fig. 3 that the distributions for sequels and non-sequels are similar, though the sequel density lies to the right of the density for non-sequels. The mean profit for sequels in the sample was about −0.8 million, while the mean profit for non-sequel movies was about −3.5 million. The highest probability on the sequel density function occurred at a profit of −2.3 million and the highest probability on the non-sequel density function occurred at a profit of −3.7 million. In panel B of the figure we have plotted the density functions with the addition of upper and lower two-standard-error bounds. In a small neighborhood around the peak of the distributions the regions within two standard errors of each density are overlapping, but over most

[10] Note that the peak of the estimated nonparametric density functions differs from the sample mean profit due to the influence of skew and heavy tails in the calculation of the mean: For star movies, the sample mean profit is about −2 million, while for non-star movies the sample mean profit is about −3.4 million. De Vany and Walls (2004) find the same result in the fitted stable Paretian model. Readers unfamiliar with the industry may find the unprofitability of movies a surprise: On average, about 80% of all movies made are unprofitable (Vogel 1998).

[11] We tested for equality of distributions using the Li et al. (2006) test with 500 bootstrap resamples. The computed test statistic for the equality-of-distributions test was 59.7315 and the upper 1% value of the distribution of the test statistic was 1.6105; the computed marginal significance level was zero. The Komogorov–Smirnov test for equality of distributions yields a marginal significance level of zero.

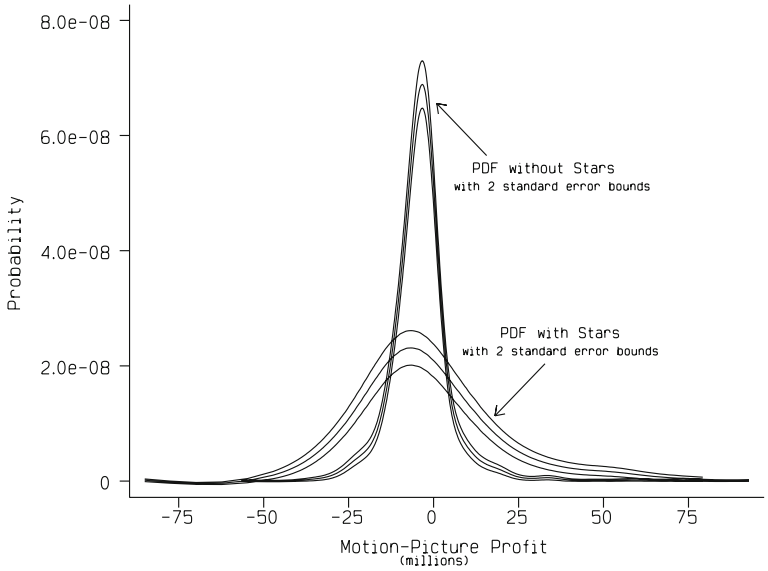

Fig. 2 Stars and the density of motion-picture profit

of the domain of profit the distributions are non-overlapping. The visual conclusion that the distributions of profit for sequel and non-sequel movies differ is supported by a formal statistical test for equality of distributions.[12]

3.2 Nonparametric regression estimates and hypothesis testing results

Before proceeding to estimate the nonparametric model, we test for a known parametric model using the Hsiao et al. (2007) consistent model specification test for mixed categorical and continuous data. Specifically we test the null hypothesis that the parametric model is correctly specified against the alternative that it is not. We used the following specification for the parametric model:

$$\begin{aligned}\text{Profit}_i = {} & \beta_0 + \beta_1 \text{Film Budget}_i + \beta_2 \text{Screens}_i \\ & + \beta_3 \text{Sequel}_i + \beta_4 \text{Star}_i + \Lambda[\text{Genre, Rating, Year}]_i' + \epsilon_i \end{aligned} \tag{7}$$

where i indexes individual movies, Star and Sequel are dummy variables equal to unity when a movie contains a star or is a sequel, respectively, and zero otherwise, Λ is a vector of coefficients conformable to the sets of explanatory variables indicating particular genres, ratings, and years of release, and ϵ_i is a random disturbance. The

[12] We tested for equality of distributions using the Li et al. (2006) test with 500 bootstrap resamples. We can reject the null hypothesis of equality of sequel and non-sequel profit distributions: The computed test statistic was 3.6850 and the upper 1% critical value of the distribution of the test statistic was 1.3173; the marginal significance level was 3.752×10^{-6}. The Kolmogorov–Smirnov test for equality of distributions yields a marginal significance level of zero.

