Joint Discussion Paper
Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

# No. 11-2021

**Maximilian Maurice Gail and Phil-Adrian Klotz**

# The Impact of the Agency Model on E-book Prices: Evidence from the UK



**Coordination: Bernd Hayo • Philipps-University Marburg**
**School of Business and Economics • Universitätsstraße 24, D-35032 Marburg**
**Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de**

# The Impact of the Agency Model on E-book Prices: Evidence from the UK

Maximilian Maurice Gail[1†] and Phil-Adrian Klotz[2*†]

[1*]Chair for Industrial Organization, Regulation and Antitrust, Department of Economics, Justus-Liebig-University Giessen, Licher Strasse 62, Giessen, 35394, Germany.
[2*]Chair for Microeconomics, Department of Economics, University of Hohenheim, Schloss Osthof-West, Stuttgart, 70599, Germany.

*Corresponding author(s). E-mail(s): philadrian.klotz@uni-hohenheim.de;
Contributing authors: maximilian.m.gail@wirtschaft.uni-giessen.de;
†These authors contributed equally to this work.

**Abstract**

This paper empirically analyzes the effect of the widely used agency model on retail prices of e-books sold in the United Kingdom. Using an unique cross-sectional data set of e-book prices for a large number of book titles across all major publishing houses, we exploit cross-genre and cross-publisher variation to identify the effect of the agency model on e-book prices. Since the genre information is ambiguous and even missing for some titles in our original data set, we also apply a Latent Dirichlet Allocation (LDA) approach to determine detailed book genres based on the book's descriptions. Using the propensity score matching, we find that retail prices for e-books sold under the agency model are approximately 20% cheaper than book titles sold under the wholesale model. This effect varies with the exact sales rank of a book, i.e. whether a book title is a bestselling or a *long tail* book. Our results are robust to different regression specifications and double machine learning techniques.

**Keywords:** e-books, agency agreements, vertical restraints, Amazon, propensity score matching

**JEL Classification:** D12 , D22 , L42 , L82 , Z11

# 1 Introduction

The rise of the internet - and platform markets more specifically - has accelerated the usage of so-called agency arrangements. Thereby, suppliers pay retailers sales royalties to distribute products at prices determined by suppliers. On the contrary, in many traditional retail markets suppliers charge retailers wholesale prices and retailers set final prices to consumers (*wholesale model*). Even though agency arrangements are also used in some of these conventional markets (e.g., newspapers sold at kiosks), this form of a vertical contract is especially prevalent in online markets (e.g., Amazon Marketplace, Apple App Store, eBay Buy It Now). It is also frequently used in the retail market examined in this paper, namely the e-book market.

In general, books are experience goods because readers can ascertain the quality only after reading a given book title (Nelson, 1970; Reimers and Waldfogel, 2021). In some countries such as Germany, France or Japan, book prices are fixed whereas countries without fixed book price (FBP) systems include the UK and the USA. Fixed book prices are a form of resale price maintenance (RPM) where publishers set retail prices and price competition between retailers is restricted or completely eliminated.[1] Mostly the motivation behind the introduction of FBP systems is the assurance for a broad and diverse supply of books, available through a geographically wide network of bookstores.

With the advent of e-books, countries with a FBP system for print books had to decide whether to extend existing legislation to e-books. It is questionable whether the same cultural policy arguments and legal considerations apply, in particular because a geographically wide network of bookstores is irrelevant for e-books. Nevertheless, eight OECD countries with fixed prices for printed books also have fixed prices for e-books, while no country is known to have such a RPM regulation for e-books but not for print books (Poort and van Eijk, 2017). However, in many countries without fixed prices for e-books (such as the UK), this digital product is partly sold under the agency model (Gilbert, 2015), which has similar effects like RPM between a manufacturer and a retailer.[2]

In 2010, Apple in co-operation with the six largest publishing houses have been the first who adopted the agency model for e-books in response to Amazon's aggressive pricing strategy to gain market share. In April 2012, the Department of Justice (DOJ) sued Apple and five of the six publishing houses for conspiring to raise e-book prices by using the agency model in conjunction with most-favored nation (MFN) clauses.[3] Three of the publishers settled shortly after the antitrust case was filed, while the

[1] Presently, 15 OECD countries have a regulation for fixing the prices of printed books. The fixed prices for printed books typically last 18-24 months after a book has been published.

[2] They are economically similar in the sense that the upstream firms control the retail prices. However, a major difference is that for agency pricing the downstream firms individually delegate retail pricing to the upstream firms, whereas under the classical case RPM is imposed at the market level for any given product.

[3] See United States v. Apple Inc., 12 Civ. 2826 (DLC). MFN clauses stipulate that the retail price of a given product set by a publisher through one retailer must be no higher than the retail price set by this publisher through a competing retailer. Hence, a MFN clause guarantees a retailer who prefers a higher commission, if it raises the commission it charges for one publisher, the retail price will remain the same relative to the other retailers. This effect encourages retailers to push for higher fees, which results in higher retail prices (Johnson, 2017).

other two followed later the same year, which meant that the five publishers could not restrict a retailer's ability to set e-book prices for a period of two years.

Empirical evidence on the price effects of RPM and fixed book prices as well as of the agency model is scarce. While systematic empirical evidence on RPM is limited to case studies (MacKay and Smith, 2017; Ippolito, 1991), the only study investigating the empirical effect of the agency model is the one from De los Santos and Wildenbeest (2017). They have used data on e-book prices of bestselling book titles for the years 2012 and 2013 and the Apple case as an exogenous shock to show that the agency model in combination with MFN clauses led to an average increase in prices between 8-18% (depending on the retailer).[4]

The goal of this study is to analyze the price effect of the agency model using a larger and more detailed data set (especially not only incorporating bestselling book titles) and to check whether similar effects also occur in the absence of an alleged conspiracy as in the Apple case. The internet allows consumers access to a larger number of book titles rather than simply the popular ones (bestselling titles) and different authors have shown the importance of those *long tail* titles in markets for creative products (e.g., Aguiar and Waldfogel, 2018; Brynjolfsson et al, 2003). For instance, Brynjolfsson et al (2003) have estimated that the benefit consumers obtain from access to *long tail* book titles may be as high as \$1.03 billion alone in the year 2000.

Our cross-sectional data set contains prices for 12,001 e-books published on Amazon UK between 2010 and 2020. Using data from Amazon ensures a high market coverage since Amazon accounted for 50 percent of the UK book sales in 2018.[5] We further use publisher- and book genre variation to estimate the effect of the agency model on e-book retail prices. The results of our propensity score matching design indicate that e-books sold under the agency model on *Amazon.co.uk* are approximately 20% cheaper than digital books sold under the wholesale model. This effect varies with the exact sales rank of a book, i.e. whether a book title is a bestselling or a *long tail* book. Various robustness checks qualitatively confirm this main finding. This result contradicts the empirical outcome from De los Santos and Wildenbeest (2017)[6], but fits into explanations put forward by the theoretical literature on agency versus wholesale models (Johnson, 2020; Foros et al, 2017; Gaudin and White, 2014).

The rest of the paper is structured as follows. In Section 2, we discuss the related literature. Section 3.1 describes the data and we present some descriptive statistics in Section 3.2. Section 4 presents our main estimation strategy and results. In Section 5,

[4] In a second paper, De los Santos et al (2018) analyze the switch back from wholesale to agency pricing (after a two-year ban on agency pricing following the Apple case) and find qualitatively similar results for the Amazon platform.

[5] See Nielsen (2018), "Books & Consumers - UK Industry Standard Report Q4 2018", p. 13.

[6] Our paper differs from the study of De los Santos and Wildenbeest (2017) in three important ways. First, while De los Santos and Wildenbeest (2017) use the court decision in the Apple Case (see Footnote 3) as an exogenous shock in their approach, our findings do not rely on an alleged conspiracy. Second, we do not only incorporate bestselling book titles into our empirical analysis but also *long tail* book titles. In our empirical analysis, we show that the agency effect is heterogeneous with respect to bestselling and *long tail* book titles. And third, De los Santos and Wildenbeest (2017) cannot measure the "pure" price effect stemming from agency agreements since those were used in conjunction with MFN clauses at that time. Our price effect can be attributed only to the agency agreements because Amazon has settled with the EU Commission in 2017 not to include MFN clauses in respect of any e-book distributed in the EEA for the next five years (See AT.40153 E-book MFNs and related matters (Amazon), Decision dated May 4, 2017. We have scraped the Amazon data in 2020).

our robustness checks are outlined. We conclude and outline the contributions of our paper in Section 6.

## 2 Related Literature

Our article contributes to several strands of literature. First and foremost, it is related to studies which investigate the competitive effects of the agency model. While the empirical literature on the economic effects of the agency model is rather scarce (apart from the two studies by De los Santos and Wildenbeest (2017) and De los Santos et al (2018)), several recent theoretical papers have analyzed differences in retail prices between the agency and the wholesale model. Lu (2017) uses a bilateral duopoly model with product differentiation in the upstream- and downstream market to show that the agency model benefits consumers relative to the wholesale model with lower retail prices due to the elimination of double marginalization. Johnson (2020) finds that when publishers set retail prices instead of retailers (agency model), prices may be higher in early periods but lower in later periods since in the wholesale model retailers initially set low prices to lock in consumers, but find it optimal to raise prices once a sufficient number of consumers are locked in.

Another strand of the theoretical literature on agency models assumes that complementary devices are necessary for the enjoyment of the main products (e.g., an e-book reader in the case of e-books). Gaudin and White (2014) point out that the incentive of a retailer to set high prices is larger when she has monopolistic control over a complementary device, as it was the case in the e-book market when e-books from Amazon could only be read on a Kindle device. In another model-theoretical setup, Abhishek et al (2016) show that agency selling is more efficient than the wholesale model and leads to lower retail prices, even though retail prices may be higher under the agency model if there are positive externalities from sales of associated products (such as e-readers in the case of e-books).

Foros et al (2017) show that the agency model is always anti-competitive (leads to higher retail prices) when it is adopted by the platforms on a market-by-market basis. To be more specific, they find that upstream firms (publishers) will set higher retail prices than downstream firms (retailers) would set if they were in control as long as competition is greater among retailers than among publishers. Moreover, they (ibid.) point out that a retailer who sets retail prices independently (wholesale model) benefits when a horizontal rival is restricted by the agency model since the latter creates a price umbrella, which makes it profitable for the independent price-setting retailer to increase prices. MFN clauses induce industry-wide adoption of agency pricing in their model when such adoption would not otherwise have occurred and can thus be seen as anti-competitive. Condorelli et al (2018) present a theory that makes the decision whether to use agency or wholesale models endogenously in an environment where the retailer has privileged information about the valuations of consumers and show that retailers prefer the agency model.

More generally, our article is also related to the broader literature on RPM. RPM can lead to lower retail prices due to the internalization of vertical externalities such as double marginalization (Spengler, 1950; Tirole, 1988), it can be used to correct for service externalities (Mathewson and Winter, 1984; Perry and Porter, 1986; Winter, 1993) and RPM regimes can also lead to a larger number of brick-and-mortar (B&M) stores compared to regimes with free prices (Dearnley and Feather, 2002; Davies et al, 2004). However, Rey and Stiglitz (1988, 1994) point out that vertical restraints that eliminate intra-brand competition can also be used to mitigate inter-brand competition and then would be anti-competitive.

Finally, our article also contributes to the newer literature on machine learning (ML) and text mining approaches. Varian (2014) and Athey and Imbens (2019) provide an overview of important ML methods. Wang et al (2019) use the *Learning to Place* ML approach to predict book sales and find that a strong driving factor of book sales across all genres is the publishing house. We will use double machine learning (DML) techniques similar to the approach of Knaus (2021), who has used DML in the case of musical practice of children on cognitive skills and school performance. For a broad overview on text mining approaches see Gentzkow et al (2019). In our article, we will use a latent Dirichlet allocation (LDA) model to determine book genres by analyzing the book descriptions and expert reviews (e.g., Larsen and Thorsrud, 2019).

# 3 Data

In this section, we present our data set containing a large number of book titles. We first describe the construction of our data set in Section 3.1. Descriptive statistics including information on prices, ratings, reviews and the digital size of e-books (in KB) are presented in Section 3.2.

## 3.1 Data Set Construction

The data generating process is structured as follows. We have scraped the *Amazon.co.uk* webpage for books by utilizing a list of publisher and imprint names starting mid February 2020 taking two weeks to get e-book prices as well as further book characteristics available on the Amazon website. Therefore, we use *a priori* a list of publishing houses, publishers and imprints, which is taken from a historical *Sunday Times* bestseller list. This procedure ensures that our sample only contains books from publishers with a relatively high market share and that we cover as many books as possible.[7]

This proceeding also incorporates books into our data set which have been published before long before 2019 since we have done the publisher search on *Amazon.co.uk* independently of the format. Thus, it may have happened that for a certain book title, which we have found within our observation period, another format of the same title has already been published a few years ago. However, as the analysis focuses on e-books we have ensured that no book is included in our working data set which has been published earlier than 2010.[8]

Our raw data set consists of roughly one million observations, whereby one observation point contains several information on different prices, formats, descriptions,

[7] The used bestseller list contains entries from January, 2006 until the end of March, 2019.

