Journal of Media Business Studies

ISSN: 1652-2354 (Print) 2376-2977 (Online) Journal homepage: https://www.tandfonline.com/loi/romb20

Sales estimations in the book industry – comparing management predictions with market response models in the children's book market

Cord Otten, Michel Clement & Dominik Stehr

To cite this article: Cord Otten, Michel Clement & Dominik Stehr (2019): Sales estimations in the book industry – comparing management predictions with market response models in the children's book market, Journal of Media Business Studies

To link to this article: https://doi.org/10.1080/16522354.2019.1623436

Published online: 12 Jun 2019.

JOURNAL OF MEDIA BUSINESS STUDIES
https://doi.org/10.1080/16522354.2019.1623436

Routledge
Taylor & Francis Group

Sales estimations in the book industry – comparing management predictions with market response models in the children's book market

Cord Otten[a], Michel Clement[b] and Dominik Stehr[a]

[a]Institute for Marketing, University of Hamburg, Hamburg, Germany; [b]Marketing and Media, Institute for Marketing, University of Hamburg, Hamburg, Germany

ABSTRACT

Estimating the demand of books is an important business-planning task for publishing companies. This study identifies, quantifies, and generalizes the sales drivers of children's and young adult literature books and investigates whether quantitative prediction models improve prerelease sales predictions. We compare the model-based predictions with prerelease management sales predictions of a subsample provided by a large German publisher. Based on a sample of 542 titles that were published in Germany, we examine (a) the relevant drivers (elasticities) for economic success and (b) analyze the prediction performance of the models using two specifications of multiplicative market response models for this highly relevant market. The quantitative model approach outperforms management heuristics, reducing the prediction error by up to 45%. We further highlight the practical feasibility of simple market response models.

ARTICLE HISTORY
Received 20 February 2018
Accepted 15 May 2019

KEYWORDS
Book industry; sales response models; sales forecasts

Introduction

The optimal allocation of scarce resources to produce and market an innovation such as a new media product requires a deep understanding of marketing elasticities and a thorough sales forecast (Dorfman & Steiner, 1954). Although it is difficult to predict sales for new products (Hirschman & Holbrook, 1982), business planning requires media managers to provide sales predictions for upcoming new releases (Eliashberg, Hui, & Zhang, 2007; Hofmann-Stölting, Clement, Wu, & Albers, 2017). In the publishing industry, sales predictions for physical books are especially relevant in determining the number of copies to be printed. Overly optimistic forecasts lead to excessive production volumes, which are associated with excess printing, inventory and disposal costs.[1] However, underestimating the demand means forgoing sales in a competitive market. Consequently, a thorough understanding of the impact of the various success drivers of book sales is an important task in business planning in terms of (1) identifying the elasticities of the various sales drivers as well as (2) forecasting sales for a new media product.

CONTACT Michel Clement michel.clement@uni-hamburg.de

In this paper, we address both aspects of a very interesting market that has received surprisingly little attention although its size is substantial. We focus on the children's book market for the following reasons:

First, the market size is large: Children's and young adult literature generates 16.5% of the total market revenue, which is only surpassed by fiction books (31.5%). As such, the children's' book segment is of significant economic importance in the nine billion Euro German book market (Börsenverein des deutschen Buchhandels, 2017). The book market is also highly competitive: in 2016, customers were offered 85,486 new publications of which 85% were first editions (Börsenverein des deutschen Buchhandels, 2017). As customers do not perceive these books as perfect substitutes for one another, they are confronted with a large choice set, which makes it difficult to select the right books (Barrot, Becker, Clement, & Papies, 2015).

Second, the children's book market is theoretically interesting as there is very high uncertainty with respect to book quality due the separation of customers (parents) and consumers (children) – children and teenagers are financially dependent on adults. This setting may affect the respective elasticities in empirical settings and may thus result in substantially different elasticities for this market compared to the findings of the previous studies that have typically relied on fiction books (e.g. Barrot et al., 2015; Beck, 2007; Hofmann-Stölting et al., 2017). In addition, books are experience goods, and consumers can only assess the quality of books after consumption (Shapiro & Varian, 1999). Since customers do not perceive books as perfect substitutes for one another, they are confronted with a large choice set, which makes it difficult to select the right books (Barrot et al., 2015). This effect is of particular interest due to the prevalence of physical copies as well as gift giving as a purchasing motive in this segment (Leitão, Amaro, Henriques, & Fonseca, 2018).

Third, the diffusion pattern of children's books potentially differs from the diffusion processes of other book segments studied by Beck (2007), Hofmann-Stölting et al. (2017) and Yucesoy, Wang, Huang, and Barabási (2018). A new children's book can be read by multiple cohorts of children over several years. Therefore, the diffusion may take longer than for fiction books. This effect leads to a longer book life cycle, which may result in different elasticities.

Fourth, very little research has studied prediction models for new media products. The existing market response models commonly focus on optimal profits depending on marginal spending changes on the firm level (e.g. Kanuri, Mantrala, & Thorson, 2017; Mantrala, Naik, Sridhar, & Thorson, 2007; Sridhar, Mantrala, Naik, & Thorson, 2011) – a notable exception is the diffusion study by Hofmann-Stölting et al. (2017). However, they did not focus on children's books, and they compared diffusion modeling with simple OLS. Interestingly, the authors find that OLS outperforms management predictions for fiction books. However, their sample does not include books that sell small numbers of copies and is therefore limited. Our sample includes books with low sales numbers as well as books with a large sales volume. Thus, managers may need to consider count data models (e.g. Poisson models or negative binomial models) that capture the sales structure of children's books better than OLS models. However, this approach has not been tested in the previous research. Other prediction approaches

heavily rely on machine learning techniques (Castillo et al., 2017; Yucesoy et al., 2018). Their generalizability remains limited as the impact of individual predictors (features) remains opaque, and comparisons across other models are difficult.

Fifth, most prediction models rely on a large number of variables that are actually not always known at an early stage of the new book's life cycle (see, for example, Hofmann-Stölting et al., 2017). For example, when the number of preorders for a certain book is known, then the short-term prediction of the sales is easier, compared to predictions that are conducted months before the release. In addition, Twitter is a helpful data source for prediction models after the new product has been released (Hennig-Thurau, Wiertz, & Feldhaus, 2015). Generally, there is limited knowledge about the prediction performance of new models with limited input data (Eliashberg et al., 2007). We specifically test the data restrictions with respect to prediction performance and find that the models predict rather well – even with limited input variables.

Sixth, most prediction studies compare their forecasts with original sales. However, the usefulness of a model depends on whether it adds any value compared to the organization's status quo. Therefore, we follow Hofmann-Stölting et al. (2017) and compare our predictions with management forecasts of a subset of our data. This approach allows us to show the increased performance of the models compared to the presales estimations of the management. Thus, our test approach is a benchmark against management predictions.

To conclude, the number of studies addressing the book market in general is limited, and we are not aware of any empirical study related to the children's books segment. Thus, we provide new insights with respect to two major research areas (elasticities and forecasting techniques) in the field of media business studies.

In addition, we make a methodological contribution. Currently, most media management predictions rely on analogies instead of formal prediction models (Hofmann-Stölting et al., 2017). Sales response models have been shown to outperform management predictions in the movie, publishing and music industries (Eliashberg, Swami, Weinberg, & Wierenga, 2001; Hofmann-Stölting et al., 2017). Interestingly, comparisons of quantitative prediction models and simple management heuristics for media products such as movies have led to mixed results in the contemporary academic literature. While, for example, Wübben and Wangenheim (2008) highlight the simple application of heuristics and show that it delivers more accurate predictions, Ainslie, Drèze, and Zufryden (2005) achieve better predictions for movies with the support of quantitative models. The prior research in the book industry has focused on the drivers of commercial success (Schmidt-Stölting, Blömeke, & Clement, 2011) and on timing and pricing decisions (Burmester, Eggers, Clement, & Prostka, 2016). However, sales response models have not been tested with respect to their prediction performance; consequently, the managerial relevance remains limited. We strive to explore relatively simple market response models that are applicable and feasible in practice (Albers, 2012).

We rely on a unique dataset that comprises the sales data of 542 children's and young adult literature books from the German speaking market. We analyze a random sample of 459 titles from the top 25,000 books tracked by Media Control and added 83 titles from one large German publishing company to benchmark our predictions as we have access to their internal prerelease sales predictions. As a first step, we examine the

relevant drivers for economic success using two specifications of multiplicative market response models. We specify a log-log model estimated by OLS as is frequently applied in the research and practice. Additionally, we specify a negative binomial (NB) model estimated by maximum likelihood to explicitly account for the count nature of sales. Subsequently, we split the data into a training sample and a validation sample. We use the former to calibrate prediction models, which we then test in comparison to the management predictions on the validation sample. Comparisons of the quantitative models to the management predictions across two success metrics and across four book success categories show that the quantitative model approach is able to outperform management heuristics. That is, it reduced the prediction error by up to 45%.

