# The Impacts of Professional Reviews and Crowd Ratings on the Book Market

2023-07-18

**Author:** Leonard Pöhls

Hello readers and welcome to my Bachelor Thesis on the impact of professional reviews and crowd ratings on the book market. Do not wonder about the way this thesis is constructed. The entire file is the output of a so-called Rmarkdown file, which is based on the R programming language. In the meantime, you will encounter some lines of code which are either for data transformation or graph generation. These process a large data set, which is stored in the background of the Rmarkdown file in memory. This type of presentation is intended to replicate the study of another paper in order to reproduce and, if necessary, extend its research. This thesis is a replication of the paper: **"Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings"** by Imke Reimers and Joel Waldfogel, that has been published in 2021 in the American Economic Journal. This paper examines two main aspects: First, how pre-purchase information in the form of crowd ratings from other individual purchaser and professional reviews from daily newspapers affect sales ranks and quantities. Second, to conclude inferences due to welfare effects with and without the presence of pre-purchase information. The base of those mentioned welfare effects got determined through the transformation from sales ranks into sales quantities. The Paper can be opened here.

For this thesis there is also an interactive problem set, which is characterized by tasks and quizzes, which can also be found in this file. The problem set is published here:

- `Github`: …

In 1995, one year after Amazon was incorporated, more than 61 percent of all book sales in the U.S. were generated by physical bookstores and book clubs, while only ten percent were made through other channels, including Amazon (Curcic, 2023).

In the following 28 years, we have witnessed the growth of digitization and the development of crowd rating infrastructure on online sites. Users are thus able to obtain important non-professional pre-purchase information from other users that can influence purchasing behavior and economic welfare effects. As a result, online retailers have largely replaced the trade in physical books. From an economic point of view, pre-purchase information also has an impact on personal quality expectations and thus on demand.

Today, the distribution of print book market share has changed significantly, and Amazon has taken the lead as the largest retailer of print books in the entire world. For instance, the Amazon share of the U.S. book market accounts more than 40 percent and around 50 percent of the Great Britain market share (McLoughlin, 2022).

To examine these effects, this problem set aims to reorganize and replicate part of the study from Reimers and Waldfogel by retyping and extending their research on the effects of professional reviews and crowd ratings on sales ranks and sales quantities.

# Contents

## Overview

Initially, basic knowledge about the book market and related terminology related to Amazon will be explained. In order to be capable to classify the background of pre-purchase information in economic terms, the economic added value of this information and the price elasticity are presented. I then focus on the underlying data set to show how it is structured and what attributes are important for further investigations. On the basis of this data set, descriptive analyses of the occurrence of crowd ratings, professional reviews, price and sales rank developments are carried out in order to obtain an overview and to be able to make initial assumptions about potential effects. Subsequently, the empirical part of this problem set begins. First, a naive regressions gets implemented to illustrate the endogenous problem. By gradually adding several methods for extended regressions, it is shown how the coefficients change and become more precise. These methodologies are then used in the following to replicate the main regression from the authors. Afterwards, a so-called cross-sectional event study is replicated to graphically display these estimated effects. Finally, I explain how the authors convert relative sales ranks into quantities to estimate elasticities, which we then implement in R. The summary of the results of the problem set and similar results of other authors are presented in the final section.

## An Instruction how to work with problem sets

As mentioned above, the purpose of this problem set is to create an interactive environment for the readers. Thus, sometimes different types of exercises appear within a so-called chunk (window) below. There are essentially three different types of exercises:

- An empty code chunk without any Information. Consequently, you have to find the solution by yourself.
- Code sections with gaps like _____ , which need to be replaced with the correct code.
- Those where all the code is already given. This code is ready to be executed.

If you need some advice while solving the task, just press `hint` to get help. Pressing `run` will execute the code. The `solution` button is a shortcut, which immediately provides the correct code. To verify the task, click on `check`. If your code was not correct, you will receive a corresponding report.

Next to code chunks, you can also work on some multiple choice quizzes to test your prior knowledge or to check your own text comprehension. Guessing the correct answer can also lead to better understanding and maintaining attention.

Press `Go to next exercise...` to continue and work through further tasks of this problem set.

# 1. Motivation

In this study, the book market was selected to determine the impact of crowd ratings and professional reviews.

The following chapter will first explain what is meant by "pre-purchase information," what the situation is on the book market, and why specifically this market is predestined for measuring review effects. To find an answer, we focus on newspaper magazines and on the role of Amazon concerning crowd ratings and so-called sales ranks.

In the second part of this chapter, I illustrate possible welfare effects from access to pre-purchase information. In addition, I explain the transition from sales ranks and sales prices to quantity elasticities.

After editing this chapter, you will gain an insight of the initial situation on the book market and you will receive deeper knowledge about basic economic issues. Thus you will be well prepared to continue with chapter two.

**Structure**

1.1. Book Market, Professional Reviews and Crowd Ratings

1.2. Introduction to Welfare, Demand and Price Elasticity

## 1.1. Book Market, Professional Reviews and Crowd Ratings

In general, economists distinguish goods concerning their characteristics, their occurrence, or other conditions. With respect to the degree of uncertainty, we differentiate between three different goods, **Search Goods**, **Experience Goods** and **Credence Goods**. Buyer of **Search Goods** already have an accurate perception and a high degree of certainty about what they want to buy (Wieneke, 2019). For instance, sugar or computer are typical search goods. **Experience Goods** are associated with a lower degree of certainty. Without information advantage, buyers are unable to assess the quality of the good until they consume it, as in the case cinema visits or whine (Wieneke, 2019). By the consumption of **Credence Goods** the buyer never gets into the situation of evaluating the quality of the underlying good, as the degree of uncertainty is the highest here. Common credence goods are services such as lawyers or surgeons (Wieneke, 2019).

#< quiz "Books_As_Goods" question: What do you think books belong to? sc: - Search Goods. - Experience Goods.* - Credence Goods. success: Great, your answer is correct! failure: Try again. #>

The underlying paper focuses on the book market to examine the impacts from crowd ratings and professional reviews on sales ranks. So why is the book market particularly well suited for this study? In their paper, Reimers and Waldfogel (2021) enumerated three main reasons for this assertion: First, books are among experience goods. For the other two goods, pre-purchase information is less or not relevant. Second, the number of professional reviews (in high visible media) is relatively small and distributed among a few major newspapers. The third reason relates to the data set on which the entire study is based. The high frequently data on book demand at Amazon should contain about 45% of the US physical book market, roughly comparable to the figures of Mcloughlin in 2022.

As explained, professional reviews are periodic reviews in daily newspaper articles. The data set includes information on the appearance of professional reviews from the New York Times, the Chicago Tribune, the Boston Globe, the Wall Street Journal, the Los Angeles Times, and the Washington Post. However, quantitative information as well as star ratings are not available. Additionally, The New York Times recommends nine books every week (New York Times, 2023). Professional reviews enable access to pre-purchase information.

#< quiz "Big_Newspapers" question: What is your suggestion, which of these newspapers has the most impact on sales? sc: - The New York Times.* - The Chicago Tribune. - The Boston Globe. - The Wall Street Journal. - The Los Angeles Times. - The Washington Post. success: Great, your answer is correct! failure: Try again.

#>

As well as professional reviews, the Amazon data set also contains information about star ratings of the buyers. Identified as a buyer on Amazon, anyone may rate their purchased product on a five-point scale. In principle, people benefit from the ratings of other customers because they optimize their buying behaviour as a result. In contrast to professional ratings, however, crowd ratings tend to generate less trust. Crowd ratings are prone to fake content for defamatory or fraudulent purposes, while professional reviews are created by objective and professional raters. Therefore, the more ratings exist for the particular product, the more likely the calculated average star rating approximates the "actual" quality. Thus, the crowd ratings represent the second piece of information before the purchase.

Hence, Reimers and Waldfogel (2021) claim that consumers interact with these two types of pre-purchase information in different ways, potentially leading to different or overlapping effects. One possible reason for this could be the different accessibility of these types of information. While audience reviews are visible to every Amazon user, professional reviews are accessible but not automatically visible to everyone. Also, people who happen to find a book on Amazon are less likely to check after the fact to see if a professional review is available. Conversely, people who have found a book reviewed by experts automatically get access to the reviews of the masses during the purchase process. Incidentally, word-of-mouth can also be understood as pre-purchase information. Nevertheless, the consumer has access to at least one type of pre-purchase information.

#< quiz "Fakereviews_Amazon" question: How many reviews on Amazon are fake or unreliable? sc: - 61%. - 23%. - 9%. - 42%.* - 15%. success: Great, your answer is correct! (Stieb, 2022) failure: Try again.

#>

We mentioned earlier that we want to make an estimate based on so-called sales ranks. Sales ranks are the numerical representation of how your products sell compared to other products in the same category (Wisniach, 2022). Since we do not have information about the quantities sold, these values are replaced by the sales rank. Amazon defined their sales rank as hourly updated calculation to "reflect recent and historical sales of every item sold on Amazon". Hence, sales ranks are relative numbers to compare sales activities. To make assumptions for a welfare analysis, the authors of the underlying paper collected additional data from the 2018 New York Times weekly top 100 bestsellers to convert ranks into quantities. This allows us to examine price elasticities.

**Summary**

In summary, we have obtained an explanation of the goods in terms of their level of certainty and have assigned the books to the experience goods. We have also identified three different types of pre-purchase information, two of which we use for estimation. Further, We have recognized that crowd ratings and professional reviews have different effects on sales rankings. Finally, we got an introduction to Amazon sales ranks and their importance for the further course.

The following chapter 1.2 will provide you with the basic economic knowledge you need to understand the analysis and the way of this audit.

## 1.2. Introduction to Welfare, Demand and Price Elasticity

Colloquially, the term "welfare" is associated with many contexts, such as unemployment benefit or other social assistance. The actual origin of this term lies in the economy. Mathematically, welfare is the sum of producer surplus and consumer surplus. The consumer surplus is the difference between the price for a good that consumers are willing to pay and the actual price of this good. Against this, producer surplus is the difference between the price that suppliers would be willing to charge for their goods and the actual price of this good. Actually, the economic reality is much more complex. For simplicity, we focus on polypol markets (markets with many suppliers) under perfect conditions, which the model requires to work. Without even one of these assumptions, the model is invalid.

#< quiz "perfectMarket_Conditions" question: Which of these conditions is **not** relevant for a perfect polypol market? sc: - Perfect information availability (Knowledge about every price for the underlying good). - No personal preferences (Preferences, that prevent you from acting rationally). - Homogeneous goods (Equal goods). - Fast reaction velocity (Changing market conditions are quickly recognized from every market participant). - Every good has the same quality.* - A large number of demanders and suppliers. success: Great, your answer is correct! Equality of quality and homogeneity of goods do not from an equivalent. Homogeneity only includes the physical condition and the substitutability. failure: Try again.

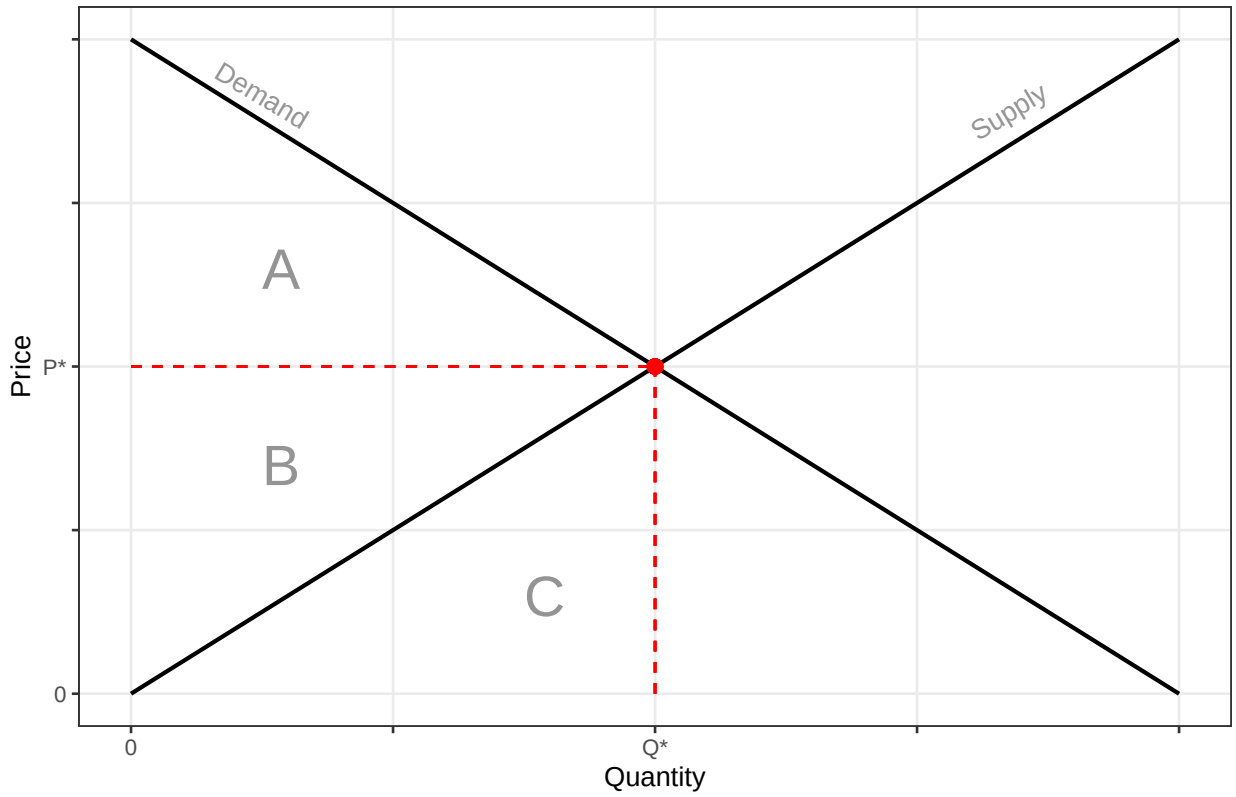
#>

Under these conditions, we illustrate the situation between demand and supply with a diagram to better understand the added value of pre-purchase information.