[8] We focused on the introductory years of Amazon's Kindle Reader, https://www.aboutamazon.eu/news/innovation/a-look-back-at-10-years-of-the-amazon-kindle.

ratings, reviews etc. being available on the Amazon website. For every book title there are three entries if all formats (hardcover, paperback, e-book) are available for a certain book title. However, due to the usage of web scraping methods the data set contains of some entries that are duplicates or not of interest for our analysis. Hence, after the data cleansing process our working data set consists of 77,629 observations, respectively 47,161 unique book titles.[9] We only use book titles in our estimation approach for which all explanatory variables (book characteristics) are available (see Table 1). Thus, for our empirical analysis 12,001 e-book titles remain in the final working data set.

Our variables of interest are the retail price, which is the price a consumer must pay for a certain e-book, and the treatment variable *Agency*, which takes the value one if there is a text field on the Amazon webpage of a book title expressing *'This price was set by the publisher'* and zero otherwise.[10] Beyond, we have data on several control variables for our empirical analysis. These variables comprise book characteristics as the book format, the book genre, the size of an e-book in KB, variables on book reviews as the star rating and the number of consumer and expert reviews, variables containing information on the author or publisher of a book title and other variables as the publication date or the recommended retail price (RRP). Table 1 summarizes the descriptions for all variables included in our data set.

| Variables | Information |
|---|---|
| Price | Retail price from the upper right *Buy-Box* |
| Format | Hardcover, paperback, Kindle |
| Star rating | Average rating normalized to be between 0 and 1 |
| No. customer reviews | Number of consumer reviews |
| No. expert reviews | Number of expert reviews on Amazon |
| Series | Dummy variable whether book is part of a series |
| Description and reviews | Detailed text-information on the book and by different reviewers |
| Genre | Constructed by LDA from the descriptions and reviews (see Appendix B) |
| RRP | Recommended retail price which is the print RRP. For Kindle it is either related to the hardcover or paperback RRP |
| Agency | Dummy variable to be one if the price was set by the publisher and zero otherwise. Only possible for e-books |
| Seller | Sample is restricted to be sold by Amazon |
| Author | Information on the author of a book |
| Title | Information on the title of a book |
| Kindle_Size | Kindle file size (in KB) |
| Publisher | Name of the publisher. We have different levels of aggregation (Imprint,Publisher,Publishing House) |
| Amazon rank | Uncategorized Amazon bestseller rank for either print books or e-books |
| Bestsellers | Number of bestsellers in the Sunday Times Bestseller List conditional on the Author's name |
| WeekInChart | Average number of weeks in the bestseller charts conditional on the Author's name |
| Identifier | Aggregation of ASINs to verify the books |
| Date Retail | Period of time since the publication of a book title (in years) |

**Table 1**: Relevant Variables per book title and the information content they provide.

We have also matched the data set obtained from Amazon with a historical *Sunday Times* bestseller list to identify authors who have already written a bestselling book title in the past. This variable is important for our empirical analysis (in which we estimate the retail price on an e-book) since the name of a bestseller author is an important quality signal for the book readers.

Each book is a unique product written by an author and mostly published by one publisher. Thus, books are heterogeneous goods which makes it impossible to actually compare the value of one specific book with one another. In order to provide

[9] In the former number all three format types (hardcover, paperback, and e-book) are included.
[10] See Appendix Figure 1 for an example from the Amazon webpage.

an acceptable analysis, it is therefore also necessary to control for the genres of the several books. However, the genre information is ambiguous and even not available for some book titles on the Amazon webpage.[11] Thus, we use a Latent Dirichlet Allocation (LDA) approach to derive book genres from the descriptions and reviews of the individual books available on the Amazon webpage. This control variable should be able to capture specific effects between the individual genres. We describe this text mining approach in Appendix B.

### 3.2 Descriptive Statistics

Our sample consists of 47,161 book titles that have been published on *Amazon.co.uk* by the publishers Bloomsbury, Faber, Hachette, HarperCollins, Oxford, Pan Macmillan, Penguin Random House, Scholastic, Simon & Schuster, and a group of smaller publishers in the period between 2010 and 2020.[12] However, in overall there are 77,629 observations in our data set since there are several formats available for some book titles. Even though the focus of our empirical analysis is on the price of e-books, in this section we also present some descriptive statistics for the book formats hardcover and paperback to show the relationship between those three book formats.

Table 1 in Appendix A offers descriptive statistics on the variables we use for our empirical analysis, summarized by publishers. In addition to e-book retail prices and the RRP, we also observe several characteristics for each book title such as the sales rank of a title on Amazon, the customer ratings, the number of customer and expert reviews, and the number of pages. As shown in Table 1, e-books from Scholastic exhibit the lowest average retail price, while the e-books from Bloomsbury have the highest mean prices. Beyond, the titles from Hachette have the lowest average book rank and the book titles published by Simon & Schuster exhibit the highest average number of customer reviews. Most of the other book characteristics are very similar across publishers.

Figure 1 represents the frequency distribution of the retail prices for e-books (top), paperback (centre) and hardcover books (bottom) below £100. It is obvious that e-book prices are in a range between £0.25 and £10, paperback prices concentrate mostly in the £10-£20 interval and hardcover prices are even more expensive. While the distributions of e-books and paperbacks are more compressed, the hardcover book prices exhibit a higher volatility. Finally, all three price distributions have significant mass points at candidate focal points (e.g., £0.49 (e-books), £9.99 (paperback) and £15.99 (hardcover)).

Table 2 presents the descriptive statistics for Figure 1. With an average price of £8.40, e-books are the cheapest of the three book formats, followed by paperback (£12.35) and hardcover books (£28.66). The high standard deviation for hardcover books confirms its high volatility, which we have already detected in Figure 1. In overall, the descriptive statistics on book formats suggest that hardcover books exhibit the highest quality of the three formats and confirm the results from Li (2019), who has

[11] In Section 4.2, we also present one specification with the Amazon genre information as a control variable.

[12] Remember that our empirical analysis is based on 12,001 e-book titles because we exclude book titles with missing information on their characteristics from our estimation approach (see Section 3.1).

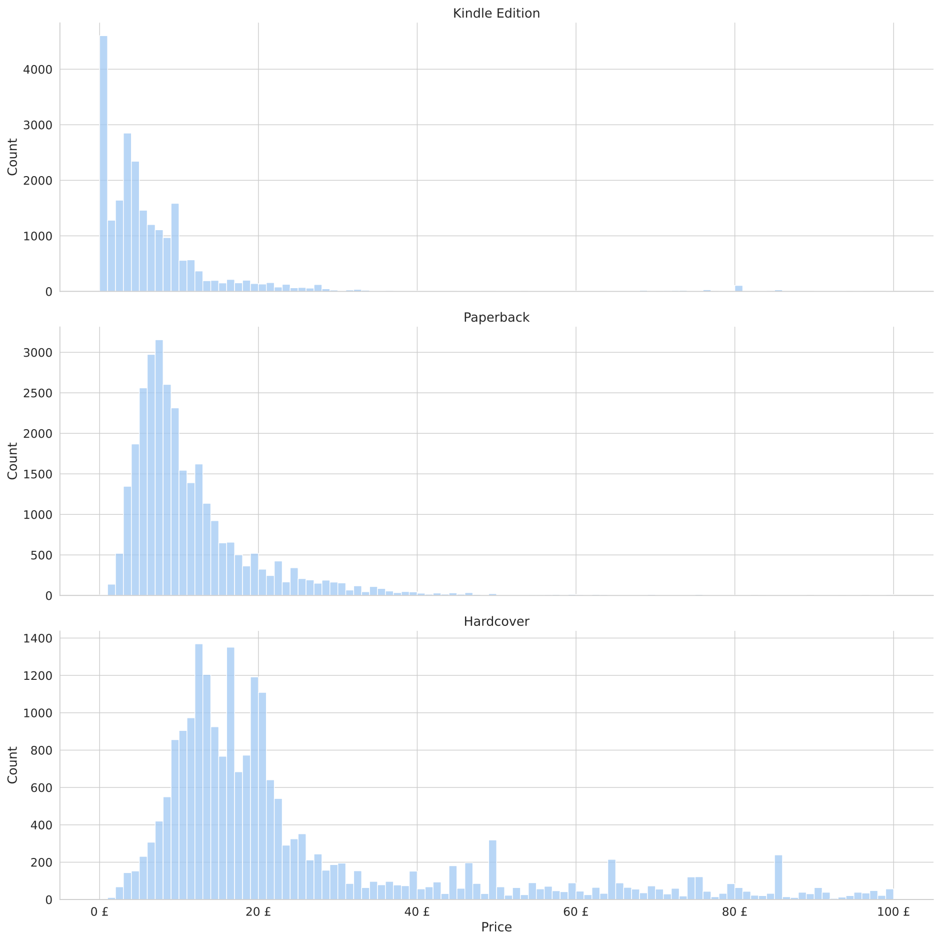


**Figure 1**: Distribution of retail prices by book format.

found that e-books and paperbacks are closer substitutes than e-books and hardcover books.

The e-books of the individual publishers are sold under different pricing arrangements. Amazon states on its product pages whether the respective publisher has set the price of an e-book. The Figures 1 and 2 in Appendix A present examples of this by showing screenshots for the e-book *Elon Musk: How the Billionaire CEO of SpaceX and Tesla is Shaping our Future* as well as for the e-book *Pulse*. In Figure 1, it can be seen in the first box on the right hand side of the Amazon webpage that the 'price was set by the publisher' so that this is an example for the usage of the agency model. On the contrary, in Figure 2 this information is missing, which means that Amazon sets the retail price for this e-book and it represents an example for the wholesale model.

Figure 2 visualizes the distribution of e-book retail prices by the different publishers. Prices are obviously more dispersed for book titles published by Bloomsbury

| Format | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Hardcover | 22,359 | 28.663 | 42.385 | 1.050 | 12.990 | 18.800 | 29.190 | 1,575.000 |
| Kindle Edition | 23,481 | 8.401 | 14.020 | 0.250 | 1.990 | 4.990 | 9.300 | 467.800 |
| Paperback | 30,515 | 12.348 | 18.932 | 0.310 | 6.550 | 9.010 | 13.950 | 1,899.000 |

**Table 2**: Retail prices grouped by book format.

and Oxford, whereas the other major publishers mostly have books in the range up to £20. The group of smaller publishers has a significantly larger fraction of e-books in the cheapest price range (about £0.49). In Appendix B, we will present this figure combined with assigned book genres.

In Table 3, we illustrate the distribution of the pricing arrangements used by the individual publishers for the e-books in our sample. If the number in the column 'Agency' takes the value one, the respective book titles are sold under the agency model, otherwise the titles are sold under the wholesale model. The table shows that all e-books in our data set published by the large publishing houses Hachette, Harper-Collins, Penguin Random House and Simon&Schuster are sold under the agency model on Amazon. Pan Macmillan has some American imprints which still use the wholesale model for their e-books. For the e-books in our sample published by Bloomsbury, Faber, Oxford, Scholastic and most of the smaller publishers, only the wholesale model is used.

| Publisher | Agency | Amount | Percentage | Mean Price |
|---|---|---|---|---|
| Bloomsbury | 0 | 397 | 100% | 10.37 |
| Faber | 0 | 223 | 100% | 5.84 |
| Hachette | 1 | 2,073 | 100% | 5.31 |
| Harper Collins | 1 | 1,469 | 100% | 5.42 |
| Small Pub. | 0 | 4,155 | 77.37% | 6.81 |
| | 1 | 1,215 | 22.63% | 2.32 |
| Oxford | 0 | 161 | 100% | 16.71 |
| Pan Macmillan | 0 | 285 | 50.35% | 9.75 |
| | 1 | 281 | 49.65% | 5.81 |
| Penguin Random House | 1 | 1,240 | 100% | 6.56 |
| Scholastic | 0 | 126 | 100% | 5.01 |
| Simon & Schuster | 1 | 376 | 100% | 7.45 |

**Table 3**: Distribution of the agency variable by publishers.

Finally, we want to illustrate the relationship between retail prices for e-books and their book sales rank on Amazon, which is illustrated in Figure 3 by using a scatter plot with a simple regression line. Obviously, there is a positive relation between the retail price and the rank of an e-book in our sample since the regression line has a positive slope. This finding in our sample is in line with the study of Fishwick (2008), who states that 'substantial discounts' (p. 370) have become prominent for bestselling books in the British book market after the abandonment of the Net Book Agreement in 1997.