This study advances the current academic literature on predictions in the context of books, with a focus on children's and young adult literature, by validating and generalizing sales drivers and illustrating the practical application of a quantitative prediction model. We find the sales drivers for fiction books from the extant literature to be applicable, and therefore generalizable, for the large market segment of children's and young adult literature. Our comparisons of the findings provide implications for empirical generalizations. From a methodological perspective, we demonstrate the superior performance of simple quantitative models in comparison to common practice management heuristics, and thus we advance the initial findings of Hofmann-Stölting et al. (2017) because the NB-model is a well-suited alternative approach for predicting sales. Publishing companies can easily adapt our proposed models for their prediction tasks.

Next, we will discuss the most important challenges for sales response and prediction models in the children's and young adult literature market and elaborate on our respective approaches. Subsequently, we introduce the individual variables, formally state our models and report the empirical results.

Modeling sales in the book market

The previous literature with respect to the diffusion of new books is sparse – the notable exceptions are Beck (2007) and Hofmann-Stölting et al. (2017). However, the diffusion of books differs somewhat from other media products as books sell over longer periods. This issue is of key interest for the segment of children's books.

The total sales quantity per title is highly positively skewed. Whereas the mean total sales of the sample are 19,879 copies sold, the median is only 5,265 copies. Similarly, only 4% (N = 21) of the titles in the sample sold a total of more than 100,000 copies. Consequently, the bulk of the titles achieve only very limited commercial success. A high variance of sales success among books results in valuation uncertainty for consumers as well as long-tail effects resulting from niche market products.

In Figure 1, we display the sales pattern for the 542 books of our sample. Due to the large dispersion of success among titles, we scaled the weekly sales quantity on the vertical axis by the standard deviation per title. Thus, the titles are visually comparable. The sales diffusion pattern is characteristic of entertainment media. Weekly sales reach the maximum within the first weeks after release (Burmester, Becker, van Heerde, & Clement, 2015). Whereas sales decrease even more sharply for many other entertainment products, the decline for this market segment of children's and young adult

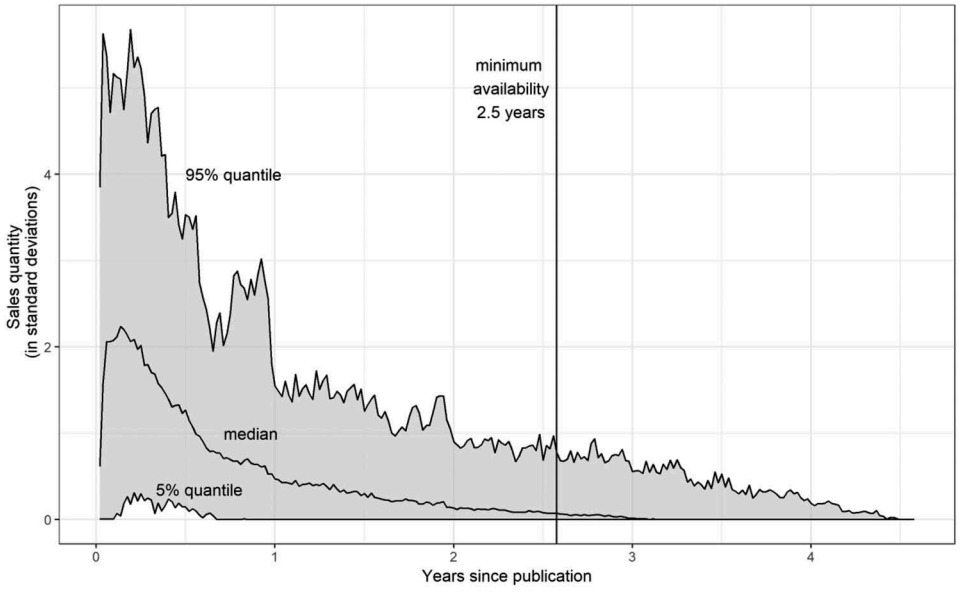

Figure 1. Large changes in sales frequently occur during the full first year after publication.
Note: Vertical axis shows weekly sales quantity per title divided by the standard deviation per title.

literature is less abrupt. Large changes in sales quantity are common even after multiple months. Most marketing expenditures occur around publication. Consequently, managers are interested in timely and meaningful sales predictions. In practical terms, that means prior to publication.

We address these challenges in modeling total sales by (1) systemizing the industry's requirements with respect to market response models, (2) selecting the relevant variables that have been identified to influence sales in the book industry, and (3) choosing the respective models to (a) analyze and (b) forecast sales.

Industry requirements

Information availability of input variables for prediction purposes: Hedonic media products such as books have a relatively short product life cycle with the majority of sales within the first few weeks (Beck, 2007). Thus, many marketing decisions are due prior to publication, and corrections over time are often not feasible. This is a major reason why managers typically request total sales estimations prior to launch. Weekly sales calculations that would require diffusion modeling are the exception rather than the norm. Hofmann-Stölting et al. (2017) further show that diffusion models only provide marginal improvements over total sales predictions in media industries. Consequently, we focus on the total quantity of sold books to customers as the dependent variable in our market response and prediction models. Furthermore, prior to publication, some market-related information is not yet available. Therefore, the proposed prediction model is nested within the driver analysis model, excluding those variables that are available only after publication.

Outcome variable – total sales: The outcome variable is the total number of sold hardcover books for a period of at least 2.5 years per title. Total sales per title range from a minimum of 39 to 898,229 sold books (mean = 19,879; median = 5,265, SD = 64,012). The hardcover format is the most interesting for managers because it is typically the first format to be published in a sequential distribution strategy. We begin by estimating a log-log model using OLS to retrieve elasticities. This approach is widespread in practice and allows us to compare our results with the previous research. That is, due to the large counts per title, we technically treat the outcome variable as continuous, which allows us to apply relatively simple and widely adopted OLS estimation techniques. We transform the dependent as well as the continuous independent variables to their respective natural logarithms, which results in distributions that are closer to normal. Additionally, the transformation also implies a multiplicative as opposed to an additive model. Multiplicative models are widely used in the entertainment industry, where the information flow is often characterized by network effects.

Count data and ease of use: To explicitly acknowledge the count nature of the outcome variable, we propose a negative binomial (NB) model as an alternative. Due to the large variation of the outcome variable, we find evidence of overdispersion and opt for a model that is based on a negative binomial distribution instead of a Poisson distribution. That is, we allow for additional heterogeneity after conditioning on our control variables (Cameron & Trivedi, 1986). We estimate this specification with a maximum likelihood method.

Evaluation of prediction: To compare the performance of competing predictions, we calculate how far the predictions are off in the number of copies and by percent (see AD and MAPE in chapter 4). Moreover, the interviewed managers were interested in predictions of success groups ranging from bestsellers to market failures. Consequently, we divide the sample into respective success categories and predict them. This exercise provides managers with actionable information.

Prediction variables

First, we identify the potential drivers for sales success (1) to extract elasticities for an overall understanding of the market dynamics and (2) as input variables for subsequent prediction models. Following the notion of marketing productivity frameworks (Lehmann, 2004), this study focuses on the impact of the firm activities of media companies and respective consumer reactions to the product-market impact in the form of sales. We categorize the constructs from the firm activities and consumer reactions to be product-, price-, distribution-, or award-related. The inclusion of variables from the empirical research in the book industry by Caliendo, Clement, and Shehu (2015), Clement, Proppe, and Rott (2007), Clerides (2002), Hofmann-Stölting et al. (2017), and Schmidt-Stölting et al. (2011) facilitates the comparison of our results with the published findings. Moreover, we consider suggestions from the managers whom we interviewed to identify potential blind spots. Table 1 lists an overview of the variables that are used in this study.

Product: Product-related attributes include quality, author power, genres, book sequels, schoolbooks, target age and quantity. First, the evaluation of quality for experience goods is inherently difficult. We use reviews such as the Amazon star ratings

Table 1. Measures and descriptive statistics (N = 542).

Variable	Description	Source	Mean/ rate (SD)	Minimum/ maximum
Quantity	Total sales quantityin DACH[a] region	Media Control	19,879 (64,012)	39/898,229
Product				
Amazon is Rated	1 = has reviews, 0 = no reviews	amazon.de	85%	0/1
Amazon Count	Count of reviews on amazon.de	amazon.de	19.58 (64.28)	0/1,096
Amazon Star	Amazon star rating: 5 = max, 1 = min	amazon.de	4.47 (0.58)	1/5
Author Power	Mean cumulative sales of first 52 weeks of last three books by author	Media Control	11,754.37 (45,462.10)	80/794,310
Genre	1 = falls into genre,	amazon.de, thalia.de,	10%	0/1
Children's Characters	0 = does not fall into genre,	buchhandel.de	21%	0/1
Fantasy and Sci-Fi	genres are not mutually exclusive		14%	0/1
Crime and Thriller			95%	0/1
Novel and Narrative			27%	0/1
Adventure			12%	0/1
Animal Stories			2%	0/1
Sport			9%	0/1
Love and Friendship			26%	0/1
Miscellaneous				
Is Series	Is part of series: 1 = yes, 0 = no	Media Control, amazon.de buchhandel.de	69%	0/1
Series Power	Mean of sales (cum. 52 weeks after resp. publication) by latest 2 predecessors	Media Control	9,810.70 (40,621.10)	0/739,193
Is Schoolbook	Is school book: 1 = yes, 0 = no	publisher websites[b]	5%	0/1
Teenager	1= young adult literature, 0 = children's book	Media Control	30%	0/1
Pages	Number of pages	amazon.de	240.91 (133.13)	40/912
Price				
Price	Mean of prices in DACH[a] region	amazon.de	12.02 (4.20)	3.01/46.51
Distribution				
Month of	1 = published in January,	Media Control	21%	0/1
publication	0 = otherwise		10%	0/1
January			13%	0/1
February			4%	0/1
March			4%	0/1
April			8%	0/1
May			8%	0/1
June			10%	0/1
July			10%	0/1
August			7%	0/1
September			4%	0/1
October			1%	0/1
November				
December				
Publisher Power	Hardcover market share per year in children and young adult literature market	Media Control	6.38% (2.88)	1.1%/12.1%
Audiobook	1 = audiobook version exists, 0 = otherwise	Media Control	32%	0/1
Awards				
Prize Nominated	1 = nominated, 0 = not nominated	djlp.jugend literatur.org	1%	0/1
Miscellaneous				
Random[c]	1 = randomly selected, 0 = added manually		85%	0/1