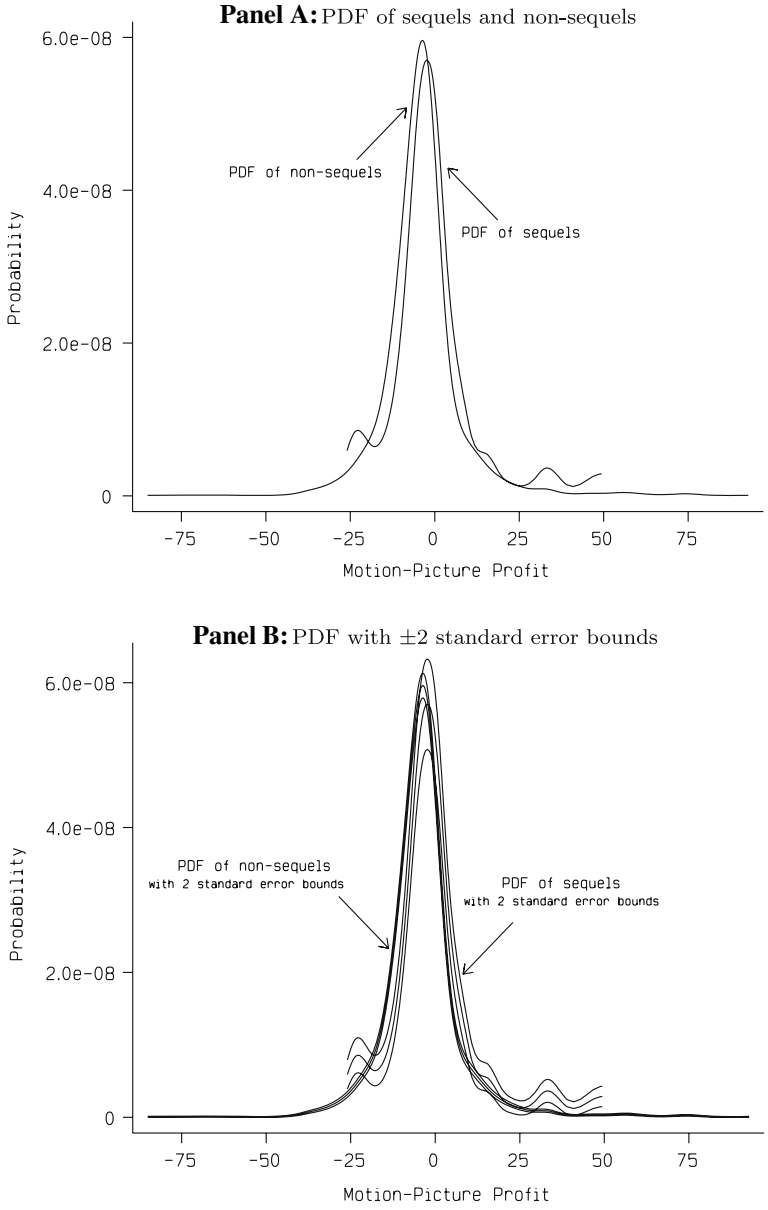

Fig. 3 Sequels and the density of motion-picture profit

computed marginal significance level for the specification test was 0.00204, rejecting the null hypothesis of correct parametric specification at all conventionally used significance levels.[13]

[13] The Hsiao–Li–Racine test was performed with 500 bootstrap resamples. The computed test statistic was 1.53577 and the upper 1% critical value was 0.159249.

Table 2 Summary of estimated model fit

Metric	Model	
	Nonparametric	Least-squares
R^2	0.825	0.209
MAE	3.038	6.203
RMSE	5.285	10.766
Sign	87.280	76.018

Both models are estimated on the full sample of 1989 observations. The within-sample predicted values are compared to the actual values to quantify model fit. R^2 is the squared correlation between actual and fitted values, *MAE* mean absolute error in millions, *RMSE* root mean squared error in millions, *Sign* percentage with the correct sign