**Task:** Run the following chunk to create this model. For performance purposes, I saved the essential part `fdemandsupplyfinal` in `fdemandsupply`. Press `check` to collect your points.

```
#load the package "ggplot2"
library(ggplot2)
demand <- c(2,1.5,1,0.5,0)
xAxis <- c(0, 1,2,3,4)
supplyvec <- c(0,0.5,1,1.5,2)
xGroup <- c(1,1,1,1,1)
DatasetTest <- data.frame(Price = demand, Quantity = xAxis, Supply = supplyvec,
                          Group = xGroup)
#read in fdemandsupply
f1 <- readRDS("material/f1.RDS")
#create the graph `fdemandsupplyfinal`
fdemandsupplyfinal <- f1  +
geom_text(aes(x=0.5, y=1.83, label = "Demand"), color = "Black", angle = 329,
          size= 4, alpha=0.1) +
geom_text(aes(x=3.35, y=1.78, label = "Supply"), color = "Black", angle = 31,
          size= 4, alpha=0.1) +
  theme_bw() +
  labs(title = "Figure 1 - Linear Demand And Supply Model Under
       Perfect Conditions",
       x = "Quantity",
       y = "Price") +
  scale_x_continuous(labels = c("0", "", "Q*", "", "")) +
  scale_y_continuous(labels = c("0", "", "P*", "", "")) +
  theme(
  panel.grid.minor=element_blank(),plot.background=element_blank())
fdemandsupplyfinal
```

Figure 1 − Linear Demand And Supply Model Under Perfect Conditions




The linear demand curve from figure 1 displays how the consumers behave on a perfect market. In this model, a maximum price exists at which no more is consumed. This model also unrealistically assumes that free goods are consumed infinitely often. On the other side, the linear supply curve from figure 1 illustrates the suppliers point of view. The model concludes that every supplier can offer his good for the maximum price. Vice versa, the more the price falls, the fewer suppliers can still offer their good. In natural competition, suppliers are forced to align their prices. This is due to the fact that each supplier wants to maximize its turnover, so it is not worthwhile for the suppliers who can offer the lowest price to actually offer the lowest price. In long term, the actual market price will converge to P* and every supplier that can not bid for P* will disappear. Finally, P* defines the equilibrium price and Q* the equilibrium quantity.

#< quiz "Mark_Surplus" question: Which of these marked triangles represents the consumer surplus? sc: - A.* - B. - C. success: Great, your answer is correct! failure: Try again. #>

**Price Elasticity**

As mentioned earlier, figure 1 represents a severe simplification of a complex market. For instance, the fiscus also impacts market activity by implementing price caps, by subsidizing various branches or by ensuring that no so-called price cartels are formed. Price cartels are an association of organizations closing price agreements for particular goods to bypass the competition. Usually, the course of the supply and demand curve is not linear and the conditions for a perfect market are not satisfied. The slope of the logarithmized demand or supply curve depends on the so-called **Price Elasticity**. For the demand side, price elasticity indicates how demand reacts on changes in prices relatively. The formula looks as follows:

$$\epsilon = (\Delta Q/Q_{\text{old}})/(\Delta P/P_{\text{old}}) = \text{Relative change in the quantity of goods demanded/Relative change in price}$$

Hence, price elasticity delivers a value $\varepsilon := [0, \infty]$ that can also show whether price increases would be

worthwhile. An elasticity of $\varepsilon = 1$ tells us that a price increase of one percents implies a demand decrease of one percent and represents the maximum turnover point for suppliers. As a result, elasticities of $\varepsilon < 1$ are inelastic and otherwise elastic.

#< quiz "Elastic_orInelastic" question: A bar owner wants to maximize his turnover. In a particular period of time, you have found out that the demand on beer is linear and a price increase from three Units to four Units implies a decreasing sales quantity of 1000 quantity units to 875 quantity units. Which of the following responses is correct? sc: - The price is inelastic - You should not increase the price. - The price is elastic - You should increase the price by 47 percent. - The price is inelastic - You should increase the price by 34 percent. - The price is inelastic - You can double the price.*

success: Great, your answer is correct! failure: Try again.

#>

#< info "How to create graphs with the package `ggplot2`."

The `ggplot` package allows you to create diagrams based on a data set you provide. The structure for the most important functions is as follows:

```
#General function to initialize ggplot2 and aes() is used to define mappings
#between variables and visual properties.
ggplot(data = data_set, aes(y = y, x = x)) +
#Family function for geometric objects to be added (e.g. geom_bar, geom_line,
#geom_density, geom_histogramm, geom_text,...)
geom_line(aes(y=y, x=x), color = "color", size = 2, ...) +
#To create multiple panels within one plot
facet_grid(~variable_to_split) +
#Function to set the labels of the plot as x-axis, y-axis, title
labs() +
#Define the scales for the axis
xlim()/ylim() +
#Function to modify the overall appearance of the plot
theme() +
#Customizing legends and guides for color, size, ...
guides() +
#Change the limits or designations for x-axis or y-axis
scale_x_continuous() +
scale_y_continuous()
```

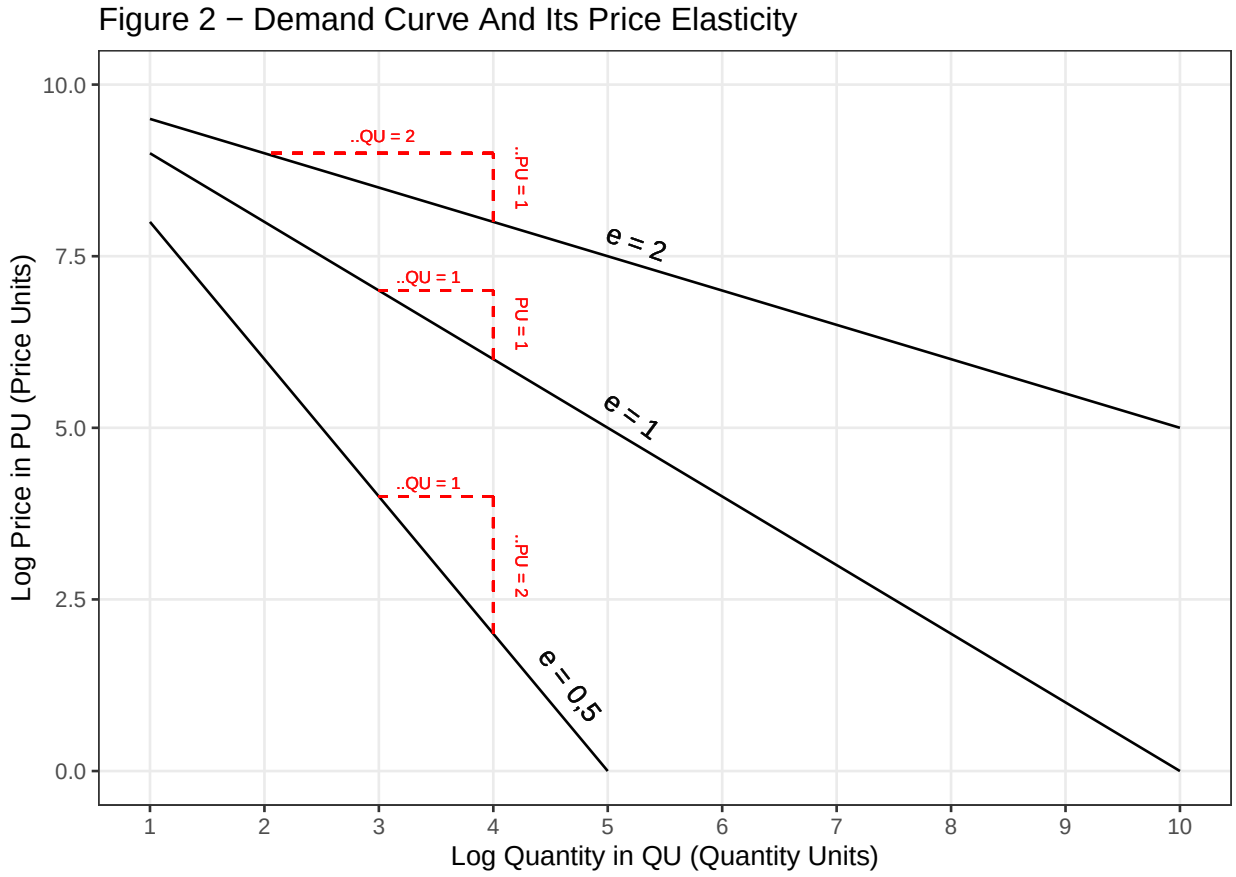
**Note:** Add a `+` after these functions to connect them (except of the last function). #>

Price elasticity can be represented graphically. The following graph illustrates the difference between elasticities in demand.

**Task:** Create the fictive data set `DatasetPE`. Then assign the axes correctly, using the top graph as a guide. `Check` your results.

```
library(tidyverse)
#create a fictive data set
DatasetPE <- data.frame(
  Price = 1:10,
  Demand_05 = 10 - 0.5*1:10,
  Demand_1 = 10 - 1*1:10,
  Demand_2 = 10 - 2*1:10
)
colnames(DatasetPE) <- c("Price", "  > 1", "  = 1", "  < 1")
```

```
#change to long format
DatasetPEN <- DatasetPE %>%
  pivot_longer(cols = starts_with(" "), names_to = " ",
               values_to = "Quantity_Value")
#read in the pre-created graph `PE`
readRDS("material/PE.RDS") +
  labs(title = "Figure 2 - Demand Curve And Its Price Elasticity",
       x = "Log Quantity in QU (Quantity Units)",
       y = "Log Price in PU (Price Units)") +
  xlim(0, 10) +
  ylim(0, 10) +
  scale_x_continuous(breaks = 0:10) +
  theme(panel.grid.minor=element_blank(),plot.background=element_blank())
```

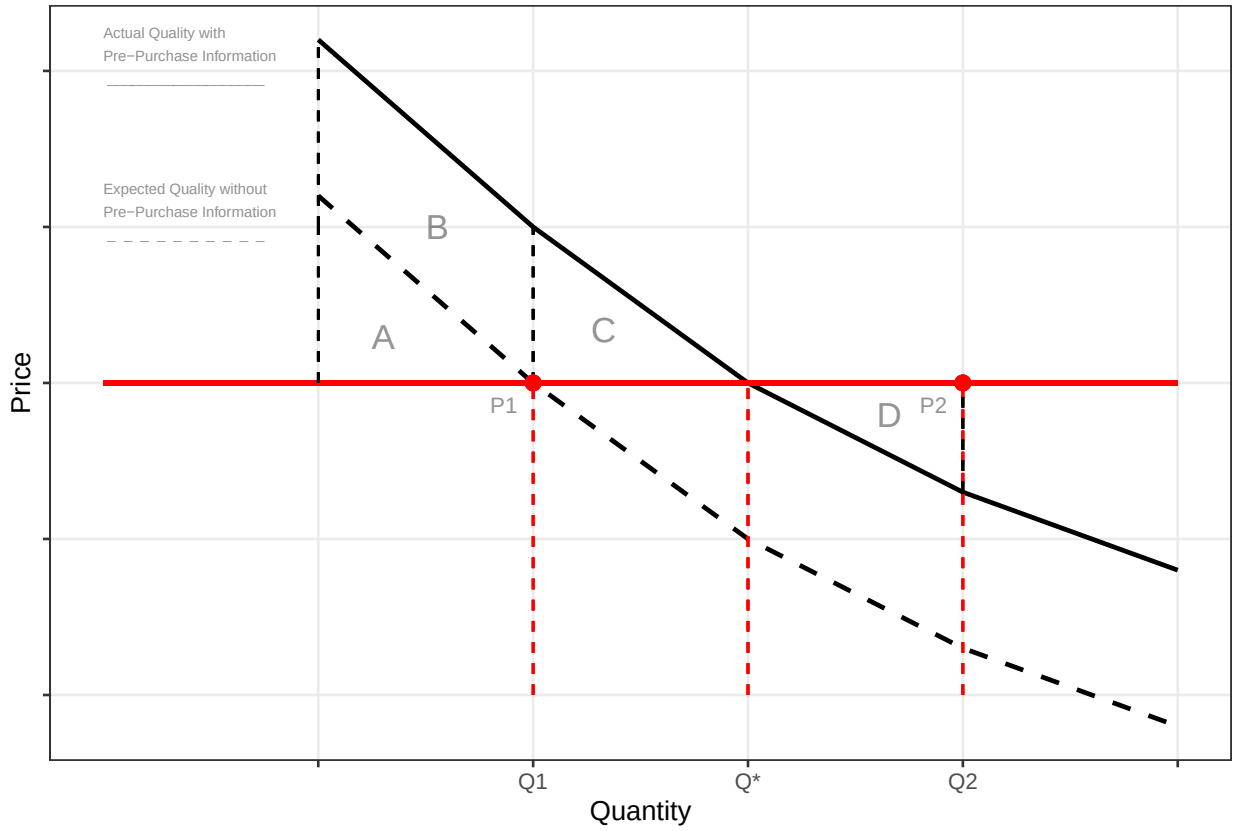



Returning to the subject of pre-purchase information, a clear classification of these is needed. Alan T. Sorensen (2007) and Kenneth Train (2015) contend that pre-purchase information alters perception of the goods quality and that a distinction between anticipated ex ante utility and experienced ex post utility is necessary to measure the effect of this information on welfare. In simpler terms, we assume that a consumer has a different utility depending on the availability of pre-purchase information. According to the authors (Reimers and Waldfogel, 2021, pp. 1949), consumers may face three different situations depending on the expected quality of the books.. The consumer can expect lower quality than the actual quality (1) ( $\bar{R}_j < R_j$ ), the same quality (2) ( $\bar{R}_j = R_j$ ) or a higher expected quality than the actual quality (3) ( $\bar{R}_j > R_j$ ).

**Task:** Run the following chunk to illustrate how pre-purchase information relates to welfare.

```
#Create a fictive data set
yPredQualityPrice <- c(1.6,1.0,0.50,0.15,-0.10)
xQuan <- c(1,2,3,4,5)
yRealQualityPrice <- c(2.1,1.5,1,0.65,0.4)
xGroup <- c(1, 1, 1, 1, 1)
DatasetTest <- data.frame(Pred_Price = yPredQualityPrice, Quan = xQuan,
                          Real_Price = yRealQualityPrice, Group = xGroup)
#Read in the plot
WE <- readRDS("material/WE.RDS")
WE
```

Figure 3 − The Economic Value Added Of Pre−Purchase Information




#< quiz "Quality_Line" question: What quality expectation does the dashed line reflect? sc: - 1.* - 2. - 3. success: Great, your answer is correct! failure: Try again. #>

As visualized in the figure 3, the consumer would choose quantity Q1 for ( $\bar{R}_j < R_j$ ), quantity Q* for ( $\bar{R}_j = R_j$ ) and quantity Q2 for ( $\bar{R}_j > R_j$ ). For choosing Q1, the consumer would expect a surplus of triangle A while getting an actual surplus of A + B (ex post recognition of actual quality). A consumer with access to pre-purchase information would always choose Q* with a surplus of A + B + C. Overestimating the books quality, the consumer would buy quantity Q2 and obtains a surplus of A + B + C - D. The quality expectations affect the overall utility of the consumer and thus the demand curve appears to shift without impacting the price elasticity. Hence, the curve merely adjusts to the actual demand.

