The effect of macro variables on textbook downloads

Analyzing 5 months of downloads from LibGen

Zoltán Puha

Abstract

This thesis analyses download data obtained from one of the biggest online shadow library, LibGen. I only include textbooks in the analysed data. I research, whether price of the textbook or macro variables of different countries have effect on the number of downloads. I use regression analysis, using both OLS and hurdle regressions. I find that prices either have no or negative effect on downloads. A higher GDP, internet penetration or number of enrolled students into tertiary education all significantly increase the chance of downloads.

Contents

1 Introduction

Unauthorized downloading of copyrighted works is a common and well-known phenomenon in today's Internet society. The broad downloading of different type of goods started at the end of the 20th century, but the sudden growth of Internet led to its everyday usage. Today software, songs, films, books and nearly anything what exists online can be downloaded from different sites.

While the piracy of software, music and films are relatively well-studied (Gopal et al. (2004), Rodman and Vanderdonckt (2006), Koklic, Kukar-Kinney, and Vida (2014)), the research regarding (scientific-)book downloading is quite scarce. In this thesis, my aim is to look at the driving forces and behavior behind downloading books illegally from a Russian shadow-library, Library Genesis, especially analyzing the behavior of University students. I distinguish university students by only analyzing books that have rent prices. These books are textbooks used by university students, as rent price is only available for textbooks. The analysis is different from other paper's approach, as it is based on observed data from downloads and not survey data.

I use a rich dataset, with more than 1 million downloaded books for a time span of 5 and a half months. This allows me to get robust results and to draw conclusions about the real users of these websites. I believe this area is really interesting and should be understood more deeply, as the explanation behind why students are downloading textbook is not clear. It also poses heavy copyright issues and publishers realize significant losses due to piracy. On the other hand, some argue, that knowledge should be free and scientific papers or books should be made available free of charge.[1]

In this thesis, I am looking at a specified part of the book industry: scientific books and textbooks, for which the demand is the highest among University students. I have chosen this subtopic because as a current Master's student, I would like to analyze my fellow students. Downloading from file sharing sites is a common thing among students, as one of their primary source for knowledge are the expensive textbooks. Students are really sensitive for prices, as the cost of University is high without the books as well (Heller (1997)). Moreover, students have the free time and willingness to find these sites and often all of heir sources of entertainment, series, films or music are downloaded illegally.

---

[1]https://www.torproject.org/

With the help of this thesis, I can give insights to the driving forces of graduate downloaders and shed light on different interesting questions. I will first take a look at the descriptive statistics of the data and check, if there are any discrepancies in it. After that, I will use OLS regressions and correlations to see if the price has an effect on the number of downloads. After analyzing the prices, I will use a hurdle regression, to see how macro variables or the topic of the book effects the number of downloads.

The thesis have two main questions: First, whether the price of a book has an effect on the number of downloads. Second, I investigate if a difference can be seen between different countries based on their region, GDP or growth.

The setup of the thesis follows like this: In Part 2 I will describe how piracy is defined, take a short look at its history and summarize some findings from recent papers in the topic of illegal Internet downloading. In Part 3 I will describe the available dataset and the website it was acquired from, present the data's possible shortcomings and differences from previously used datasets. I present basic exploratory data analysis to have a good overview of the dataset. In Part 4 I have a look at some control OLS and hurdle regressions, trying to understand the possible explanation behind downloading. Part 5 discusses the thesis and gives recommendations for further development.

2 Literature review

2.1 Piracy

Piracy is a part of human societies since the origins of civilization (Johns (2010)) as people were copying intellectual content from other's without their permission. The meaning of piracy is still controversial in academic research. For this reason, there is no date in the past, for what scholars agree as the start of piracy.

There are few definitions of piracy circulating in the academic field, but according to Karaganis (2011), piracy is "ubiquitous, increasingly digital practices of copying that fall outside the boundaries of copyright law". We can see that piracy is an illegal act, where somebody copies or steals an intellectual property of somebody else.

People can have three driving forces to pirate content. First, it is a utility maximizing behavior, as the price of the pirated content is zero and the probability of getting

caught is very small. Thus people have no fear of the consequences of downloading and avoid paying a price for copyrighted content. The second, is the commercial war, where people make money from making pirate content available. They exploit the weakness of the system (too high price, too little supply). The third one is when countries do not enforce the law, and let the piracy flourish. A good example for this is China, where counterfeited phones or books without copyrights exist nationwide, despite the laws of the Chinese government. This behavior helps them catch up to more developed countries.

These three forces can be interpreted as a development vehicle on individual, company and nation-wide level, respectively. In this thesis, I am analyzing both the individual and nation-wide level.

2.2   Piracy of books

The start of book trading gave birth to copyright and developed concepts relevant today as well. The piracy of the books started with Gutenberg's revolution (Johns (2010)), as people were able to copy books within a reasonable time, without the author's permission - most of the time, without the knowledge of the original writer. During the (Pre-)Modern Era, the copyrights of the books were only valid for a city or a little area and other printers could freely print books from other cities. The owners of the copyright's of these books realised this issue, and started to establish connections through cities and ask help from their state to protect their rights. This co-operation lead to the exclusive right to print and sell books, limited in time, space and scope.

One institution enforcing these copyrights was "The Stationers' register" in London, which contained printing right of books' (today a patent register has the same aim). This process was close to law and politics in an era, when people were trying to get independent from the government, the topic became a well-known one in the whole society. Also, for authors and readers, this system was not beneficiary and pirated books started to circulate in the book shops.

Pirated texts in this time were not only competitors on their own market (Bodó (2011)), but foreign printed copies entered other markets as well. Pirates were always there to exploit the shortcomings of the law and provide a cheaper, yet legal way to fulfill the demand for books. Moreover, small publishers started to surface who were really

effective on the market: they were selling these books for a lower price, not taking the regulations of the government into account and believed in mass-printing: profiting little on one book, but selling lots of books.

A law-battle with the English monarch, led by a writer named Richard Atkyns (Johns (1998)) led later to the shutting down of the absolutistic book market and a more democratic came into its place, redefining piracy. The heart of this revolution was London, but it quickly span across Western Europe and led to a more free market. Up until 1886, there was no inter-country agreement on copyright, when the Berne Convention[2] was signed (Bodó (2011)).

2.2.1   The Russian way and the birth of LibGen

From this point on, pirates had a harder task to distribute their copies, however in some countries i.e. Russia, this treaty was not in effect, so piracy was more alive there than ever before. The demand for books was skyrocketing in The Soviet Union, because this was the primary tool for entertainment. On the other hand, the Soviet regime applied a very strict censorship which lead to a shortage of available books. As the USSR had not entered any treaties, people were also "free" to distribute translations of books - however, these were mainly released into the shadow market, as the censorship banned these books. With the advance of technology, other methods entered the pirates' tools, such as Xerox machines and CD-ROMs .

As the Internet started to establish itself, these shadow libraries, floating around on CDs were uploaded and gathered into online databases, where everyone was available to download them. One of the biggest site storing these books was Gigapedia, later called library.nu. This sites' aim was to gather all the offline, but digitally available books and organize them into a giant, online library. Library.nu closed down after a successful injection by several publishers, however its library was merged into another page, Library Genesis, also known as LibGen. LibGen's catalogue is mirrored to more sites, who can easily add their own collection to the existing one. The mission of the site is to:

- collect valuable academic literature

---

[2]http://www.wipo.int/treaties/en/text.jsp?file_id=283698

- build and maintain a community, who helps the inflow of the books and who can improve the quality of the uploaded documents
- make this service available for free.

LibGen's collection was Russian by default, but after the merge of library.nu, English books became dominant. (Bodó (2014))

2.3 Piracy of software, music, films and books

Piracy on the Internet started in the 90's with illegal downloading of software. In that era, the prices of software were high, but a lot of people wanted to use them - for free. Christensen and Eining (Christensen and Eining (1991)) in an early study, asked university students about their knowledge about piracy laws. They found, that the majority of the students were using pirated content, even though they were aware of the laws. They stated, that they don't think, that the law would be enforced against them. Givon, Mahajan, and Muller (1995) also investigated the software piracy, but with an other approach. They used a diffusion model to track the transition from pirated content to legal copies. Although they found, that over 90% of the users utilized illegal software, they generated more than 80% of the profit of new software. They argued, that software piracy is not necessarily bad, as the shadow diffusion created the base of the customers.

The next industry where piracy became relevant was music. Bhattacharjee, Gopal, and Sanders (2003) call illegal music downloading as the 2.0 of software piracy. The authors in this study also used a survey to look at downloaders attitude towards online music downloading. They found that price and bandwidth had a significant effect on the choice to utilize piracy. They also suggest, that the well-known music has more downloads. Today, with the presence of Spotify or Tidal it is also interesting to see, how they suggested that music providers should switch to subscription based services, as the respondents were positive about that. In a more recent study, Podoshen (2008) explores the relation of numerous effects to student's download decision. Podoshen also chose the survey approach, where he found, that avoiding payment is one of the key factors that students are downloading music. The survey data also revealed, that students were not afraid of the consequences of downloading, just like in the case of Christensen's paper.

The third big industry affected by piracy was the film industry. As films are bigger files compared to songs, the introduction of online movie piracy only came with the upcoming of peer-to-peer (P2P) networks (Danaher and Waldfogel (2012)). In his paper Fetscherin (2005) introduces a model, which shows, why do people choose to download films. The model gives evidence, that people download because of the low probability of being caught, while the users can reach very high quality products. Fetscherin also shows, that the perceived value of the films play an important role when people decide on downloading. In their paper, Bodó and Lakatos (2012) investigate the case of Hungary with movie downloads. They take a different approach and track the traffic of three P2P networks. They found, that the biggest shaping factor of download choices is the failure of the markets to supply the demand. According to them , people download films because they do not find enough movies in traditional channels and this way they are forced to download.