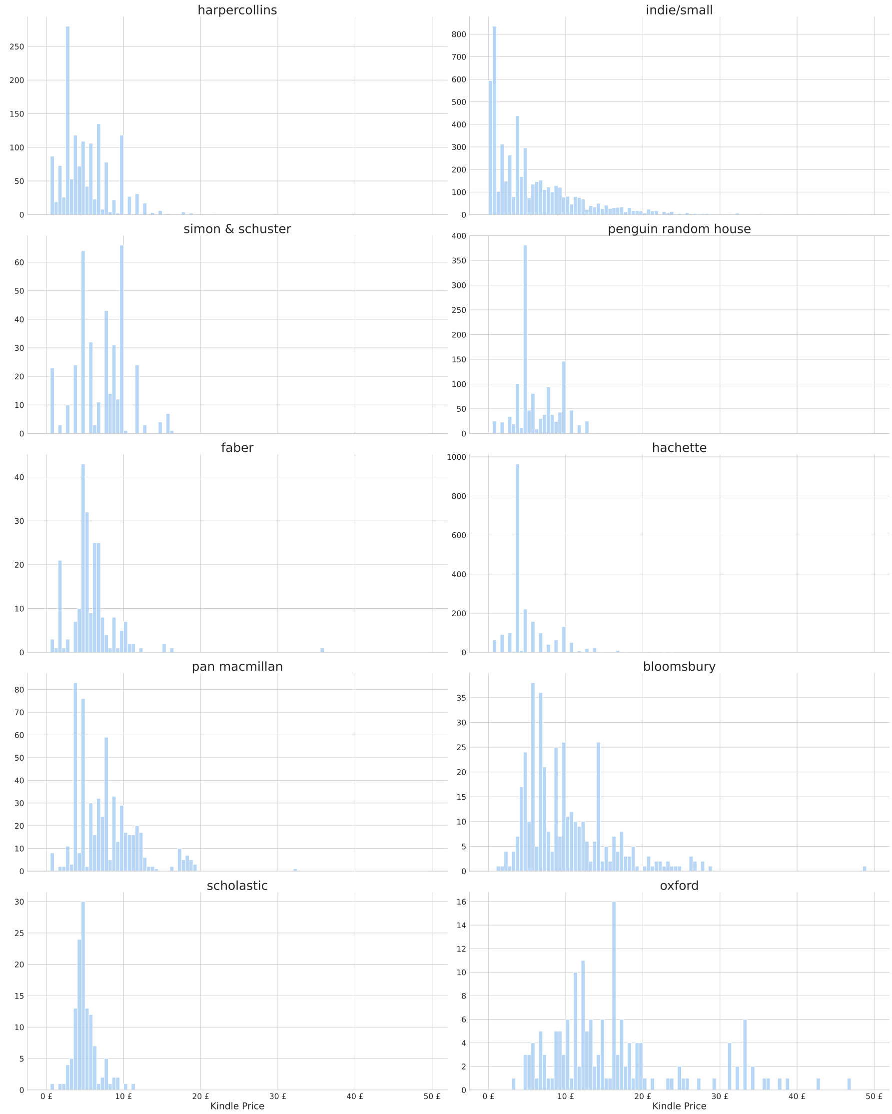


**Figure 2**: Prices for e-books grouped by publishers. The interval size for each bar is 1 Pound. For illustration purposes the figures are censored at 50 Pound.

With respect to the book sales rank, we have to stress how the rank on Amazon is determined. According to sources of Amazon, the ranks are internally updated hourly, but it does not appear immediately. The rank includes current and *all* past sales with higher weights on current sales.[13] This information on the definition of book ranks on Amazon will be important for our estimation approach because the price today might be affected by current sales or past sales, but not necessarily vice versa.

[13]See https://kdp.amazon.com/en_US/help/topic/G201648140. (Last accessed: July 15, 2023)

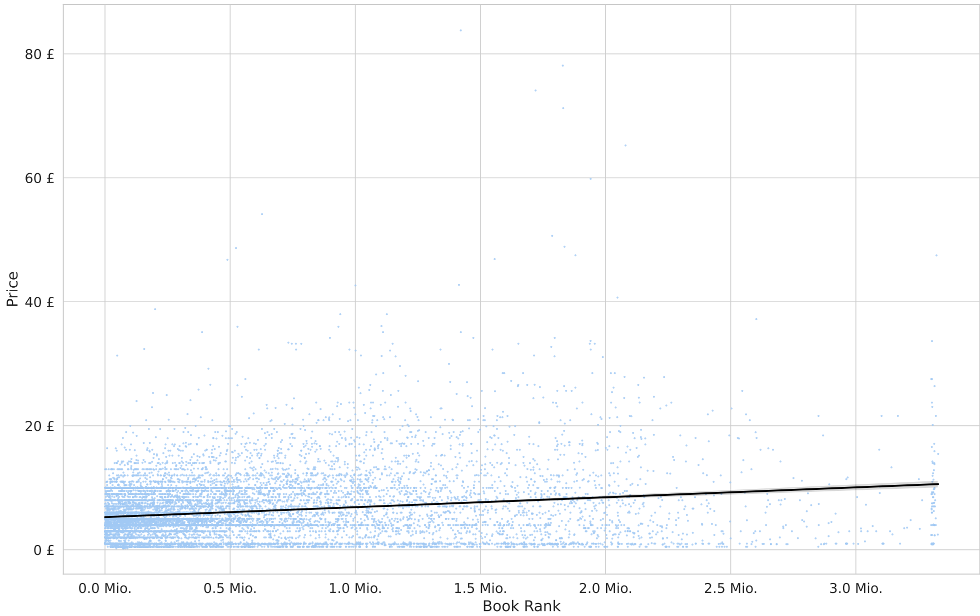


**Figure 3**: Relation of Amazon book ranks for e-books and retail prices.

# 4 Empirical Analysis

In this chapter, we present our empirical analysis. We first describe our estimation strategy in Section 4.1. In Section 4.2, we present the results of our OLS estimation, in which we estimate the effect of the pricing arrangement on the retail price of an e-book. The propensity score matching is outlined in Section 4.3.

## 4.1 Estimation Strategy

The goal of our study is to analyze the effect of the pricing arrangement on the retail prices of e-books sold on *Amazon.co.uk*. Therefore, we use publisher and book genre variation to estimate the price effect of e-books sold under the agency model. Before turning to the presentation of our estimations, we formalize the hypothesis that is to be tested. If there was no difference between the two types of vertical contracts (agency and wholesale) with respect to the retail price of an e-book, the ability for a publisher to set the final consumer price should (*ceteris paribus*) not have any impact on the final consumer prices. Hence, the hypothesis to be tested is:

**Hypothesis** ($H_0$)**.** *The retail price of an e-book is independent of the used vertical contract.*

If there will be a positive correlation between the agency model and the price of e-books, $H_{4.1}$ can be falsified and e-books sold under the agency model are more expensive on average. Observing a negative correlation would also lead to a falsification of $H_{4.1}$, but then e-books sold under the agency model would be cheaper on average.

In our baseline estimation approach, we use the standard hedonic modeling approach in the spirit of Rosen (1974), which relies on observing differences in market prices to infer the value or implicit price of underlying characteristics. Thus, we estimate the following log-log OLS model with heteroscedasticity-consistent standard errors:

$$p_i = \alpha_0 + \alpha_1 A_i + \alpha_2 G_i + \alpha_3 R_i + \alpha_4 RRP_i + W\theta \ + \eta_i. \tag{1}$$

In equation (1), the dependent variable $p_i$ is the logarithm of the retail price for an e-book $i$ sold on *Amazon.co.uk*. The treatment variable $A_i$ is a dummy variable which takes the value one if an e-book is sold under the agency model and zero otherwise. The variable $G_i$ is a categorical variable for the book genres and $R_i$ is a continuous variable for the e-book sales ranks on Amazon. The variable $RRP_i$ reflects the recommended retail price of book title $i$. All other book-specific covariates are included in the matrix $W$ (see Table 1 in Section 3.1).

However, there may be two endogeneity issues in regression equation (1), which might distort our estimated coefficients. First and foremost, our treatment variable $A_i$ might be endogenous due to selection effects. Specifically, e-books sold under agency contracts might differ from e-books sold under wholesale contracts in ways that were directly related to their retail prices. Hence, it was probably not a random process whether an e-book is sold under the agency or the wholesale model. To eliminate this potential selection bias, we apply a matching procedure in Section 4.3 and perform two robustness checks in Section 5.

Second, the rank of an e-book $R_i$ does not only affect the retail price of an e-book $p_i$ but also does the retail price reflect the demand side and, therefore, affect the rank of an e-book which mirrors its sold quantity. As already explained, e-book ranks on *Amazon.co.uk* are internally determined by overall weighted sales ranks. Thus, ranks might be driven by the quantities sold today but the relation is ambivalent since the rank is also affected by quantities sold in the past. The e-book prices can be affected by current and past sales but the impact of prices on total (current and past) sales is not so clear. Nevertheless, we cannot clearly reject the endogeneity issue due to a potential reverse causality. To resolve this potential source of endogeneity between the e-book price and its sales rank, we will also present an instrumental variable approach in Appendix C.

## 4.2 OLS Estimation

The results of our log-log OLS regression model are outlined in Table 4. We estimate three different specifications in this estimation approach: in the first column of Table 4, we present a naive OLS regression model without genre effects, in column (2) we include the LDA genre fixed effects[14], and in the third column we use the Amazon genre information as a control variable.

It is obvious that there is a negative and significant effect of the pricing arrangement *Agency* on the retail price of e-books in all three specifications. According to the amount, the effect is between 18.29% and 18.94% depending on the exact specification.

[14] We describe the process to generate book genres by using an LDA approach in Appendix B. Thereby, we have identified a topic for every e-book title based on the largest probability assigned by the LDA.

For the regression in column (1), an e-book which is sold under the agency model on *Amazon.co.uk* is approximately 18.53% cheaper than an e-book which is sold under the wholesale model on average.[15] Including the LDA genre fixed effects into the regression (see column (2) in Table 4) increases the agency effect to 18.94% (on amount), while the effect gets slightly lower when using the Amazon genre information (see column (3)).[16]

| | Dependent Variable: log Price | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Agency | $-0.205^{***}$ | $-0.210^{***}$ | $-0.202^{***}$ |
| | (0.011) | (0.011) | (0.011) |
| log sales rank | $0.053^{***}$ | $0.059^{***}$ | $0.051^{***}$ |
| | (0.004) | (0.004) | (0.004) |
| log Kindle Size | $0.045^{***}$ | $0.050^{***}$ | $0.027^{***}$ |
| | (0.004) | (0.004) | (0.004) |
| log star rating | $0.531^{***}$ | $0.484^{***}$ | $0.436^{***}$ |
| | (0.049) | (0.048) | (0.048) |
| No. expert reviews | $0.015^{***}$ | $0.015^{***}$ | $0.018^{***}$ |
| | (0.004) | (0.004) | (0.004) |
| log RRP | $1.091^{***}$ | $1.071^{***}$ | $1.080^{***}$ |
| | (0.011) | (0.012) | (0.012) |
| Date Retail | $0.010^{***}$ | $0.012^{***}$ | $0.008^{***}$ |
| | (0.003) | (0.003) | (0.003) |
| Bestsellers | $0.001^{***}$ | $0.001^{***}$ | $0.001^{***}$ |
| | (0.0004) | (0.0004) | (0.0004) |
| WeekInChart | $0.007^{***}$ | $0.007^{***}$ | $0.007^{***}$ |
| | (0.002) | (0.002) | (0.002) |
| Constant | $-1.856^{***}$ | $-2.204^{***}$ | $-1.745^{***}$ |
| | (0.066) | (0.084) | (0.074) |
| Genre | No | Yes | Amazon Genres |
| Inclusion | All | All | All |
| Robust F Statistic | 1285.3022 | 662.2041 | 627.4533 |
| Observations | 12,001 | 12,001 | 12,001 |
| Adjusted $R^2$ | 0.602 | 0.610 | 0.615 |

Robust Standard Errors in parentheses  $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

**Table 4**: OLS estimation approach. Dependent variable is the logarithm of the retail price for e-books sold on Amazon.

The estimated coefficients for the other explanatory variables shown in Table 4 are very similar for the three specifications. The sign of the variable *log sales rank* indicates that e-books with higher sales ranks are sold at higher prices, which we have already demonstrated descriptively in Figure 3 of Section 3.2 and which confirms the results of Fishwick (2008), whereupon bestsellers are sold cheaper in the UK due to 'substantial discounts'. However, we will discuss potential endogeneity issues

[15] To calculate the exact effect of the dummy variable *Agency* on the dependent price variable, the formula $100 \times (e^{\beta} - 1)\%$ must be used.