#< quiz "ValueAdded_Triangle" question: Which triangle represents the value added of pre-purchase information? sc: - A - B - C* - D success: Great, your answer is correct! failure: Try again. #>

**Summary**

After looking at figure 1, we have learned to distinguish between demand and supply, to categorize the economic meaning of welfare and the composition of consumer / supplier surpluses and the formation of equilibrium values. Focusing the demand side, we got an insight into the topic of price elasticity to better understand the following examination of price elasticities on books. Hence, after collecting the necessary knowledge, we returned to pre-purchase information and their affects on welfare. We found out that pre-purchase information impacts the expectation for quality and leads to an adjustment of the demand curve.

In the following chapter 2, we will get an introduction to the Amazon data set followed by descriptive analyses.

# 2. Data and Descriptive Insights

The underlying examination by Reimers and Waldfogel is based on a data set provided by Amazon.

In the first part, we deepen our understanding and the origin of the underlying data set and define important attributes that we later use for descriptive and empirical measurements.

Second, we focus on professional reviews and crowd ratings to investigate superficial contexts. To illustrate them, we use descriptive tables.

Finally, we create an entire overview about descriptive analyses to detect potential effects between professional and non-professional reviews on sales ranks and prices to create a transition to the following empirical part of my Thesis.

After working through this chapter, you are surefooted in dealing with the underlying data set and thus well prepared to continue with the empirical part in chapter 3.

**Structure**

2.1. Introduction to the data set

2.2. Analysis of Amazon Star Ratings

2.3. Analysis of Professional Reviews

## 2.1. Introduction to the Data Set

As previously mentioned, the entire quantitative calculations are based on a data set provided by Amazon. The data set includes slightly less than 8.8 million observations of non-professional crowd ratings. One observation represents one sales rank for one day between 02-01-2018 and 31-12-2018 for a individual book in one country. The sales ranks are incomplete for some books, which means that sales ranks are not available for every day for these books. Observations were made in the USA, in Great Britain and in Canada. In addition, the data set was merged with newspaper data from every magazine listed in 1.1 with information on whether the newspaper reviewed the particular book and whether the New York Times recommended the book. In fact, only 3.22 million observations were taken into consideration due to missing values, which reduces the actual market share we have examined. Reimers and Waldfogel claim that Amazon covered about 44.5 percent of the physical book market share in 2017, which I cannot accurately confirm. After removing about 63% of the data set, 44.5 percent cited by Reimers and Waldfogel means that about 16 percent of the book market is still covered by the Amazon data set. To improve the problem sets performance, we have already removed observations that are not relevant to our examination.

To better understand the data set, we categorize some important variables and take a look at an excerpt of it.

#< info "How to read in data sets with readRDS()."

`readRDS()` enables you to read in data sets from defined working directories. The advantage of RDS files is that they are R files, which can be read much faster and are more compact than, for example, csv or dta files. Furthermore, RDS files can also store alternative data, such as regressions or three-dimensional data. To read in, the structure is as follows:

```
data_set <- readRDS("find/your/path/data_set.RDS")
```

The data set is stored in `data_set`. If you want to save RDS files, the structure is as follows:

```
data_set <- saveRDS(data_set, file = "find/your/path/data_set.RDS")
```

**Note:** Originally, the Amazon data set was five gigabytes in size as a dta file. After converting this data set, I saved this file as RDS and now this record is less than 100 megabytes.

#>

**Task:** Use the function `readRDS` to load the data set called `dataEst.RDS`. Save this data set under the name `data`.

**Note:** The data sets are stored in the file `material`.

```
data <- readRDS("material/dataEst.RDS")
```

You successfully loaded the data set. Now, we want to see how the data set is organized and which values the single attributes can take.

**Task:** Use the function `head` to show the first rows from `data`. In addition, use the function `colnames` to list every column name from `data`. I only show the first ten columns to avoid an unsightly illustration.

```
#show the first ten columns and the first five rows
head(data[,c(1:10)])
```

```
##         asin country      ddate pelapse            genre numbooks est_sample
## 1 0001712713      GB 2018-01-02   13335 JUVENILE FICTION       76          1
## 2 0001712713      GB 2018-01-17   13350 JUVENILE FICTION       76          1
## 3 0001712713      GB 2018-01-18   13351 JUVENILE FICTION       76          1
## 4 0001712713      GB 2018-01-19   13352 JUVENILE FICTION       76          1
## 5 0001712713      GB 2018-01-26   13359 JUVENILE FICTION       76          1
## 6 0001712713      GB 2018-01-27   13360 JUVENILE FICTION       76          1
##     rank pamzn pnew
## 1  29374  4.11 2.98
## 2  38581  3.97 2.81
## 3 113860  3.96 2.79
## 4  57788  3.95 2.78
## 5  57582  4.81 2.97
## 6  37539  4.80 2.97
```

```
# show the column names of the data set
colnames(data)
```

```
##   [1] "asin"          "country"       "ddate"         "pelapse"
##   [5] "genre"         "numbooks"      "est_sample"    "rank"
##   [9] "pamzn"         "pnew"          "pused"         "R"
##  [13] "review"        "lrank"         "lpamzn"        "lR"
##  [17] "lreview"       "lrR"           "DUSAT"         "DNYT"
##  [21] "NYTDATE"       "NYT_elapse"    "drecommended"  "DBG"
##  [25] "BGDATE"        "BG_elapse"     "DCHI"          "CHIDATE"
##  [29] "CHI_elapse"    "DLAT"          "LATDATE"       "LAT_elapse"
##  [33] "DWAPO"         "WAPODATE"      "WAPO_elapse"   "DWSJ"
##  [37] "WSJDATE"       "WSJ_elapse"    "DPW"           "DGR"
##  [41] "OTHDATE"       "OTH_elapse"    "dnytpost1_5"   "dnytpost6_10"
##  [45] "dnytpost"      "dnytpost10"    "dnytpostpre"   "dothpost"
##  [49] "dothpost10"    "dothpostpre"   "epos"          "epos2"
##  [53] "epos3"         "eneg"          "eneg2"         "eneg3"
##  [57] "ano"           "cno"           "titleno"       "canum"
##  [61] "gno"           "pubno"         "L1.lrank"      "dnytpost1"
##  [65] "dnytpost6"     "DNYT0"         "DNYT1"         "DNYT2"
##  [69] "DNYT3"         "DNYT4"         "DNYT5"         "DNYT6"
##  [73] "DNYT7"         "DNYT8"         "DNYT9"         "DNYT10"
##  [77] "DNYT11"        "DNYT12"        "DNYT13"        "DNYT14"
##  [81] "DNYT15"        "DNYT16"        "DNYT17"        "DNYT18"
##  [85] "DNYT19"        "DNYT20"        "DNYT21"        "DNYT22"
##  [89] "DNYT23"        "DNYT24"        "DNYT25"        "DNYT26"
##  [93] "DNYT27"        "DNYT28"        "DNYT29"        "DNYT30"
##  [97] "DNYT31"        "DNYT32"        "DNYT33"        "DNYT34"
## [101] "DNYT35"        "DNYT36"        "DNYT37"        "DNYT38"
## [105] "DNYT39"        "DNYT40"        "DNYTm1"        "DNYTm2"
## [109] "DNYTm3"        "DNYTm4"        "DNYTm5"        "DNYTm6"
## [113] "DNYTm7"        "DNYTm8"        "DNYTm9"        "DNYTm10"
```

```
## [117] "DNYTm11"        "DNYTm12"        "DNYTm13"        "DNYTm14"
## [121] "DNYTm15"        "DNYTm16"        "DNYTm17"        "DNYTm18"
## [125] "DNYTm19"        "DNYTm20"        "DOTH0"          "DOTH1"
## [129] "DOTH2"          "DOTH3"          "DOTH4"          "DOTH5"
## [133] "DOTH6"          "DOTH7"          "DOTH8"          "DOTH9"
## [137] "DOTH10"         "DOTH11"         "DOTH12"         "DOTH13"
## [141] "DOTH14"         "DOTH15"         "DOTH16"         "DOTH17"
## [145] "DOTH18"         "DOTH19"         "DOTH20"         "DOTH21"
## [149] "DOTH22"         "DOTH23"         "DOTH24"         "DOTH25"
## [153] "DOTH26"         "DOTH27"         "DOTH28"         "DOTH29"
## [157] "DOTH30"         "DOTH31"         "DOTH32"         "DOTH33"
## [161] "DOTH34"         "DOTH35"         "DOTH36"         "DOTH37"
## [165] "DOTH38"         "DOTH39"         "DOTH40"         "DOTHm1"
## [169] "DOTHm2"         "DOTHm3"         "DOTHm4"         "DOTHm5"
## [173] "DOTHm6"         "DOTHm7"         "DOTHm8"         "DOTHm9"
## [177] "DOTHm10"        "DOTHm11"        "DOTHm12"        "DOTHm13"
## [181] "DOTHm14"        "DOTHm15"        "DOTHm16"        "DOTHm17"
## [185] "DOTHm18"        "DOTHm19"        "DOTHm20"        "dnytpost1_1"
## [189] "dnytpost6_1"    "dnytpost10_1"   "dnytpostpre_1"  "dothpost_1"
## [193] "dothpost10_1"   "dothpostpre_1"  "dnytpost1r_1"   "dnytpost6r_1"
## [197] "dnytpost10r_1"  "dnytpostprer_1" "dnytpost1_2"    "dnytpost6_2"
## [201] "dnytpost10_2"   "dnytpostpre_2"  "dothpost_2"     "dothpost10_2"
## [205] "dothpostpre_2"  "dnytpost1r_2"   "dnytpost6r_2"   "dnytpost10r_2"
## [209] "dnytpostprer_2" "dnytpost1_3"    "dnytpost6_3"    "dnytpost10_3"
## [213] "dnytpostpre_3"  "dothpost_3"     "dothpost10_3"   "dothpostpre_3"
## [217] "dnytpost1r_3"   "dnytpost6r_3"   "dnytpost10r_3"  "dnytpostprer_3"
## [221] "DOTH"           "DALL"           "A"              "B"
## [225] "sigma"          "sigma_se"       "gamma1"         "alpha1"
## [229] "e1"             "q1_1"           "L1.lR"
```

As can be seen, the main data set consists of 231 variables. Not every one of them is essential for the following course and will not be considered further. For those to whom this does not apply, there is an explanation below.

**Identification Variables**

The variable `asin` represents the corporate Amazon ID to differentiate between different products. `Country` indicates whether the underlying valuation belongs to the US market (US), the UK market (UK) or the Canadian market (CA). `canum` represents the main ID and combines these variables to create a powerful country-dependent identifier. A title-author identifier is stored in the variable `titleno`. Identification variables are indispensable for each data set to clearly distinguish each observation. In this data set, the combination of `ddate` (see below) and `canum` defines each unique observation.

**Chronological variables**

`ddate` indicates the main date on which the crowd ratings were created, to which all other chronological variables refer. `NYT_elapse` (New York Times), `BG_elapse` (Boston Globe), `CHI_elapse` (Chicago Tribune), `LAT_elapse` (Los Angeles Times), `WAPO_elapse` (Washington Post), `WSJ_elapse` (Wall Street Journal) and `OTH_elapse` (all magazines except NYT) indicate how many days have passed since/until the publication of the individual professional review. As a reference date serves here `ddate`. `epos` (if `pubno` $> 0$) and `eneg` (if `pubno` $< 0$) display counting days since publication.

**Value Variables**

The variable `rank` represents the Amazon sales rank, based on which we will do most of the calculations. `pamzn` represents the price of the particular book while `R` provides the given star rating on a five-point scale. `review` delivers the number of reviews in total, so this variable delivers the same number for each

specification of `canum`. Each of these attributes occur twice, once each with a preceding "l". This means that the values are logarithmized.

**Dummy Variables**

Dummy variables are binary variables that can only take two values (here: 1 and 0). Any variable starting with `dnytpost` indicates "1" if the New York Times has published reviews within the period defined by the following numbers after `dnyt`. For instance, `dnytpost1_5` takes the value 1 if [0 < `NYT_elapse` < 6]. For `dnytpost` and `dnytpost10`, [0 <`NYT_elapse` < 10] and [11 <`NYT_elapse` < 20] holds. `dnytpostpre` takes the value 1 if [-10 < `NYT_elapse` < 20]. The same principle applies to all variables starting with `dothpost` where the variable `OTH_elapse` defines the base instead of `NYT_elapse`.

In order to convert the sales figures from our main data set into quantities, the authors provided confidential data for determining the quantities. These data were provided by the Nielsen Company (a market research firm), and Reimers and Waldfogel published them only in an already edited form. The aforementioned confidential data sets provide information on the weekly 100 best-selling books between 2015 and 2018, while the accessible confidential data only provide intermediate calculations to determine price elasticities.

**Note:** Many of the transformations from the raw data to the data on which the studies are based have already been added to the data set to improve the performance of this problem set.

**Summary**

With the underlying data set, we have not only a very large set of observations in a given time period, but also a large amount of information in the form of variables. Particularly in this case, where an already merged and edited data set has been published, this large amount of information may also have a detrimental effect on the correct replication of this study. Nevertheless, we continue with the data set provided by Reimers and Waldfogel. In this chapter, we have gained insight into the data set by running initial code chunks in R and the key variables that will guide us throughout the study. Finally, we have learned how to define dummy variables and the importance of uniquely identifying variables.

In chapter 2.2, we will perform initial calculations on the data set by constructing descriptive statistics that focus on Amazon ratings.

## 2.2. Analysis of Amazon Star Ratings

In the following part, we will analyse data to gain a deeper understanding for Amazon star ratings. For comparison purposes, the average prices and sales rank are shown to illustrate how they change based on different star ratings on Amazon.

**Task:** Check the following chunk to read in the main data set.

```
data <- readRDS("material/dataEst.RDS") %>%
  arrange(canum, ddate)
```

#< info "How to work with `group_by`, `summarise` and other functions using the `dplyer` package."