Given the formal rejection of the parametric model, we proceed to estimate the nonparametric model using the same set of explanatory variables set out above.[14] The local linear nonparametric model was estimated using the product kernel with constant Nadaraya (1965); Watson (1964) bandwidths selected by the AICc method outlined above.[15] We calculated several goodness-of-fit measures to compare the nonparametric model to the parametric model; these metrics are summarized in Table 2. The R^2 of the kernel regression—calculated as the squared correlation between the fitted and actual values—is 0.825, compared to a value of 0.209 for the linear model. The mean absolute error from the linear model (6.203) was about twice as large as from the nonparametric model (3.038 million). The mean absolute percentage error for the linear model (4.056%) was also about double value for the nonparametric model (2.079%). The various goodness-of-fit metrics support the evidence from the Hsiao–Li–Racine specification test that the nonparametric model is preferred to the linear model.[16]

The estimation output consists not of a single parameter and associated standard errors but of an estimate of the dependent variable and the marginal effects for each data point. For this reason, the output of a fully nonparametric estimation can be difficult to interpret. Graphical display of the output is instructive. In Fig. 4 we plot the marginal effects for the production budget and their associated 95% confidence bands. The average marginal effect is −0.283, indicating that on average an increase

[14] The statistical tools used to perform the nonparametric analysis in this paper are available for most computing platforms, and they are free. The interested reader is referred to the R computing environment Ihaka and Gentleman (1996) and specifically to the nonparametric kernel smoothing package developed by Hayfield and Racine (2006).

[15] We recomputed the bandwidth selection algorithm ten times, using different starting values each time, to ensure that we obtained bandwidths that are a global optimum. After obtaining the optimal bandwidths by the AICc method we computed the standard errors for the fitted values $m(\cdot)$ and the gradients $\beta(\cdot)$ using 500 bootstrap resamples.

[16] It should be mentioned that the fixed bandwidth estimator used here is perhaps not the best choice for data with heavy tails where it would be possible for the tails to be sparesely populated; a variable bandwidth may be preferable in the presence of heavy tails. In the current application, we have nearly 2,000 motion pictures resulting in sufficient observations in the tails for estimation.

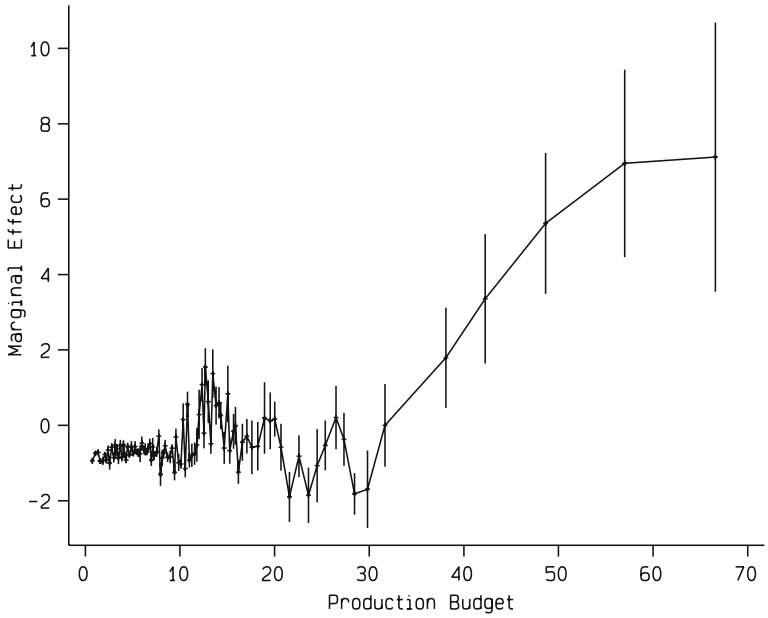

Fig. 4 Marginal effect of production budget on profit. Note: Estimate marginal effects plotted with 95% confidence bands

in production budget of one dollar leads to a loss of about twenty-eight cents. But in the nonparametric regression model the marginal effect need not be constant. The actual pattern of marginal effects in the figure makes clear that the response of profit to changes in production budget does differ substantially as the production budget increases, with the marginal profitability increasing substantially over a 40–80 million dollar range of production budgets; this result still holds when one accounts for the sometimes wide confidence bands. It would be a serious mistake to estimate the impact of production budget on film profitability using a statistical model that constrained the impact to be constant across budget levels.

In Fig. 5 we plot the marginal effects and associated 95% confidence bands for opening screens on film profitability. The average marginal effect is calculated to be 5468. But the plot of the marginal effects over differing levels of the explanatory variable shows that the marginal effect is not constant, even accounting for the sometimes wide confidence bands. The impact of opening screens on film profitability increases rapidly for openings in the range of 1,500–2,200 screens, and then decreases for larger openings. Again, it would be a mistake to quantify the impact of opening screens on profitability using a model that constrained the impact to be constant over all levels of opening screens. The average marginal effect is not a very informative quantity.