[16] Table 3 in Appendix A illustrates that a varying number of topics derived from our LDA text mining approach does not qualitatively changes our estimation results.

concerning this control variable in Appendix C. Moreover, there is a significant and positive relation between the RRP of an e-book and its retail price (see variable *log RRP*). We can further observe that the memory space of an e-book (given in KB) has a positive and significant price effect (see variable *log Kindle Size*).

The results presented in Table 4 also indicate that consumer and expert recommendations seem to drive the price of an e-book because the consumer star rating (*log star rating*) as well as the number of expert reviews (*No. expert reviews*) have a positive and significant effect in our regressions. Moreover, the time since the publication of a book title (in years) (variable *Date Retail*) seems to have a positive effect on e-book prices.

Lastly, two covariates remain in Table 4 which control for the author's quality. The explanatory variable *WeekInChart* reflects the average number of weeks former bestsellers of an author have last in the bestseller charts of the *Sunday Times* and the continuous variable *Bestsellers* exhibits the number of bestselling book titles an author has written historically. As expected, both variables have a positive and significant effect on the retail price of an e-book in all three specifications, which can be interpreted as quality signals increasing the price of a book title.[17]

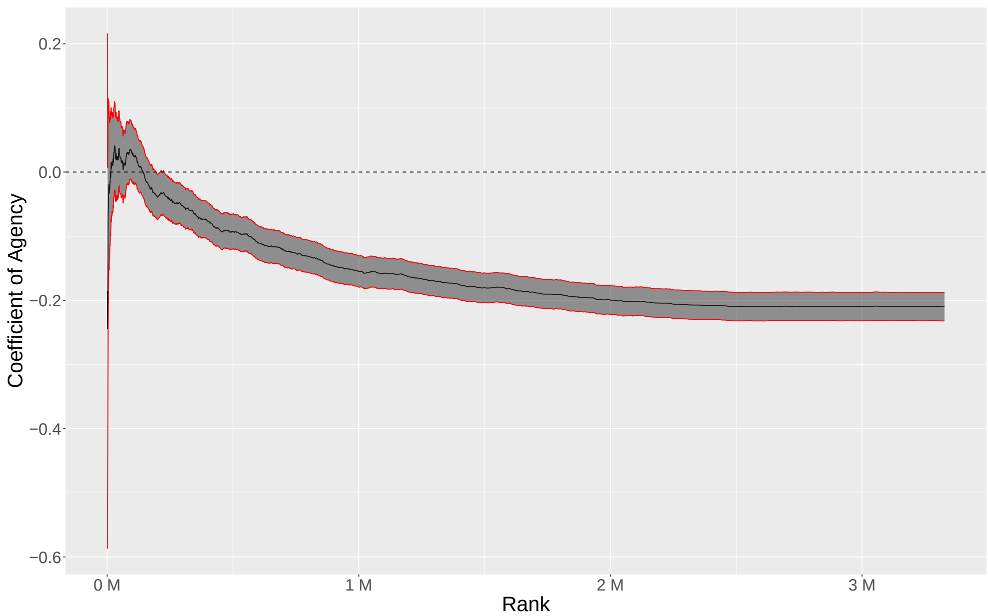


**Figure 4**: Results from equation (1) if higher sales ranks are gradually included. The Figure starts at sales rank of 1000 which contains circa 100 observations.

It is quite possible that the agency effect varies with the sales rank of a book title, so whether we observe a bestselling book title or books sold in higher sales ranks

[17]The variables *WeekInChart* and *Bestsellers* are based on a historical *Sunday Times Bestseller* list. The matching process was conducted via Python's Fuzzy Matching.

(*long tail* books). To illustrate this relation, we again estimate equation (1) by only including a small subset of the data and then gradually include only books up to a specific sales rank. For instance, we include up to rank 1000, then up to rank 2000, and so on, until we reach the largest sales rank. Thus, we receive many estimated agency coefficient based on different e-book subsets. Figure 4 illustrates how the coefficients of the agency variable change if we gradually include higher book sales ranks. The black line gives the point estimator, while the shaded area up to the red lines represents the 95% confidence interval. For the lowest range of sales ranks (bestselling books), the coefficient is not statistically different from zero. By including book titles from higher sales ranks (*long tail* books), the coefficients start getting significantly and converge to −0.21, the outcome from column (2) of Table 4. Overall, Figure 4 implies that the agency effect strongly depends on the book sales rank, which can also explain the divergence between the results of De los Santos and Wildenbeest (2017) (only based on bestselling titles) and our results.

However, the results of our OLS estimation in Table 4 and Figure 4 might be biased and inconsistent due to endogeneity issues regarding our treatment variable *Agency* (see Section 4.1 for a discussion). Hence, we will apply a matching procedure in Section 4.3 to reduce this potential selection bias. Beyond, we will follow an instrumental variable approach in Appendix C to resolve the potential source of endogeneity of our important control variable book sales rank.

## 4.3 Propensity Score Matching

A simple OLS estimation would allow us to identify the effect of different pricing arrangements if the treatment was random. Yet, it was probably not a random process whether an e-book is sold under the agency or wholesale model so that there is a selection bias when using a standard OLS estimation approach. Hence, in a first step we seek to reduce this selection bias by relying on the propensity score matching (Rubin, 1977; Rosenbaum and Rubin, 1983). Thereby, we identify appropriate treated and control e-books through a matching procedure.

In particular, we identify those e-books that are not sold under the agency model but that had ex-ante the same probability of being sold under the agency model as those that are actually sold under the agency model. To implement the propensity score matching, we first run a logistic regression to recover the likelihood that an e-book is sold under the agency model based on its observable characteristics and use the predicted values from that estimation to collapse those covariates into a single scalar called the propensity score. Second, we match an e-book sold under the wholesale model which is as similar as possible to the considered e-book sold under the agency model based on this propensity score.

The propensity score is the selection probability conditional on the confounding variables ($p(X) = Pr(D = 1|X)$) and relies on two identifying assumptions. The first assumption is the conditional independence assumption, which requires that the outcome variable is independent from the treatment, conditional on the propensity score. This means we should only include variables that are expected to simultaneously influence both the treatment and the outcome. The second assumption is called common support assumption and simply means that for any probability there must be units in both the treatment group and the control group.

Then, the model we estimate through a logit regression as a first step of the matching procedure is:

$$A_i = \alpha + X_i\beta + \eta, \quad (2)$$

where $A_i$ is the probability that an e-book $i$ is sold under the agency model and $X_i$ is a vector of e-book characteristics (see the explanatory variables in equation (1)).

| | (1) Logit |
|---|---|
| Agency Set | |
| log RRP | -0.143*** |
| | (-3.55) |
| log sales rank | -0.175*** |
| | (-12.50) |
| AverageR | -0.184 |
| | (-1.37) |
| No. expert reviews | 0.370*** |
| | (20.85) |
| log Kindle Size | 0.119*** |
| | (6.94) |
| WeekInChart | 0.0110 |
| | (1.23) |
| Date Retail | -0.00530 |
| | (-0.52) |
| Bestsellers | 0.0192*** |
| | (3.53) |
| Constant | 0.839** |
| | (2.70) |
| N | 12001 |
| r2_p | 0.103 |
| chi2 | 1691.4 |
| Genre | Yes |

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

**Table 5**: Logistic regression with *Agency* as the dependent variable.

In Table 5, we report the estimates for equation (1) on which the propensity scores are estimated. The frequency distributions of the propensity scores for the treated and untreated e-books are presented in Figure 5. As the Figure shows, the frequency distributions of the propensity scores for the two groups of e-books are very similar, indicating that there is a good set of e-books sold under the wholesale model that can be matched with e-books sold under the agency model based on their characteristics.

Only a few e-book titles drop out of the sample after the matching process (indicated by the blue and yellow bars in Figure 5).

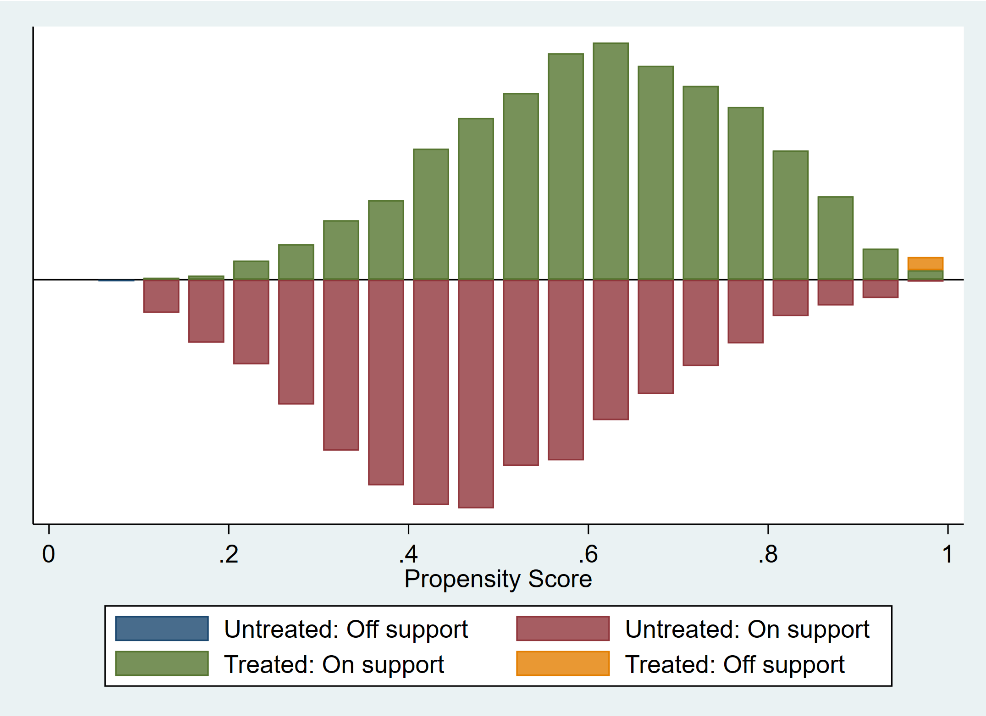


**Figure 5**: Propensity scores of a logistic model with Gaussian kernel fitting model. We have 11,953 observations which are on support. We present the covariate balance in Appendix Figure 3.

In a second step, we use the estimated propensity scores to produce a sample where only the matched pairs of e-books remain to estimate the average treatment effect (ATE) of the agency model on the retail price of e-books. The empirical model that we employ is reported in equation (3), where $p$ indicates two paired e-books (given by the matching procedure):

$$\Delta p_p = \alpha_0 + \alpha_1 A + \alpha_2 \Delta X_p + \epsilon_p. \quad (3)$$

Thus, $\Delta p_p$ is the difference in the retail price between a treated e-book and a non-treated (control) one. The treatment variable $A$ is equal to one if an e-book is sold under the agency model and zero otherwise. $\Delta X_p$ gives the differences in the e-book characteristics between the treatment and the control group.

Table 6 reports the average treatment effect (ATE) from equation (3). We apply four different matching methods (two nearest neighbour covariate matching procedures, kernel and caliper) to deal with the issue of possible non-randomness in the treatment. The agency model still has a significant and negative effect on the retail

| | (1) Neighbor5 | (2) Neighbor10 | (3) Kernel | (4) Caliper |
|---|---|---|---|---|
| Agency | -0.228*** | -0.223*** | -0.230*** | -0.224*** |
| | (-13.31) | (-14.85) | (-18.67) | (-15.09) |
| OnSupport | 11953 | 11953 | 11953 | 11953 |

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

**Table 6**: Propensity Score Matching. Standard Errors calculated by bootstrap with 101 random draws. The different models Neighbor5, Neighbor10, Kernel and Caliper refer to a 5-nearest neighbor, 10-nearest neighbor, a Gaussian kernel and a caliper matching, respectively. The caliper radius is determined by 0.2 times one standard deviation from the propensity scores derived from the logistic regression in Table 5 (see Austin, 2011, for optimal caliper length).

price of e-books. The effect ranges between 19.9% and 20.6% depending on the used matching procedure.[18]

# 5 Robustness Checks

To check the robustness of our results presented in the previous section, we apply two further estimation approaches. In Section 5.1, we use data on the paperback format to apply an alternative OLS estimation procedure. Following, we present a double machine learning (DML) approach in Section 5.2.

## 5.1 Format Estimation

The coefficient for the agency variable in our OLS estimation approach might be biased since we compare different book titles with each other. An ideal analysis would compare the same title offered to similar consumers under two separate vertical contracts. However, a specific e-book title is either sold under the agency or the wholesale model so that the counterfactual scenario is unknown. As an alternative, we run a robustness check in which we compare price differences between the digital and the paperback format for the same book title. Since print books are generally sold under wholesale contracts in the UK, this allows us to identify the price effect of the agency model on e-book prices (by also controlling for format differences).