The package `dplyer` is part of the package `tidyverse` and is used for data manipulation. Several functions from this package are used in the further course of this problem set. First, load the tidyverse package

```
library(tidyverse)
```

Then, some important functions with individual explanation are listed below.

```
# Starting sequence
data_set  <- data_set_manipulated %>%
 # Use filter() to filter out observations according to certain conditions
  filter(x < 1 & x > 3)
 # Use group_by() to aggregate the data set to one or more variables
  group_by(variable1, variable2, ...)
# Use summarise() associated with aggregation functions such as mean() or
# sum() to calculate new attributes
  summarise(new_column = mean(x),
            New_column2 = sum(x),
            ...)
 # Use mutate() to add customized columns
  mutate(new_column_Ratio = x / y)
 # Use arrange() to arrange rows by specific variables (use a "-" to order them
# in ascending order)
  arrange(column_id, column_date)
```

More information about the `dplyer` package you find here.

**Note:** The function `rbind()` is not part of the `dplyer` package but also used for data manipulation. This function connects the rows of two or more data sets with each other. The condition is matching column names.

#>

Focusing the Amazon star ratings in general, let us first turn to the structure of these ratings. In this context, we create a density plot to display how the number of books is distributed on this five-point scale.

**Task:** The aim is to create a line chart showing the considered distribution. In this context, try to fill the gaps. Do not forget to press `Check`.

```
#Create the data set `DesCRDensity`
DesCRDensity <- data %>%
  group_by(R) %>%
  summarise(Number_Ratings = n_distinct(titleno))
```

```
#Create a density plot `DesCRDensity`
DesCRDensity %>%
  ggplot(aes(x = R, y = Number_Ratings/sum(Number_Ratings))) +
  geom_line(linetype = "dashed") +
  geom_ribbon(aes(ymin = 0, ymax = Number_Ratings/sum(Number_Ratings)),
              fill = "blue", alpha = 0.2) +
  theme_bw() +
  labs(title = "Figure 4 - Distribution Of Books Among Amazon Star Ratings",
       x = "Amazon Star Ratings",
       y = "Book Density") +
  guides(fill=guide_legend(title="")) +
  theme(legend.position = "bottom",
  panel.grid.minor=element_blank(),plot.background=element_blank())
```


Figure 4 shows a leftward skew, which means that less well-rated books are much less common than the other way around. Apparently, books are generally well-rated on Amazon leading to the assumption that purchaser are more likely to buy their preferred books due to an information advantage.

#< info "How to use `ifelse()` to generate new columns."

The `ifelse` function is predestined to generate dummy-variables. This function allows you to create nested values that assume certain expressions for different conditions. The structure is as follows:

```
condition1 <- x >= 1
data_set$new_column <- ifelse(condition1, Value_if_condition_is_met (1),
                              Value_if_condition_not_met (0))
```

In this case, the value of `new_column` takes the value 1 if $x \geq 1$. Otherwise, the value for `new_column` take the value 0.

**Note:** The symbol `|` between several conditions stands for the logical OR and `&` for the logical AND. Use `==` instead of `=` within the condition formulation.

#>

Now, let us divide Amazon star ratings into three categories: less than three stars, between three and four stars, and more than four stars.

**Task:** Create three dummy variables that show `1` when the star rating is between one and three, three and four and more than five. Orientate on exercise `2.2.1` to create a dummy variable. Then, create three filtered data sets to differentiate between these three categories. Name these data sets `dataCR1_3`, `dataCR3_4` and `dataCR4_5`.

**Note:** Professionally reviewed books gets filtered out to focus on star ratings that could not affected by professional reviews.

```
#create three dummy variables to differentiate every category
data$starRating1_3 <- ifelse(data$R >= 1 & data$R < 3, 1, 0)
data$starRating3_4 <- ifelse(data$R >= 3 & data$R < 4, 1, 0)
data$starRating4_5 <- ifelse(data$R >= 4 & data$R < 5, 1, 0)

dataCR1_3 <- data %>%
  filter(data$starRating1_3 == 1 & DALL == 0 & cno == 3) %>%
  summarise(AVGPrice = mean(pamzn),
            AVGSalesRank = mean(rank, na.rm = TRUE))

dataCR3_4 <- data %>%
  filter(data$starRating3_4 == 1 & DALL == 0 & cno == 3) %>%
  summarise(AVGPrice = mean(pamzn),
            AVGSalesRank = mean(rank, na.rm = TRUE))

dataCR4_5 <- data %>%
  filter(data$starRating4_5 == 1 & DALL == 0 & cno == 3) %>%
  summarise(AVGPrice = mean(pamzn),
            AVGSalesRank = mean(rank, na.rm = TRUE))
```

In the following, we continue to summarize the three data sets into one data set.

**Task:** Execute the following chunk to create the needed data set. Press `check` to confirm.

```
#create the data set to aggregate the price and sales rank information
dataCRPrice <- data.frame(Price = c(dataCR1_3$AVGPrice, dataCR3_4$AVGPrice,
                                    dataCR4_5$AVGPrice),
                          SalesRank = c(dataCR1_3$AVGSalesRank,
                                        dataCR3_4$AVGSalesRank,
                                        dataCR4_5$AVGSalesRank),
                          Category = c("1-3 stars", "3-4 Stars",
                                       "more than 4 stars"))
dataCRPrice
```

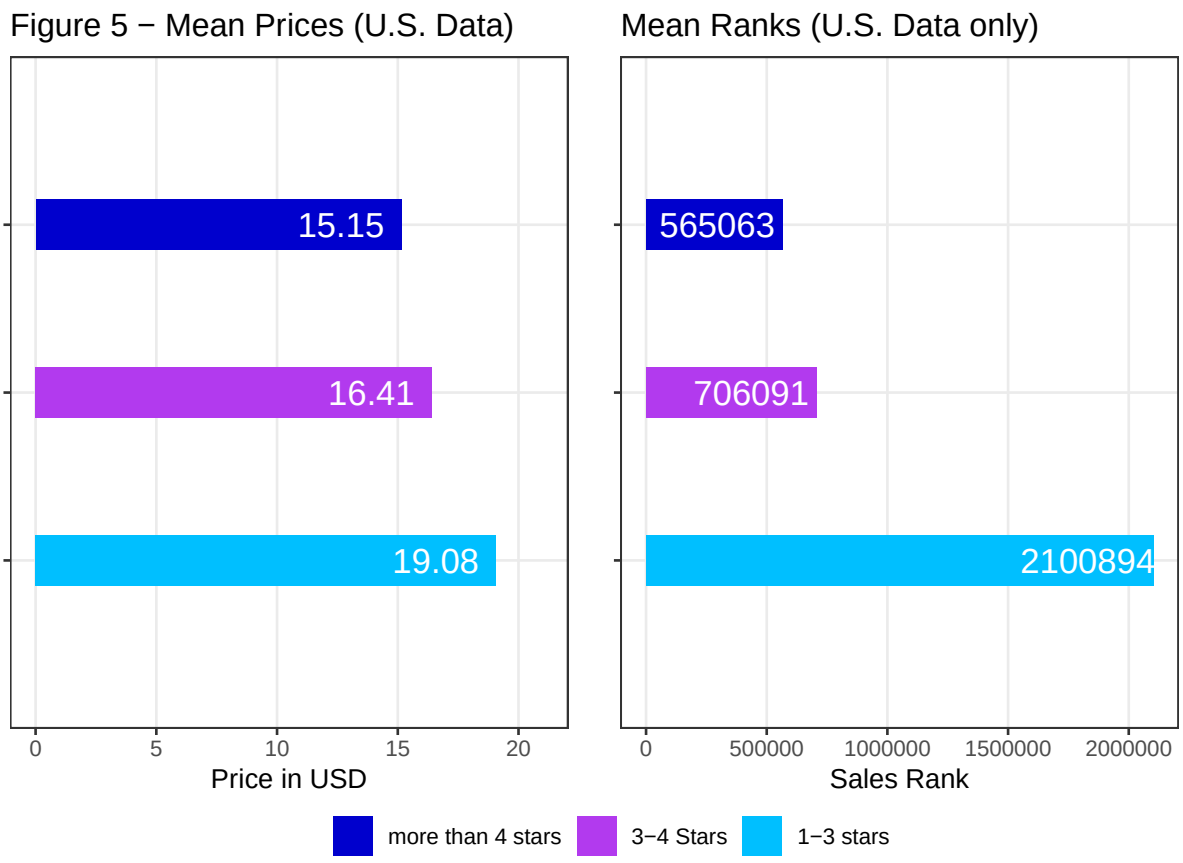
```
##      Price SalesRank          Category
## 1 19.07730 2100894.4         1-3 stars
## 2 16.41147  706090.9         3-4 Stars
## 3 15.15168  565062.9 more than 4 stars
```

As seen, we have calculated three different mean prices and sales ranks for the three different categories. Edit the next chunk to visualize these differences graphically.

**Note:** Data including information about the occurrence of professional reviews has been filtered out due to not contaminate the statistics.

**Task:** Create `f5` to create a chart that shows the average prices according to their star ratings. Use `ylim` to define the y-axis scale between 0 and 21 (note that the diagram has been flipped). Do not forget to press `Check`. For more information about the `patchwork` package you find here.

```
#load the package `patchwork`
library(patchwork)
#figure 5: Mean price
f5 <- dataCRPrice %>%
   ggplot() +
   geom_histogram(aes(y = Price, x = Category,  fill = Category),
                  stat = "identity", width = 0.3) +
  geom_text(aes(x = Category, y = Price - 2.5, label = round(Price, 2)),
            color = "white", size = 5) +
   coord_flip() +
   scale_x_discrete(expand = c(0, 1)) +
   theme_bw() +
   ylim(0, 21) +
 scale_fill_manual("", values = c("1-3 stars" = "deepskyblue",
                                  "3-4 Stars" = "darkorchid2",
                                  "more than 4 stars" = "blue3")) +
  labs(title = "Figure 5 - Mean Prices (U.S. Data)",
       x = "",
       y = "Price in USD") +
  guides(fill = guide_legend(reverse = TRUE)) +
   theme(legend.position = "none",
     axis.text.y=element_blank(),
   panel.grid.minor=element_blank(),plot.background=element_blank())

#figure 6: Mean Sales Ranks
f6 <- readRDS("material/f6.RDS")

combined_f5_f6 <- f5 + f6 + plot_layout(ncol = 2) +
  theme(legend.position = c(-0.03, -0.16), legend.direction = "horizontal") +
  theme(plot.margin = margin(t = 0, b = 1, unit = "cm"))
combined_f5_f6
```




The "better" the books are rated, the more the price of those books goes down. One possible explanation would be that higher crowd ratings on Amazon could potentially resulting in long-term increase in demand, which leads to more contested competition, where everyone has to reduce their prices. In fact, book pricing depends on more factors than are included in our data set.

The mean sales rank decreases the more stars were given. As we noticed in figure 4, more than 75 percent of all observations were rated with four stars or higher, so we can assume a similar percentage of observations in the upper bar (565063). Nevertheless, the mean sales rank for books recommended by the New York Times is significantly lower than the mean sales rank of books rated more than four stars by the crowd.

**Summary**

In summary, we have gained deeper insights into Amazon star ratings and their descriptive impact on average prices and average sales ranks. First, a density plot (figure 4) was considered in order to evaluate the distribution of the books among the star ratings based on it. Overall, the ratings turned out to be relatively high, as a few books were rated worse than 3.5 stars. Below, we have created three filtered data sets for the median price and median sales rank of books offered on Amazon that received a rating between one and three stars, three and four stars, and more than four stars. We have combined the data into one data set and plotted the statistics in two bar charts (figure 5 and 6). Finally, We have found higher prices for less well-rated books and lower sales ranks for well-rated books

In chapter 2.3, we will focus on the analysis of professional reviews.

## 2.3. Analysis of Professional Reviews

In this chapter 2.3, we focus on descriptive statistics concerning the occurrence of professional reviews. In contrast to crowd ratings from Amazon, the data set only provides information about the occurrence of these reviews and about the information the review contains. Without this information, we can use dummy variables to track these professional reviews and put them in context over time.

**Task:** Check the following chunk to read in the main data set.

```
data <- readRDS("material/dataEst.RDS") %>%
  arrange(canum, ddate)
```

Two dummy variables are added below to distinguish between the occurrence of all reviews from all magazines and from non-New York Times magazines.

**Task:** Create these variables that indicate `1` when an observation belongs to a book that has been reviewed. Also, read in the data set `DesRat` and save it accordingly, and then `check` the chunk.

**Note:** In sum, the underlying data set includes data about **8770 books**. `DOTH` should contain reviews from at least one non-New York Times magazine and `DALL` from at least one of all magazines.

```
#create two new dummy variables to indicate professional reviews from
#non-NYT Magazines and from all Magazines
data$DOTH <- ifelse(data$DBG == 1 | data$DCHI == 1 | data$DLAT == 1 |
                      data$DWAPO == 1 | data$DWSJ == 1 , 1, 0)
data$DALL <- ifelse(data$DBG == 1 | data$DCHI == 1 | data$DLAT == 1 |
                      data$DWAPO == 1 | data$DWSJ == 1 | data$DNYT == 1 , 1, 0)
DesRat <- readRDS("material/DesRat.RDS")
```

**Task:** Press `check` to visualize the percentage of books reviewed by magazines.

```
library(kableExtra)
#Create row names
rownames(DesRat) <- c("Relative share", "Absolut share")
#use of kbl() function to create a table
t1 <- DesRat  %>%
kbl(col.names = c("Share_of_NYT_Ratings" = "New York Times",
                  "Share_of_BG_Ratings" = "Boston Globe",
                  "Share_of_CHI_Ratings" = "Chicago Tribune",
                  "Share_of_LAT_Ratings" = "Los Angeles Times",
                  "Share_of_WAPO_Ratings" = "Washington Post",
                  "Share_of_DWSJ_Ratings" = "Wall Street Journal",
                  "Share_of_OTH_Ratings" = "Non New York Times",
                  "Share_of_ALL_Ratings" = "All"), caption =
      "Share Of Professional Reviewed Books") %>%
  kable_paper(full_width = TRUE)
t1
```

In total, 12.66 percent of all books (1521) included in the data set have been professionally reviewed. 11 percent (1315) of these books were reviewed by the New York Times, representing the largest percentage of books reviewed by professionals. The second largest percentage provides the Chicago Tribune with only 1.36 percent (139), which is approximately ten times less than the number of reviews published by the New York Times. Overall, non-New York Times magazines account for nearly 3.1 percent of the share.