The fourth industry where pirates became a real concern were books. Hoorebeek (2003) shows, that the option to download books has been available since the early 2000's. Scanned versions were circulating on the web, on sites nearly identical to Napster. However, Rohde and others (2001) argues that the market for e-readers was not well established in that time and thus book downloading had no significant effect on the industry. The uprising of e-book piracy, came with the new versions of handheld devices (Kindle, and later tablets) capable of displaying books in a user-friendly format.

Also, e-book piracy became really important in the academic field (Zimerman (2011)), as students create a continuous demand for textbooks and scholarly articles. In his paper, Zimerman provides evidence, that the e-book piracy in the field of academia is clearly because of the low availability of books and articles.

2.4 Textbook piracy

Textbooks are always wanted products by University students, as they are one of the primary sources of knowledge. Also, after the education boom after the II. World War, in developed countries, like Western Europe and USA, a significantly bigger upturn in education started, with countries from Asia and Eastern Europe.

Several websites or groups across the world were formed to transfer second-hand textbooks. The response for this from the bookmakers is the constant updating of the

books - they make a new version every year, so students are forced to always buy the new ones. With the emerge of e-readers, the downloading of textbooks became more popular.

Rebelo (2015) was looking at this effect in a recent paper, where she looked at survey data from a Portuguese University and found, that the price of the book does not play a significant role, whether a book will be downloaded or not. In the study, she also found, that the downloading of the book is connected to the perceived usefulness of a book - books, that are considered to be more useful, are less likely to be downloaded.

Another study, by Scorcu and Vici (2012) also researches the illegally obtained books. They concentrate on the individual and social characteristics of downloaders through a survey conducted in Italy and find that males are more likely to download books. Also, they suggest, that income or additional costs of living, such as travel expenses have an effect on the decision towards downloading a book.

My thesis builds upon the above described literature and use real-life data to check, whether people really behave how they report it in surveys. It is interesting to pinpoint, that in the earlier literature about piracy, the authors described price as a significant factor that affects people's behavior, while in the recent ones, the contents unavailability and low supply is shown as a significant one. In the next part of the thesis I describe the dataset and the additional resources I am using and test if price has an effect on the number of downloads.

3   Data

I will analyze a database acquired from one of the biggest peer-to-peer books sharing sites. The database consists of two parts:

- All of the available scientific books from LibGen's catalogue,
- All the downloads from a mirror of the website (IP log).

Library Genesis (also known as LibGen) is one of the biggest sites, where people can download books freely. The database contains mostly scientific books and text books, but there are other books that can be found in the library of the site, everyday literature, comics and scientific papers as well. In this analysis, I only use the database of the

scientific books. The IP log data contains information about both the downloader and the book: the IP-address from where the book was downloaded and an ID of downloaded book. The catalogue lists all of the available books, with their ID.

LibGen's scientific book database at the end of the analyzed period contained a total of 1 987 987 books. This means it nearly doubled its size during a year, as in 2014 it is reported to contain a little over one million books. (Cabanac (2016))

In order to be able to research only my selected sub-group of books, the textbooks, I needed to restrict the database (explained later). The final, analyzed database contained a total of 4196 books and 77 560 downloads.

I selected textbooks if a rent price is available for it on Amazon. I used the rent price as a proxy for books, that are primarily targeted for graduates and most probably textbooks. Amazon's website describes this service as one made for college students[3] . Students can rent in two different ways: by renting the paper version, which is delivered by post and needs to be returned on time or renting the e-book version, what would allow them to read it on e-book readers. The e-book will be made unavailable after the rental period is over. This type of approach allows me to select textbooks from the database, however one shortcoming of the data is that not all of the textbooks have rent prices on Amazon. Unfortunately, with the available data, this can't be tested.

3.1 Additional data

Besides the already described database, I connected several other resources in order to gain more insight from the data and be able to answer more complex questions. Here, I describe these data resources, show how could they help in the analysis and also discuss their possible shortcomings.

First, I connected the prices of the books from Amazon from the period of analysis. I used the prices of the USA Amazon, as it contains the majority of the books, but it lacks the prices of most of the Russian language books. However, I do not think this affects the analysis in a drastic way, as demand for Russian books is close to 0 outside of Russia.

The prices from Amazon come in many format: prices of *paperback* books, *hardcover* books, *e-books* also the first two in new and used format. In the analysis, I used different

---

[3]https://www.amazon.com/Rent-Textbooks/b?ie=UTF8&node=5657188011

of type of prices: list price of new paperback books and the rent price of e-books. I used this two, because of the lack of rent price for paperback books.

As some books occurred in the original database more than once but with different IDs, such as different editions of the e-book version appeared as another book or a newer print of the book appeared with another ID, I selected always the lowest available price for the different editions. I chose this solution, because I assume the demand side is really price sensitive.

There are two considerable shortcomings of the Amazon data. First, it only contains prices for the USA market as it would be really hard to connect all of the countries' prices to the database, so I use the best available approach. The prices in different countries are not the same, one real life example for this is the case *Kirstaeng v. John Wiley and sons*[4] . Kirstaeng realized that the textbooks of Wiley were significantly cheaper in his home country, Thailand. He bought the rights to sell textbooks in Thailand and then shipped them back to the USA and sold them for an alleged $1.2 million profit. This case provides evidence, that prices are not uniform throughout the world. Thus, when using the prices of Amazon, the effect will be probably overestimated.

The other problem is that the prices are only from Amazon. Books are available at several places such as online and offline bookstores or second-hand shops. As the biggest influencer of the textbook market is USA, I assume that the prices of other shops do not differ significantly from the prices of Amazon.

Secondly, I added the missing metadata to the books, as the original database was quite imperfect. I matched at least the title and author of the book, as in lot of cases the full metadata was unavailable. However, where it was available, it contains several features of the book: publisher, length, date of publication or number of the edition of the book.

Also, I matched another database to the book's, the classification of the books. I used the Library of Congress' system, the Dewey Decimal Classification (DDC)[5] . The DDC is the most frequently used classification system. As the official overview of the system says : "*The DDC is built on sound principles that make it ideal as a general knowledge organization tool: meaningful notation in universally recognized Arabic numerals, well defined categories, well-developed hierarchies, and a rich network of relationships among topics. In the DDC, basic classes are organized by disciplines or fields of study. At the*

---

[4]https://www.supremecourt.gov/opinions/15pdf/15-375_4f57.pdf

[5]http://www.oclc.org/content/dam/oclc/dewey/versions/print/intro.pdf

broadest level, the DDC is divided into ten main classes, which together cover the entire world of knowledge. Each main class is further divided into ten divisions, and each division into ten sections (not all the numbers for the divisions and sections have been used)."

This type of approach gives me the chance to be able to analyze the effects of different disciplines of books, and see if the variables have different effects for books from different backgrounds.

The disciplines and their DDCs are:

Table 3.1: DDC categories

# of top category	Name of category
0	General works, Computer science and Information
1	Philosophy and psychology
2	Religion
3	Social Sciences
4	Language
5	Pure Science
6	Technology
7	Arts and recreation
8	Literature
9	History and geography

The original database contains another aspect of the available books, the attributes of the files. These attributes are mainly dummies, indicating the format of the book. It contains, among other things, whether a book is paginated, the file is scanned version of the book or an original e-book release and extension of the available copy. The extension can have a significant role, when people decide to download a book or not, as a PDF version is easy to open on any device without additional programs, the e-book version (epub, pdb) require additional conversions or specific programs on computers. Tablets and e-books are not always compatible with all e-book extensions.

I connected another database to the downloaded books' data, which contained the exit nodes for TOR addresses. TOR is an open network, where users can hide their real IP address and block network surveillance in order to keep their privacy protected. I acquired the IP addresses, that were serving as exit nodes for the TOR network between the date of the first and last download and set up a flag, if a download was coming

from that IP-address[6] .

I also connected an other IP-address related database, the identified address range of Universities from all over the world.[7] I used this data to identify downloaders who are downloading the books from their Universities. The database is freely available and is using several resources but 2 years old, so it might lack some data and some of the ranges in it might be outdated. Despite this, I use this dataset as this was the only available database with this type and amount of data. It contains more than 5800 different Universities.

3.2   Exploratory data analysis

First, I check, whether my assumption that books with available rent price are mainly for University students is true or not.

To begin the analysis, I excluded some of the downloads based on different aspects.

1. The downloads that were marked as spam downloads from the admins of the site. These downloads were marked as "false" downloads, because for example previous downloads from that IP-addresses were spams.

2. I also excluded those downloads, that had more than two download request from the same IP-address for the same book in a one-hour time span. This is required, because there are some crawlers from Eastern-Asian IP addresses that are copying the whole database to their own and ask later money for the books downloaded from LibGen. Some of these crawlers are probably stuck at some books and are re-downloading the books every second. On the other hand, if a download request came from the same IP-address only 2 times, it is possible that it was only by mistake.

3. As mentioned above, I excluded the downloads that were coming from TOR exit nodes.

When getting the data on downloads, I also excluded the books with outlying prices. I set the thresholds from 1 dollar to 200 dollars. The threshold proved to be robust, as changing it did not change the results significantly.

---

[6]https://www.torproject.org/
[7]http://ruleofthirds.de/analyzing-scihub-data/

The analysis is two-fold. First, I will look at only the books and their attributes, while in the second part of the analysis I will include macro statistics of different countries. The latter will allow me to study the difference between developed and developing countries.

3.2.1 Books

The most downloaded books with rent price can be seen in the following table (3.2).