To implement this robustness analysis, we only keep e-books sold under the agency model with an equivalent paperback version in our sample, which reduces our data set to 7,446 observations. Then, we estimate the following OLS model:

$$p_{i,j} = \alpha_0 + \alpha_1 A_{i,j} + \alpha_2 F_{i,j} + W\theta \ + \epsilon_{i,j}. \quad (4)$$

The dependent variable $p_{i,j}$ in equation (4) is the logarithm of the retail price for a book title $i$ in format $j$. Again, the treatment variable $A_{i,j}$ is a dummy variable with value one if a book is sold under the agency model and zero when the wholesale

[18] Even though this seems to be a relatively large effect at first glance, it is put into perspective when looking at an example in absolute terms. The median retail price of an e-book in our data set is £4.99 (see Table 2). Supposing that this median e-book is sold under the wholesale model (at £4.99), a negative agency effect of 20% in our estimation approach implies that the same e-books would be sold for £3.99 under the agency pricing model in a counterfactual scenario.

| | Dependent Variable: log Price | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Agency | −0.174*** | −0.261*** | −0.236*** | −0.166*** |
| | (0.017) | (0.021) | (0.021) | (0.018) |
| Paperback | 0.485*** | 0.439*** | 0.447*** | |
| | (0.014) | (0.015) | (0.015) | |
| log star rating | 0.219*** | 0.215*** | 0.182*** | 0.281*** |
| | (0.045) | (0.046) | (0.045) | (0.078) |
| No. expert reviews | 0.038*** | 0.032*** | 0.031*** | 0.039*** |
| | (0.005) | (0.006) | (0.006) | (0.007) |
| log RRP | 0.888*** | 0.931*** | 0.888*** | 0.839*** |
| | (0.014) | (0.013) | (0.014) | (0.021) |
| Date Retail | 0.004 | 0.005 | 0.003 | 0.014*** |
| | (0.004) | (0.004) | (0.004) | (0.005) |
| Bestsellers | −0.0002 | −0.001*** | −0.001** | −0.00002 |
| | (0.0002) | (0.0002) | (0.0002) | (0.0005) |
| Constant | −0.479*** | −0.398*** | −0.257*** | −0.462*** |
| | (0.058) | (0.037) | (0.059) | (0.101) |
| Publisher | No | Yes | Yes | No |
| Genre | Yes | No | Yes | Yes |
| Robust F Statistic | 497.1137 | 627.141 | 381.3814 | 181.6775 |
| Observations | 7,446 | 7,446 | 7,446 | 3,723 |
| $R^2$ | 0.607 | 0.609 | 0.623 | 0.482 |
| Adjusted $R^2$ | 0.606 | 0.609 | 0.622 | 0.479 |
| Residual Std. Error | 0.406 (df = 7424) | 0.405 (df = 7429) | 0.398 (df = 7415) | 0.503 (df = 3702) |
| F Statistic | 546.200*** (df = 21; 7424) | 724.590*** (df = 16; 7429) | 409.270*** (df = 30; 7415) | 172.310*** (df = 20; 3702) |

Robust Standard Errors in parentheses *p<0.1; **p<0.05; ***p<0.01

**Table 7**: Robustness check with paperback books (log-log OLS). Dependent variable is the logarithm of the retail price for digital and paperback versions of book titles sold on Amazon.

contract is used. The format fixed effect $F_{i,j}$ is equal to one for the paperpack version and zero for the digital one. All other covariates are captured by the matrix $W$.

The results of equation (4) are reported in Table 7 controlling for genre fixed effects (column (1)), publisher fixed effects (2), and both types of fixed effects (3). In column (4), we estimate equation (1) by only incorporating the e-book format so that we drop all the observations for the paperback format (see the number of observations and the empty cell for the paperback coefficient in column (4)). Comparing the estimation results to those reported in Table 6 we observe that the agency effect is very similar, even though it is partly smaller in amount here (between 15.3% and 23%). The estimated coefficients for the variable *Format* imply that the paperback versions are significantly more expensive than the digital ones, confirming our descriptive statistics in Section 3.2 (see Figure 1 and Table 2). The coefficients of the remaining covariates in Table 7 are very similar to our main estimations in Section 4.

## 5.2 Double Machine Learning (DML)

There are further methods beyond the established approaches in the standard econometric analysis for causal inference. We have already applied standard econometric methods as the OLS estimation and the propensity score matching to deal with econometric issues. Recent advances in machine learning approaches also offer a larger toolbox for empirical analyses in economics (e.g., Athey and Imbens, 2019; Athey, 2018, for a broad overview).

Due to recent developments in the machine learning literature there are many approaches, e.g. the DML technique, that gives the possibility to deal with common econometric issues as confounding variables or variable selection by using cross-validation and non-parametric models. This technique also allows to make use of non-standard modelling for relations between variables, like the independent variables have a specific, often assumed linear or quadratic effect on the dependent variable. It permits to use any arbitrary machine learning technique relying on algorithms to find a fitting model for some chosen score functions like the mean squared error.

The reason for relying on further non-parametric/semi-parametric regressions is to circumvent the imposition of a specific model structure. The algorithm of the machine learning models will choose the best fitting model under some restrictions or parametrization. We will then use regularized linear regression techniques like the least absolute shrinkage and selection operator (Lasso) or regression trees/forests. These methods help to compare our standard econometric approaches with models that are able to ignore irrelevant variables or include non-linear effects (Athey and Imbens, 2019). Moreover, this methods can help in a similar way like propensity score matching to deal with the underlying selection issue (e.g., Lee et al, 2010; Westreich et al, 2010; Knaus, 2021).

Therefore, we apply DML techniques and compare the results to our previous estimations to provide further robustness checks. From a prediction's perspective, the estimations will be split into three different regression models, which is proposed by Chernozhukov et al (2017, 2018). The models use the DML framework to deal with high dimensional variables, non-parametric functional forms, or unobserved confounding variables which have to be addressed by, e.g., an instrumental variable approach, based on the *DML-Conditional-Average-Treatment-Effect-Estimator*.[19]

The equation system we estimate has the following general form (we drop the index $i$ for each book), which stems from the partial linear regression model of Robinson (1988):

$$\begin{aligned} p &= \theta A + q(W) + \varepsilon \\ A &= f(W) + \eta \\ s.t.\ \mathbf{E}[\varepsilon|W] &= \mathbf{E}[\eta|W] = \mathbf{E}[\varepsilon \cdot \eta|W] = 0, \end{aligned} \tag{5}$$

where $p$ denotes the price of an e-book depending on the agency dummy variable $A$ and function $q(W)$ depending on the covariates $W$. The agency dummy variable $A$ is explained by a function $f(W)$ of these covariates $W$ (similar to the logit specification in equation (2)). The variables $\eta$ and $\varepsilon$ represent stochastic error terms. The conditional expectation functions then can be solved by some non-parametric regressions:

$$\begin{aligned} q(W) &= \mathbf{E}[p|W] \\ f(W) &= \mathbf{E}[A|W]. \end{aligned} \tag{6}$$

[19] For the estimation implementation we follow the Python Module *econml* provided by Battocchi et al (2019) (see also https://econml.azurewebsites.net/spec/estimation/dml.html#overview-of-formal-methodology).

In a next step, the residuals of the price, $\tilde{p}$, and the residuals of the agency dummy variable, $\tilde{A}$, are computed by subtracting the fitted values (given by the regression tasks in (6)) from the actual price $p$ and the actual agency dummy variable $A$:

$$\begin{aligned} \tilde{p} &= p - q(W) \\ \tilde{A} &= A - f(W) \end{aligned} \tag{7}$$

Finally, we use these residuals for the prices $\tilde{p}$ and the agency dummy variable $\tilde{A}$ to estimate a linear treatment effect $\theta$ that is unbiased based on the assumptions of Chernozhukov et al (2018):

$$\tilde{p} = \theta\tilde{A} + \epsilon. \tag{8}$$

In Table 8, the column *Model* represents the applied functional form. The first entry in this column refers to the functional form of computing and predicting $\tilde{p}$ and the second one for classifying $\tilde{A}$. Therefore, *Lin-Logit* relies on OLS and a logistic regression, *Lin-Lasso* relies on an OLS and a logistic regression including $L_1$ penalty (called Lasso), *Lasso-ElasticNet* uses a Lasso and a logistic regression with a combination of $L_1$ and $L_2$ penalties (*Elastic Net)*, *Lasso-RFC* refers to a Lasso regression and a random forest classifier, *RFR-ElasticNet* combines a random forest and an *Elastic Net*, *RFR-RFC* uses a random forest for both stages and XGBoost relates to Extreme Gradient Boost for both stages.[20]

The column *Score* in Table 8 displays the mean squared error of the final stage. In the final stage, a simple linear regression is used to get the conditional average treatment effect. The hyperparameters for each model are chosen from a reasonable set and then we use $3-5$ cross-fold validation within Python's Sklearn GridSearch. Besides, we also do another 5-fold splitting in each estimation. The presented results outline the best estimation (lowest score) for each model class.

| | Agency | Std. Error | p-value | Score | Perc. Change |
|---|---|---|---|---|---|
| Lin-Logit | -0.2138 | 0.0115 | 0.0000 | 0.3263 | -19.2519 |
| Lin-Lasso | -0.2138 | 0.0115 | 0.0000 | 0.3263 | -19.2485 |
| Lasso-ElasticNet | -0.2137 | 0.0115 | 0.0000 | 0.3264 | -19.2443 |
| Lasso-RFC | -0.2417 | 0.0132 | 0.0000 | 0.3266 | -21.4719 |
| RFR-ElasticNet | -0.1714 | 0.0094 | 0.0000 | 0.2160 | -15.7488 |
| RFR-RFC | -0.2057 | 0.0114 | 0.0000 | 0.2153 | -18.5962 |
| XGBoost | -0.1688 | 0.0119 | 0.0000 | 0.2314 | -15.5280 |

**Table 8**: Double Machine Learning Approach. The dependent variable is the e-book retail price and the treatment variable is *Agency*. The price effect of the agency arrangement for the respective model is given in the column *Agency* and represents the ATE. Column *Score* refers to the mean squared error. Note: The data has been mean centered with unit variance as Lasso requires this normalization.

[20] There are many more possible estimation techniques but these are sufficient to highlight the stability of our results.

The point estimates of the individual regression models are given in the column *Agency* and the relative percentage changes in relation to the intercept are presented in the column *Perc Change* of Table 8.[21] In overall, the DML techniques confirm the results of our main estimations presented in Section 4 and prove the robustness of our regressions, even if we rely on more flexible methods. For instance, e-books sold under the agency model are 18.6% cheaper (on average) than digital books sold under the wholesale model when using the regression model *RFR-RFC*.

# 6 Conclusion

In this paper, we provide evidence that e-books sold under the agency model on *Amazon.co.uk* are significantly cheaper than e-books sold under the wholesale model on average. Our results are based on an unique data set containing many characteristics of an e-book. To measure the relationship between the retail price of an e-book and the used pricing arrangement, we rely on classical econometric techniques as well as on newer methods as the DML approach. We find an robust and statistically significant effect that e-books sold under the agency model are approximately 20% cheaper than digital books sold under the wholesale model. The exact effect depends on the sales rank of a book, i.e. whether a book title is a bestselling or a *long tail* book.

The results of our empirical analysis are in line with many theoretical papers studying the price effect of the agency model. Those theoretical analyses argue that retail prices for e-books sold under the agency model are lower due to the elimination of double marginalization (Lu, 2017), a lock-in effect exploited by retailers (Johnson, 2020), the monopolistic power of retailers over a complementary device as it is the case for Amazon when e-books could only be read on a Kindle device (Gaudin and White, 2014) or because agency selling is just more efficient than the wholesale model and leads to lower retail prices (Abhishek et al, 2016). Our results also match to the model-theoretical analysis from Foros et al (2017) if one assumes that competition should be greater among publishers than among retailers, which most likely is the case due to the quasi-monopolistic power of Amazon.

To the best of our knowledge, this paper is the first empirical analysis estimating the price effect of the agency model for e-books by not only incorporating bestselling titles but also *long tail* book titles. Besides, in contrast to previous empirical analyses regarding the retail price of e-books, we apply an LDA approach to determine book genres. Nevertheless, a limitation of our approach is that we use cross-sectional data instead of panel data so that we cannot control for any dynamic effects on the retail price of e-books. Moreover, we only include one online platform, namely Amazon, instead of comparing various online retailers. Even though Amazon has a relatively high market share for e-books in the UK (see footnote 5), competition between the individual retailers very likely also has an effect on e-book prices.

The dynamic effect of the agency model on e-book prices including bestselling as well as *long tail* book titles remains an open question. Future research should concentrate on panel data to address such dynamic effects of different vertical contracts. Thereby, also other online platforms selling e-books should be included in such an

[21] Again, one can calculate the exact percentage change by using the formula $100 \times (e^{Agency} - 1)\%$.

analysis to identify effects between the online retailers. Finally, the long-run effect of the agency model on consumer welfare is an interesting research area. Consumer welfare does not only depend on the e-book prices, but also on other factors such as the number, variety, and quality of book titles written and published.