Notes: [a] DACH = German speaking market of Germany, Austria and Switzerland

[b] arena-verlag.de, luebbe.de, beltz.de, carlsen.de, randomhouse.de, dressler-verlag.de, schneiderbuch.de, fischerverlag.de, kosmos.de, hanser.de, loewe-verlag.de, thienemann-esslinger.de, ravensburger.de

[c] Dummy variable to control for potential selection bias. Some titles where manually added from the portfolio of the cooperating publisher to increase the sample size for the success prediction analysis.

as an important approximation of quality (Babić Rosario, Sotgiu, De Valck, & Bijmolt, 2015; Chang & Ki, 2005; Floyd, Freling, Alhoqail, Cho, & Freling, 2014; You, Vadakkepatt, & Joshi, 2015). Second, authors are human brands and serve as stars in their market. For customers, a recognized author is a signal of quality that is easy to identify. The success of previous books serves as an approximation of an author's power. This approach follows from the intuition that prior success generates satisfaction with consumers, which in turn increases the purchasing probability of later publications (e.g. Schmidt-Stölting et al., 2011). We operationalize author power as the mean total sales of the last three books by the respective author. Per prior title, we take the first year of sales into account. The distribution of the resulting author power metric is plotted in Figure 2. Please note that the horizontal scale is log transformed in the figure since the measure is highly skewed and this depiction is congruent with the model specification. All authors have at least one prior publication. Mean prior sales range from 80 to 790,310 copies. The most frequent mean sales are between 1,000 and 10,000 copies with a median of 3,539. Third, genres provide consumers with an orientation in a highly differentiated market such as the publishing industry (Kamphuis, 1991; Leemans & Stokmans, 1991). Genre membership is not mutually exclusive, and genre classification is often inconsistent across platforms. However, Schmidt-Stölting et al. (2011) find, for example, that biographies are positively correlated with the hardcover sales of fictional literature. Fourth, being part of a series signals a certain type of book. Moreover, it may lead to a lock-in effect for the reader. Subsequent titles may profit from the success of an established author, title, or topic (Schmidt-Stölting et al., 2011). Fifth, school curricula lead to the use of books in classes,

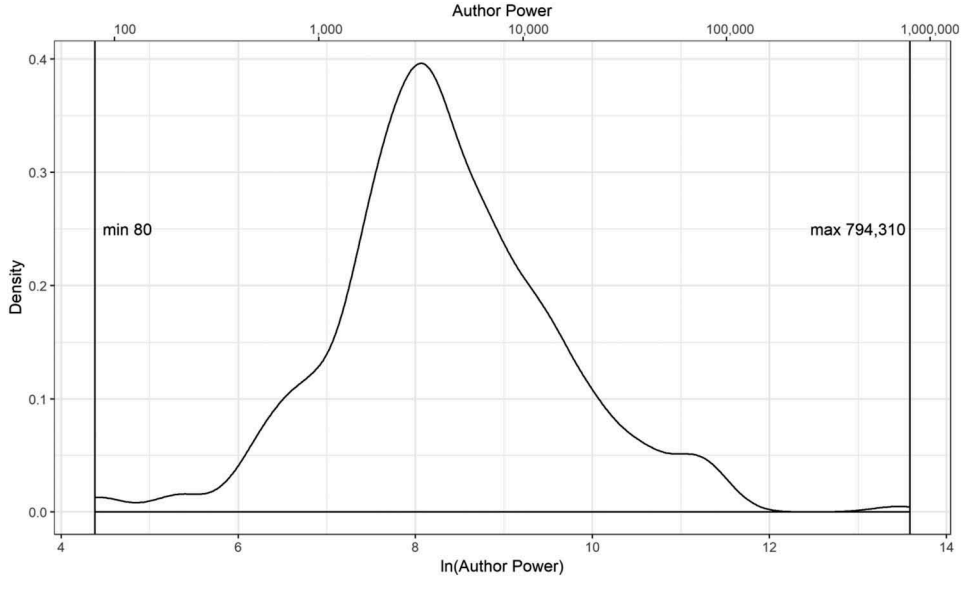

Figure 2. Most authors sold between 1,000 and 10,000 copies per prior publication.

Note: Author Power is the mean total sales of the last three books by the respective author. Per prior title, the first year of sales data are taken into account. The horizontal scale is log-transformed, coherent with the specification in the model.

which may generate an additional demand effect. Sixth, in terms of age, younger and older target audiences may have diverging demand and supply dynamics. In particular, teenagers are likely to be more autonomous in terms of book purchases, whereas children depend heavily on adults for book acquisitions. Seventh, the number of pages of a book influences the reading time and, therefore, serves as a quantity measure for books.

Price: The price of a product is a critical factor. Customers can easily observe it; it serves as a quality indicator; however, it also limits demand due to limited customer budgets (Bijmolt, van Heerde, & Pieters, 2005). Some countries, such as Germany, regulate book prices (fixed book price agreement) so that publishers set constant consumer prices (A) for all retailers and (B) over time. The previous studies report price elasticities that range from 0 (Clerides, 2002) to −9.80 (Brynjolfsson, Dick, & Smith, 2009) in unregulated markets and from −1.34 to −1.44 in regulated markets (Hjorth-Andersen, 2000). Recently, Barrot et al. (2015) reported a price elasticity of −3.7 for fictional hardcover books in the price-regulated German book market.

Distribution: Larger publishing houses have a stronger influence on the success of a title since they profit from prior experience, higher budgets and stronger negotiation positions towards retailers (Farris, Olver, & De Kluyver, 1989). Moreover, they ensure that there is a minimum of marketing support for each title (Spencer, 2017). Schmidt-Stölting et al. (2011) report no significant effect for hardcover books; however, they find a significant effect for paperbacks. The timing of publication within a year is another distribution decision. High seasons such as Christmas are likely to affect the sales of a book (Schmidt-Stölting et al., 2011). In addition to timing, successive formats such as audio and paperback books also influence the sales trend of hardcover books. The publication of the paperback version usually marks the end of the hardcover product's life cycle, the latter being the focus of this study. The timing of audiobook releases is less rigid. We control for the existence of an audiobook version with a binary variable.

Awards: Literature awards and the associated increase in visibility may affect the success of books at later points in their life-cycle. The impact of awards is a much discussed topic in the media industry (Caliendo et al., 2015; Nelson, 2001). The awarding of a prize serves as a signal for the quality of a product and increases its visibility. In this study, we consider a category-specific award, namely, the *German Young Adult Literature Award* (*Deutscher Jugendliteraturpreis)*.

Modeling challenges

Success driver analysis models

For the initial step of identifying success drivers, we regress the potential success drivers on the total hardcover quantity sold. We include a control variable to alleviate concerns for a potential sampling bias, which we discuss in more detail in section 3.1. First, we rely on a log-log formulation, which results in a multiplicative model where the resulting estimated parameters can be interpreted as elasticities (or in the case of dummy variables, as multipliers). Formally, we model the expected total sales as follows:

$$\begin{aligned}\ln quantity = {} & \beta_0 + \beta_1\, isRated \;+ \beta_2 ReviewVolume \\ & + \beta_3 ReviewValence + \beta_4 \ln authorPower + \gamma_G GENRE \\ & + \beta_5 isSeries + \beta_6 \ln seriesPower + \beta_7 isSchoolbook \\ & + \beta_8 teenager + \beta_9 \ln pages + \beta_{10} \ln price \\ & + \gamma_M MONTH + \beta_{11} \ln publisherPower \\ & + \beta_{12} audiobook + \beta_{13} prizeNominated + \beta_{14} random + \varepsilon\end{aligned} \tag{1}$$