#< info "How to calculate with descriptive formulas in R"

R provides multiple formulas to calculate simple descriptive statistics. The R implementations is as follows:

Table 1: Share Of Professional Reviewed Books

| | New York Times | Boston Globe | Chicago Tribune | Los Angeles Times | Washington Post | Wall Street Journal | Non New York Times | All |
|---|---|---|---|---|---|---|---|---|
| Relative share | 11% | 0.58% | 1.36% | 0.4% | 0.67% | 0.56% | 3.06% | 12.66% |
| Absolut share | 1315 | 83 | 139 | 45 | 87 | 73 | 370 | 1521 |

- mean(x) - Calculates the mean of vector or column $x$.
- sum(x) - Calculates the sum of vector or column $x$.
- quantile(x, probs = 0.25) - Calculates the 25 percent quantile of vector or column $x$.
- NROW(data_set) - Calculates the number of rows of a data set.
- length(x) - Calculates the length of vector $x$ (or the number of columns of data_set).
- n_distinct(x) - Calculates the distinct number of observations of vector or column $x$.

**Note:** The `rownames()` function adds or edits the row names of a data set.

#>

Continuing with an overview distinguishing different countries, we want to create a table including simple descriptive values distinguished by `country`.

**Task:** Create `dataDesKript` to create a descriptive chart of specific variables. When creating this chart, distinguish between data from the USA, Canada and the UK. Also read in `dataDesKript2` to add information about all countries. `Check` your result.

```
#create `dataDesKript` to aggregate statistics based on different countries
dataDesKript <- data %>%
   group_by(country) %>%
   summarize(Price = round(mean(pamzn),2),
             Star_Rating = round(mean(R),2),
             Sales_Rank = round(mean(rank),2),
             Number_of_Ratings = round(mean(review), 2),
             Teenth = quantile(R, probs = 0.1, na.rm = TRUE),
             Tweentyfifth = quantile(R, probs = 0.25, na.rm = TRUE),
             Fiftith = quantile(R, probs = 0.5, na.rm = TRUE),
             Seventyfifth = quantile(R, probs = 0.75, na.rm = TRUE),
             Ninetith = quantile(R, probs = 0.9, na.rm = TRUE),
             Titles = n_distinct(titleno),
             Observations = NROW(country),
             Editions = n_distinct(asin))
#read in dataDesKript2.RDS
dataDesKript2 <-readRDS("material/dataDesKript2.RDS")
#merge Dataframes
DataDes <- rbind(dataDesKript, dataDesKript2)
#transform Dataframe into a clearer schema
DataDesTest <- t(DataDes)
colnames(DataDesTest) <- rownames(DataDes)
DataDescriptive <- as.data.frame(DataDesTest)
colnames(DataDescriptive) <- unlist(DataDescriptive[1,])
DataDescriptivefinal <- DataDescriptive[-1,]
#change Row Names
row.names(DataDescriptivefinal) = c("Price", "Star rating", "Sales rank",
                                    "Number of ratings", "10th", "25th",
```

```
                                    "50th", "75th", "90th", "Titles",
                                    "Observations", "Editions")
DataDescriptive[c(2:13),]
```

```
##                          CA       GB       US      All
## Price                 21.07    13.12    15.85    16.42
## Star_Rating            4.35     4.36     4.40     4.38
## Sales_Rank        347955.1 781660.7 562232.4 562084.6
## Number_of_Ratings    96.34   199.48  1220.59   745.51
## Teenth                 3.7      3.7      3.9      3.8
## Tweentyfifth           4.1      4.1      4.2      4.1
## Fiftith                4.4      4.5      4.5      4.5
## Seventyfifth           4.7      4.7      4.7      4.7
## Ninetith               5.0      5.0      4.9      5.0
## Titles                4747     5021     8274     8770
## Observations        722335   703209  1795265  3220809
## Editions              6800     6339    12201    13652
```

The first four rows return the mean values of the book price, the Amazon Star Rating, the sales rank and the number of ratings per book. The book markets in all three considered countries do not have fixed book prices. Fixed book price (FBP) means that the publisher has the exclusive right to set the price of his book. The retailer is not permitted to discount more than five percent from this set price (Nakayama, 2015).

#< quiz "Fixed_Prices" question: Comparing UK (without FBP) and Germany (with FBP), which country accounted a higher price increase between 1996 (end of FBP in UK) and 2018? sc: - United Kingdom (UK).* - Germany. success: Great, your answer is correct! The UK accounted a price increase of 80 percent after this period, while Germany accounted an increase of 29 percent (Fuchs, Sprang, Beurich, Götz, 2019). failure: Try again. #>

The data set only includes countries data from countries without FBP. Comparing the prices of the three available countries, we recognize a large price difference between Canada (21.07) and Great Britain (13.12) while the US account an average price of 15.86. Possible reasons for those expensive book prices could be that Canada imports many of these books where fees and other costs are incurred (Kwan, 2013), transportation costs over large land masses and a loss of economies of scale due to a smaller book market. The star ratings and their percentiles are quite even while large differences occur in the sales ranks and the number of ratings. Considering the fact that sales ranks are generated on their individual market place and less transparent, we cannot list any specific reasons for this. The high differences in the number of ratings may be due to the level of awareness of Amazon in the individual countries. The US market accounts for more than twice as many observations as Canada or Great Britain. Hence, we gained insight into general descriptive values over the entire data set. To elaborate this, let us review the same values for observations where the availability of professional reviews is guaranteed.

#< info "How to use `kbl()` to generate tables."

The `kbl` function belongs to the package `kableExtra` and creates tables based on data sets. Next to the basic structure that we use within this problem set, there are multiple more methods to style these tables. We adhere to a uniform structure within the problem set. More details about styling possibilities you find here. The structure is as follows:

```
library(kableExtra)
```

```
#for multiple functions we use so-called pipes `%>%` to connect these functions
data_set  %>%
kbl(col.names = c("old_column_name = new_column_name", "..."),
```

```
    caption = "title") %>%
  kable_paper("striped", full_width = TRUE) %>%
#creates a separate area with a caption within the table
  pack_rows("caption", starting_row, ending_row) %>%
  kable_styling(bootstrap_options =
                  c("striped", "hover", "condensed", "responsive"))
```

`col.names` changes the column names so that column names with multiple words are also available. `kable_paper` and `kable_styling` influence the style and color formatting of the table.

#>

Table 2: Statistics Among Countries

| | Overall Data | | | | Professional Reviewed Data | | | |
|---|---|---|---|---|---|---|---|---|
| | Canada | Great Britain | United States | All | Canada | Great Britain | United States | All |
| Price | 21.07 | 13.12 | 15.85 | 16.42 | 26.34 | 15.36 | 18.38 | 18.91 |
| Star rating | 4.35 | 4.36 | 4.40 | 4.38 | 4.34 | 4.26 | 4.28 | 4.28 |
| Sales rank | 347955.1 | 781660.7 | 562232.4 | 562084.6 | 157228.6 | 493568.4 | 332210.0 | 343134.0 |
| Number of ratings | 96.34 | 199.48 | 1220.59 | 745.51 | 17.56 | 29.06 | 162.39 | 107.93 |
| **Star rating percentiles** | | | | | | | | |
| 10th | 3.7 | 3.7 | 3.9 | 3.8 | 3.5 | 3.4 | 3.6 | 3.5 |
| 25th | 4.1 | 4.1 | 4.2 | 4.1 | 4 | 4 | 4 | 4 |
| 50th | 4.4 | 4.5 | 4.5 | 4.5 | 4.5 | 4.4 | 4.4 | 4.4 |
| 75th | 4.7 | 4.7 | 4.7 | 4.7 | 5.0 | 4.8 | 4.6 | 4.7 |
| 90th | 5.0 | 5.0 | 4.9 | 5.0 | 5.0 | 5.0 | 4.9 | 5.0 |
| | | | | | | | | |
| Titles | 4747 | 5021 | 8274 | 8770 | 826 | 901 | 1473 | 1521 |
| Observations | 722335 | 703209 | 1795265 | 3220809 | 63994 | 96998 | 246704 | 407696 |
| Editions | 6800 | 6339 | 12201 | 13652 | 885 | 1009 | 1824 | 1964 |

**Task:** Create a table by using the `kbl()` function according to the table above.

```
#read out the data
DataDescriptiveJournals <- readRDS("material/DataDescriptiveJournals.RDS")
DataDescriptiveOverall <- cbind(DataDescriptivefinal, DataDescriptiveJournals)
t2 <- DataDescriptiveOverall %>%
 kbl(col.names = c("CA" = "Canada", "GB" = "Great Britain",
                   "US" = "United States", "All" = "All", "CA" = "Canada",
                   "GB" = "Great Britain",
                   "US" = "United States", "All" = "All"),
     caption = "Statistics Among Countries") %>%
  kable_paper("striped", full_width = TRUE) %>%
  add_header_above(c("Overall Data" = 5, "Professional Reviewed Data" = 4)) %>%
  column_spec(5, border_right = TRUE) %>%
  pack_rows("Star rating percentiles", 5, 9) %>%
 pack_rows("", 10, 12)
t2
```

Without differentiating between magazines or knowing whether the review turned out "positive" or "negative", we find an increase of price of about 15 percent, a decrease in average sales rank of about 39 percent, and the average number of ratings of about 86 percent, as well as significantly higher variance in percentiles of star ratings. The Amazon star rating also drops by a smaller percentage of about 2.3 percent. This leads to the assumption that the occurrence of professional reviews extensively affects the price, the sales rank and the number of ratings.

The variance in percentiles of star ratings could be due to the fact that we have no information about how it has been reviewed, with which a polarization of the reviews may have taken place. For the New York Times reviews, the main data set contains information about whether the New York Times recommended a book using the variable `drecommended`. In the following, we generate two chronological graphs to show the

descriptive affects from professional reviews on sales ranks and prices.

**Note:** For further examinations we distinguish between the US data and the entire data, as the U.S. market is the largest and has the most average crowd ratings (`cno` == 3 -> U.S.).

**Task:** Execute the following chunk to create three different filtered data sets. Press `check` to collect your points.

```
#U.S. data recommended by New York Times
dataReviewRec <- data %>%
  filter(drecommended == 1 & cno == 3) %>%
  summarise(AVGPrice = mean(pamzn),
            AVGSalesRank = mean(rank, na.rm = TRUE),
            AVGCR = mean(R, na.rm = TRUE))

#U.S. data reviewed by the New York Times but not recommended
dataReviewNRec <- data %>%
  filter(drecommended == 0 & DNYT == 1 & cno == 3) %>%
  summarise(AVGPrice = mean(pamzn),
            AVGSalesRank = mean(rank, na.rm = TRUE),
            AVGCR = mean(R, na.rm = TRUE))

#U.S. data not recommended or reviewed by The New York Times.
dataReviewNNYT <- data %>%
  filter(drecommended == 0 & DNYT == 0 & cno == 3) %>%
  summarise(AVGPrice = mean(pamzn),
            AVGSalesRank = mean(rank, na.rm = TRUE),
            AVGCR = mean(R, na.rm = TRUE))
```

After creating these three data sets, proceed to transform the data by creating the final data set to visualize the results.

**Task:** Use `data.frame` to create the data set for the following graph. Press `check` to execute your code chunk.

```
#create the data set to aggregate the price and sales rank information
dataNYTPrice <- data.frame(Price =
                             c(dataReviewRec$AVGPrice, dataReviewNRec$AVGPrice,
                               dataReviewNNYT$AVGPrice),
                           SalesRank = c(dataReviewRec$AVGSalesRank,
                                         dataReviewNRec$AVGSalesRank,
                                         dataReviewNNYT$AVGSalesRank),
                           Category = c("Recommended", "Not Recommended",
                                        "Not NYT"))
dataNYTPrice
```

```
##      Price SalesRank        Category
## 1 18.53470  244874.4     Recommended
## 2 18.31643  382957.7 Not Recommended
## 3 15.51289  592154.3         Not NYT
```

Similar to figure 5 and 6, mean prices and sales ranks based on are generated two show differences in prices and sales ranks among the expressions of New York Times reviews.

**Task:** Try to fill in the gaps to create the following plot `f3`. The x-axis should represent the `Category` and the y-axis should represent the `Price`. Use `geom_text` to label the bars. Then, read in and the save the plot `f4` correspondingly. Finally, use `grid.arrange` to plot these graphs side by side. Press `check` to confirm.

```
#figure 7: Mean Prices
f7 <- dataNYTPrice %>%
   ggplot() +
   geom_histogram(aes(y = Price, x = Category,  fill = Category),
                  stat = "identity", width = 0.3) +
  geom_text(aes(x = Category, y = Price - 2.5, label = round(Price, 2)),
            color = "white", size = 5) +
   coord_flip() +
   scale_x_discrete(expand = c(0, 1)) +
   theme_bw() +
   ylim(0, 21) +
 scale_fill_manual("", values =
                     c("Not NYT" = "deepskyblue",
                       "Not Recommended" = "darkorchid2",
                       "Recommended" = "blue3")) +
  labs(title = "Figure 7 - Mean prices (U.S. Data)",
       x = "",
       y = "Price in USD") +
  guides(fill = guide_legend(reverse = TRUE)) +
   theme(legend.position = "none",
     axis.text.y=element_blank(),
   panel.grid.minor=element_blank(),plot.background=element_blank())
#figure: Mean Sales Ranks
f8 <- readRDS("material/f8.RDS")
#Side by side plot
combined_f7_f8 <- f7 + f8 + plot_layout(ncol = 2) +
  theme(legend.position = c(-0.03, -0.16), legend.direction = "horizontal") +
  theme(plot.margin = margin(t = 0, b = 1, unit = "cm"))
combined_f7_f8
```

Figure 7 − Mean prices (U.S. Data)