## Acknowledgement(s)

We would like to thank participants at the MaCCI Annual Conference 2021, EARIE Annual Conference 2021 and CRESSE Conference 2021. In particular, we thank Georg Götz, Daniel Herold, Jan Thomas Schäfer, Jona Stinner, Xiang Hui, Franco Mariuzzo and Matthew Olczak for helpful comments. The authors alone are responsible for the content.

## Declarations

**Funding:** No funding was received to assist with the preparation of this manuscript.
**Conflict of Interests:** The authors have no conflict of interests to declare that are relevant to the content of this article.
**Ethical Approval:** N/A.
**Informed Consent:** N/A.
**Authors' contributions:**
Maximilian Maurice Gail: Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualization, Writing – Review & Editing.
Phil-Adrian Klotz: Conceptualization, Methodology, Validation, Formal analysis, Writing – Original Draft, Supervision.
**Data Availability Statement:** Data will be made available on request.

## References


Abhishek V, Jerath K, Zhang ZJ (2016) Agency selling or reselling? channel structures in electronic retailing. Management Science 62(8):2259–2280

Aguiar L, Waldfogel J (2018) Quality predictability and the welfare benefits from new products: Evidence from the digitization of recorded music. Journal of Political Economy 126(2):492–524

Athey S (2018) The impact of machine learning on economics. In: The economics of artificial intelligence: An agenda. University of Chicago Press, p 507–547

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annual Review of Economics 11:685–725

Austin PC (2011) Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical statistics 10(2):150–161

Battocchi K, Dillon E, Hei M, et al (2019) EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. URL https://github.com/microsoft/EconML, version 0.12.0

Brynjolfsson E, Hu Y, Smith MD (2003) Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. Management science 49(11):1580–1596

Chernozhukov V, Chetverikov D, Demirer M, et al (2017) Double/debiased machine learning for treatment and causal parameters. Tech. rep., Massachusetts Institute of Technology

Chernozhukov V, Chetverikov D, Demirer M, et al (2018) Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. The Econometrics Journal 21(1)

Condorelli D, Galeotti A, Skreta V (2018) Selling through referrals. Journal of economics & management strategy 27(4):669–685

Davies S, Coles H, Olczak M, et al (2004) Benefits from competition: Some illustrative uk cases. DTI ECONOMICS PAPER NO 9

Dearnley J, Feather J (2002) The uk bookselling trade without resale price maintenance an overview of change 1995–2001. Publishing research quarterly 17(4):16–31

Fishwick F (2008) Book prices in the uk since the end of resale price maintenance. International journal of the economics of business 15(3):359–377

Foros Ø, Kind HJ, Shaffer G (2017) Apple's agency model and the role of most-favored-nation clauses. The RAND Journal of Economics 48(3):673–703

Gaudin G, White A (2014) On the antitrust economics of the electronic books industry. Available at SSRN 2352495

Gentzkow M, Kelly B, Taddy M (2019) Text as data. Journal of Economic Literature 57(3):535–74

Gilbert RJ (2015) E-books: A tale of digital disruption. Journal of Economic Perspectives 29(3):165–184

Ippolito PM (1991) Resale price maintenance: Empirical evidence from litigation. The journal of law and economics 34(2, Part 1):263–294

Johnson JP (2017) The agency model and mfn clauses. The Review of Economic Studies 84(3):1151–1185

Johnson JP (2020) The agency and wholesale models in electronic content markets. International Journal of Industrial Organization 69:102581

Knaus MC (2021) A double machine learning approach to estimate the effects of musical practice on student's skills. Journal of the Royal Statistical Society: Series A (Statistics in Society) 184(1):282–300

Larsen VH, Thorsrud LA (2019) The value of news for economic developments. Journal of Econometrics 210(1):203 – 218. https://doi.org/https://doi.org/10.1016/j.jeconom.2018.11.013, URL http://www.sciencedirect.com/science/article/pii/S0304407618302148

Lee BK, Lessler J, Stuart EA (2010) Improving propensity score weighting using machine learning. Statistics in medicine 29(3):337–346

Lenz D, Winker P (2020) Measuring the diffusion of innovations with paragraph vector topic models. PLOS ONE 15(1):1–18. https://doi.org/10.1371/journal.pone.0226685, URL https://doi.org/10.1371/journal.pone.0226685

Li H (2019) Intertemporal price discrimination with complementary products: E-books and e-readers. Management Science 65(6):2665–2694

Lu L (2017) A comparison of the wholesale model and the agency model in differentiated markets. Review of Industrial Organization 51:151–172

MacKay A, Smith DA (2017) Challenges for empirical research on rpm. Review of Industrial Organization 50(2):209–220

Mathewson GF, Winter RA (1984) An economic theory of vertical restraints. The RAND Journal of Economics pp 27–38

Nelson P (1970) Information and consumer behavior. Journal of political economy 78(2):311–329

Perry MK, Porter RH (1986) Resale Price Maintenance and Exclusive Territories in the Presence [of] Retail Service Externalities. Department of Economics, State University of New York at Stony Brook

Poort J, van Eijk N (2017) Digital fixation: the law and economics of a fixed e-book price. International Journal of Cultural Policy 23(4):464–481

Řehůřek R, Sojka P (2010) Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp 45–50

Reimers I, Waldfogel J (2021) Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings. American Economic Review 111(6):1944–71

Rey P, Stiglitz J (1988) Vertical restraints and producers' competition. Tech. rep., National Bureau of Economic Research

Rey P, Stiglitz J (1994) The role of exclusive territories in producers' competition. Tech. rep., National Bureau of Economic Research

Robinson PM (1988) Root-n-consistent semiparametric regression. Econometrica: Journal of the Econometric Society pp 931–954

Rosen S (1974) Hedonic prices and implicit markets: product differentiation in pure competition. Journal of political economy 82(1):34–55

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rubin DB (1977) Assignment to treatment group on the basis of a covariate. Journal of educational Statistics 2(1):1–26

De los Santos B, Wildenbeest MR (2017) E-book pricing and vertical restraints. Quantitative Marketing and Economics 15(2):85–122

De los Santos B, O'Brien DP, Wildenbeest MR (2018) Agency pricing and bargaining: Evidence from the e-book market. Kelley School of Business Research Paper pp 1–51

Spengler JJ (1950) Vertical integration and antitrust policy. Journal of political economy 58(4):347–352

Steyvers M, Griffiths T (2007) Probabilistic topic models. Handbook of latent semantic analysis 427(7):424–440

Thorndike RL (1953) Who belongs in the family? Psychometrika 18(4):267–276

Thorsrud LA (2020) Words are the new numbers: A newsy coincident index of the business cycle. Journal of Business & Economic Statistics 38(2):393–409

Tirole J (1988) The theory of industrial organization. MIT press

Varian HR (2014) Big data: New tricks for econometrics. Journal of Economic Perspectives 28(2):3–28

Wang X, Yucesoy B, Varol O, et al (2019) Success in books: predicting book sales before publication. EPJ Data Science 8(1):31

Westreich D, Lessler J, Funk MJ (2010) Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. Journal of clinical epidemiology 63(8):826–833

Winter RA (1993) Vertical control and price versus nonprice competition. The Quarterly Journal of Economics 108(1):61–76

# A Additional figures and tables

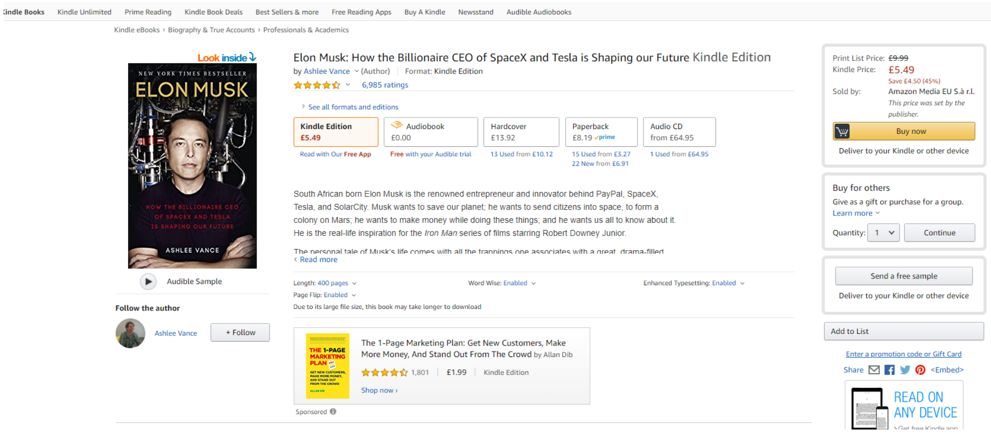


**Figure 1**: Screenshot of *Elon Musk: How the Billionaire CEO of SpaceX and Tesla is Shaping our Future* (*Amazon.co.uk*).

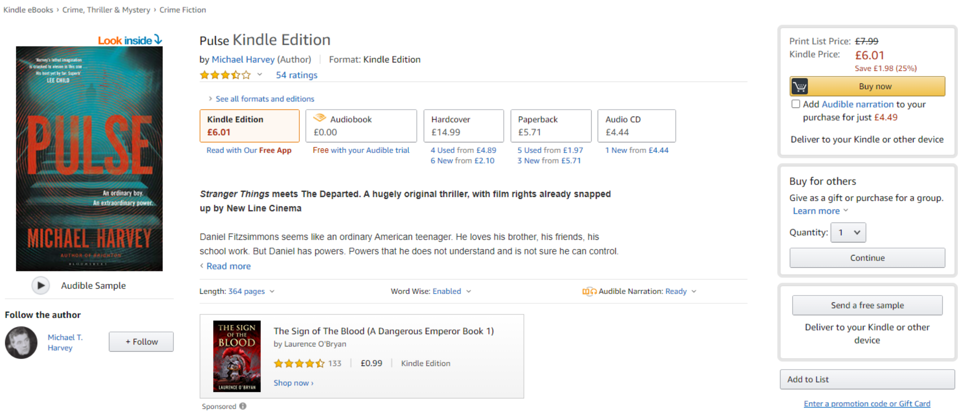


**Figure 2**: Screenshot of *Pulse* (*Amazon.co.uk*).

| | Publisher | Bloomsbury | Faber | Hachette | HarperCollins | Small Pub. | Oxford | Pan Macmillan | Penguin Random House | Scholastic | Simon & Schuster |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Retail Price | mean | 10.37 | 5.84 | 5.31 | 5.42 | 5.80 | 16.71 | 7.79 | 6.56 | 5.01 | 7.45 |
| | std | 7.02 | 3.21 | 2.86 | 3.20 | 6.07 | 10.12 | 3.86 | 2.70 | 1.58 | 3.27 |
| Sales Rank | mean | 554,409.87 | 326,982.37 | 262,402.04 | 427,905.75 | 778,429.39 | 937,724.43 | 497,575.36 | 365,604.06 | 597,692.75 | 555,114.93 |
| | std | 594,258.21 | 432,500.62 | 383,524.36 | 514,257.98 | 728,316.54 | 653,612.74 | 551,601.82 | 533,573.80 | 655,036.93 | 649,582.93 |
| Star Rating | mean | 0.90 | 0.85 | 0.89 | 0.89 | 0.86 | 0.90 | 0.89 | 0.88 | 0.93 | 0.90 |
| | std | 0.10 | 0.13 | 0.08 | 0.09 | 0.12 | 0.10 | 0.10 | 0.08 | 0.07 | 0.08 |
| No. Customer Reviews | mean | 51.10 | 60.80 | 119.85 | 106.88 | 80.50 | 17.19 | 124.70 | 134.59 | 76.33 | 142.89 |
| | std | 110.04 | 113.83 | 178.01 | 168.94 | 161.48 | 42.95 | 210.71 | 195.05 | 138.88 | 210.61 |
| Pages | mean | 292.57 | 280.60 | 388.78 | 342.59 | 286.86 | 412.85 | 336.34 | 319.01 | 231.84 | 367.91 |
| | std | 124.35 | 191.93 | 1949.17 | 330.58 | 212.83 | 207.28 | 121.63 | 140.63 | 113.11 | 256.70 |
| Kindle Size | mean | 13,078.55 | 2,849.00 | 14,134.92 | 8,288.26 | 8,626.52 | 9,995.83 | 12,118.73 | 20,466.37 | 35,684.97 | 15,118.97 |
| | std | 27,346.77 | 10,529.46 | 46,813.98 | 28,600.24 | 30,096.99 | 16,416.27 | 37,011.08 | 48,498.42 | 42,097.93 | 27,885.95 |
| RRP | mean | 15.50 | 10.28 | 13.07 | 12.04 | 12.08 | 35.29 | 13.61 | 14.59 | 8.31 | 13.85 |
| | std | 9.76 | 4.88 | 5.32 | 5.90 | 10.22 | 30.15 | 5.68 | 5.91 | 3.15 | 5.29 |
| Date Retail | mean | 1.64 | 2.41 | 1.81 | 2.09 | 1.93 | 2.97 | 1.66 | 2.18 | 2.11 | 2.13 |
| | std | 1.76 | 2.62 | 2.06 | 2.36 | 1.91 | 1.81 | 1.78 | 2.25 | 2.11 | 2.29 |
| No. Expert Reviews | mean | 1.74 | 2.43 | 2.33 | 1.25 | 1.34 | 1.14 | 2.24 | 1.71 | 0.65 | 1.35 |
| | std | 0.96 | 1.24 | 1.06 | 1.12 | 1.51 | 0.77 | 1.57 | 1.40 | 0.85 | 1.04 |