Second, we test an alternative formulation of the model to account for the data type of the available sales data, specifically count data. Specifically, we estimate a negative binomial (NB) model. NB type models are well grounded in statistics and have been widely applied in media, communication, and marketing studies (e.g. Ateca-Amestoy, 2008; Chen, Fay, & Wang, 2011; Frisbie, 1980; Yan & Napoli, 2006). NB models have rarely been applied to sales data in the media industry. These models are designed to address the common issue of overdispersion in the application of Poisson models to count data (Cameron & Trivedi, 1986). In the Poisson regression model, the number of events y (number of copies sold) for book i is Poisson distributed with conditional mean λ depending on the characteristics x of the book:

$$\lambda_i = E(y_i|x_i) = e^{x_i\beta} \tag{2}$$

The probability of y given x is:

$$Pr(y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!} \tag{3}$$

For Poisson distributed data, the variance is equal to its mean. However, we find overdispersion in the data, likely due to unobserved individual heterogeneity rooted in the valuation uncertainty of hedonic media products (Hirschman & Holbrook, 1982). In this case, Poisson regressions will be inefficient. NB regression models address unobserved heterogeneity by adding an error term (ε) to the conditional mean of the Poisson distribution:

$$NB_i = E(y_i|x_i) = e^{(x_i\beta+\varepsilon_i)} \tag{4}$$

It is commonly assumed that $\exp(\varepsilon_i)$ is gamma distributed with mean 1 and variance $1/\theta$. Thus, the conditional mean of y_i is still λ_i and the conditional variance of y is:

$$Var(y_i) = E(y_i) + \frac{E(y_i)^2}{\theta} \tag{5}$$

Similar to the first model, we enter the explanatory variables in their natural logarithms. Consequently, this model is also multiplicative, and the estimated parameter values can be interpreted as elasticities (e.g. Ayyagari, Deb, Fletcher, Gallo, & Sindelar, 2013). We model the expected total sales with the NB specification with the following explanatory variables:

$$\begin{aligned} quantity \sim NB(&\beta_0 + \beta_1 isRated + \beta_2 ReviewVolume \\ &+ \beta_3 ReviewValence + \beta_4 \ln authorPower + \gamma_G GENRE \\ &+ \beta_5 isSeries + \beta_6 \ln seriesPower + \beta_7 isSchoolbook \\ &+ \beta_8 teenager + \beta_9 \ln pages + \beta_{10} \ln price \\ &+ \gamma_M MONTH + \beta_{11} \ln publisherPower \\ &+ \beta_{12} audiobook + \beta_{13} prizeNominated + \beta_{14} random + \varepsilon) \end{aligned} \tag{6}$$

Prediction models

We rely on our log-log specification Equation (1) as a starting point for prediction purposes. First, we drop the variables that were unknown prior to publication. These are whether the title is (1) a schoolbook; (2) review volume (and its valence); (3) an award, such as in the nomination for the *German Young Adult Literature Award*, and whether (4) an associated audiobook exists. Second, we exclude the control variable that was originally introduced to account for a potential selection bias as the objective of sales predictions does not require a clean identification of the individual drivers. Third, we systematically drop individual variables and compare the AIC to increase the parsimony of the model. By means of this process, for the prediction models, we further exclude the prior sales of sequels, whether the book is marketed to teenagers, the number of pages, and price. Moreover, we additionally exclude the month of the year indicators from the log-log specification. To conclude, we formulate the log-log prediction model as follows:

$$\ln quantity = \beta_0 + \beta_1 \ln authorPower + \gamma_G GENRE + \beta_2 isSeries + \beta_7 \ln publisherPower + \varepsilon \tag{7}$$

We formulate the NB prediction model as follows:

$$quantity \sim NB(\beta_0 + \beta_1 \ln authorPower + \gamma_G GENRE + \beta_2 isSeries + \gamma_M MONTH + \beta_3 \ln publisherPower + \varepsilon) \tag{8}$$

Empirical analysis of success drivers

Sample

We base our empirical analysis on the sales data of 542 children's and young adult literature books in Germany, Austria and Switzerland, tracked by the market research institute *Media Control*. The data reflect consumer purchases of printed books, both across online and offline channels. In 2015, the channel market shares were 48.2% for book retailers, 20.9% for direct sales by publishers, 17.4% for online retailers, and 13.5% for other channels (Börsenverein des deutschen Buchhandels, 2016).

The sample criteria for our study include the book format, the type of title in terms of category, and the publisher. First, the sample focuses on hardcover and softcover editions. These two formats generate 74.8% of all book revenues in Germany and are typically released before respective paperback editions (Börsenverein des deutschen Buchhandels, 2017). Second, we focus on the two subcategories of children's books up to the age of 11 and young adult literature books starting from the age of 12. These subcategories cover more than 50% of the children's and young adult literature book

market. That is, we deliberately exclude subcategories such as picture and learning books, books in foreign languages, book boxes, miscellany books and manga or comic books. Third, we consider the top 15 publishers in terms of market share of the quantity sold in our 4.5-year study period (CW1 2010 throughout CW 30, 2014). These publishers make up 85% of the total volume in the respective period. Self-publishing is excluded from this study based on the expected difference of prevailing market dynamics for self-publishing, particularly in the market of children's and young adult literature (Waldfogel & Reimers, 2015).

In terms of the sampling strategy, we drew a random sample of 500 titles from the top 25,000 books tracked by Media Control that fit the postulated criteria. A subsequent in-depth check of the sampling criteria disqualified 41 titles. We cooperated with one of the major publishers of children's and young adult literature in Germany, which made its internal management prerelease sales predictions available to us. Correspondingly, we added 83 titles from this publisher to expand our potential validation sample. These titles met the sampling criteria. To control for potential systematic differences between the original random sample and these additional titles from the specific publisher, we code a dummy variable to identify the 459 books that are part of the original random sample.

To set up the quantitative prediction model and subsequent comparisons to management predictions, we split the sample into a training sample and a validation sample. The validation sample is a random draw (N = 84) from the 96 titles by the cooperating publisher. For these titles, the publisher made internal management predictions available to us. The managers reported in personal interviews that predictions are done on a case-by-case basis. They typically have one or two similar titles in mind that were previously published and use the respective success recollected from memory as a proxy. We used the remaining 458 titles as the training sample. The sample split ensures that information from the validation sample does not enter the calibration phase of the prediction model (out of sample prediction). Table 2 provides an overview of the sampling strategy and descriptive statistics of the sales for the two subsamples that are used in the prediction analysis.

The sample comprises total sales from publication in 2010 or 2011 up to calendar week 30 in 2014. Thus, a period of at least 2.5 years is available per title. Due to the typical life cycle of books as plotted for our sample in Figure 1, we are confident that these data include the relevant sales per title (Beck, 2007).

Measures

In Table 1, we provide an overview of the variables and measures that are used in this study.

Product: We code the count of Amazon reviews with a mean of 19.59 (SD = 64.28). Fifteen percent of the titles did not receive any Amazon reviews. The mean Amazon star rating is 3.80 (SD = 1.69). Amazon reviews approximate a book's quality, which is inherently difficult to quantify. We operationalize author power by calculating the mean sales volume of books published by the author prior to the title under consideration. Here, we use the total sales volume of the first year of an author's three latest books, which results in mean prior sales of 11,754.37 (SD = 45,462.10) books. In this study, nine nonmutually exclusive dummy variables

Table 2. Descriptive statistics of subsamples.

	Sample/ change	Count		
	Tracked and matching criteria	25,000		
	Initial random sample	500		
	Exclusion criteria detail	*- 41*		
	Checked random sample [a]	459		
	Additional titles fromcooperating publisher	*+ 83*		
	Success driver analysis sample	**542**		
	Of which fromcooperating publisher	96		
Split for prediction model	**Estimation sample**	**458**	Mean sales	19,831
			Std. Dev.	68,217
			Min sales	39
			Max sales	898,229
	Validation sample	**84**	Mean sales	20,137
			Std. Dev.	32,950
			Min sales	166
			Max sales	156,133

Notes: [a] dummy variable as control in success driver analysis

indicate genres. Another dummy variable captures the 69% of the books that were part of a series at the time of publication. We code the success of the series as the mean of the total first year of sales of the first and second predecessors. The predecessors sold 9,810.70 books on average (SD = 40,621.10). Moreover, we code dummy variables for the titles of school curricula (5%) and for the titles that are targeted towards young adults (30%) as opposed to children up to the age of 11 (70%). The books have 241.91 pages on average (SD = 134.13), which approximates the quantity of a book.

Price: German fixed book price laws inhibit price changes over time. The average price in the German-speaking markets of Germany, Austria and Switzerland is € 12.02 (SD = 4.20) with prices that range from € 3.01 up to € 46.51.

Awards: In terms of awards, we consider the *German Young Adult Literature Award*. However, as none of the sample titles received this prize during the considered period, we resort to nominations of this award. Eight titles (1%) were respective nominees.

Distribution: We operationalize the power of a publisher by the publisher's market share (mean = 6.39, SD = 2.89) at the children's and young adult literature book market in the year of a title's publication. Additionally, we define monthly dummy variables to capture the timing effects of different publication dates throughout the year. Publications are somewhat more frequent in the first and third quarters as opposed to the second and fourth quarters. This pattern is congruent with the spring and fall book fair cycles that are prevalent in Germany. Furthermore, we coded a dummy variable for the 32% of the titles where an audiobook version exists.

Estimation of market response models

Table 3 shows the coefficients and standard errors for the log-log model as well as for the NB model. In terms of fit, the R^2 of the log-log model is 0.71 (adjusted R^2 = 0.70), which suggests a high model fit. The closest benchmark is the market response model by Schmidt-Stölting et al. (2011), who report an R^2 of 0.40 (adjusted R^2 = 0.38) based on their hardcover specification.

There are a few titles with extreme values; however, they either score high on standardized residuals or on leverage, which means that the observations with unusual properties do not seem to affect the model in extreme ways. Consequently, we did not opt to make further adjustments. This approach also keeps the process the most basic for potential implementation.