Table 3.2: Most downloaded books

#	Title
1	Mathematical Methods for Physicists, Sixth Edition
2	Introduction to Probability Models, Tenth Edition
3	Orientalism: Western Conceptions of the Orient
4	Data Mining. Concepts and Techsumniques, 3rd Edition
5	Instrument Engineers' Handbook, Volume 1
6	JavaScript: The Good Parts
7	Medical Secrets, Fifth Edition
8	Models and Methods in Social Network Analysis
9	Introduction to Fuzzy Logic using MATLAB
10	Black holes and time warps: Einstein's outrageous legacy

These books are nearly exclusively relevant for University students or scholars. We can find books in the field of Physics (1, 2, 10), Philosophy (3, 9) or Engineering (5). Only 2 out of the 10 aforementioned books can be directly connected to outside of academy (4, 6) as they would be useful for professionals working in Information Technology. Also Table 6.6 in the Appendix contains the three most downloaded books by DDCs. These books were downloaded 400-600 times, which is 20 times higher, than the mean of downloads.

Table 3.3: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Rent price	4,189	25.247	22.436	1.500	189.710
List price	4,189	55.976	25.815	3.500	199.950

The summary statistics show that the mean of the rent prices is less than the half of the mean of the list prices. The standard deviation of both type of prices are quite high, especially for rent prices.

The following plot (Fig. 3.1) shows the frequency of the number of downloads. It can be seen, that the frequency is shifted towards the left.
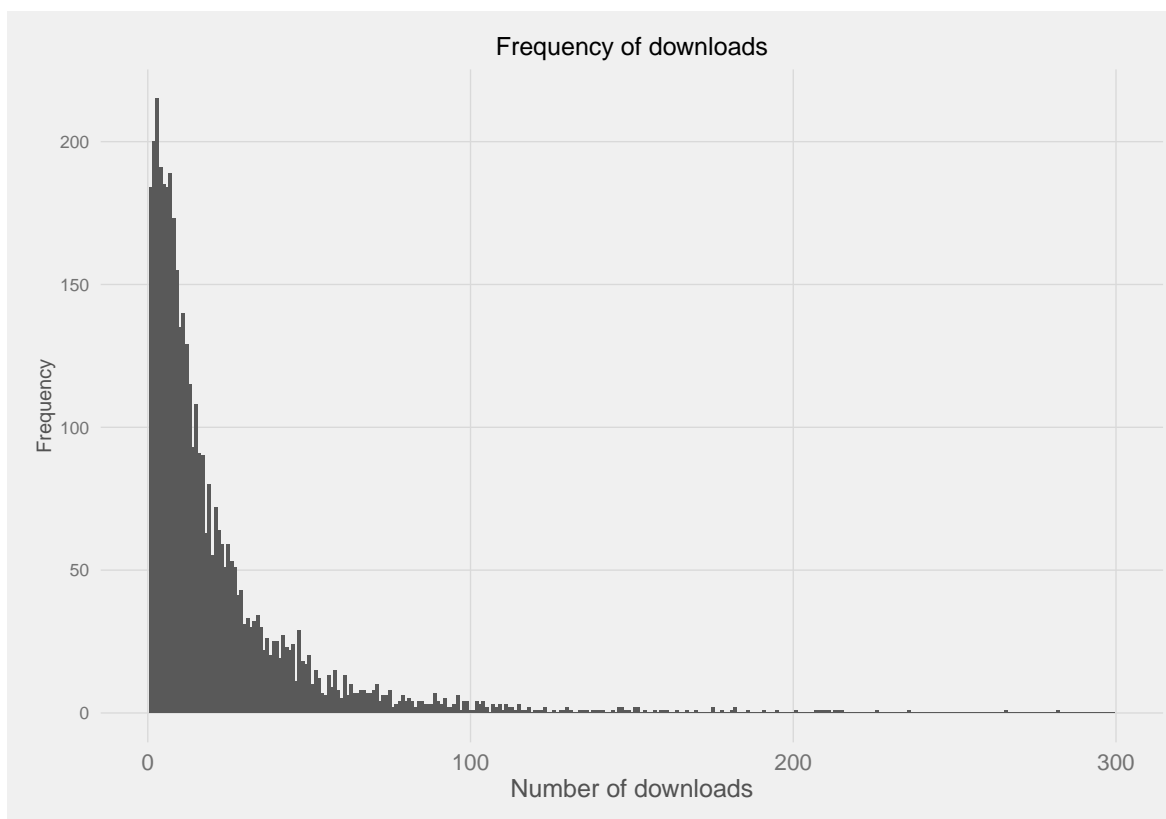
Figure 3.1: Frequency of downloads

On the next graph, we can see the plots of the frequencies of the list and rent prices of the books. We see, that the rent prices are more concentrated, compared to the frequency of the list prices.

Now, look at the correlation between the different prices and number of downloads. The correlation between the number of downloads and rent price is 0.0568109. However, here all countries are included, yet the rent price is only available in the US. The correlation of the number of downloads and rent price in the US is -0.0616014, which is negative and close to the global value of correlation. The correlation between the list price is
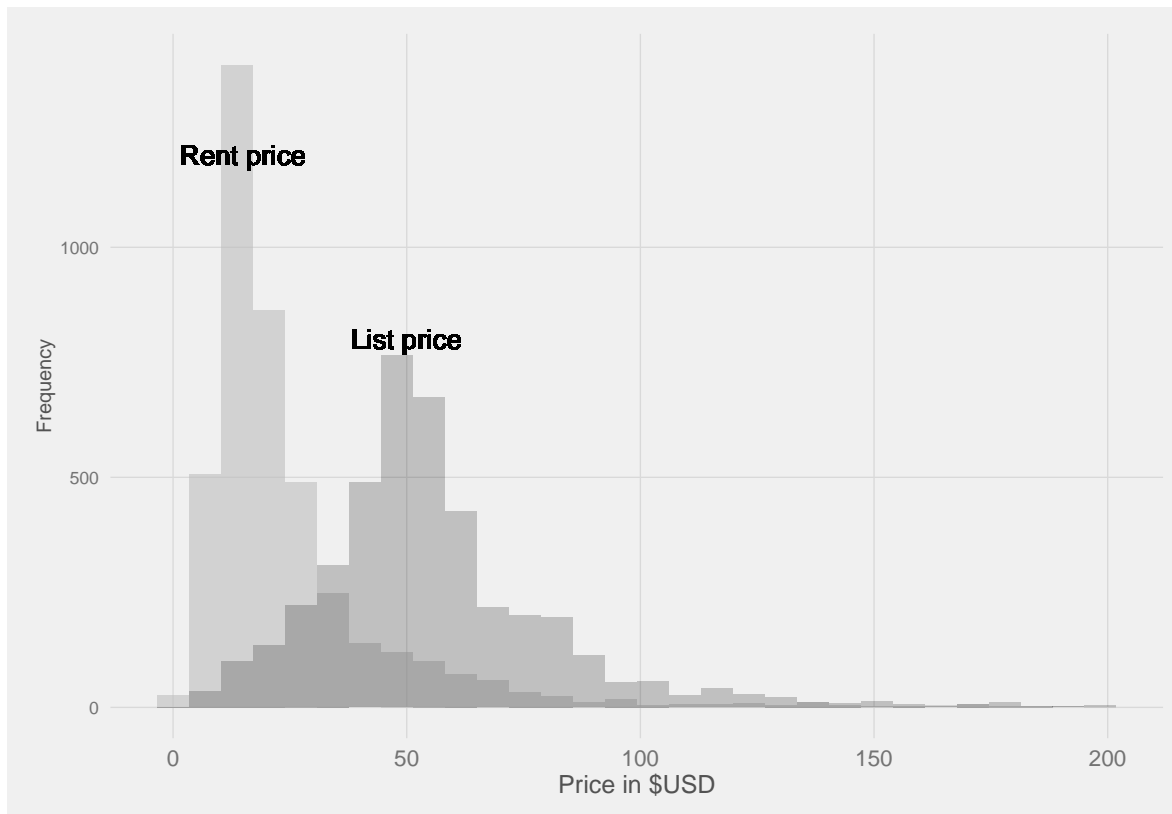
Figure 3.2: Frequency of prices

-0.0031576. We can see that the price of the book and the times it was downloaded does not move together, as the correlation is close to 0.

In order to check my assumptions from the correlation results, I ran a regression (3.4). As the distribution of both the prices and number of downloads were not normal, I used the log of all three variables. The coefficients of both regressions are really low. In the case of the regression with list price, the coefficient means that, ceteris paribus, a 1% increase in the price of a book would mean 3% decrease in the number of downloads. Also the explanatory power of the regressions are really small, as I only used one dependent variable.

Table 3.4:

	Number of downloads	
	(1)	(2)
Rent price	0.566	
	(0.662)	
List price		−2.952***
		(0.919)
Constant	20.225***	33.489***
	(2.024)	(3.630)
N	4,189	4,189
R^2	0.0002	0.002
Adjusted R^2	−0.0001	0.002
Residual Std. Error (df = 4187)	28.287	28.255
F Statistic (df = 1; 4187)	0.730	10.318***

Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.
	All variables in log

The negative coefficient can seem counter intuitive, as this would mean that lower priced books are frequently downloaded. However, this can be due to the fact, that there are several books, that have really high prices but are only being downloaded just a few times. If the majority of price's of the textbooks are lower, this phenomenon can occur.

I checked, whether the different DDCs and their interactions with prices have (Appendix, Table 6.1) effects on the coefficient of the price. The baseline category were the IT and General books (DDC - 0) When including the different categories, the effects of the price decreased, but remained significant. Only 3 categories had a significantly lower number of downloads. However, when including the extensions of the books (Appendix, Table 6.1.) , the list price became non-significant but the coefficient of the PDF version showed, that my assumptions were correct and it significantly increases the number of downloads. This means, people are looking at the extension when downloading a book. This behavior means, that the easy access and instant readability is important for the downloaders. The baseline was the *.chm* extension.

3.3   Downloaders

3.3.1   Dates of downloads

I am also interested in the behavior of the downloaders as the dataset allows me to gain some insight about their habits. The dataset is from 2014-09-27 to 2015-03-01. This means, I have data for more than 5 months of downloads. I plotted out the total number of downloads per day, to see if the data has any issues.