We perform a number of hypothesis tests within the nonparametric framework that correspond to the standard exclusion restrictions one would routinely test in the context of a linear regression model, equivalent to testing the null hypothesis that a regression coefficient (or a group of coefficients) equals zero in a linear regression model. In the context of the nonparametric regression model, if the conditional mean $E(Y|X)$

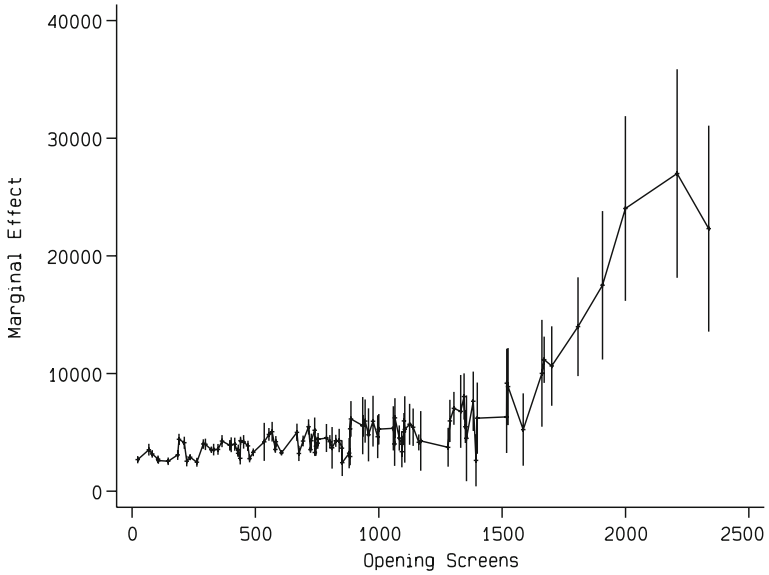

Fig. 5 Marginal effect of opening screens on profit. Note: Estimate marginal effects plotted with 95% confidence bands

is independent of a variable or group of variables $x_{(j)}$, then the partial derivative of the conditional mean with respect to these variables will be zero. Because the partial derivatives in the nonparametric model can vary over the domain of the variable, the standard statistical tests used in parametric models cannot be applied. The null hypothesis to be tested is specifically

$$H_0 : \frac{\partial E(Y|X)}{\partial x_{(j)}} = 0 \quad \forall\, x \in X \tag{8}$$

where the partial derivative must equal zero over the entire domain of every explanatory variable; the alternative hypothesis is that the equality does not hold. Racine (1997) provides a treatment of this approach to hypothesis testing including details of the bootstrapping procedure required to generate the test statistic and its distribution under the null hypothesis.

The statistical tests for the hypotheses that the marginal effects of certain variables or groups of variables are equal to zero—corresponding to the standard t- and F-tests reported in regression output—are summarized in Table 3. We first test the null hypothesis that the budget is orthogonal to the conditional mean, and for this test we obtain a marginal significance level of 0.021. Second, we test the null hypothesis that the conditional mean is orthogonal to the opening screens, and the marginal significance level is 0.028. Third, we test the null hypothesis that the conditional mean is orthogonal to the presence of star talent, and we obtained a marginal significance level of zero. Next, we test the null hypothesis that the conditional mean is orthogonal to the film being a sequel (or prequel), and find that the marginal significance level is 0.024.

Table 3 Summary of hypothesis testing results

Variable	p value
Production budget	0.021
Opening screens	0.028
Star	0
Sequel	0.024
Ratings (group)	0.005
Genre (group)	6.87×10^{-10}

Summary of statistical tests for the null hypothesis that the variable (or group) is orthogonal to the conditional mean of film earnings

Finally, we tested the null hypothesis that the groups of variables representing genres and ratings were orthogonal to the conditional mean and the marginal significance levels for these statistical tests were 6.87×10^{-10} and 0.005, respectively.