Endogeneity concerns remain plausible even when controlling for an extensive set of control variables. Reviews approximate quality; in line with existing literature, we do not disentangle the effect of quality itself and the marginal word-of-mouth effect of an additional (positive) review. Author and publisher power may be endogenous since we may not be able to fully capture market size effects with the other control variables. Genres, series, target audience, month, and audiobook are rather controls and approximations of market size where we refrain from causal interpretations. Prices may be set strategically by management in expectation of success and may therefore also be

Table 3. Full market response model – results.

		Log-log model		NB model	
		Coefficient	Std. error	Coefficient	Std. error
	Constant	3.643	0.614***	4.048	0.514***
Product	Amazon Is Rated	0.035	0.419	−0.033	0.350
	Amazon Volume [a]	0.594	0.050***	0.567	0.041***
	Amazon Valence [a]	0.085	0.264	0.094	0.221
	Author Power [a]	0.414	0.039***	0.477	0.032***
	Genre (not mutually exclusive)				
	Beloved Children's Character	−0.027	0.152	−0.014	0.127
	Fantasy and Science-Fiction	−0.151	0.108	−0.161	0.090*
	Crime and Thriller	−0.137	0.122	−0.077	0.102
	Novel and Thriller	−0.047	0.188	−0.071	0.157
	Adventure	0.000	0.089	0.031	0.074
	Animal Stories	−0.135	0.123	−0.062	0.103
	Sport	−0.340	0.270	−0.259	0.225
	Love and Friendship	0.013	0.130	−0.013	0.109
	Miscellaneous	0.199	0.090**	0.140	0.076*
	Is Series	0.211	0.116*	0.137	0.097
	Series Power [a]	0.023	0.012*	0.016	0.010
	Is Schoolbook	0.141	0.189	0.142	0.158
	Teenager	−0.411	0.115***	−0.341	0.096***
	Pages [a]	0.105	0.111	0.094	0.092
	Price[a]	−0.355	0.185*	−0.504	0.155***
Distribution	*Month of publication*				
	February	−0.115	0.143	−0.142	0.120
	March	−0.186	0.130	−0.165	0.109
	April	−0.469	0.196**	−0.387	0.164***
	May	−0.029	0.206	0.092	0.172
	June	0.102	0.151	0.093	0.126
	July	−0.296	0.152*	−0.217	0.127*
	August	−0.171	0.144	−0.142	0.120
	September	−0.212	0.146	−0.252	0.122**
	October	−0.500	0.163***	−0.428	0.136***
	November	−0.102	0.208	0.020	0.174
	December	−0.343	0.432	−0.415	0.361
	Publisher Power [a]	0.238	0.074***	0.165	0.062***
	Audiobook	0.326	0.090***	0.244	0.075***
	Prize Nominated	0.229	0.322	0.219	0.269
	Random	0.221	0.111**	0.285	0.092***
	Dispersion parameter			2.070	0.117

[a] Logarithmic values of variables. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. R^2 of log-log model = 0.714 (adj. R^2 = 0.695)

endogenous (Barrot et al., 2015). To alleviate these concerns about author power, publisher power and price, we test for the robustness of the results by means of the copula approach proposed by Park and Gupta (2012). The Shapiro test for normality rejects the null hypothesis for all three variables. None of the copula terms is statistically significant at common cut-off levels; the results of the model including copula terms do not change substantively and are given in the Table Appendix 1. Due to some minor variation for the price elasticity estimate, we remain cautious with the interpretation of the effect. Overall, we conclude that the introduction of the extensive set of explanatory variables is effective in addressing reasonable endogeneity concerns.

The results include valuable information for the management of the marketing mix instruments of children's and young adult literature books and extend the previous studies by broadening their generalizability. The direction and magnitude of the sales drivers are coherent with the study by Schmidt-Stölting et al. (2011). Author power has a particularly strong impact, and the time of year plays a role. Publisher power is more important in the children's and young adult literature market. Together with the high share of explained variance, the results indicate that the response elasticities from the general fictional literature market are generalizable to the specific segment of children's and young adult literature. This finding is not trivial due to the specific purchasing process in this market. Moreover, the NB model generates similar results in terms of statistically significant variables and the direction of their coefficients. The generalizability across market segments and the robustness of substantive elasticities between model specifications lay the foundation for subsequent prediction models.

In the following, we focus on the NB model and highlight the estimated elasticities (multipliers). Amazon valence as measured by the star average is not statistically significant, whereas the volume of reviews is significant and positive with estimates of 0.57. This is coherent with Babić Rosario et al. (2015), who report in their meta-analysis that volume has a stronger impact on sales than valence. We predominantly consider Amazon reviews as a quality measure from the reader's perspective, although we cannot rule out network effects due to enhanced discoverability. Schmidt-Stölting et al. (2011) similarly find valence to be statistically indistinguishable from zero.

Concerning product management, the author's star power has a significant and positive effect on the sales of new books with an estimated elasticity of 0.48. That is, an author who has published commercially successful books before will also have higher sales for newly published books.

Somewhat surprisingly, the impact of genres as part of a series that is used in schools, and the number of pages, display no strong association with sales. The binary variable for the young adult literature age group is statistically significant and negative, which suggests that children's books have 34% more sales than young adult literature books.

The price variable has an intuitively correct negative sign; however, in the log-log model, it is only statistically significant at the 10% level. This finding is again coherent with Schmidt-Stölting et al. (2011). Note that the price of a book is likely to be endogenous (Barrot et al., 2015). That is, managers likely set prices with some expectation of future success. In the robustness check using Gaussian copulas, the estimate in the NB specification remains similar but changes signs in the log-log specification. They continue to be statistically indistinguishable from zero in both specifications. For our objective of predicting sales, we follow Ebbes, Papies, and van Heerde (2011), who find

that if both the holdout sample and the estimation sample are similar in terms of endogenous regressors, then OLS approaches that are not corrected for endogeneity are favored over IV estimations that control for endogeneity.

With reference to distribution, the timing strategies, strength of the publisher and audiobooks are considered. Monthly dummy variables capture the publication month with January as a baseline. Books that are published in April, July and October show lower sales compared to books that are published in January. In the NB model, publications in September are also associated with lower sales. The two largest book fares in Germany, namely, the Leipzig book fare and the Frankfurt book fair, take place in spring and fall, respectively. Hence, during these months, competition may be particularly fierce. Therefore, the publication month can be seen as an approximation for seasonality with associated competitiveness in the market. The estimated elasticity for publisher strength is 0.17. However, although the estimates remain stable and the copula term is statistically not significant, the elasticities are not significant in the robustness model. The sign of the estimates is coherent with the hypothesis that larger publishers benefit from more experience and stronger negotiation positions in the market. The existence of an audio book is statistically significant and positive. The decision to record an audiobook is likely to be endogenous. That is, successful titles are arguably more likely to be recorded as an audiobook than less successful ones.

Nominations for the *German Young Adult Literature Award* are not statistically significant, which may be due to its low visibility compared to prize winners.

Note that the dummy variable for the original random sample is positive and statistically significant at the 5% level. That is, titles from the random sample sold on average 28.5% more books than manually added titles.

Estimation of prediction models

In addition to the sales driver analysis, we estimate the models without variables unknown at the time of publication using the estimation subsample. Table 4 displays the respective results of the log-log regression and the NB model based on Equations (7) and (8). Even with this reduced set of predictors, the R^2 for the log-log model remains with 0.52 (adj. R^2 = 0.50) high compared to the market response model by Schmidt-Stölting et al. (2011) and the prediction model by Hofmann-Stölting et al. (2017). The coefficients remain generally similar to the market response model.

Comparison of model and management predictions

We compare the prediction performance in two ways. First, we compare our estimations with observed sales. Second, we compare our predictions with the presales predictions of the managers and analyze whether we predicted better than the managers.

Comparison with observed sales: The metrics of the absolute difference (AD) and the mean absolute percentage error (MAPE) serve as the main indicators to evaluate the prediction performance on the holdout sample. The absolute difference is the difference between the prediction and the actual quantity sold per title summed as absolute values over all titles. From a practical perspective, this metric is useful due to its simple

Table 4. Prediction model – results.

		Log-log model		NB model	
		Coefficient	Std. error	Coefficient	Std. error
	Constant	1.893	0.390*	2.094	0.346***
Product	Author Power [a]	0.757	0.040*	0.841	0.034***
	Genre (not mutually exclusive)				
	Beloved Children's Character	−0.281	0.193*	−0.203	0.162
	Fantasy and Science-Fiction	0.422	0.132*	0.335	0.111***
	Crime and Thriller	−0.067	0.155	−0.076	0.130
	Novel and Thriller	−0.448	0.248*	−0.324	0.207
	Adventure	−0.036	0.116	−0.110	0.097
	Animal Stories	−0.461	0.151*	−0.380	0.126***
	Sport	−0.120	0.335	−0.213	0.284
	Love and Friendship	−0.134	0.171*	−0.175	0.144
	Miscellaneous	0.063	0.123	−0.069	0.103
	Is Series	0.266	0.117*	0.194	0.099**
Distribution	*Month of publication*				
	February			−0.092	0.154
	March			−0.169	0.144
	April			−0.670	0.261**
	May			−0.080	0.260
	June			−0.011	0.165
	July			−0.174	0.164
	August			0.070	0.165
	September			−0.315	0.168*
	October			−0.409	0.194**
	November			0.588	0.227**
	December			−0.344	0.450
	Publisher Power [a]	0.380	0.088	0.151	0.079*
	Dispersion parameter			1.302	0.078

[a] Logarithmic values of variables. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. R^2 of log-log model = 0.517 (adj. R^2 = 0.504)

interpretation. The second metric, specifically, the mean absolute percentage error, is the absolute difference divided by the actual quantity sold. The resulting average percentage deviance can thus be interpreted independent of the underlying data.