Besides two big drops in the number of downloads (the data analysed came in 3 chunks and some days are missing) in November and January, the data seems to behave as expected. I think, these drops do not have a significant effect on my results, as besides these dates, the data still contains 5 months of good data. Moreover, both of the missing data is nearly exactly one week, which does not distort the observations.

2 things can be seen instantly:

1. During the weekends (weekends have light red background) the number of downloads were lower than during the start of the week.

2. After a sharp increase in downloads from Sunday to Monday, starting from Tuesdays there is a downward trend per week, with a break at Thursdays. Nearly every week has their highest number of downloads on Tuesdays.
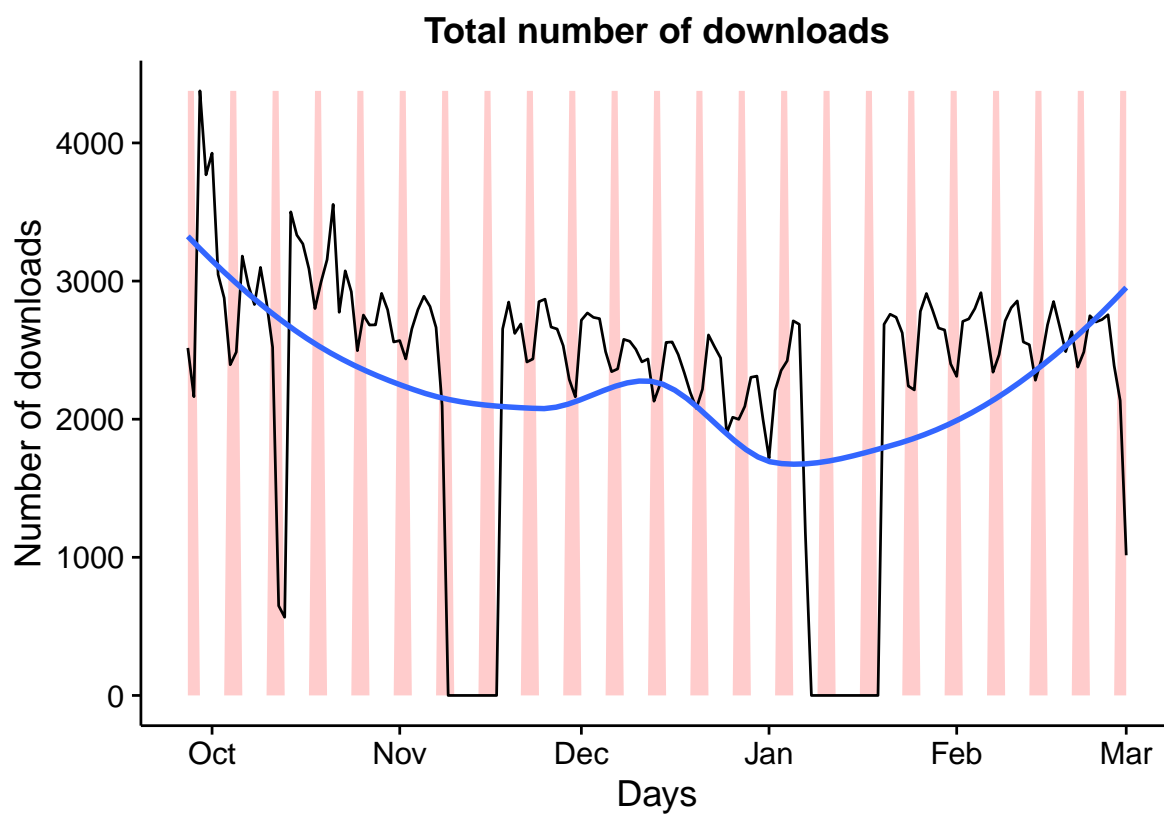
Figure 3.3: Number of download per day. Weekends with red background

3.3.2 Location of downloads

After having a look at the time of the downloads, I now turn to the location of these downloads. The following map shows the top 50 IP-addresses, from where download requests are registered in the database. The red points show the locations, which are identified as University IP-addresses. Moreover, a bigger dot represent more downloads from that IP-address. The top locations are on 4 continents, only Africa does not represent itself with big downloader country.

The locations correspond to a blog post by the administrators of Sci-Hub, on of the biggest scientific paper sharing website[8] . They found that most of the downloads are coming from developing countries: on the map, the biggest dots are from Asia.
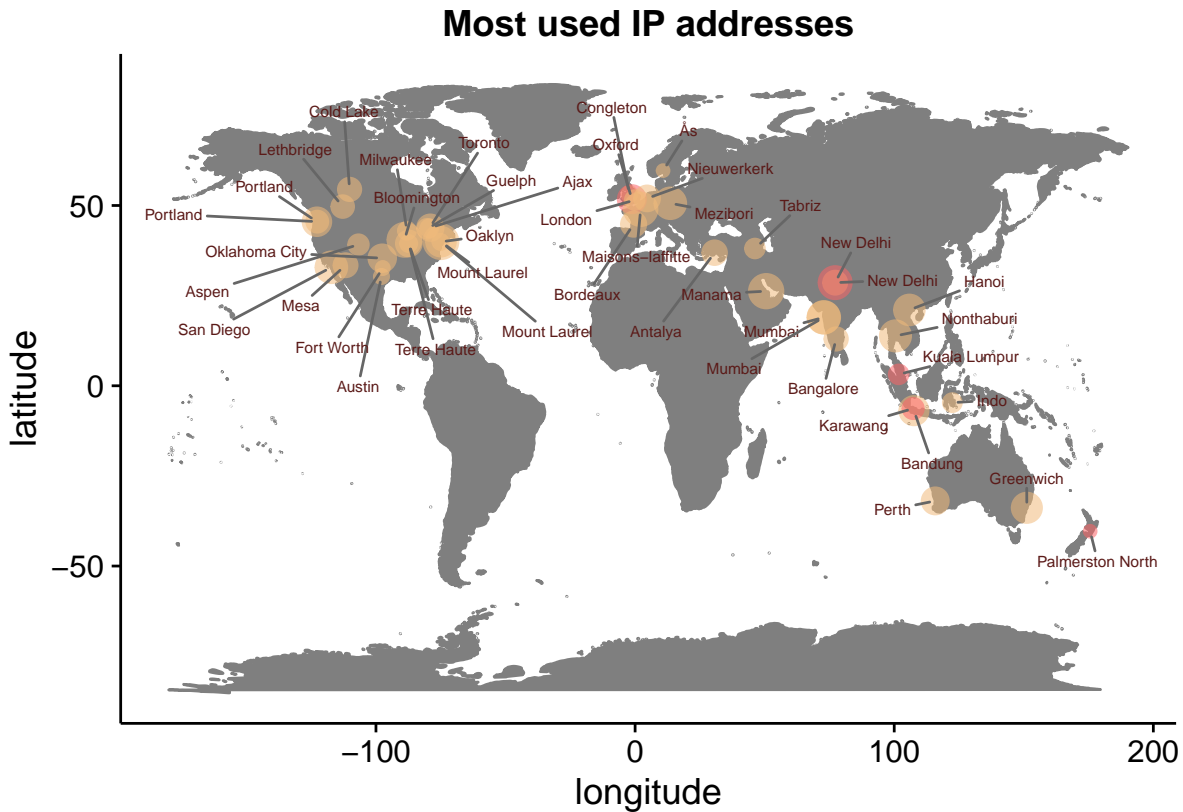
Figure 3.4: Number of downloads per city. Red dots are University IP addresses.

---

[8]https://blog.datadryad.org/2016/04/28/sci-hub-stories/

3.4 Introducing country & book specifics

In the first part of the data analysis, I only used the rent or list price of the books and their extensions. They were significant, but had really small effects, and the models had really low explanatory power due to the fact of missing independent variables.

In this part of the thesis, I introduce more regressors, that I assume have significant effect on the number of downloads.

The dependent variable is the same as in the previous regressions, the count of downloads per book. However, now I also aggregate on country level and use country attributes in order to deal with differences between different countries. When aggregating on country level, I will control for both the size of the country (population) and the number of people who are interested in the books (number of people enrolled into tertiary education)

I chose this approach, because this way I can also include the books that were not downloaded. This is important, because people simultaneously decide which books they will download or not and leaving out the books with 0 downloads would not reflect their real decisions.

In this approach, I included four country attributes for the observed countries:

- GDP per capita in purchasing power parity (PPP) in current USD,
- Internet penetration (number of people using Internet per 100 people) and
- GDP growth, as it can indicate, if a country is in a development state or not,
- The percentage of young population (18-24) enrolled into tertiary education.

I obtained all of this data from the database of the Worldbank.[9] The GDP PPP measures GDP per capita, normalized to a common consumer basket. The Internet penetration and the GDP are closely moving together, the correlation is 0.8522633, as the richer countries have better infrastructure and possibilities to provide Internet service throughout the country. I set some countries[10] as developed, where I believe the education is good enough, that students does not have problems accessing a book or

---

[9]http://data.worldbank.org/

[10]The developed countries: United States, Canada, France, Germany, UK, Netherlands, Belgium, Austria, Denmark, Finland, Luxembourg, Sweden. The OECD has no established convention for the developed vs developing countries term. https://stats.oecd.org/glossary/detail.asp?ID=6326

scientific article through their libraries. I include the mean of the last 16 (8 years before and after the recession) years' GDP growth to capture the countries' behavior during the last 10 years. The percentage of young population enrolled into tertiary education is important, because not everybody in the country is interested in the books, thus comparing downloads only to the population size can be misleading.

I also attached the classification of the books (DDC), as I think there is difference between the demand of religious or IT books. As DDC has 10 main classes, I made dummies for each 10 classes. My benchmark were the General and IT (0. class) books and for the regions, the benchmark was Australia and New Zealand.