**Table 1**: Summary Statistics

| | Agency == 0 | Agency == 1 | t-stat | p-value |
|---|---|---|---|---|
| Price | 7.447 | 5.162 | 23.207 | 0.000 |
| log Price | 1.644 | 1.346 | 17.882 | 0.000 |
| Series | 0.101 | 0.081 | 3.677 | 0.000 |
| Kindle Size | 10,409.946 | 12,322.658 | -2.958 | 0.003 |
| Star Rating | 0.876 | 0.875 | 0.393 | 0.694 |
| No. customer reviews | 70.036 | 120.327 | -16.507 | 0.000 |
| Np. expert reviews | 1.227 | 1.905 | -27.624 | 0.000 |
| Amazon Rank | 726,015.678 | 440,894.733 | 23.921 | 0.000 |
| Date Retail | 1.937 | 1.984 | -1.262 | 0.207 |
| RRP | 14.075 | 12.108 | 10.913 | 0.000 |
| Bestsellers | 0.518 | 2.260 | -8.632 | 0.000 |
| WeekInChart | 0.478 | 0.783 | -6.906 | 0.000 |
| SReader | 0.792 | 0.791 | 0.185 | 0.853 |
| TtS | 0.959 | 0.960 | -0.095 | 0.924 |
| XR | 0.189 | 0.297 | -13.964 | 0.000 |

**Table 2**: t-Tests

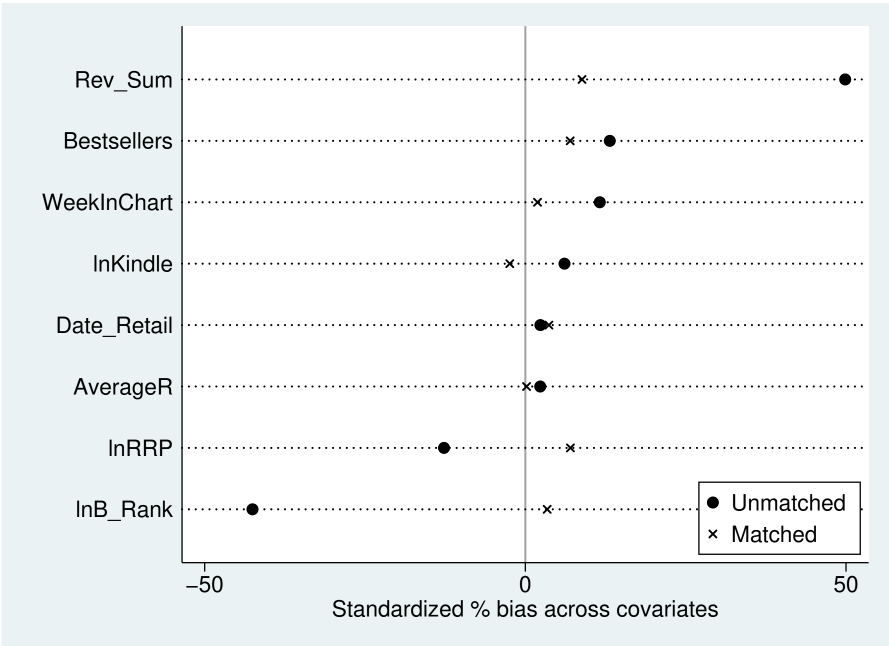


**Figure 3**: Covariate Balance.

# B Latent Dirichlet Allocation (LDA)

In the recent past, new technologies have made it possible to use text as data and, therefore, as an input to economic research. Text data, which is inherently high-dimensional, can capture relevant economic concepts not covered by "hard" economic data. In the last years, there has been an explosion of empirical economics research using text as data (e.g., see Larsen and Thorsrud (2019) for an Latent Dirichlet Allocation (LDA) approach or Lenz and Winker (2020) for paragraph vector topic modelling). We have decided to use an LDA approach to generate book genres and to assign every single book title from our data set into one of these genres. Such a text mining approach is necessary because on the Amazon webpage the genre information is ambiguous and even not available for some book titles. For this purpose, we use the descriptions and expert reviews from the individual books in our data set as text data input. We further rely on natural language processing (NLP) to extract the relevant information.

We apply several Python-Modules to clean and prepare the raw data set.[22] Thereby, we remove common words and surnames, eliminate stop words, remove punctuation and pronouns as well as reduce all words to their respective word stems. We note here that around 45,819 unique tokens are kept after this filtering process.

This cleaned descriptions corpus is decomposed into book genres using the already mentioned LDA model. The LDA provides a statistical framework for the generation of documents based on topics. It is an unsupervised topic model that clusters words into topics/ genres, which are distributions over words, while at the same time classifying descriptions as mixtures of topics/ genres. The term "latent" is used because the words are intended to communicate a latent structure, namely, the subject matter (topic) of the description. The term "Dirichlet" is used because the topic mixture is drawn from a conjugate Dirichlet prior (Thorsrud, 2020).

The structure of the LDA model is as follows: the whole corpus is represented by $M$ distinct documents (descriptions) and $N = \sum_{m=1}^{M} N_m$ is the total number of words in all documents. Assuming $K$ latent topics/ genres, each topic is given by a probability vector $\boldsymbol{\phi_k} = (\phi_{k,1}, ..., \phi_{k,N})$ with $\sum_{n=1}^{N} \phi_{k,n} = 1$ indicating the probability that each word shows up in this topic. Further, each document $m \in \{1, ..., M\}$ contains all topics with different probabilities (weights) $\boldsymbol{\theta_m} = (\theta_{m,1}, ..., \theta_{m,K})$ with $\sum_{k=1}^{K} \theta_{m,k} = 1$. Both $\boldsymbol{\phi_k}$ and $\boldsymbol{\theta_m}$ are assumed to have conjugate Dirichlet distributions with hyper parameters (vectors) $\alpha$ and $\beta$, respectively.

Given $\boldsymbol{\phi_k}$ and $\boldsymbol{\theta_m}$, a document is generated by drawing for each word a topic $k \in \{1, ..., K\}$ according to the probabilities $\boldsymbol{\theta_m}$ and one word from the selected topic according to its distribution $\boldsymbol{\phi_k}$. This procedure is repeated until the length of the document is reached. To solve the LDA model, we a priori set $\alpha = 50$ and $\beta = 0.01$. The hyper parameter optimization is executed by using Gibbs simulations. Gibbs sampling (also known as alternating conditional sampling) is a specific form of Markov chain Monte Carlo and simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of all others (e.g., Steyvers and Griffiths, 2007).

[22] Base module is gensim by Řehůřek and Sojka (2010) which is a open-source NLP text analytics tool.

| | Dependent Variable: log Price | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Agency | −0.204*** | −0.210*** | −0.210*** | −0.210*** | −0.202*** |
| | (0.011) | (0.011) | (0.011) | (0.011) | (0.011) |
| log sales rank | 0.056*** | 0.055*** | 0.060*** | 0.059*** | 0.051*** |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| log Kindle Size | 0.045*** | 0.045*** | 0.050*** | 0.050*** | 0.027*** |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| log star rating | 0.502*** | 0.492*** | 0.491*** | 0.484*** | 0.436*** |
| | (0.048) | (0.049) | (0.048) | (0.048) | (0.048) |
| No. expert reviews | 0.013*** | 0.014*** | 0.015*** | 0.015*** | 0.018*** |
| | (0.004) | (0.004) | (0.004) | (0.004) | (0.004) |
| log RRP | 1.090*** | 1.094*** | 1.076*** | 1.071*** | 1.080*** |
| | (0.011) | (0.011) | (0.011) | (0.012) | (0.012) |
| Date Retail | 0.011*** | 0.011*** | 0.013*** | 0.012*** | 0.008*** |
| | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Bestsellers | 0.001*** | 0.001*** | 0.001*** | 0.001*** | 0.001*** |
| | (0.0004) | (0.0004) | (0.0004) | (0.0004) | (0.0004) |
| WeekInChart | 0.007*** | 0.007*** | 0.007*** | 0.007*** | 0.007*** |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Constant | −1.987*** | −1.960*** | −1.970*** | −2.204*** | −1.745*** |
| | (0.074) | (0.074) | (0.075) | (0.084) | (0.074) |
| No. Genre Topics | 9 | 10 | 12 | 15 | Amazon Genres |
| Observations | 12,001 | 12,001 | 12,001 | 12,001 | 12,001 |
| Adjusted $R^2$ | 0.608 | 0.608 | 0.610 | 0.610 | 0.615 |

Robust Standard Errors in parentheses *p<0.1; **p<0.05; ***p<0.01

**Table 3**: OLS estimation approach for different genres generated by LDA or given by Amazon. Dependent variable is the logarithm of the retail price for e-books sold on Amazon.

The sampling is done sequentially and proceeds until the sampled values approximate the target distribution. We set the number of sampling iterations equal to 1,000. Then, based on different coherence values across the estimated LDA models using smaller numbers of genres, we find that 15 topics/genres provide an optimal statistical decomposition of our book description corpus and give a comprehensible set of topics that can be attributed to officially available genres.[23] A detailed list of all 15 genres is presented in Table 4.

A caveat of the LDA estimation procedure is that it does not give the topics/genres any names or labels. Thus, labels are subjectively given to each genre based on the most important words associated with each topic. In the most cases, it is conceptually simple to classify the genres. Besides, the exact labeling plays no material role in our empirical approach, it is just used as a convenient way of referring to the different topics instead of only using topic numbers.

[23] For 15 different genres, the coherence value exhibits a local optimum when the elbow method is applied (Thorndike, 1953). We have also considered 9, 10 and 12 different genres, but in the end the agency effect remains quite robust when controlling for a different number of genres (see Table 3).

| Topic | Genre |
|---|---|
| 0 | Philosophy |
| 1 | Textbooks |
| 2 | Childen and Youth |
| 3 | History |
| 4 | Guidebook |
| 5 | General Fiction |
| 6 | Fantasy Fiction |
| 7 | Personal development |
| 8 | Cultural |
| 9 | Self-Guidance |
| 10 | Thriller |
| 11 | Historic Novel |
| 12 | Social Novel |
| 13 | Crime Novel |
| 14 | Politics |

**Table 4**: 15 different genres identified by our LDA approach.

It is more important that the LDA decomposition gives a meaningful and easily interpretable genre classification of the book descriptions, which it does because our LDA approach identifies all important book genres and clearly delineates the topics. This is shown by the Figures 4, 5 and 6, which illustrate three examples of our 15 word clouds to visualize the genre distribution of words by assigned probabilities through the LDA. We have labelled the topic in Figure 4 *Crime Novel*, in Figure 5 *Thriller* and the genre in Figure 6 *Politics*. The larger the size of a word in these clouds is, the higher is its weight within the respective topic.