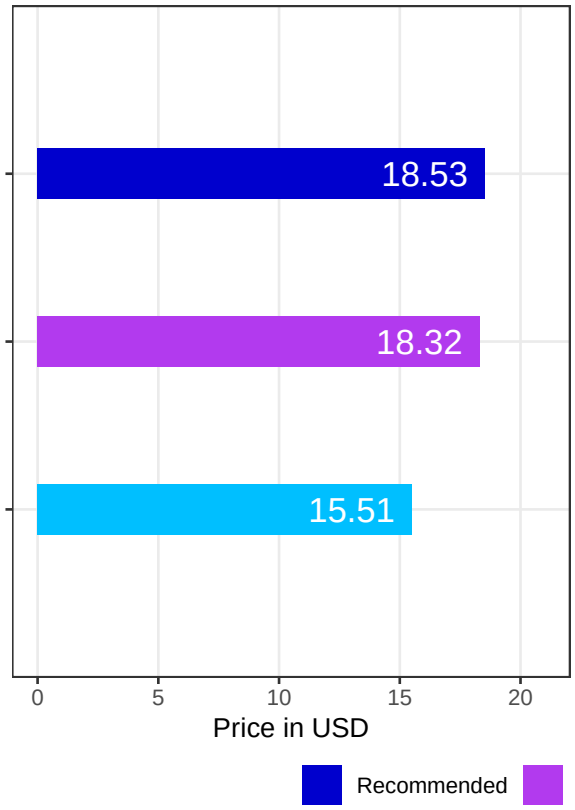
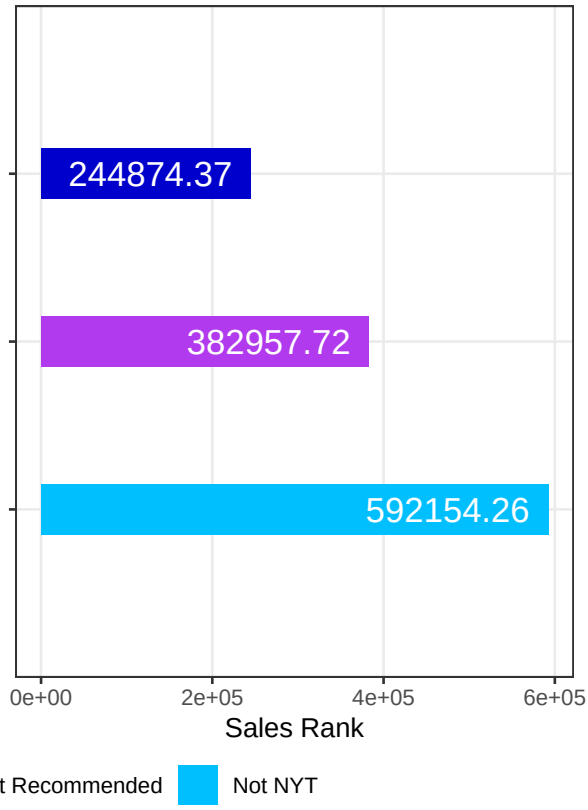



Figure 8 − Mean Ranks (U.S. data)







The price differences between New York Times recommended and non-recommended books is not significant while there is a big price gap between New York Times data and non-New York Times data in figure 7. The fact that the price difference between recommended and non-recommended books is quite small could be due to the fact that a non-recommendation does not always equate to a poor rating. On the other hand, books not mentioned in the New York Times are on average about 2.8 USD cheaper than books mentioned in the New York Times. Observing the average sales ranks in figure 8, we determine clearer differences between these three categories. Since the sales ranks are not direct quantity data, they can still serve as a quite useful comparative value. We see a decline of approximately 35 percent in sales rank from category to category.

#< quiz "Prices_and_SalesRanks" question: Which of the following statements may be made? sc: - With the information of these graphs we can assume a price elasticity $< 1$ for the US book market. - None of those listed.* - With the information of these graphs we can assume a price elasticity $> 1$ for the US book market. - The effect of professional reviews on sales ranks (quantities) is higher than the effect on prices. success: Great, your answer is correct! failure: Try again. #>

To verify whether absolute book prices remain constant over time, we create a timeline for the average book price. We only monitor mean prices of books that have been listed at least once at sales rank 10000 or better or 1000 or better.

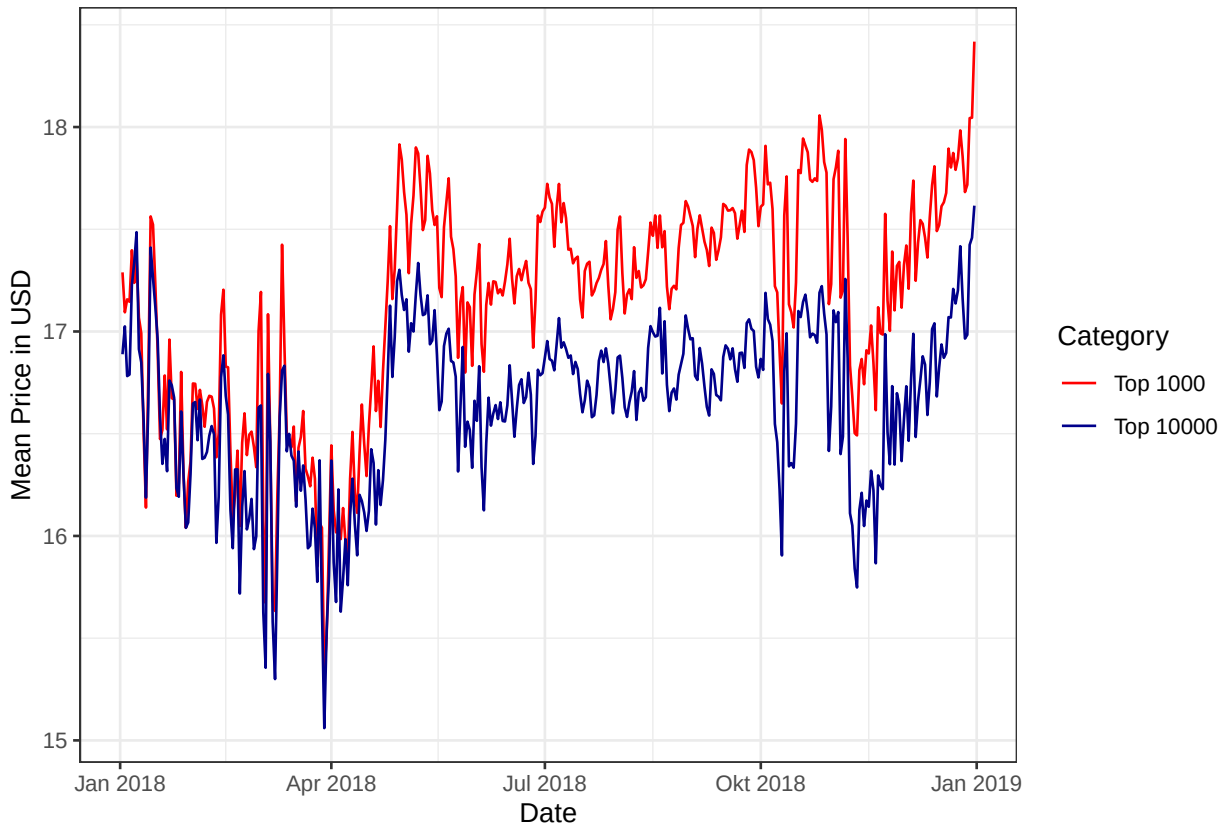
**Task:** Create two according vectors. Save these vectors under `books10k` and `books1k`. Press `check` to confirm your results.

```
#generate the top 10.000 books
books10k <- data %>%
  filter(rank <= 10000) %>%
  distinct(titleno) %>%
  pull(titleno)
#generate the top 1.000 books
books1k <- data %>%
  filter(rank <= 1000) %>%
  distinct(titleno) %>%
  pull(titleno)
```

**task:** Press `check` to create and merge the data sets `dataTop10000` and `dataTop1000` by `ddate`, to visualize the price development for observations for which applies $SalesRank < 10.001$ and $SalesRank < 1.001$.

```
#create aggregated data sets for both categories that are required for figure 9
dataTop10000 <- data %>%
  filter(titleno %in% books10k) %>%
  group_by(ddate) %>%
  summarise(price = mean(pamzn),
            Group = "Top 10000")
dataTop1000 <- data %>%
  filter(titleno %in% books1k) %>%
  group_by(ddate) %>%
  summarise(price = mean(pamzn),
            Group = "Top 1000")
dataTop <- rbind(dataTop10000, dataTop1000) %>%
  group_by(ddate, Group) %>%
  summarise(price = mean(price))
#figure 9: Mean Prices by Sales Rank Category (U.S. Data only)
dataTop %>%
   ggplot() +
   theme_bw() +
   geom_line(aes(x = ddate, color = Group,  y = price), stat = "identity",
             width = 0.3,  position = position_dodge()) +
   scale_color_manual("Category", values = c("Top 10000" = "darkblue",
                                             "Top 1000" = "red"))  +
   labs(title = "Figure 9 - Mean Prices by Sales Rank Category (U.S. Data only)",
        x = "Date",
        y = "Mean Price in USD")
```

Figure 9 − Mean Prices by Sales Rank Category (U.S. Data only)




Overall, prices rise within one year by approximately 1 USD. Price collapses can be observed in the spring, and the fall months from October to the end of November see a massive drop in prices, with average prices (top 1000) falling by around 8.5 percent. Obviously, prices in both categories fluctuate in a similar way due to seasonality, while the Top 1000 category records higher prices. Higher prices in summer could be due to lower demand, as people prefer to read during the summer months. For the following event study in chapter 3.3 we have to take into account that there are fluctuations in book prices within a year and that these can be attributed to various factors.

**Summary**

Using descriptive analysis, we have gained deeper insight into the data and were able to establish initial associations between variables related to pre-purchase information and price or quantity (sales ranks). Starting with an overview of the occurrence of professional evaluations, we have created a table displaying the relative and absolute proportion of these reviews among all observations in the data set. Then, we have created an overview over general values of the underlying data set and compared them with the same value filtered on professional reviewed data. We have recognized differences at all values. We have delved deeper into our research regarding price and sales rank and found that both professional ratings (NYT and non-NYT) appear to have a negative impact on sales ranks (lower rank means higher sales). Finally, we have checked whether book prices fluctuate within a year and found that the price varies seasonally within a year.

In the following chapter, we will start working through the empirical part of this problem set.

# 3. Empirical Strategies on Sales Ranks and Prices

In following chapter *Empirical Strategies on Sales Ranks and Prices* we focus on empirical methods in general and their application to real data from Amazon.

First, we implement a naive regression around Amazon star ratings, where we continuously add multiple methods. In general, it deals with control variables, fixed effects, robust standard errors, and the use of logarithmic estimates.

In chapter 3.2, we focus on the replication of the main regression originally created by Reimers and Waldfogel. We discuss different of those effects and deepen the examination with further regressions. Based on these regressions on real data, we also focus on the validity of regressions and the explanation of values that inform them.

Chapter 3.3 explains the subject of event studies. We also conduct two event studies based on the underlying data set to identify long- and short-term reactions of sales ranks and prices to professional reviews. Finally, We focus on intersections between different magazines and specific distributions to uncover issues in the study structure.

Working through this chapter will provide fundamentals to advanced empirical strategies that are used to make predictive statements in economic science.

**Structure**

3.1. Initial econometric studies on crowd ratings

3.2. Estimation of the Effects on Sales Ranks and Prices

3.3. Introduction and Implementation of Event Studies

## 3.1. Regressions, Robust Standard Errors and Fixed Effects

Before we start into the econometric part with a mature regression, let us first implement a "naive" regression to address the endogeneity problem as well. The formula for this regression is as follows:

$$SalesRank = \beta_0 + \beta_1 AmazonStarRating + \varepsilon$$

where $y$ indicates the dependent sales rank, $x_1$ represents the Amazon star rating on the five-point scale. $\varepsilon$ is the error term and shows the sum of the residuals from the regression. Finally, $\beta_1$ indicates the average increase or decrease of the sales rank for every unit increase of Amazon star rating.

#< quiz "Simple_Regression" question: Now, We want to estimate the effect from Amazon star ratings $x1$ on the sales rank $y$. We determined the following formula for 100 books with $\hat{y}$ = 500.000 - 50.000 $*$ $AmazonStarRating$. What is the correct answer? sc: - With three stars on Amazon of you are exactly placed on rank 350.000. - Within the 100 books, each additional Amazon star results in a 50.000 place lower ranking. - We only consider 100 books - we cannot make any statement. - Within the 100 books, each additional star rating is associated in a 50.000 place lower ranking -> to make statements beyond our sample, we rather need more data and more variables.*
success: Great, your answer is correct! failure: Try again. #>

In their paper, Reimers and Waldfogel used logarithmic values for their estimation. The logarithmization of dependent or explanatory variables can be associated with many advantages. First, logarithmization can be used to stabilize the variance of the residuals to avoid heteroskedasticity. Further, outlier values are mitigated in their effect so that they have less influence on the magnitude of the regression coefficient. The third advantage is due to interpretation of the regression coefficients. Logarithmized coefficients allow us to interpret these coefficients as percentages for simplicity. The possible interpretation are as follows (Kranz, 2022):

- **log - log**: $\log \hat{y} = \hat{\beta}_0 + \log \beta_1 x_1$, for a one percent increase in $x_1$, the predicted value of $y$ increases by approximately $\beta_1$ percent.
- **log - level**: $\log \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$, for a one unit increase in $x_1$, the predicted value of $y$ increases by approximately 100 * $\beta_1$ percent.
- **level - log**: $\hat{y} = \hat{\beta}_0 + \log \hat{\beta}_1 x_1$, for a one percent increase in $x_1$, the predicted value of $y$ increases by approximately 0.01 * $\beta_1$ units.

The purchasers decision on the amount of star ratings also depends on other components of error term $\varepsilon$. This leads to the assumption that there exists an endogeneity problem and accordingly the variable `R` is endogenous. One possible effect on star ratings could be the individual quality of each book. By implementing *fixed effects* this underlying endogeneity problem can be reduced.

**Fixed Effects**

First, it must be considered which variable is suitable as fixed effects. When using fixed effects, we select certain variables to control for. These fixed effects are kept constant and do not affect the estimation. Using this method, we attempt to reduce the variation within explanatory variables by minimizing the potential for bias from omitted variables (Hill, Davis, Roos and French, 2020). Reimers and Waldfogel, among others, chose the variable `canum` as a fixed effect to control for, providing information on a country-dependent "asin" (Amazon product identifier).

**Task:** Check the following chunk to read in the main data set.

```
data <- readRDS("material/dataEst.RDS") %>%
  arrange(canum, ddate)
```

Let us now run two regressions comparing the results due to the effects of country-specific title fixed effects. Before that, we sample the main data set to improve the performance of this study.

#< info "How to draw samples in R"

To randomize examinations in R, random samples can be drawn. Likewise, it can be useful not to count on all data all the time for performance reasons. The following chunks shows how to draw samples in R.

```
# set.seed() is used to fix a sample (without this command there would
# be another sample every time after execution)
set.seed(123)
# create sample data using the function sample()
sample_data <- data_set[sample(nrow(data_set),
                               size = number_of_targeted_observations,
                               replace = TRUE),]
```

As a result, we draw a sample in amount of `number_of_targeted_observations`.