Before analyzing the data, I looked at the correlation between GDP and number of books downloaded by country. I found, that the correlation was 0.4549957, which is much higher than the correlation with the prices of the books. As above, I used logarithms. The correlation between the number of downloads per 100 000 enrolled people and GDP is 0.4922532, which is a higher than the first correlation. This means, that the number of university in a country increases the correlation between the number of downloads and GDP.

Plotting the (log-)GDP and (log-)count of downloads (Fig. 3.5) shows, that developed countries download more. The developed countries have light blue dots.

Looking at the correlation between the Internet penetration and number of downloads in total population shows a higher correlation with 0.5025761. The correlation with the number of downloads controlled for number of university students is 0.4908647.

This shows, that downloading a book is not necessarily because people do not have enough money to buy the textbooks. Countries with higher GDP have better education, students are most likely have easy access (especially in the developed countries on the graph) for books either through a library or by having enough money to buy it.

When looking at the number of downloads divided by the population of a country (in millions) (Fig. 3.7), Southern- and Eastern-European countries are over-represented. In these countries, education changed a lot in the recent years and there was an impactful regime change, which led to the uprise of higher education. The introduction of the Bologna-process tried to unify the university education through the countries. Most of the universities had to update their reading lists and thus they moved closer to each other. Also, the European Union's open job market requires a unified knowledge for job seekers to be able to compete with each other.

Correlation between GDP/capita and downloads (log)
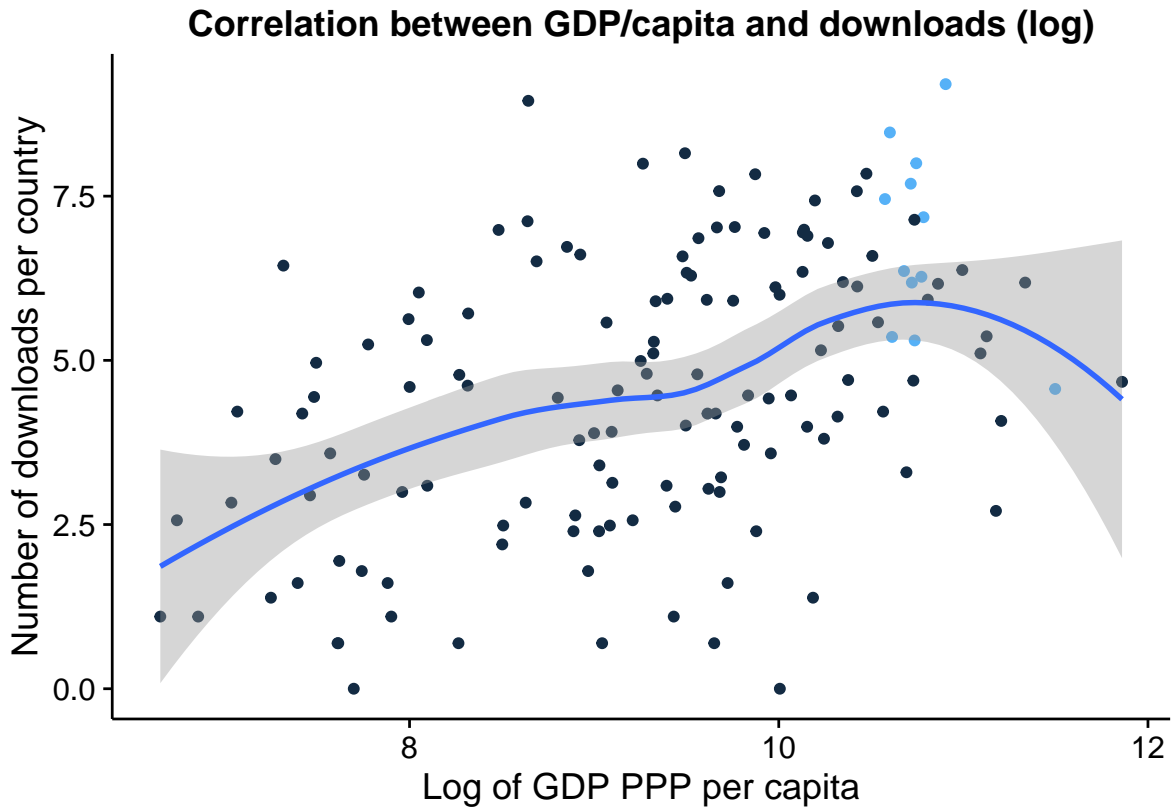
Figure 3.5: GDP and Number of downloaded books per country

Table 3.5: Top 10 countries with downloads/population (in 1 000 000)

	Country	Number of downloads
1	Luxembourg	172.563
2	Greece	155.429
3	Slovenia	121.243
4	Croatia	106.644
5	Singapore	88.670
6	Barbados	88.221
7	Portugal	85.184
8	Iceland	82.471
9	Trinidad and Tobago	81.212
10	Ireland	81.001

Correlation between internet penetration and downloads (log)
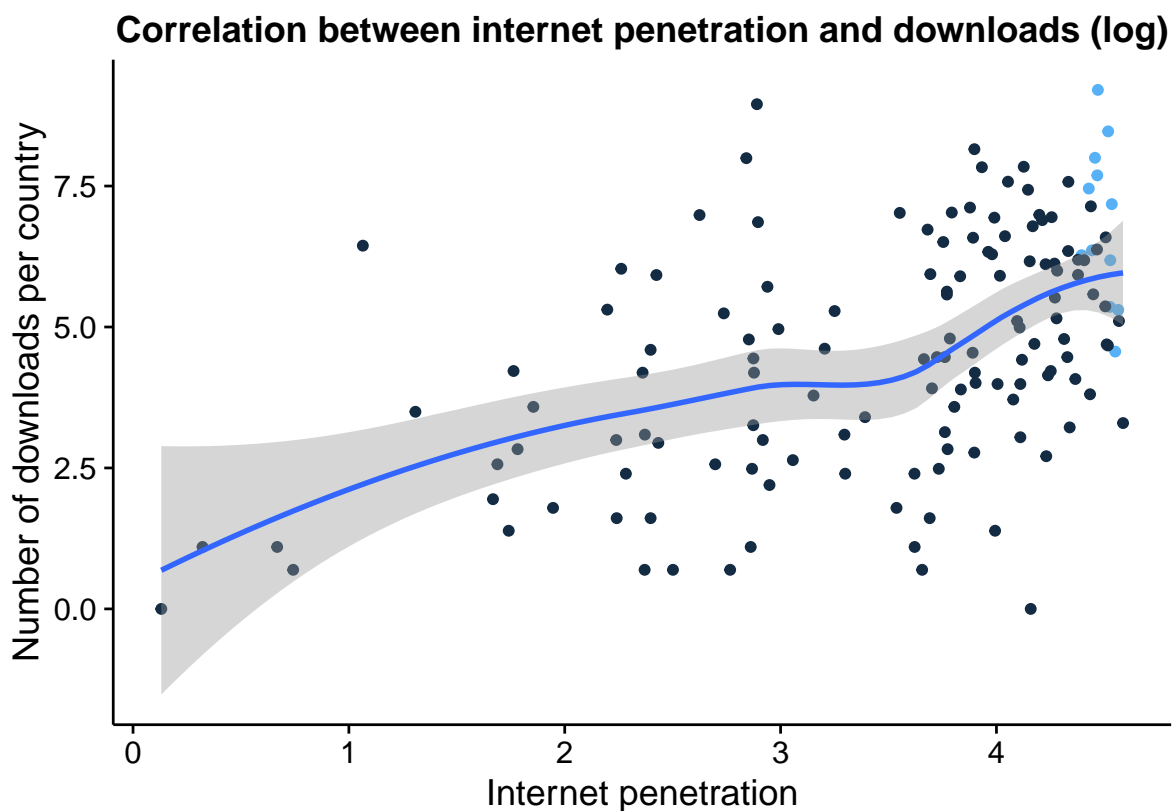
Figure 3.6: Internet penetration and Number of downloaded books per country. Developed countries with light blue

Table 3.6: Top 10 countries with downloads/enrolled people into tertiary education (in 100 000)