Following Hofmann-Stölting et al. (2017), we align our predictions with common management practice of thinking in success categories. We split the titles into four different success groups ranging from market failures with less than 10,000 sold copies to bestsellers with more than 100,000 sold copies. The publishers' management confirmed the face validity of this classification. Table 5 depicts the respective cutoff points. Additionally, the table also shows the count and quantity sold for a sample that is split due to books being part of a series or not.

We calculate BINGO and WINNER as supplementary metrics. BINGO is the percentage of correct hits per size category (Sharda & Delen, 2006). For easy comparisons, the WINNER metric shows the percentage of titles for which the respective prediction model outperforms the competing predictions in the absolute difference metric.

In Figure 3, we plot predicted versus actual sales for the holdout sample. Actual sales are on the horizontal axis, and predictions are on the vertical axis. The top left panel depicts the management predictions: almost all predictions are lower than actual sales. Interviews with the publisher substantiated that predictions may be systematically biased towards fewer sales. Consequently, we show scaled management predictions

Table 5. Categories of books in validation sample.

Category		Cut off	Count (in %)	Quantity sold
1	Worst performers	< 10,000	50 (60%)	192,967
2		10,000–49,999	25 (30%)	559,275
3		50,000–99,999	6 (7%)	492,876
4	Best performers	≥ 100,000	3 (4%)	446,357
Is Sequel			51 (61%)	1,493,787
Not sequel			33 (39%)	197,688
Total			84 (100%)	1,691,475

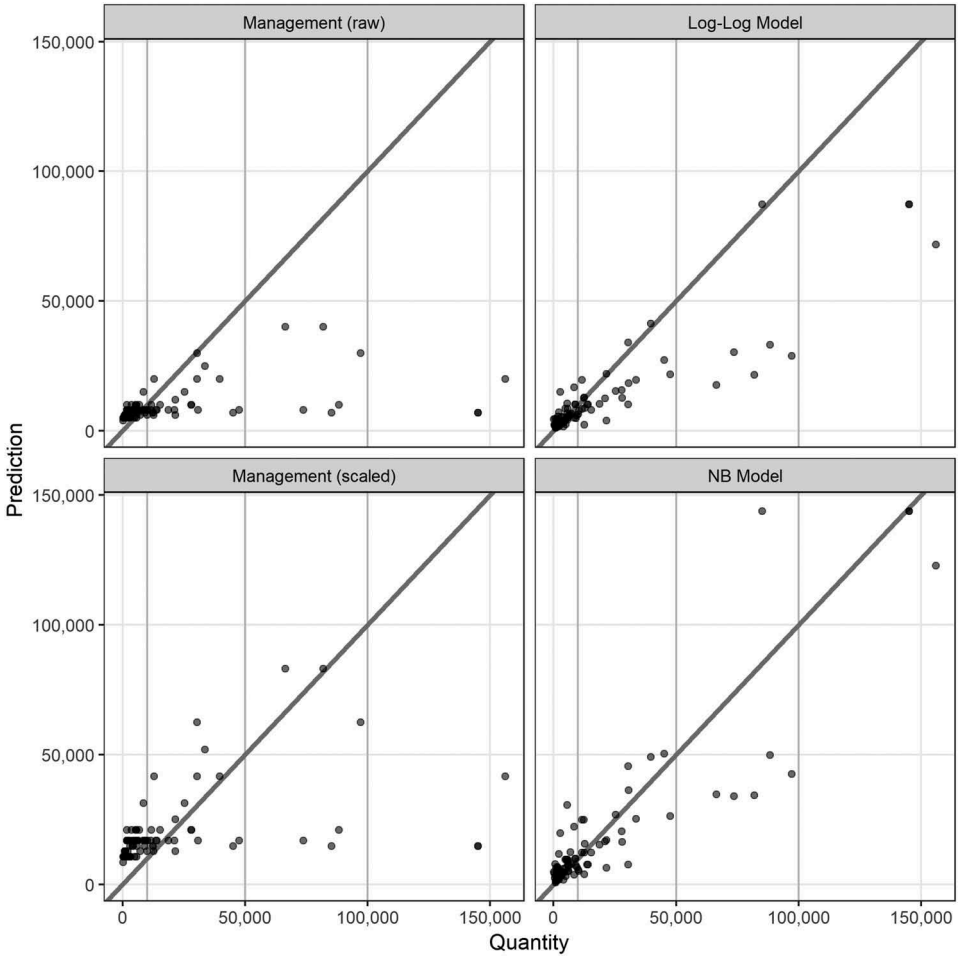

Figure 3. Prediction errors are smallest for NB model predictions.

Note: Actual sales quantity on the x-axis, predicted sales quantity on the y-axis; diagonal is the perfect prediction line

on the lower left panel. That is, management predictions are corrected by an intercept (274.501) and a slope (2.070) fitted by OLS regression (R^2 0.192). The panels on the right show the log-log model predictions on top and the negative binomial predictions below. Points on the diagonal would be perfect predictions.

Table 6 shows these different metrics for a comparison of the models. In addition to the overall performance (total), the table includes splits of the validation sample by success groups (1 = worst to 4 = best) as well as by whether the book is part of a sequel. To enhance readability, we highlight the best performance per row in bold.

Overall, the log-log and negative binomial models outperform management predictions across almost all the metrics and sample splits that are considered. The only exceptions are the BINGO metric for low-performing titles and titles that are sequels.

Comparison with management predictions: Compared to management predictions, the log-log predictions are superior by 36% and the NB predictions by 45% based on the AD metric. This translates into a 74.5 percentage point lower MAPE score of the log-log predictions. The NB model produces MAPE scores between management predictions and log-log predictions, particularly due to discrepancies for titles with few sales. However, the NB results perform significantly better for successful titles. Lastly, the WINNER metric highlights the superior performance of the statistical models over the management predictions.

Given the uncertainties of success predictions, management is particularly interested in the success category of a book as opposed to discrete predictions. Most books fall into the low performing group with fewer than 10,000 sales. Whereas management predictions are relatively low across all titles, the NB model is able to identify the bestseller titles in the holdout sample. Similar to the study of Hofmann-Stölting et al. (2017), we find that a log-log specification performs rather well for titles with few sales.

Table 6. Overview of prediction comparison.

	Group (N)		Management prediction	Log-log prediction	NB prediction
AD	Total (84)	1	1,233,755	794,024	**677,051**
	Worst (50)	2	174,821	**100,390**	170,749
	(25)	3	288,701	215,530	**199,603**
	(6)	4	357,876	278,148	**270,931**
	Best (3)		412,357	199,956	**35,767**
	Is sequel (51)		1,066,523	692,344	**563,094**
	Not sequel (33)		167,232	**101,680**	113,957
MAPE %	Total (84)	1234	175.0	**100.5**	139.0
	Worst (50)		256.3	**140.6**	207.0
	(25)		47.3	**37.7**	38.8
	(6)		71.6	56.9	**54.7**
	Best (3)		92.5	44.6	**7.7**
	Is sequel (51)		131.4	**103.7**	124.1
	Not sequel (33)		242.5	**95.6**	161.9
BINGO %	Total (84)	1	67	76	**77**
	Worst (50)	2	**98**	90	90
	(25)	3	28	**72**	68
	(6)	4	0	**17**	**0**
	Best (3)		0	0	**100**
	Is sequel (51)		**96**	78	87
	Not sequel (33)		56	**75**	74
WINNER %	Total (84)	1	23	38	**39**
	Worst (50)	2	24	**50**	26
	(25)	3	20	24	**56**
	(6)	4	33	17	**50**
	Best (3)		0	0	**100**
	Is sequel (51)		29	29	**41**
	Not sequel (33)		12	**52**	36

Note: best performance per row highlighted in **bold**

Overall, we find that the quantitative models outperform management predictions with the log-log specification leading for the large number of titles with few sales and the NB model for medium successful to bestselling titles.

Conclusion

Managers in the publishing industry for children's and young adult literature are interested in (A) drivers of sales success as well as (B) predicting sales success before publication to manage first-copy cost and related marketing activities. These tasks are not trivial due to the high variation in the data, the nature of books as experience goods, the particular separation of customers and consumers for children's and young adult literature books as well as the sheer number of new titles that are published each year.

In our study, first, we estimate a sales response model using a simple log-log formulation to model sales for a sample of 542 children's and young adult literature books. The transformation of the dependent and independent variables to their respective natural logarithms implies a multiplicative model and allows for the interpretation of the resulting coefficients as elasticities. In addition to a log-log model, we estimate an NB specification. The independent variables enter again in their natural logarithms allowing for the interpretation of estimated coefficients as elasticities. Second, we calibrate a prediction model on an estimation sample of 458 books limiting variables to information that is available prior to publication. Subsequently, we predict the sales for a holdout sample of 84 books and compare model-based predictions with management predictions. The quantitative models are robust in the qualitative results, provide easily interpretable results and outperform management predictions on the holdout sample.