#>

**Task:** Prepare the data set `dataLite` with 100.000 observations. `Check` your results.

```
#fix a sample
set.seed(123)
# draw 100.000 observations from `data`
dataLite <- data[sample(nrow(data), size = 100000, replace = TRUE),]
```

#< info "How to use `feols()` from the package `fixest`"

To create advanced regressions using robust standard errors, fixed effects, or instrumental variables, the `fixest` package provides the `feols()` function to implement them. Currently, `fixest` is the fastest software to perform these estimations (Berge and McDermott, 2023). The structure is as follows:

```
#First, load the package `fixest`
library(fixest)
```

```
# we want to regress x on y using z as fixed effects. We also apply robust
# standard errors over "vcov = "hetero".
reg <- feols(y ~ x | 0 | z, data = data_set)
```

**Note:** The 0 indicates a placeholder for instrumental variables. Without the use of instrumental variables, this value can also be omitted and serves only for illustrative purposes.

#>

#< quiz "Fixed_Effects" question: What is your suggestion, does the implementation of fixed effects significantly affect the value of the coefficients of the explanatory variable? sc: - Yes.* - No. success: Great, your answer is correct!
failure: Try again. #>

**Task:** Replace the ____ gaps to create a regressions with fixed effects. Compare this regression without fixed effects `regWFE` to a regression with fixed effects `regFE`. Finally, we display the results with `modelsummary`. Press `check` after solving this exercise.

Table 3: Simple Regression (1) Versus Fixed Effects (2)

| | (1) | (2) |
|---|---|---|
| (Intercept) | 14.518*** | |
| | (0.078) | |
| Amazon Star Rating | −1.981*** | −0.319** |
| | (0.053) | (0.154) |
| Num.Obs. | $1 \times 10^5$ | $1 \times 10^5$ |
| R2 | 0.014 | 0.914 |
| F | 1415.499 | |
| Std.Errors | | by: canum |
| FE: canum | | X |

* p < 0.1, ** p < 0.05, *** p < 0.01

```
library(modelsummary)
library(fixest)
#regression with fixed effects
regFE <- feols(lrank ~ lR | canum, data = dataLite)
#regression without fixed effects
regWFE <- lm(lrank ~ lR, data = dataLite)
#create table 3
modelsummary(list(regWFE, regFE), statistic = "({std.error})",
             coef_rename = c("lR" = "Amazon Star Rating"),
             stars = c('*' = .1, '**' = 0.05 ,'***' = .01),
             title = "Simple Regression (1)
             Versus Fixed Effects (2)",
             gof_omit = "BIC|R2 Adj.|RMSE|AIC|R2 Within|Log.Lik.")
```

Column two in table 3 shows the fixed effects estimate, while column one shows the regular regression. It is noticeable that the regressions estimate completely different values for the coefficients. The regular regression seems to show exactly that endogeneity problem, that the Amazon star rating also depends on the country and title specific effects. Using fixed effects, we reduced these effects and obtained the information that the Amazon star rating and the sales rank are not as strong related as we originally assumed. Without fixed effects, we obtained a coefficient of -1.981 and with fixed effects -0.319. Finally, we recognize massive differences in the $R^2$ value with 0.014 and 0.914 with fixed effects. This phenomenon reinforces the assumption that especially on title and country level large variations of the error term regarding star ratings occur.

#< quiz "Compare_FE_NFE" question: What do these coefficients tell us? sc: - For a one percent increase in Amazon star ratings (without fixed effects), the predicted sales rank on average decreases by approximately 198 percent. - For a one percent increase in Amazon star ratings (without fixed effects), the predicted sales rank on average decreases by approximately 1.98 percent.* success: Great, your answer is correct!
failure: Try again. #>

However, the addition of fixed effects can also lead to inaccuracies and biases. The disadvantage is that the variables used as fixed effects can no longer be included in the regression as explanatory variables. This model also assumes that the explanatory variable is not collinear (perfectly correlated) to the fixed effects, otherwise this explanatory variable would be determined purely by the fixed effects. Finally, the endogeneity problem can not get completely eliminated by a fixed-effects model. Better suited for this purpose is the Difference-in-Difference approach or Instrumental Variable Estimation, which are not further discussed in the course of this problem set.

### Control Variables

Another useful method is the use of *control variables*. After adding fixed-effects to reduce dependencies between the main estimator and the error term, the sales rank is not determined exclusively by the star ratings. Hence, the price, the number of Reviews and other factors also affects the sales rank. In following we add the variables `lpamzn` and `lreviews` as control variables.

$$SalesRank = \beta_0 + \beta_1 AmazonStarRating + \beta_2 Price + \beta_3 NumberOfRatings + \varepsilon$$

Now, we control for the effect of `lpamzn` and `lreview` and on the sales rank, so that these variables no longer affect the coefficient for Amazon star ratings. However, adding to many control variables could lead to overfitting, which means that the model is overfitted to the specific sample, so that the model reacts to random variation instead of actual contexts.

#< quiz "Multiple_Regression" question: What is the correct answer? sc: - Adding the variables `lreview` and `lpamzn` to the regression increases the value of 1. - Adding the variables `lreview` and `lpamzn` to the regression decreases the value of 1.* - Adding the variables `lreview` and `lpamzn` to the regression does not influence the coefficient of 1.
success: Great, your answer is correct! failure: Try again. #>

The ideal number of control variables generally depends on the research question and the number of observations (Kranz, 2022). But even with a high frequently data set like the Amazon data set it would be recommended not to control for too many variables.

#< info "How to work with the function `modelsummary` and the eponymous package?"

The function `modelsummary` creates tables to visualize regression tables among others. This function provides the ability to add coefficients of determination, standard errors, significance levels, and more. Likewise, the `modelsummary()` function can rename coefficient names, omit variables and add significance stars from the view. The Structure is as follows:

```
library(modelsummary)
```

```
modelsummary(list(Regression_Names),
             statistic = "({std.errors})" / "({p.value})" / "...",
             coef_rename = "old_coeff_name = new_coef_name", "...",
             coef_omit = c("coeff_name1", "..."),
             stars = c('*' = .1, '**' = 0.05 ,'***' = .01), gof_omit =
               "BIC|and_other_variables_to_omit_from_below", title = "title")
```

More information about the package `modelsummary` you find here.

#>

In the following, a linear regression with fixed effects is compared with the same regression with additional control variables.

**Task:** Replace the ____ gaps with the correct code. Use the function `feols` to estimate the regression with control variables. Do not forget to press `check`.

```
#regression with control variables
regContrVar <- feols(lrank ~ lR + lpamzn + lreview | canum, data = dataLite)
#create table 4
modelsummary(list(regContrVar, regFE), coef_rename =
               c("lR" = "log Amazon Star Rating",
                 "lpamzn" = "log Amazon Price",
                 "lreview" = "log Number of Ratings"),
```

Table 4: Control Variables (1) Versus Simple Regression (2)

| | (1) | (2) |
|---|---|---|
| log Amazon Star Rating | −0.416*** | −0.319** |
| | (0.150) | (0.154) |
| log Amazon Price | 0.388*** | |
| | (0.029) | |
| log Number of Ratings | 0.367*** | |
| | (0.018) | |
| Num.Obs. | $1 \times 10^5$ | $1 \times 10^5$ |
| R2 | 0.918 | 0.914 |
| Std.Errors | by: canum | by: canum |
| FE: canum | X | X |

* p < 0.1, ** p < 0.05, *** p < 0.01

```
stars = c('*' = .1, '**' = 0.05 ,'***' = .01),
title = "Control Variables (1)
Versus Simple Regression (2)",
gof_omit = "BIC|R2 Adj.|RMSE|AIC|R2 Within|Log.Lik.")
```

The first column in table 4 contains the estimated coefficients with added control variables. The estimate value of `Amazon Star Rating` indicates that an increase of price by one percent on average results in an decrease of the sales rank by 0.416 percent which represents a stronger negative effect than without the addition of control variables (-0.319). We also detect a positive effects from `Amazon Price` and `Number of Ratings` on the sales rank and thus negative effects on the sales quantity.

The $R^2$ Value indicates large numbers of 0.918 and 0.914, while the standard error is larger for the multiple regression. In a regression analysis, the standard error provides information on the precision of the estimated regression coefficients. To reduce these and make them even more precise, *robust standard errors* are one method to accomplish this.

Table 5: Robust Stand Errors (RSE) Versus Without RSE

| | (1) | (2) |
|---|---|---|
| log Amazon Star Rating | −0.416*** | −0.416*** |
| | (0.121) | (0.150) |
| log Amazon Price | 0.388*** | 0.388*** |
| | (0.022) | (0.029) |
| log Number of Ratings | 0.367*** | 0.367*** |
| | (0.013) | (0.018) |
| Num.Obs. | $1 \times 10^5$ | $1 \times 10^5$ |
| R2 | 0.918 | 0.918 |
| Std.Errors | Heteroskedasticity-robust | by: canum |
| FE: canum | X | X |

* p < 0.1, ** p < 0.05, *** p < 0.01

**Robust Standard Errors**

To apply regressions correctly, certain assumptions must be made. One of these assumptions is **homoskedasticity**. To fulfill this assumption, the residuals of the regression must be uniformly distributed.**Robust Statistics** addresses making estimates that are insensitive to small changes in the basic assumptions like outlier values falsifying the regression residuals (Rachev, 2007). In R, there are different methods to implement Robust Standard Errors, all of which follow a similar approach. Basically, this algorithm determines the coefficients by disregarding outlier values and other disruptive factors.

We want to compare the regression above with the same regression using robust standard errors.

**Task:** Solve the following chunk to run a regression with robust standard errors. Do not forget to press `check`.

**Note:** Look at the `feols`- Info box to implement robust standard errors.

```
#regression using robust standard errors
regRSE <- feols(lrank ~ lR + lpamzn + lreview | canum, vcov = "hetero",
                data = dataLite)
#create table 5
modelsummary(list(regRSE, regContrVar), coef_rename =
               c("lR" = "log Amazon Star Rating",
                 "lpamzn" = "log Amazon Price",
                 "lreview" = "log Number of Ratings"),
             stars = c('*' = .1, '**' = 0.05 ,'***' = .01),
             title = "Robust Stand Errors (RSE) Versus Without RSE",
             gof_omit = "BIC|R2 Adj.|RMSE|AIC|R2 Within|Log.Lik.")
```

There are almost no differences in the coefficients, while differences in standard errors occur. If the assumption of homoskedasticity is fulfilled, estimators without robust standard errors tend to have lower standard errors. This could be due to the fact, that Robust Standard Errors take uncertainties into consideration which makes the estimate somewhat less imprecise. However, Robust Standard Errors should still be used to guarantee homoscedasticity.

#< quiz "Robust_standard_errors" question: We assume that (for the same regression) the standard errors for the regular regression are lower than for the regression where robust standard errors were used. Whats is the correct answer? sc: - You can omit the robust standard errors, as they reduce the accuracy of the coefficients. - Robust standard errors should generally be added to any regression, as regressions without them are generally inaccurate. - It is advisable to add robust standard errors but not necessary. Even with

higher standard errors, it may be important to avoid heteroscedasticity.*
success: Great, your answer is correct!
failure: Try again. #>

**Note:** Heteroscedasticity is the opposite of homoskedasticity.

**Summary**

To sum up, this chapter has given us a overview about the methods and tools for the following analysis with the regressions used by the authors. First, we compared a "naive" regression with a linear regression using fixed effects. As a result, we have found large differences in coefficients and in the $R^2$ value in table 3. Further, we have distinguished between linear and multiple regressions and how the addition of the control variables `pamzn` and `lreview` potentially affect the main coefficient for Amazon star ratings. Multiple differences in coefficients have been found and explained. Finally, the effect of Robust Standard Errors have been discussed and implemented in R. We have extracted the information that inserting Robust Standard Errors can be quite useful to ensure homoskedasticity. Application on the Amazon data set underscored the utility and potential of fixed effects, robust standard errors, and control variables for subsequent examinations.

In the following chapter, we apply the methods and tools from this chapter to examine the overall effects on sales ranks and prices.

## 3.2. Estimation of the Effects on Sales Ranks and Prices

After the introduction for regressions, we focus on the research question to estimate the effect from professional reviews and crowd ratings on the book market. Replicating the underlying regressions by Reimers and Waldfogel, the data set must first be transformed.

**Task:** Check the following chunk to read in the main data set.

```
data <- readRDS("material/dataEst.RDS")
```

**Task:** Create the data set `dataUS` that only includes observations from the U.S. market. To enable this, use the condition `cno == 3`. Finally, use `arrange()` to sort the data by `canum` and then by `ddate`. Press `check` to confirm your solution.

```
dataUS <- data %>%
    filter(cno == 3) %>%
    arrange(canum, ddate)
```

### Lagged Sales Rank

In their main regression table, the authors added `L1.lrank` as explanatory variable. The variable `L1.lrank` specified the lagged sales rank from the last date documented in the data set. According to the Reimers and Waldfogel (2021), the addition of this variables should enable us to consider also long-term effects beyond a time period of one year. However, in order to measure long-term effects using this methodology, it must be assumed that (considering `lR` as the main regressor) `L1.lR`, which is the logarithmized star rating of the antecedent date, must not correlate with `lrank` (Kranz, 2023). Otherwise, a high coefficient for `L1.lrank` might rather be due to the dependence on an exogenous time trend, which is less related to the actual effect of the regressor.

#< quiz "Dependent_L1lR" question: What is your guess, are `L1.lR` and `lrank` correlated with a minimum coefficient of $> 0.5$ or $< -0.5$? sc: - Yes.* - No. success: Great, your answer is correct! failure: Try again. #>

Let us check whether this assumption is correct.

**Task:** Press `check` to compare these regressions and see to what extent these variables are correlated.

```
#regression for correlation check
regDCheck <- feols(lrank ~ L1.lR + lreview + lpamzn | canum, data = dataUS)
#create table 6
modelsummary(list(regDCheck), stars = c('*' = .1, '**' = 0.05 ,'***' = .01),
             title = "Correlation Check",
             gof_omit = "BIC|R2 Adj.|RMSE|AIC|R2 Within|Log.Lik.")
```

As can be seen in table 6, the lagged log Amazon star rating is strongly correlated with `lrank`, so we must assume an exogenous time trend. However, adding `L1.lrank` as a control variable may be useful to account for seasonal variations.

### Main Regression

Now, let us replicate the main regression table from the underlying paper by using `L1.lrank` as control variable.