Country	Number of downloads (rounded)
Zimbabwe	160
Italy	66
Greece	141
Portugal	129
Jamaica	67
Japan	9
Denmark	44
Brunei Darussalam	148
Germany	60
France	42
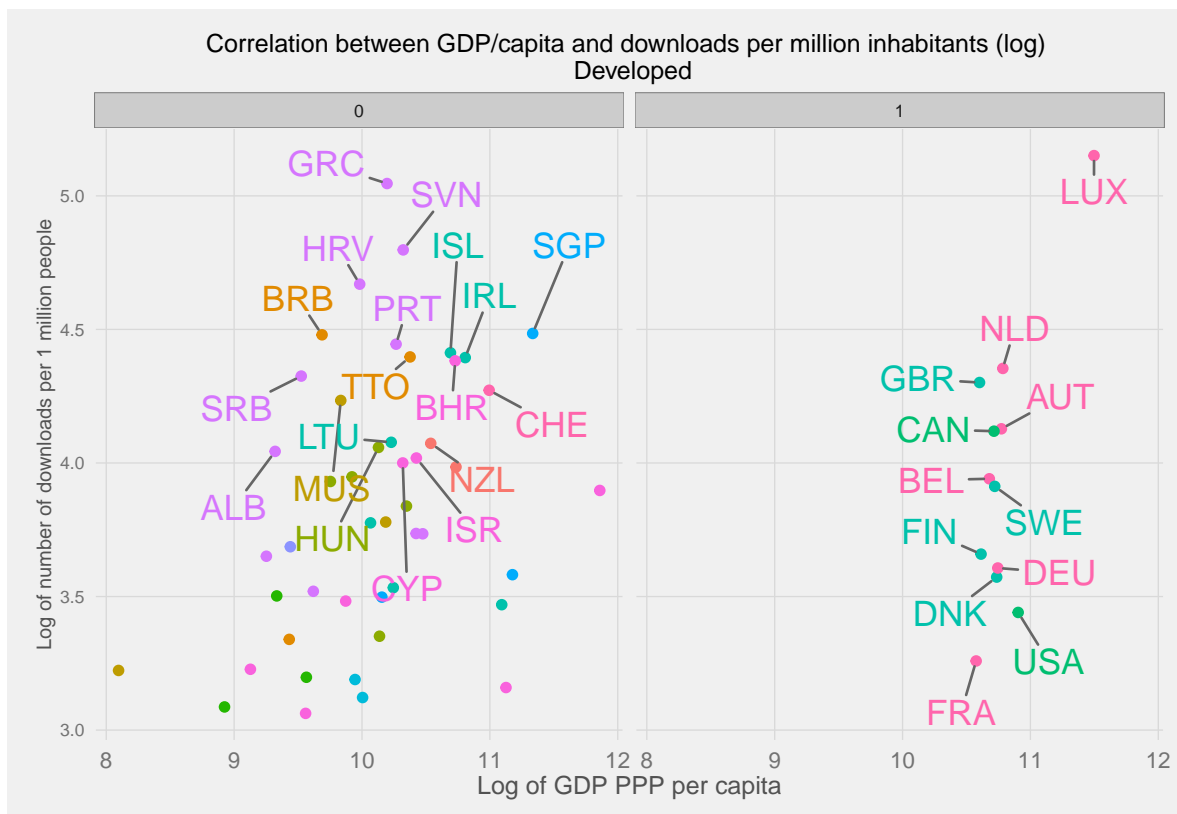
Figure 3.7: GDP and Number of downloads per million people. Developed countries with light blue

To see the exact effect of including the books with zero downloads, I ran a preliminary regression, which I could use for comparison.

As I mentioned above including books with 0 downloads is necessary, as it reflects the real choices of the downloaders. However, with all of the non-downloaded books in the dataset there will be lot of excess zeroes for the counts of downloads. In this case, the data is not normally distributed (and because of the zeros, logarithm can not be used), so OLS regression would not give a good and reliable estimation.

Because of this, I will use a hurdle (Zeileis, Kleiber, and Jackman (2008)) model. This type of regression allows for zeros and contains two estimations, which are in this case:

- Estimating if a book will be downloaded or not,
- If the book was downloaded, estimate the effects of the independent variables on the number of downloads.

The hurdle model can be applied in this case, as it differentiates between two types of behavior. At first, it describes the behavior between choosing to download or not (a binomial regression with logit link) and then estimates the number of downloads. Because of the long time span, we can assume, that if there would be demand for a book from a country, there would be at least one download (This is an assumption, we should make to use the hurdle regression).

The regressors in the two phase do not have to be the same ones. The expected number of downloads look like this:

$E(number of downloads) = P(not downloaded book) * 0 + P(downloaded book) * E(downloaded book)$

4 Hurdle model and control regressions

Before proceeding to the hurdle regression, I ran some basic OLS regressions, to compare the coefficients in the end. First, I only included the country's GDP to the regression and took all the downloads.

It can be seen (4.1), that the coefficient of GDP in both regressions are positive. As I used the logarithm of both the count of downloads and both the GDPs, the coefficient

Table 4.1:

	No. of downloads (log)	No. of downloads/million ppl (log)
	(1)	(2)
GDP (log)	0.816***	1.057***
	(0.134)	(0.074)
Constant	−2.997**	−7.498***
	(1.267)	(0.696)
N	143	143
R^2	0.207	0.592
Adjusted R^2	0.201	0.590
Residual Std. Error (df = 141)	1.878	1.031
F Statistic (df = 1; 141)	36.811***	204.948***

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

means (4.1: Model 1), that a 10% increase in a country's GDP would boost the number of downloads by 8%. Comparing the two models, the latter has a higher adjusted Rˆ2, when the independent variable was adjusted by the number of people living in the country.

To continue, I excluded the countries that only have few downloads (4.2 / Model 1.) (less than 75, which would mean on average less than 15 per month. I excluded 67 countries out of the 149 in the data), as they are not adding explanatory power to the models. In this models, the coefficient of the GDP was only significant on a 10% level, however, the coefficient of the intercept became significant. I also included a regression (4.2 / Model 2.), when I used a count divided by the number of university students in the country. Here, the GDP is significant, meaning that a higher GDP would increase the downloads per university students in a country, ceteris paribus. I included all countries, where there was available data for enrollment to tertiary education.

In these regression, the data was aggregated on only country level, so I could use the number of university students and population as a proxy for downloads. In the following section, the aggregation happens on both country and book level. With this aggregation, the majority of the books have around 1 to 5 downloads per country, which

Table 4.2:

	Count of downloads (log)	Count of downloads per university students
	(1)	(2)
GDP (log)	0.036*	
	(0.020)	
GDP (log)		0.472***
		(0.089)
Constant	1.447***	−0.840
	(0.201)	(0.854)
N	82	91
R^2	0.037	0.242
Adjusted R^2	0.025	0.234
Residual Std. Error	0.187 (df = 80)	0.930 (df = 89)
F Statistic	3.058* (df = 1; 80)	28.463*** (df = 1; 89)

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
1: Only countries with more than 75 downloads.

makes the population and university student numbers not suitable to adjust the data with, however, including them as as a dependent variable is feasible.

I also added all of the downloaded books to the database with zero downloads for the countries, where it was not downloaded. This means, every book that had at least one download from the analyzed countries, is in the dataset 82 (number of countries) times, each time with a different country. This way, the data had excess number of zeroes, that allowed me to use the hurdle regression.

As the rent prices are only available in the US, I also checked (6.3), whether the interaction between the rent price and USA has an effect on downloads. In the regression, I also included GDP as an regressor. The interaction between the rent price and the USA dummy has a significant negative effect. Out of the 4 regression, the one with the number of people as an independent variablee

4.1 Hurdle regressions

In these models, I analyzed 4 159 books, from 82 countries, meaning that the N was 341 038. I used several different hurdle models, because using only one big model would make it hard to interpret and analyze the results.

Table 4.3: Hurdle regressions

	Model 1	Model 2	Model 3	Model 4	Model 5
CM: Intercept	−39.56	−13.49	−10.68	−9.70**	−12.85
	(19398.13)	(22.65)	(6.04)	(3.63)	(23.14)
CM: Log GDP	0.41***				
	(0.01)				
CM: Log Theta	−36.13	−14.13	−11.36	−10.47**	−14.16
	(19398.13)	(22.65)	(6.03)	(3.63)	(23.14)
ZM: Intercept	−3.92***	−1.54***	−2.86***	−2.59***	−1.57***
	(0.05)	(0.01)	(0.04)	(0.03)	(0.04)
ZM: Log GDP	0.20***				
	(0.01)				
CM: Log GDP growth		0.01			
		(0.01)			
ZM: Log GDP growth		−0.34***			
		(0.01)			
CM: Tertiary education			0.02		
			(0.02)		
ZM: Tertiary education			0.28***		
			(0.01)		
CM: Internet penetration				−0.03*	
				(0.01)	
ZM: Internet penetration				0.17***	
				(0.01)	
CM: Log List price					−0.17***
					(0.02)
ZM: Log List price					−0.09***
					(0.01)
AIC	354023.06	356843.17	295635.93	361013.49	361333.48
Log Likelihood	−177006.53	−178416.58	−147812.97	−180501.75	−180661.74
Num. obs.	341038	336879	253699	341038	341038

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Every hurdle model has two parts. The first is the zero part, where a logit is used to estimate the effects of several different dependent variables. The logit looks at if somebody will download or not a book. The count model estimates the effects on the number of downloads per book. First, I included (Table 4.3) 4 country specifics and 1 book related variable. The zero model of the four country related variables showed significantly positive effects, except the growth. This means, that a higher GDP, internet panetration or tertiary education ratio means, ceteris paribus, that a book will have a higher chance to be downloaded. However, in those countries, where the historic growth (dating from 2000) was higher, books have a lower chance to be downloaded.

The macro variables have a coefficient around 0.2, which means that a 10 percent increase the GDP (4.3/Model 1) would increae the log odds of a book to be downloaded by 2 percent. The other macro variables all predict a 2-3% log odds change with a 10% GDP increase, ceteris paribus. In contrast the count models of the hurdle regressions with macro variables show non-uniform results. The coefficient of the GDP is significant and positive, the one of the percentage enrolled to tertiary education is positive, but

not significant, while the coefficient of the internet usage per 100 people is negative (but very close to 0) and significant on a 10% level. This is really interesting, because the correlation between the GDP and internet usage is quite high, one would expect the same direction for the coefficient. The coeofficient's of the model with the list prices included (4.3/Model 5) are negative in both models, meaning a higher price both decreases the log odds and number of downloaded books. Comparing the five models, the Model 3., the one where the percentage of young adults enrolled into tertiary education is the dependent variable, had the lowest AIC, which means that it explains the most out of the data.

In the next model (6.4) I introduced the region's of the countries and analyzed their effects. The baseline region was Australia and New Zealand. In Africa, the Caribbean, Southern America and West-Asia, the log odds of downloading a book is significantly lower, than in Australia and New Zealand. However, the number of downloads in West-Asia is significantly higher, which can be because the demand for books there is more unified, and people are downloading the same books. In Eastern-Asia, US and Europe, people have a higher log-odds to download a book, and also download significantly more books in the end.

The next hurdle model (6.5) shows the effects of the different categories of books. The baseline again is the DDC - 0, General and IT books. All of the other caterogies of books have a significantly higher log-odds to be downloaded, only the Arts and Recreation books are not significant. This can be due the fact, that lots of books can be categorizied as General books, and the scope of this categeory is quite wide. The count model of the regression shows, that per book, the number of downloads are usually significantly lower, than the General books, only History books have significantly more downloads per book.