3.3 Discussion of results

Films featuring movie stars have profit distributions that differ substantially from those of movies without stars. Not only is the location of the distribution higher, but it has a fundamentally different shape as shown in Fig. 2; these results from nonparametrically estimated distributions confirm the earlier findings of De Vany and Walls (2004) who fit a skew-stable model to profits for movies with and without stars. In addition to modeling the density function of profit only conditioning on star presence, we also model the conditional expectation of profit using the nonparametric regression model. We find that the use of stars is systematically related to the level of profit and that the average impact of including a star is a $6.5 million increase in profit. Our finding is similar to the result from the stable Paretian model of De Vany and Walls (2004) that stars increase expected movie profits by about $7.7 million. The results of the nonparametric model confirm the 'curse of the superstar' result that the expected profit is positive even though star movies actually earn negative profit.

It is common in the movie business that a single film will lead to a series of films such as the *Star Wars, Batman, Police Academy,* and *Die Hard* films. We find that profits for sequels are higher than profits for non-sequels, and that the density functions of profit for these groups—as set out graphically in Fig. 3—are statistically different. We also find that sequels are systematically related to the expected profit even when accounting for the attributes of a film and its theatrical release. We estimate that on average sequels earn $0.88 million more in profit than non-sequel films. Of course, a sequel is no guarantee of success, and this is seen graphically in the way that the shape of the profit distribution is nearly the same for sequel and non-sequel movies. There is no evidence of a 'curse' for sequels, since the expected increase in movie profits and the mean increase in movie profits are both positive for sequel movies compared to non-sequels.

Big and growing budgets are a feature of the Hollywood movie industry. In recent years, fewer films are made while average budgets are rising. The results of our nonparametric regression analysis indicate that this may be a way of capturing the increasing marginal returns of budget. The marginal return to an additional dollar of budget, plotted above in Fig. 4, increases substantially over a budget range beginning at 40

million and continuing through about 70 million. The industry's increasing focus on making fewer movies with larger budgets is consistent with our results, though one would be puzzled if using the results from a least-squares-regression analysis to interpret the industry's actions. A least-squares analysis would estimate that the marginal profit returned to an additional dollar spend on the budget is -0.64 and that it is constant across budget levels; this would lead one to conclude that the industry is operating in a region where there are negative returns to budget.

Another feature of the motion-picture industry is the move in recent years toward wider openings of films in contrast to platform releases. Again, the results of the nonparametric regression analysis reveal why this move may be a way in which the film industry is able to capture increasing marginal returns to opening screens. In Fig. 5 we plot the marginal return to an additional opening screen. The marginal return to screens is not constant as the level of opening screens varies; instead it rises until about 2,200 screens after which it falls. If there are rising marginal returns to opening a film on more screens—due, for example, to the dynamics of movie demand—then it is sensible that the industry has moved toward wider distribution of feature films. Again, if one were to use a least-squares model to estimate returns to opening screens, the estimate would be that an additional screens increases profits by 4,625 across all levels of opening screens.

Finally, we should acknowledge that the nonparametric model and the estimation method applied in this paper is intended to be an incremental step in enhancing our understanding of the film industry. The features of the movie industry that challenge standard statistical models can also pose problems for nonparametric models, especially the characteristic that a handful of extremely successful films can drive industry profitability. The sample of nearly 2,000 motion pictures used in this research has ensured that there are a large number of high-grossing films; however, the empirical methods used here might be less useful with few a small data set because there would be few observations in the tail. Thus, while the results of the present nonparametric analysis help us to understand more about the economics of the film industry, this should be seen as an incremental step in modeling motion-picture profit.

4 Conclusion

Big budgets, wide openings, stars, and sequels are palpable features of the motion-picture industry. This paper has investigated empirically how these features are related to film profitability using recently-developed nonparametric statistical tools. The nonparametric methods used in this research account for the extreme uncertainty of the movie industry—including heavy tails, skewness, and infinite variance—that has been documented in other recent research. Our approach generalizes other work on motion picture profit by making no assumptions of functional form or distribution as well as modeling expected film profitability in a completely nonparametric regression model that accounts for the attributes of a film and its theatrical release.

The results of our nonparametric analysis confirm the 'curse of the superstar' result found in earlier research, namely that the expected profitability of including a star in a film is positive while the mean profitability is negative. We find no such result

for sequels, where expected profitability and mean profitability are both positive. Our results indicate that there are strongly increasing marginal returns to both budget and opening screens over the domain of these variables, a result that is consistent with the industry's move toward films with bigger budgets and wider releases.

Acknowledgments The author would like to thank three anonymous referee for comments that have help to improve the statistical analysis and the exposition in the paper.