Despite being well-established in econometrics, negative binomial models have rarely been applied to predict sales quantities for media products. This model family addresses the count nature of the data and the high variance of the response variable by design. Moreover, entering the predictors in their natural logarithms allows for the estimation of multiplicative models for which estimated parameters can be interpreted as elasticities. Applying negative binominal models may also help to reduce biases in carryover effects, thus improving marketing resource allocations (Köhler, Mantrala, Albers, & Kanuri, 2017). Our results show that the sales drivers for fiction books that have been identified in the previous studies extend to children's and young adult literature. Thus, our study enlarges the generalizability of this stream of research. Further, this study tests the predictive power of sales response models in this market and as such provides stronger evidence for the validity of the approach.

From a theoretical as well as a management perspective, we identify and quantify relevant sales success drivers. Specifically, we find such drivers to be author power, online review volume, and, to some degree, the power of the publisher to increase book sales. Moreover, the market response model allows for sales simulations under varying parameters. Most prominently, a simple quantitative prediction model is able to outperform management predictions based on heuristics. For this study, we cooperated with a large German publisher whose management verified the feasibility of the approach and highly appreciated the discovered results. The proposed negative binomial model is fairly straightforward to interpret and can be estimated with all standard statistical packages including widely available open source software (Venables & Ripley, 2002).

In summary, this study advances the literature on the sales success drivers for books in the category of children's and young adult literature. We quantify the impact of sales drivers and show how managers can use sales response models as a decision support system for business planning. Quantitative models can supplement or even substitute predictions based on management heuristics.

The limitations of this study relate to the characteristics of the relevant statistical concepts and the generalizability of the available data. First, our study made extensive use of OLS regressions that assume correct specification, strict endogeneity, and normality. Correct specification hinges on the available data; however, such data are influenced by the time of the prediction. We expect some degree of endogeneity for some of the variables – especially for price. One strategy to identify the exogenous impact of price is the use of instrumental variable approaches. Barrot et al. (2015) use costs as an instrument, and the price of other genres could also be considered. As a robustness check, we follow Park and Gupta (2012), who propose an instrument-free approach to address endogeneity using copulas to model the multivariate distribution of the endogenous predictor and the (normally distributed) error. The results show some minor variance for price and remain stable for other variables. Moreover, the publication of a sequel is more likely when the prequel was successful. Similarly, the production of an audiobook is more likely for titles that are successful in the hardcover format. This is coherent with the notion from De Vany and Walls (1996) that demand is highly dependent on the stochastic components of the diffusion process. Consequently, sequel decisions can be made once the actual demand development has been observed (Hennig-Thurau, Houston, & Heitjans, 2009; Sood & Drèze, 2006). However, prior success may also increase the bargaining power of the content creator (Ma, Huang, Kumar, & Strijnev, 2015). Second, in terms of market coverage, the dataset covers books from German-speaking countries and children's and young adult literature, with a focus on hardcover titles. Note that the relationships between formats may change over time. For example, e-books became more widespread in the period under consideration (from 2010 to mid-2014). The sales data that are used originate from the market research company *Media Control* and cover approximately 85% of the market, which it then extrapolates.

Further research may explore how new media, particularly the advancement of e-books, changes the market dynamics of sales drivers because digital content may provide more opportunities to signal quality. Additionally, field experiments may shed further light on the endogeneity considerations that are mentioned above.

Overall, this study demonstrates how theoretical concepts can be translated into practical implications for marketing management and thus advance our knowledge in the field of media economics.

Note

1. Printing costs are a major concern in the German book market whose annual value is worth more than nine billion Euro (Börsenverein des deutschen Buchhandels, 2017). In comparison, the German book market has six times the turnover of the German music industry (Bundesverband Musikindustrie, 2014).

Acknowledgments

We would like to thank Ella Schellmoser for her support in compiling the dataset.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [DFG-FG 1452].

Notes on contributors

Cord Otten is doctoral candidate, University of Hamburg, Institute for Marketing, Moorweidenstr. 18, D-20148 Hamburg, Germany and Kühne Logistics University, Großer Grasbrook 17, D-20457 Hamburg, Germany, e-mail: cord.otten@uni-hamburg.de

Michel Clement is Professor of Marketing and Media, University of Hamburg, Institute for Marketing, Moorweidenstr. 18, D-20148 Hamburg, Germany, e-mail: michel.clement@uni-hamburg.de (corresponding author)

Dominik Stehr is research assistant, University of Hamburg, Institute for Marketing, Moorweidenstr. 18, D-20148 Hamburg, Germany, e-mail: dominik-stehr@vodafone.de

ORCID

Cord Otten http://orcid.org/0000-0001-5684-215X

References

Ainslie, A., Drèze, X., & Zufryden, F. (2005). Modeling movie life cycles and market share. *Marketing Science*, *24*(3), 508–517. doi:10.1287/mksc.1040.0106

Albers, S. (2012). Optimizable and implementable aggregate response modeling for marketing decision support. *International Journal of Research in Marketing*, *29*(2), 111–122. doi:10.1016/j.ijresmar.2012.03.001

Ateca-Amestoy, V. (2008). Determining heterogeneous behavior for theater attendance. *Journal of Cultural Economics*, *32*(2), 127. doi:10.1007/s10824-008-9065-z

Ayyagari, P., Deb, P., Fletcher, J., Gallo, W., & Sindelar, J. L. (2013). Understanding heterogeneity in price elasticities in the demand for alcohol for older individuals. *Health Economics*, *22*(1), 89–105. doi:10.1002/hec.1817

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. A. (2015). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research*, *53*(3), 297–318. doi:10.1509/jmr.14.0380

Barrot, C., Becker, J. U., Clement, M., & Papies, D. (2015). Price elasticities for hardcover and paperback fiction books. *Schmalenbach Business Review*, *67*(1), 73–91.

Beck, J. (2007). The sales effect of word of mouth: A model for creative goods and estimates for novels. *Journal of Cultural Economics*, *31*(1), 5–23. doi:10.1007/s10824-006-9029-0

Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, *42*(2), 141–156. doi:10.1509/jmkr.42.2.141.62296

Börsenverein des deutschen Buchhandels, B. (2016). Buch und Buchhandel in Zahlen 2016.

Börsenverein desdeutschen Buchhandels, B. (2017). *Buch und Buchhandel in Zahlen 2017*. Frankfurt am Main, Germany.

Brynjolfsson, E., Dick, A. A., & Smith, M. D. (2009). A nearly perfect market? *QME*, *8*(1), 1–33. doi:10.1007/s11129-009-9079-7

Bundesverband Musikindustrie e.V. (2014). *Musikindustrie in Zahlen 2013* (1st Aufl.). Berlin: Bundesverband Musikindustrie.

Burmester, A. B., Becker, J. U., van Heerde, H. J., & Clement, M. (2015). The impact of pre- and post-launch publicity and advertising on new product sales. *International Journal of Research in Marketing*, *32*(4), 408–417. doi:10.1016/j.ijresmar.2015.05.005

Burmester, A. B., Eggers, F., Clement, M., & Prostka, T. (2016). Accepting or fighting unlicensed usage: Can firms reduce unlicensed usage by optimizing their timing and pricing strategies? *International Journal of Research in Marketing*, *33*(2), 343–356. doi:10.1016/j.ijresmar.2015.06.005

Caliendo, M., Clement, M., & Shehu, E. (2015). The effect of individual professional critics on books' sales: Capturing selection biases from observable and unobservable factors. *Marketing Letters*, *26*(4), 423–436. doi:10.1007/s11002-015-9391-9

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, *1*(1), 29–53.

Castillo, P. A., Mora, A. M., Faris, H., Merelo, J. J., García-Sánchez, P., Fernández-Ares, A. J., … García-Arenas, M. I. (2017). Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment. *Knowledge-Based Systems*, *115*, 133–151. doi:10.1016/j.knosys.2016.10.019

Chang, B.-H., & Ki, E.-J. (2005). Devising a practical model for predicting theatrical movie success: Focusing on the experience good property. *Journal of Media Economics*, *18*(4), 247–269. doi:10.1207/s15327736me1804_2

Chen, Y., Fay, S., & Wang, Q. (2011). The role of marketing in social media: How online consumer reviews evolve. *Journal of Interactive Marketing*, *25*(2), 85–94. doi:10.1016/j.intmar.2011.01.003

Clement, M., Proppe, D., & Rott, A. (2007). Do critics make bestsellers? Opinion leaders and the success of books. *Journal of Media Economics*, *20*(2), 77–105. doi:10.1080/08997760701193720

Clerides, S. K. (2002). Book value: Intertemporal pricing and quality discrimination in the US market for books. *International Journal of Industrial Organization*, *20*(10), 1385–1408. doi:10.1016/S0167-7187(02)00004-8

De Vany, A., & Walls, W. D. (1996). Bose-Einstein dynamics and adaptive contracting in the motion picture industry. *The Economic Journal*, *106*(439), 1493–1514. doi:10.2307/2235197

Dorfman, R., & Steiner, P. O. (1954). Optimal advertising and optimal quality. *The American Economic Review*, *44*(5), 826–836.