My findings about the price are in line with recent survey-based studies about textbooks. Students dot no care about the price of the textbook, when they decide to download a book. This result can have an influence on the pricing behavior of the book publishers. This is in contrast with the papers about other indutries, where the authors argued, that people download to avoid the cost of getting the song, software or movie. This suggests, that the textbook piracy is not like the other industries.

The results can mean that downloads, as Zimerman argued, can occur because of the lack of supply. If this is the case, punishing the downloaders with higher prices or new

versions every year will not lead to results. Book publishers should focus on making their books available everywhere. An evidence from the data for the lack of supply is that in traditionally developing regions (Norhthern-Africa, Western-Asia) books have ceteris paribus significantly lower log odds to be downloaded. However, if a book was donwloaded in this regions, the number of downloads is significantly higher. This means, people in these regions are looking for the same books and download more of them. I was able to confirm the findings of the survey data of Rebelo and found that price have either no or negative effect on the download decision, based on the model specification.

5   Summary and extensions

In my thesis I was interested in the downloading decisions of graduate students around the world and how they are affected by macro variables. I used a dataset with download data from Library Genesis and analyzed, what effects the download decisions of students. I found, that developing countries download with a higher chance due to the recent education boom in those countries. I also found, that the price of a book has no or negative effect on the number of downloads of a book.

The study could be extended in various ways. First, the data I used was not always perfect (price of the books). With the help of a more extensive dataset about price of the books more robust results could be estimated. To add, I only excluded the TOR network as a VPN possibility, however many service exists that provide VPN. Acquiring data from more of these services would also strengthen the results. An interesting addition would be, to distinguish in a level deeper in the DDC categories, however that would require intense computation power.

Combining this dataset with survey questionnaire or online behavioral data (passive metering) could also lead to interesting results. This way, the path to a download could be analyzed or downloaders could be directly asked about their decisions, why did they choose to download a book.

6   Appendix

Data analysis: I used Python 2.7 and MongoDB to build the dataset and tidy it, so it was ready for analysis. For analyzing and visualizing the data, I used R (version 3.3.1) and RStudio.

Data discrapancy

I checked for discrepancy in the data, by looking at the hour of downloads. The time in the dataset belongs to the server where the books are stored (Eastern-Europe), which makes it harder to derive the overall behavior per hour. However, the same trend can be seen through all of the countries. Here, I included three countries. Germany is the closest to the server from the three countries and it shows an expected behavior: the number of downloads are small during the night and increase during the daytime.
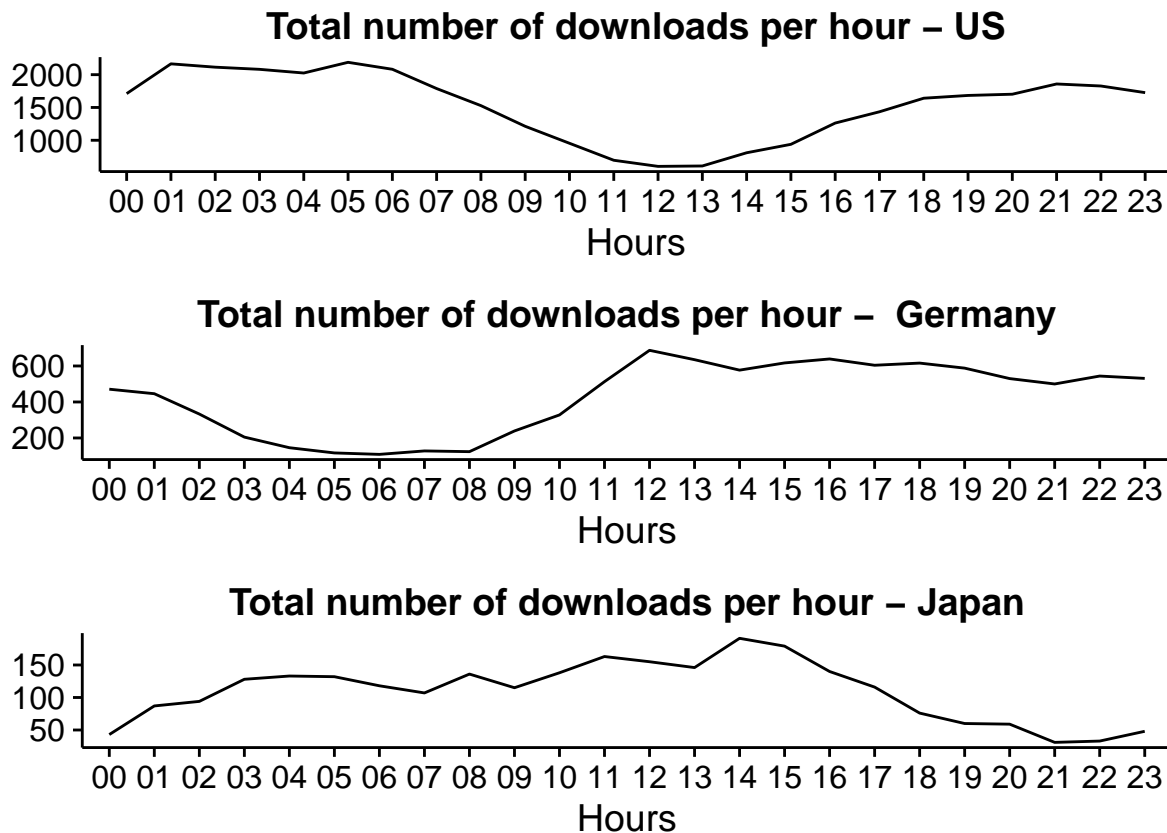
Figure 6.1: Number of downloads per hour per country. Times in own timezone

Regression with extensions:

Table 6.1:

	log_n	
	(1)	(2)
log_rentp	0.003	
	(0.026)	
log_listp		−0.085**
		(0.037)
extensiondjvu	−0.107	−0.106
	(0.233)	(0.232)
extensionepub	0.837***	0.777***
	(0.248)	(0.249)
extensiongz	−0.796	−0.731
	(1.138)	(1.138)
extensionlit	−1.198	−1.388*
	(0.819)	(0.822)
extensionmobi	−0.082	−0.141
	(0.375)	(0.376)
extensionpdf	0.635***	0.630***
	(0.212)	(0.212)
extensionPDF	0.217	0.216
	(0.543)	(0.543)
extensionrar	−0.165	−0.160
	(0.429)	(0.428)
extensionzip	0.010	0.028
	(0.503)	(0.503)
Constant	1.884***	2.233***
	(0.226)	(0.258)
N	4,189	4,189
R^2	0.020	0.021
Adjusted R^2	0.018	0.019
Residual Std. Error (df = 4178)	1.118	1.117
F Statistic (df = 10; 4178)	8.587***	9.118***

Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Regression with list price and DDC:

Table 6.2:

	log_n
List price	−0.319**
	(0.141)
DDC - 0	0.784
	(0.788)
DDC - 1	0.083
	(0.989)
DDC - 2	−1.234*
	(0.652)
DDC - 3	1.244
	(1.217)
DDC - 4	−0.080
	(0.698)
DDC - 5	−1.182*
	(0.662)
DDC - 6	−2.140**
	(0.836)
DDC - 7	−1.470**
	(0.682)
DDC - 8	2.098**
	(0.895)
DDC - 9	−0.045
	(0.203)
List price*DDC - 1	0.170
	(0.263)
List price*DDC - 2	0.397**
	(0.164)
List price*DDC - 3	−0.079
	(0.311)
List price*DDC - 4	0.174
	(0.173)
List price*DDC - 5	0.378**
	(0.164)
List price*DDC - 6	0.576***
	(0.214)
List price*DDC - 7	0.516***
	(0.180)
List price*DDC - 8	−0.447*
	(0.233)
List price*DDC - 9	3.370***
	(0.562)
N	4,189
R^2	0.047
Adjusted R^2	0.043
Residual Std. Error	1.104 (df = 4169)
F Statistic	10.876*** (df = 19; 4169)

Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Regression for controlling for USA:

Hurdle with regions

Table 6.3:

	count
log(gdp)	0.015***
	(0.002)
usa1	2.937***
	(0.065)
log(rentp)	0.004*
	(0.002)
usa1:log(rentp)	−0.258***
	(0.021)
Constant	0.041**
	(0.017)
N	341,038
R^2	0.067
Adjusted R^2	0.067
Residual Std. Error	0.896 (df = 341033)
F Statistic	6,166.825*** (df = 4; 341033)

Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

36

	Model 1
Count model: (Intercept)	−10.51
	(5.42)
Count model: regionCaribbean	−1.22***
	(0.34)
Count model: regionCentral America	0.23*
	(0.11)
Count model: regionEastern Africa	−0.51***
	(0.11)
Count model: regionEastern Asia	1.05***
	(0.08)
Count model: regionEurope	0.46***
	(0.07)
Count model: regionMiddle Africa	−1.20***
	(0.35)
Count model: regionNorthern Africa	0.20*
	(0.09)
Count model: regionNorthern America	0.71***
	(0.09)
Count model: regionSouth America	0.15
	(0.08)
Count model: regionSouth-Eastern Asia	0.55***
	(0.08)
Count model: regionSouthern Africa	−0.11
	(0.12)
Count model: regionSouthern Asia	1.58***
	(0.08)
Count model: regionUS	2.22***
	(0.08)
Count model: regionWestern Africa	−0.11
	(0.11)
Count model: regionWestern Asia	0.30***
	(0.08)
Count model: Log(theta)	−10.36
	(5.42)
Zero model: (Intercept)	−1.89***