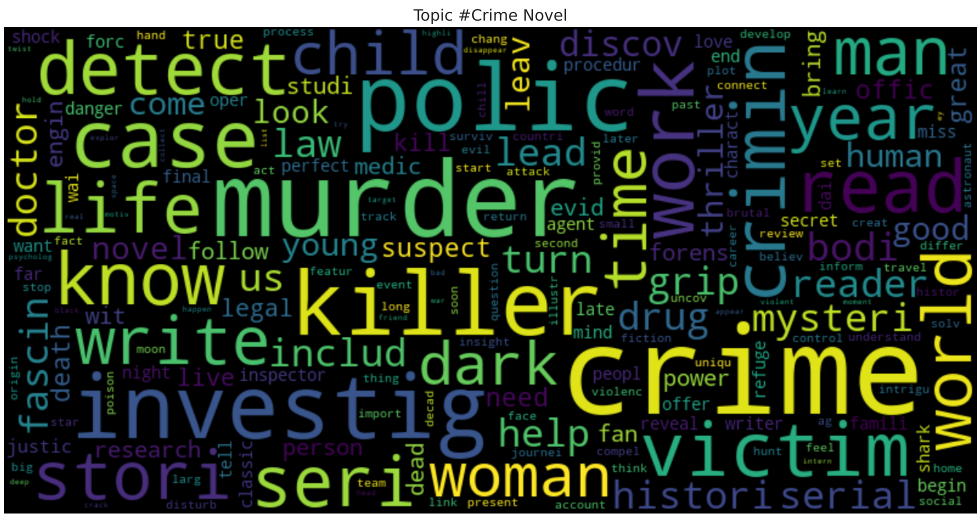


**Figure 4**: Topic Crime Novel from the LDA. Size is according to weight of word within the topic.

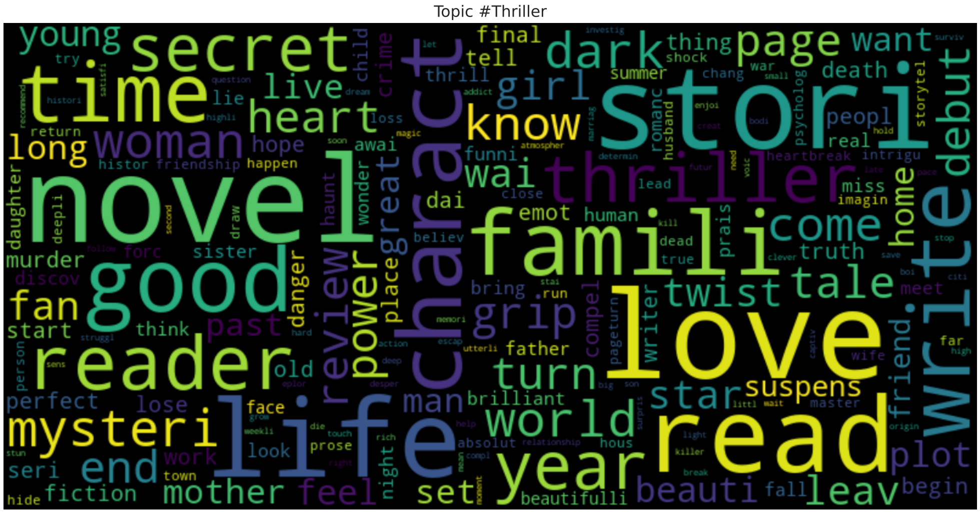


Figure 5: Topic Thriller from the LDA. Size is according to weight of word within the topic.

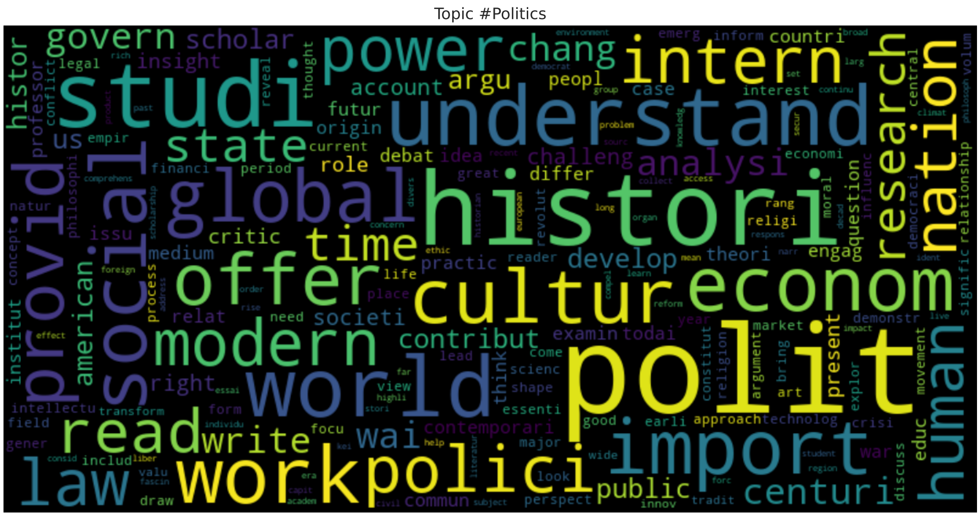


Figure 6: Topic Politics from the LDA. Size is according to weight of word within the topic.

The LDA provides the ability to give each document (book description) from our corpus (data) a probability of being a respective topic. For the visualization purpose in Figure 7, which is based on Figure 2, the largest probability value is chosen to highlight the distribution of topics over e-book prices and publishers. This distribution of prices by genres exhibits the high comparability between the several publishers in our data set because they are not specialised in certain topics, but all publishers sell book titles from different genres. This is an advantage for our empirical approach, since otherwise we would get multicollinearity issues and from an economic point of view we would fail in the sense that we could not compare publishers at all, if there were only specific topics from specific publishers.

However, it is obvious that the individual publishers have distinct main topics. For instance, Pan Macmillan primarily publishes fiction titles like crime novels, thrillers or society novels whereas HarperCollins has a focus on the genres family novel and drama. However, it is important not to take these topics at face value since the LDA assigns a probability to each individual topic.

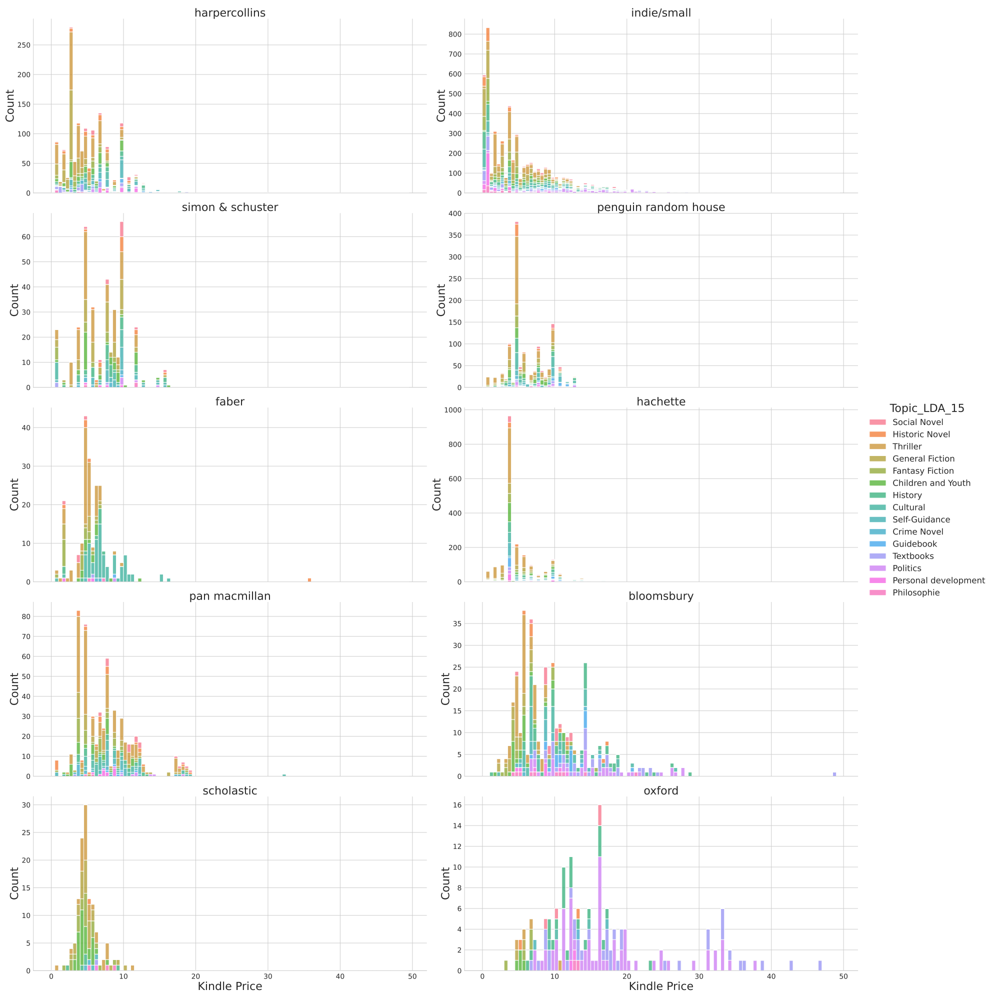


**Figure 7**: Prices for e-books grouped by publisher and genre. The ordinate is scaled differently for each subplot.

# C IV Estimation

The results of our OLS estimation in Section 4.2 might be biased and inconsistent since the important control variable book sales rank might be endogenous. Hence, we will follow an instrumental variable approach in this section to resolve this potential source of endogeneity. We use the logarithmized number of customer reviews (*log no. customer reviews*) as an instrument for the book sales rank to avoid inconsistent estimates due to reverse causality.

Our instrumental variable *log no. customer reviews* is highly correlated with our endogenous regressor book sales rank (relevance condition) but should have no partial effect on the price of an e-book (orthogonality assumption). Customer reviews can enhance the awareness and information quality for a consumer and, thus, change the tendency for a consumer to purchase a book. However, the absolute number of customer reviews does not affect the purchasing decision of a consumer for a book title, but only surprisingly positive (negative) reviews can increase (decrease) the consumption of a given good (Reimers and Waldfogel, 2021). Hence, the absolute number of customer reviews should also have no partial effect on e-book prices, even though our instrument is highly correlated with the book sales rank (as it is an indicator for past sales).

Following the approach explained above, the linear projection in the first stage regression of our 2SLS estimation can be formalized as follows:

$$R_i = \beta_0 + \beta_1 A_i + \beta_2 P_i + \beta_3 G_i + \beta_4 P_i \times G_i + \beta_5 RRP_i + \beta_6 CR_i + W\theta + \xi_i. \qquad (9)$$

In equation (9), the dependent variable $R_i$ refers to the sales rank of book title $i$ on *Amazon.co.uk*. We have already introduced most of the covariates used here in the context of our baseline estimation in equation (1). Our instrumental variable *log no. customer reviews* is displayed by $CR_i$.

The structural equation of our basic model then takes the following form:

$$p_i = \gamma_0 + \gamma_1 A_i + \gamma_2 P_i + \gamma_3 G_i + \gamma_4 P_i \times G_i + \gamma_5 RRP_i + \gamma_6 \hat{R}_i + W\theta + \varepsilon_i, \qquad (10)$$

where the dependent variable $p_i$ is the retail price of e-book $i$ and the fitted values from the first-stage are captured by $\hat{R}_i$.

The regression results based on equation (10) are presented in columns (2) and (3) of Table 5. In column (1), we again show the results of a (naive) OLS estimation when we control for the book genres. The two IV approaches in Table 5 differ in terms of the added genre control variable (LDA generated or Amazon book genres). The results of our IV approach confirm that e-books sold under the agency model on *Amazon.co.uk* are on average significantly cheaper than book titles sold under the wholesale model. Compared to the OLS estimation result in column (1), the estimated coefficients for the variable *Agency* only differ in their magnitude as we find a negative effect of agency pricing between 17.4 and 18% now. Hence, we slightly overestimate the absolute effect of agency pricing when we ignore the endogeneity issue of the variable book sales rank. Also the effects of the other explanatory variables barely differ between the OLS and the IV estimation approaches, even though the effect of the book sales rank has become larger in the IV regressions.

| | Dependent Variable: log Price | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| log sales rank | 0.059*** | 0.090*** | 0.076*** |
| | (0.004) | (0.008) | (0.007) |
| Agency | −0.210*** | −0.198*** | −0.191*** |
| | (0.011) | (0.012) | (0.012) |
| log Kindle Size | 0.050*** | 0.049*** | 0.026*** |
| | (0.004) | (0.004) | (0.004) |
| log star rating | 0.484*** | 0.505*** | 0.456*** |
| | (0.048) | (0.048) | (0.048) |
| No. expert reviews | 0.015*** | 0.021*** | 0.023*** |
| | (0.004) | (0.005) | (0.004) |
| log RRP | 1.071*** | 1.073*** | 1.082*** |
| | (0.012) | (0.012) | (0.012) |
| Date Retail | 0.012*** | 0.011*** | 0.008*** |
| | (0.003) | (0.003) | (0.003) |
| Bestsellers | 0.001*** | 0.002*** | 0.002*** |
| | (0.0004) | (0.0004) | (0.0004) |
| WeekInChart | 0.007*** | 0.009*** | 0.009*** |
| | (0.002) | (0.003) | (0.002) |
| Constant | −2.204*** | −2.612*** | −2.076*** |
| | (0.084) | (0.123) | (0.111) |
| Genre | Yes | LDA | Amazon Genres |
| Instrument | No | Reviews | Reviews |
| First Stage F-Statistic | No | 3663.8841 | 3915.4707 |
| Wu-Hausman Statistic | No | 23.179 | 17.0578 |
| Observations | 12,001 | 12,001 | 12,001 |
| Adjusted $R^2$ | 0.610 | 0.608 | 0.614 |

Robust Standard Errors in parentheses *p<0.1; **p<0.05; ***p<0.01

**Table 5**: IV estimation results.