References

Aitchison J, Aitken CGG (1976) Multivariate binary discrimination by the kernel method. Biometrika 63(3):413–420

Albert S (1998) Movie stars and the distribution of financially successful films in the motion picture industry. J Cult Econ 22:249–270

Albert S (1999) Reply: Movie stars and the distribution of financially successful films in the motion picture industry. J Cult Econ 23:325–329

De Vany AS, Walls WD (1996) Bose-Einstein dynamics and adaptive contracting in the motion picture industry. Econ J 439(106):1493–1514

De Vany AS, Walls WD (1997) The market for motion pictures: rank, revenue and survival. Econ Inq 4(35):783–797

De Vany AS, Walls WD (1999) Uncertainty in the movie industry: Does star power reduce the terror of the box office? J Cult Econ 23(4):285–318

De Vany AS, Walls WD (2002) Does Hollywood make too many R-rated movies?: risk, stochastic dominance, and the illusion of expectation. J Bus 75(3):425–451

De Vany AS, Walls WD (2004) Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. J Econ Dyn Control 28(6):1035–1057

Fan J, Gijbels I (1992) Variable bandwidth and local linear regression smoothers. Ann Stat 20(4):2008–2036

Fan J, Gijbels I (1996) Local polynomial modeling and its applications. Chapman and Hall, London

Hall P, Racine J, Li Q (2004) Cross-validation and the estimation of conditional probability densities. J Am Stat Assoc 99(486):1015–1026

Hayfield T, Racine JS (2006) np: Nonparametric kernel smoothing methods for mixed datatypes. R package version 0.12-1

Henderson DJ, Kumbhakar SC (2006) Public and private capital productivity puzzle: a nonparametric approach. South Econ J 73:219–232

Hsiao C, Li Q, Racine JS (2007) A consistent model specification test with mixed categorical and continuous data. J Econom 140(2):802–826

Hurvich CM, Simonoff JS, Tsai CL (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Soc B 60:271–293

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput Graph Stat 5(3):299–314

Li Q, Ouyang D (2005) Uniform convergence rate of kernel estimation with mixed categorical and continuous data. Econ Lett 86:291–296

Li Q, Racine J (2006) Nonparametric econometrics: theory and practice. Princeton University Press, Princeton

Li Q, Racine JS (2004) Cross-validated local linear nonparametric regression. Stat Sin 14(2):485–512

Li Q, Maasoumi E, Racine JS (2006) A nonparametric test for equality of distributions with mixed categorical and continuous data. Mimeo, McMaster University

Litman BR (1983) Predicting the success of theatrical movies: an empirical study. J Pop Cult 16:159–175

Litman BR, Ahn H (1998) Predicting financial success of motion pictures: The early '90s experience. In: The motion picture mega-industry, Allyn and Bacon, Needham Heights, Massachusetts, chap. 10, pp 172–197

Litman BR, Kohl LS (1989) Predicting financial success of motion pictures: The '80s experience. J Media Econ 2:35–50

Nadaraya EA (1965) On nonparametric estimates of density functions and regression curves. Theory Appl Probab 10:186–190

Nelson RA, Donihue MR, Waldman DM, Wheaton C (2001) What's an Oscar worth? Econ Inq 39(1):1–16
Pagan A, Ullah A (1999) Nonparametric econometrics. Cambridge University Press, Cambridge
Prag J, Cassavant J (1994) An empirical study of determinants of revenues and marketing expenditures in the motion picture industry. J Cult Econ 18(3):217–235
Racine J (1997) Consistent significance testing for nonparametric regression. J Bus Econ Stat 15(3):369–379
Racine J, Li Q (2004) Nonparametric estimation of regression functions with both categorical and continuous data. J Econ 119:99–130
Ravid SA (1999) Information, blockbusters and stars: a study of the film industry. J Bus 72:463–486
Sedgwick J, Pokorny M (1999) Comment: Movie stars and the distribution of financially successful films in the motion picture industry. J Cult Econ 23:319–323
Smith SP, Smith VK (1986) Successful movies: a preliminary empirical analysis. Appl Econ 18(5):501–507
Vogel HL (1998) Entertainment industry economics: a guide for financial analysis, 4th edn. Cambridge University Press, New York
Wallace WT, Seigerman A, Holbrook MB (1993) The role of actors and actresses in the success of films: How much is a movie star worth? J Cult Econ 17(1):1–24
Wang MC, Van Ryzin J (1981) A class of smooth estimators for discrete estimation. Biometrika 68:301–309
Watson GS (1964) Smooth regression analysis. Sankhya 26(15):175–184