	Model 1
	(0.03)
Zero model: regionCaribbean	−1.84***
	(0.11)
Zero model: regionCentral America	0.40***
	(0.05)
Zero model: regionEastern Africa	−1.21***
	(0.04)
Zero model: regionEastern Asia	0.81***
	(0.04)
Zero model: regionEurope	0.03
	(0.03)
Zero model: regionMiddle Africa	−1.95***
	(0.11)
Zero model: regionNorthern Africa	−0.44***
	(0.04)
Zero model: regionNorthern America	1.06***
	(0.05)
Zero model: regionSouth America	−0.50***
	(0.04)
Zero model: regionSouth-Eastern Asia	0.26***
	(0.04)
Zero model: regionSouthern Africa	−0.02
	(0.06)
Zero model: regionSouthern Asia	0.38***
	(0.04)
Zero model: regionUS	2.54***
	(0.05)
Zero model: regionWestern Africa	−0.47***
	(0.05)
Zero model: regionWestern Asia	−0.65***
	(0.04)
AIC	342489.09
Log Likelihood	−171211.54
Num. obs.	341038

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table 6.4: Statistical models

Hurdle with DDCs

	Model 1
Count model: (Intercept)	−15.66
	(73.41)
Count model: factor(ddc)1	−0.10*
	(0.05)
Count model: factor(ddc)2	−0.04
	(0.06)
Count model: factor(ddc)3	−0.35***
	(0.04)
Count model: factor(ddc)4	0.03
	(0.06)
Count model: factor(ddc)5	0.02
	(0.04)
Count model: factor(ddc)6	−0.21***
	(0.04)
Count model: factor(ddc)7	−0.42***
	(0.06)
Count model: factor(ddc)8	0.05
	(0.06)
Count model: factor(ddc)9	0.12*
	(0.06)
Count model: Log(theta)	−16.45
	(73.41)
Zero model: (Intercept)	−2.15***
	(0.02)
Zero model: factor(ddc)1	0.44***
	(0.02)
Zero model: factor(ddc)2	0.42***
	(0.03)
Zero model: factor(ddc)3	0.16***
	(0.02)
Zero model: factor(ddc)4	0.57***
	(0.03)

	Model 1
Zero model: factor(ddc)5	0.36***
	(0.02)
Zero model: factor(ddc)6	0.19***
	(0.02)
Zero model: factor(ddc)7	0.02
	(0.03)
Zero model: factor(ddc)8	0.29***
	(0.03)
Zero model: factor(ddc)9	0.28***
	(0.03)
AIC	360482.99
Log Likelihood	-180220.49
Num. obs.	341038

$^{***}p < 0.001,\ ^{**}p < 0.01,\ ^{*}p < 0.05$

Table 6.5: Statistical models

Table 6.6: Top 3 downloaded books per DDC category

ddc	title	author	count
8	The Long White (Iowa Short Fiction Award)	Sharon Dilworth	212
8	The Postcolonial Exotic: Marketing the Margins	Graham Huggan	208
8	Virgil and the Augustan Reception	Richard F. Thomas	160
5	Mathematics As A Constructive Activity: Learners Generating Examples (Studies in Mathematical Thinking and Learning Series)	Anne Watson, John H. Mason,	195
5	Rubberlike elasticity: A molecular primer	James E. Mark, Burak Erman	137
5	What Science Is and How It Works	Gregory N. Derry	136
3	American Legal Thought from Premodernism to Postmodernism: An Intellectual Voyage	Stephen M. Feldman	245
3	Regulating the Global Information Society (Warwick Studies in Globalisation)	C. Marsden	245
3	Mathematical Finance (Routledge Advanced Texts in Economics and Financea)	Dokuchaev	236
6	Programming 16-Bit PIC Microcontrollers in C: Learning to Fly the PIC 24	Lucio Di Jasio	214
6	Echohouse: A Design Guide	Susan Roaf, Manuel Fuentes, Stephanie Thomas	207
6	Fast Fourier Transform - Algorithms and Applications	K.R. Rao, D.N. Kim, J.-J. Hwang	159
1	Relational Ethics in Practice: Narratives from Counselling and Psychotherapy	Lynne Gabriel, Roger Casemore	325
1	A New Introduction to Modal Logic	M.J. Cresswell, G.E. Hughes	295
1	I Need Your Love - Is That True?: How to Stop Seeking Love, Approval, and Appreciation and Start Finding Them Instead	Byron Katie, Michael Katz,	145
7	If It's Purple, Someone's Gonna Die: The Power of Color in Visual Storytelling	Patti Bellantoni	172
7	Physical Theatres: An Introduction	Simon Murray	121
7	The Game Audio Tutorial: A Practical Guide to Sound and Music for Interactive Games	Richard Stevens, Dave Raybould	119
0	Adaptive Learning of Polynomial Networks: Genetic Programming, Backpropagation and Bayesian Methods	Nikolaev N., Iba H.	230
0	What Every Engineer Should Know about Developing Real-Time Embedded Products	Kim R. Fowler	161
0	Variational Methods for Crystalline Microstructure - Analysis and Computation	Shuvra S. Bhattacharyya, Ed F. Deprettere, Jürgen Teich	160
4	Questionnaires in Second Language Research: Construction, Administration, and Processing (Second Language Acquisition Research Series)	Zoltan Dornyei	294
4	Philosophy of Language - A Contemporary Introduction	William G. Lycan	215
4	A Dictionary of Grammatical Terms in Linguistics	R. L. Trask	135
9	British Armour in the Normandy Campaign	John Buckley	148
9	Early Islamic Spain: The History of Ibn al-Qutiyah (Culture and Civilization in the Middle East)	David James	148
9	Tiberius the politician	Barbara Levick	144
2	God, Chance and Purpose: Can God Have It Both Ways?	David J. Bartholomew	106
2	The Waldensian Dissent: Persecution and Survival, c.1170-c.1570	Gabriel Audisio	104
2	Fifty Key Jewish Thinkers	Dan Cohn-Sherbok	100

References

Bhattacharjee, Sudip, Ram D Gopal, and G Lawrence Sanders. 2003. "Digital Music and Online Sharing: Software Piracy 2.0." *Communications of the ACM* 46 (7). ACM: 107–11.

Bodó, Balázs. 2011. "Coda: A Short History of Book Piracy." *Media Piracy in Emerging Economies*, 399.

———. 2014. "A Short History of the Russian Digital Shadow Libraries." *Available at SSRN 2616631*.

Bodó, Balázs, and Zoltán Lakatos. 2012. "Piracy Cultures - P2P and Cinematographic Movie Distribution in Hungary." *International Journal of Communication* 6: 33.

Cabanac, Guillaume. 2016. "Bibliogifts in LibGen? A Study of a Text-Sharing Platform Driven by Biblioleaks and Crowdsourcing." *Science and Technology* 67 (4): 874–84.

Christensen, Anne L, and Martha M Eining. 1991. "Factors Influencing Software Piracy: Implications for Accountants." *Journal of Information Systems* 5 (1). American Accounting Association: 67–80.

Danaher, Brett, and Joel Waldfogel. 2012. "Reel Piracy: The Effect of Online Film Piracy on International Box Office Sales." *Available at SSRN 1986299*.

Fetscherin, Marc. 2005. "Movie Piracy on Peer-to-Peer Networks—the Case of KaZaA." *Telematics and Informatics* 22 (1). Elsevier: 57–70.

Givon, Moshe, Vijay Mahajan, and Eitan Muller. 1995. "Software Piracy: Estimation of Lost Sales and the Impact on Software Diffusion." *The Journal of Marketing.* JSTOR, 29–37.

Gopal, Ram D, G Lawrence Sanders, Sudip Bhattacharjee, Manish Agrawal, and Suzanne C Wagner. 2004. "A Behavioral Model of Digital Music Piracy." *Journal of Organizational Computing and Electronic Commerce* 14 (2). Taylor; Francis: 89–105.

Heller, Donald E. 1997. "Student Price Response in Higher Education: An Update to Leslie and Brinkman." *Journal of Higher Education.* JSTOR, 624–59.

Hoorebeek, Mark. 2003. "EBooks, Libraries and Peer-to-Peer File-Sharing." *The Australian Library Journal* 52 (2). Taylor; Francis: 163–68.

Johns, Adrian. 1998. *The Nature of the Book: Print and Knowledge in the Making.* University of Chicago Press.

———. 2010. *Piracy: The Intellectual Property Wars from Gutenberg to Gates.* University of Chicago Press.

Karaganis, Joe. 2011. *Media Piracy in Emerging Economies.* Lulu. com.

Koklic, Mateja Kos, Monika Kukar-Kinney, and Irena Vida. 2014. "Three-Level Mechanism of Consumer Digital Piracy: Development and Cross-Cultural Validation." *Journal of Business Ethics.* Springer, 1–13.

Podoshen, Jeffrey S. 2008. "Why Take Tunes? An Exploratory Multinational Look at Student Downloading." *Journal of Internet Commerce* 7 (2). Taylor; Francis: 180–202.

Rebelo, Francisca. 2015. "Understanding Textbook Piracy."

Rodman, Gilbert B, and Cheyanne Vanderdonckt. 2006. "Music for Nothing or, I Want My MP3: The Regulation and Recirculation of Affect." *Cultural Studies* 20 (2-3). Taylor; Francis: 245–61.

Rohde, Amanda, and others. 2001. "The Creation of the Enook: Ebooks in the Library." *LASIE: Library Automated Systems Information Exchange* 32 (1). State Library of New South Wales: 54.

Scorcu, Antonello Eugenio, and Laura Vici. 2012. "Economic and Cultural Factors and Illegal Copying in the University Textbook Market."

Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. "Regression Models for Count Data in R." *Journal of Statistical Software* 27 (8). Foundation for Open Access Statistics: 1–25.

Zimerman, Martin. 2011. "E-Books and Piracy: Implicationsissues for Academic Libraries." *New Library World* 112 (1/2). Emerald Group Publishing Limited: 67–75.