Ebbes, P., Papies, D., & van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of Endogeneity. *Marketing Science*, *30*(6), 1115–1122. doi:10.1287/mksc.1110.0666

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881–893. doi:10.1287/mnsc.1060.0668

Eliashberg, J., Swami, S., Weinberg, C. B., & Wierenga, B. (2001). Implementing and evaluating silverscreener: A marketing management support system for movie exhibitors. *Interfaces*, *31*(3_supplement), S108–S127. doi:10.1287/inte.31.3s.108.9685

Farris, P., Olver, J., & De Kluyver, C. (1989). The relationship between distribution and market share. *Marketing Science*, *8*(2), 107–128. doi:10.1287/mksc.8.2.107

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing*, *90*(2), 217–232. doi:10.1016/j.jretai.2014.04.004

Frisbie, G. A. (1980). Ehrenberg's negative binomial model applied to grocery store trips. *Journal of Marketing Research*, *17*(3), 385–390. doi:10.2307/3150539

Hennig-Thurau, T., Houston, M. B., & Heitjans, T. (2009). Conceptualizing and measuring the monetary value of brand extensions: The case of motion pictures. *Journal of Marketing*, *73*(6), 167–183. doi:10.1509/jmkg.73.6.167

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2015). Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. *Journal of the Academy of Marketing Science*, *43*(3), 375–394. doi:10.1007/s11747-014-0388-3

Hirschman, E. C., & Holbrook, M. B. (1982). Hedonic consumption: Emerging concepts, methods and propositions. *Journal of Marketing*, *46*(3), 92–101. doi:10.2307/1251707

Hjorth-Andersen, C. (2000). A model of the Danish book market. *Journal of Cultural Economics*, *24*(1), 27–43.

Hofmann-Stölting, C., Clement, M., Wu, S., & Albers, S. (2017). Sales forecasting of new entertainment media products. *Journal of Media Economics*, *30*(3), 143–171. doi:10.1080/08997764.2018.1452746

Kamphuis, J. (1991). Satisfaction with books: Some empirical findings. *Poetics*, *20*(5–6), 471–485. doi:10.1016/0304-422X(91)90021-G

Kanuri, V. K., Mantrala, M. K., & Thorson, E. (2017). Optimizing a menu of multiformat subscription plans for ad-supported media platforms. *Journal of Marketing*, *81*(2), 45–63. doi:10.1509/jm.15.0372

Köhler, C., Mantrala, M. K., Albers, S., & Kanuri, V. K. (2017). A meta-analysis of marketing communication carryover effects. *Journal of Marketing Research*, *54*(6), 990–1008. doi:10.1509/jmr.13.0580

Leemans, H., & Stokmans, M. (1991). Attributes used in choosing books. *Poetics*, *20*(5), 487–505. doi:10.1016/0304-422X(91)90022-H

Lehmann, D. R. (2004). Metrics for making marketing matter. *Journal of Marketing*, *68*(4), 73–75.

Leitão, L., Amaro, S., Henriques, C., & Fonseca, P. (2018). Do consumers judge a book by its cover? A study of the factors that influence the purchasing of books. *Journal of Retailing and Consumer Services*, *42*, 88–97. doi:10.1016/j.jretconser.2018.01.015

Ma, J., Huang, D., Kumar, M. V. S., & Strijnev, A. (2015). The impact of supplier bargaining power on the advertising costs of movie sequels. *Journal of Cultural Economics*, *39*(1), 43–64. doi:10.1007/s10824-014-9223-4

Mantrala, M. K., Naik, P. A., Sridhar, S., & Thorson, E. (2007). Uphill or Downhill? Locating the firm on a profit function. *Journal of Marketing*, *71*(2), 26–44. doi:10.1509/jmkg.71.2.026

Nelson, R. A. (2001). What's an Oscar worth? *Economic Inquiry*, *39*(1), 1–16. doi:10.1111/%28ISSN%291465-7295/issues

Park, S., & Gupta, S. (2012). handling endogenous regressors by joint estimation using copulas. *Marketing Science*, *31*(4), 567–586. doi:10.1287/mksc.1120.0718

Schmidt-Stölting, C., Blömeke, E., & Clement, M. (2011). Success drivers of fiction books: An empirical analysis of hardcover and paperback editions in Germany. *Journal of Media Economics*, *24*(1), 24–47. doi:10.1080/08997764.2011.549428

Shapiro, C., & Varian, H. R. (1999). *Information rules: A strategic guide to the network economy*. Boston, Massachusetts: Harvard Business Press.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, *30*(2), 243–254. doi:10.1016/j.eswa.2005.07.018

Sood, S., & Drèze, X. (2006). Brand extensions of experiential goods: Movie sequel evaluations. *Journal of Consumer Research*, *33*(3), 352–360. doi:10.1086/508520

Spencer, K. (2017). Marketing and sales in the U.S. young adult fiction market. *New Writing*, *14*(3), 429–443. doi:10.1080/14790726.2017.1307419

Sridhar, S., Mantrala, M. K., Naik, P. A., & Thorson, E. (2011). Dynamic marketing budgeting for platform firms: Theory, evidence, and application. *Journal of Marketing Research (JMR)*, *48*(6), 929–943. doi:10.1509/jmr.10.0035

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S (4th ed.). In *Statistics and computing* (4th ed.). Retrieved from www.springer.com/de/book/9780387954578

Waldfogel, J., & Reimers, I. (2015). Storming the gatekeepers: Digital disintermediation in the market for books. *Information Economics and Policy*, *31*, 47–58. doi:10.1016/j.infoecopol.2015.02.001

Wübben, M., & Wangenheim, F. (2008). instant customer base analysis: Managerial heuristics often "Get It Right". *Journal of Marketing*, *72*(3), 82–93. doi:10.1509/jmkg.72.3.82

Yan, M. Z., & Napoli, P. M. (2006). Market competition, station ownership, and local public affairs programming on broadcast television. *Journal of Communication*, *56*(4), 795–812. doi:10.1111/j.1460-2466.2006.00320.x

You, Y., Vadakkepatt, G. G., & Joshi, A. M. (2015). A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, *79*(2), 19–39. doi:10.1509/jm.14.0169

Yucesoy, B., Wang, X., Huang, J., & Barabási, A.-L. (2018). Success in books: A big data approach to bestsellers. *EPJ Data Science*, *7*(1), 7. doi:10.1140/epjds/s13688-018-0135-y

Appendix

Appendix 1. Full market response model with copulas – results.

		Log-log model		NB model	
		Coefficient	Std. error	Coefficient	Std. error
	Constant	−1.051	2.341	1.810	1.957
Product	Amazon Is Rated	0.048	0.419	0.003	0.350
	Amazon Volume [a]	0.587	0.050***	0.562	0.042***
	Amazon Valence [a]	0.079	0.264	0.078	0.221
	Author Power [a]	0.723	0.212***	0.732	0.177***
	Author Power COPULA	−0.393	0.267	−0.324	0.223
	Genre (not mutually exclusive)				
	Beloved Children's Character	−0.034	0.152	−0.014	0.127
	Fantasy & Science-Fiction	−0.167	0.109	−0.170	0.091*
	Crime & Thriller	−0.104	0.123	−0.069	0.103
	Novel & Thriller	−0.005	0.189	−0.064	0.158
	Adventure	0.000	0.089	0.037	0.074
	Animal Stories	−0.121	0.123	−0.052	0.103
	Sport	−0.293	0.270	−0.257	0.226
	Love & Friendship	0.015	0.131	−0.016	0.110
	Miscellaneous	0.189	0.091**	0.137	0.076*
	Is Series	0.254	0.118**	0.159	0.098
	Series Power [a]	0.020	0.012	0.015	0.010
	Is Schoolbook	0.146	0.189	0.155	0.158
	Teenager	−0.413	0.115***	−0.349	0.096***
	Pages [a]	0.140	0.113	0.105	0.094
	Price[a]	0.413	0.541	−0.460	0.452
	Price COPULA	−0.278	0.184	−0.019	0.154
Distribution	*Month of publication*				
	February	−0.132	0.143	−0.150	0.120
	March	−0.188	0.130	−0.170	0.108
	April	−0.445	0.196**	−0.375	0.164**
	May	−0.029	0.207	0.071	0.173
	June	0.094	0.152	0.075	0.127
	July	−0.272	0.153*	−0.225	0.128*
	August	−0.163	0.144	−0.146	0.120
	September	−0.193	0.146	−0.257	0.122**
	October	−0.489	0.164***	−0.429	0.137***
	November	−0.115	0.208	−0.002	0.174
	December	−0.338	0.432	−0.423	0.361
	Publisher Power [a]	0.239	0.252	0.123	0.211
	Publisher Power COPULA	−0.012	0.146	0.024	0.122
	Audiobook	0.332	0.090***	0.245	0.075***
	Prize Nominated	0.210	0.324	0.210	0.271
	Random	0.250	0.113	0.296	0.094***
	Dispersion parameter			2.076	0.118

[a] Logarithmic values of variables. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. R^2 of log-log model = 0.714 (adj. R^2 = 0.695)