**Task:** Run the following chunk to create `reg2` with `ano` as fixed effects and robust standard errors.

Table 6: Correlation Check

| | (1) |
|---|---|
| L1.lR | −0.700*** |
| | (0.116) |
| lreview | 0.465*** |
| | (0.017) |
| lpamzn | 0.426*** |
| | (0.027) |
| Num.Obs. | 1 795 265 |
| R2 | 0.900 |
| Std.Errors | by: canum |
| FE: canum | X |

* p < 0.1, ** p < 0.05, *** p < 0.01

```
#create regression `reg2`
reg2 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR  + dnytpost1_3 +
                dnytpost6_3 +  dnytpost10_3 + dothpost_3 + dothpost10_3 +
                dnytpostpre_1 + dnytpostpre_2 + dnytpostpre_3 + dothpostpre_1 +
                dothpostpre_2 + dothpostpre_3 + dnytpost1r_3 + dnytpost6r_3 +
                dnytpost10r_3 + dnytpostprer_1 + dnytpostprer_2 +
                dnytpostprer_3 + epos + epos2 + epos3 + eneg + eneg2 +
                eneg3 | ano, vcov = "hetero", data = dataUS)
```

Computing regressions in this high-frequency data set requires high computational power, so it usually takes time to compute the problem set. Now, we want to generate the second regression using more explanatory variables.

**Task:** Create `reg3` according to `reg2` above. Add the explanatory variables `lrR`, `dnytpost10_1`, `dnytpost10_2`, `dothpost_1`, `dothpost_2`, `dothpost10_1` and `dothpost10_2` behind the variable `eneg`. Remove the explanatory variables `dothpostpre_1`, `dothpostpre_2` and `dothpostpre_3`.

```
#create regression `reg3`
reg3 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR + dnytpost1_3 +
                dnytpost6_3 + dnytpost10_3  + dothpost_3  + dothpost10_3 +
                dnytpostpre_1 + dnytpostpre_2 + dnytpostpre_3 + dnytpost1r_3 +
                dnytpost6r_3 + dnytpost10r_3 + dnytpostprer_1 + dnytpostprer_2 +
                dnytpostprer_3 + epos + epos2 + epos3 + eneg + eneg2 + eneg3 +
                lrR + dnytpost10_1 + dnytpost10_2 + dothpost_1 + dothpost_2 +
                dothpost10_1 + dothpost10_2 | ano, vcov = "hetero",
              data = dataUS)
```

The variable `ano` indicates an Amazon identifier (asin) and the numbers `1`, `2` and `3` after the dummy variables `dnytpost` and `dothpost` provide information about which country the observation belongs to. U.S. data were used for both regressions.

Subsequently, the three remaining regressions are then read in. To the first regression `reg1` some dummy variables were added as control variables. In the interval $[-20; 40]$ for `NYT_elapse` and `OTH_elapse` all expressions were stored individually in different dummy variables, in sum 120 control variables.

**Task:** check the following chunk to run `reg1`, `reg4` and `reg5`.

```
#create regressions `reg1`, `reg4` and `reg5`
reg1 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR + DNYT + DNYT0 + DNYT1 +
                DNYT2 + DNYT3 +DNYT4 + DNYT5 + DNYT6 + DNYT7 + DNYT8 + DNYT9 +
                DNYT10 + DNYT11 + DNYT12 + DNYT13 +DNYT14 + DNYT15 +DNYT16 +
                DNYT17 +DNYT18 +DNYT19 + DNYT20 + DNYT21 + DNYT22 + DNYT23 +
                DNYT24 + DNYT25 + DNYT26 + DNYT27 + DNYT28 + DNYT29 + DNYT30 +
                DNYT31 + DNYT32 + DNYT33 + DNYT34 + DNYT35 + DNYT36 + DNYT37 +
                DNYT38 + DNYT39 +DNYT40 + DNYTm1 + DNYTm2 + DNYTm3 + DNYTm4 +
                DNYTm5 + DNYTm6 + DNYTm7 + DNYTm8 + DNYTm9 + DNYTm10 + DNYTm11 +
                DNYTm12 + DNYTm13 + DNYTm14 + DNYTm15 + DNYTm16 + DNYTm17 +
                DNYTm18 + DNYTm19 + DNYTm20 + DOTH0 + DOTH1 + DOTH2 + DOTH3 +
                DOTH4 + DOTH5 + DOTH6 + DOTH7 + DOTH8 + DOTH9 + DOTH10 +
                DOTH11 + DOTH12 + DOTH13 + DOTH14 + DOTH15 +DOTH16 + DOTH17 +
                DOTH18 + DOTH19 + DOTH20 + DOTH21 + DOTH22 + DOTH23 + DOTH24 +
                DNYT25 + DNYT26 + DNYT27 + DNYT28 + DNYT29 + DNYT30 + DNYT31 +
                DNYT32 + DNYT33 + DNYT34 + DNYT35 + DNYT36 + DNYT37 + DNYT38 +
                DOTH39 +DOTH40 + DOTHm1 + DOTHm2 + DOTHm3 + DOTHm4 + DOTHm5 +
                DOTHm6 + DOTHm7 + DOTHm8 + DOTHm9 + DOTHm10 + DOTHm11 +
                DOTHm12 + DOTHm13 + DOTHm14 + DOTHm15 + DOTHm16 + DOTHm17 +
                DOTHm18 + DOTHm19 + DOTHm20 + epos + epos2 + epos3 + eneg +
                eneg2 + eneg3 | canum, vcov = "hetero", data = dataUS)

reg4 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR  + dnytpost1_1 +
                dnytpost1_2 + dnytpost1_3 + dnytpost6_1 + dnytpost6_2 +
                dnytpost6_3 + dnytpost10_1 + dnytpost10_2 +  dnytpost10_3 +
                dothpost_1 + dothpost_2 + dothpost_3 + dothpost10_1 +
                dothpost10_2 + dothpost10_3 + dnytpostpre_1 + dnytpostpre_2 +
                dnytpostpre_3 + dothpostpre_1 + dothpostpre_2 + dothpostpre_3 +
                dnytpost1r_1 + dnytpost1r_2 + dnytpost1r_3 + dnytpost6r_1 +
                dnytpost6r_2 + dnytpost6r_3 + dnytpost10r_1 + dnytpost10r_2 +
                dnytpost10r_3 + dnytpostprer_1 + dnytpostprer_2 +
                dnytpostprer_3 + epos + epos2 + epos3 + eneg + eneg2 +
                eneg3 | canum, vcov = "hetero", data = data)

reg5 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR + lrR + dnytpost1_1 +
                dnytpost1_2 + dnytpost1_3 + dnytpost6_1 + dnytpost6_2 +
                dnytpost6_3 + dnytpost10_1 + dnytpost10_2 +  dnytpost10_3 +
                dothpost_1 + dothpost_2 + dothpost_3 + dothpost10_1 +
                dothpost10_2 + dothpost10_3 + dnytpostpre_1 + dnytpostpre_2 +
                dnytpostpre_3 + dothpostpre_1 + dothpostpre_2 + dothpostpre_3 +
                dnytpost1r_1 + dnytpost1r_2 + dnytpost1r_3 + dnytpost6r_1 +
                dnytpost6r_2 + dnytpost6r_3 + dnytpost10r_1 + dnytpost10r_2 +
                dnytpost10r_3 + dnytpostprer_1 + dnytpostprer_2 +
                dnytpostprer_3 + epos + epos2 + epos3 + eneg + eneg2 + eneg3 |
                canum, vcov = "hetero", data = data)
```

After all regressions are available, we create a regression table for comparison purposes. For this purpose, we use the function `modelsummary`, which can display multiple regressions more clearly and with more possibilities than `summary`.

**Task:** Use the function `modelsummary` from the R package `modelsummary` to display the five regressions `reg1`, `reg2`, `reg3`, `reg4` and `reg5`. Do not forget to `check` this chunk.

```
#create table 7
RegTable7 <- modelsummary(list(reg1, reg2, reg3, reg4, reg5),
                          statistic = "({std.error})",
                          coef_omit = "DOTH|DNYT|epos|eneg|postpre|_1|_2",
                          coef_rename = c("L1.lrank" = "Lagged Log sales rank",
                                          "lpamzn" = "Log Amazon price",
                                          "lreview" = "Log Number of ratings",
                                          "lR" = "Log star rating",
                                          "lrR" = "log number of ratings x
                                          log stars", "dnytpost1_3" =
                                          "NYT: 0-5 days", "dnytpost6_3" =
                                          "NYT: 6-10 days", "dnytpost10_3" =
                                          "NYT: 11-20 days", "dnytpost1r_3" =
                                          "NYT Rec: 0-5 days",
                                          "dnytpost6r_3" ="NYT Rec: 6-10 days",
                                          "dnytpost10r_3" = "NYT Rec:
                                          11-20 days", "dothpost_3" =
                                            "OTH: 1-10 days", "dothpost10_3" =
                                            "OTH: 11-20 days"), title =
                            "Main Regression Table",
                           stars = c('*' = .1, '**' = 0.05 ,'***' = .01),
                          gof_omit = "Std|BIC|R2 Adj.|RMSE|AIC|R2 Within|Log.Lik.",
                          column_widths =
                            c(0.2, 0.15, 0.1, 0.15, 0.1, 0.15, 0.15))
 RegTable7
```

The results from table 7 deviate from the underlying results of Reimers and Waldfogel. This deviation is likely due to the variable `L1.lrank`, which is originally calculated with an specific Stata (the programming language used by the authors for the original data processing) function called `L1`. For my calculation of this variable I created a manual function. It can be seen that, the coefficient for `L1.lrank` turns out to be lower than calculated by the authors (Reimers and Waldfogel, 2021, table 2). Consequently, the actual effect of seasonal effects emanating from the variable `L1.lrank` could potentially be smaller than calculated by Reimers and Waldfogel.

#< quiz "Reg_Assumptions" question: Which of these assumptions can be made? sc: - The impact of the New York Times is only short-term, as the coefficient becomes smaller over time. - The R-squared error of over 0.95 is due to overfitting after adding these amount of explanatory variables. - Log Amazon price (the coefficient) $> 0$, so an increasing price appears to increase sales rank and implies a decrease in sales volume.* success: Great, your answer is correct!
failure: Try again. #>

The coefficient for `NYT: 0-5 days` represents the occurrence of a New York Times review in the U.S. within the first five days of the publication.

#< quiz "Coeff_Interpretation" question: How should the coefficient for `NYT 0-5 days` in column 2 be interpreted? sc: - A professional review in the New York Times lowers the sales rank by an average of 0.28 percent within the first five days of publication. - A professional review from the New York Times lowers the sales rank by an average of 0.280*100 units within the first five days of publication. - A professional review from the New York Times lowers the sales rank by an average of 28 percent within the first five days of publication.* success: Great, your answer is correct!
failure: Try again. #>

The coefficient for `Log Amazon Price` is 0.083. This value indicates that a price increase on average increases the sales rank.

#< quiz "Coeff_Interpretation2" question: How should the coefficient for column 2 `Log Amazon Price` be

Table 7: Main Regression Table

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Lagged Log sales rank | 0.784*** | 0.785*** | 0.785*** | 0.747*** | 0.746*** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Log Amazon price | 0.084*** | 0.083*** | 0.083*** | 0.095*** | 0.096*** |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| Log Number of ratings | 0.086*** | 0.087*** | 0.168*** | 0.070*** | 0.169*** |
| | (0.001) | (0.001) | (0.007) | (0.001) | (0.006) |
| Log star rating | −0.122*** | −0.126*** | −0.033** | −0.091*** | −0.021** |
| | (0.011) | (0.011) | (0.014) | (0.008) | (0.010) |
| NYT: 0-5 days | | −0.280*** | −0.280*** | −0.300*** | −0.300*** |
| | | (0.015) | (0.015) | (0.015) | (0.015) |
| NYT: 6-10 days | | −0.068*** | −0.069*** | −0.091*** | −0.092*** |
| | | (0.012) | (0.012) | (0.012) | (0.012) |
| NYT: 11-20 days | | −0.018* | −0.019* | −0.026*** | −0.027*** |
| | | (0.010) | (0.010) | (0.010) | (0.010) |
| OTH: 1-10 days | | −0.007 | −0.152*** | −0.020 | −0.016 |
| | | (0.016) | (0.014) | (0.017) | (0.017) |
| OTH: 11-20 days | | 0.042*** | −0.108*** | 0.044*** | 0.046*** |
| | | (0.012) | (0.008) | (0.013) | (0.013) |
| NYT Rec: 0-5 days | | −0.336*** | −0.335*** | −0.361*** | −0.361*** |
| | | (0.021) | (0.021) | (0.021) | (0.021) |
| NYT Rec: 6-10 days | | −0.188*** | −0.186*** | −0.222*** | −0.223*** |
| | | (0.017) | (0.017) | (0.017) | (0.017) |
| NYT Rec: 11-20 days | | −0.074*** | −0.072*** | −0.091*** | −0.093*** |
| | | (0.013) | (0.013) | (0.014) | (0.014) |
| log number of ratings x log stars | | | −0.055*** | | −0.067*** |
| | | | (0.004) | | (0.004) |
| Num.Obs. | 1 795 265 | 1 795 265 | 1 795 265 | 3 220 809 | 3 220 809 |
| R2 | 0.969 | 0.969 | 0.969 | 0.959 | 0.959 |
| FE: canum | X | | | X | X |
| FE: ano | | X | X | | |

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

interpreted? sc: - By increasing the price for one unit, the sales rank increases by an average of 0.084*100 units. - By increasing the price for one percent, the sales rank increases by an average of 0.084 percent. There is no information about price elasticity.* - By increasing the price for one percent, the sales rank increases by an average of 8.4 percent. As a result, the price is definitely elastic. success: Great, your answer is correct! Concrete quantities are needed to measure price elasticity.
failure: Try again. #>

The New York Times recommendations are associated with the largest effect on sales ranks. The coefficient accounts approximately -0.34 for U.S. data and about -0.36 for all data. In column three, an *interaction term* between `lR` and `lreview` was used. Generally, interaction terms are used to control for bias when an explanatory variable also depends on another independent variable. In this case, the addition of these interaction term implies a shrinking effect of the star ratings. This suggests that star ratings have more influence depending on multiple underlying ratings (Reimers, Waldfogel, 2021).

In the following, price is examined as dependent variable. In chapter 2.3, the assumption was made that New York Times reviews could increases the price in the short-term. To investigate this further, we regress price on similar explanatory variables as in the regression above.

**Task:** Create an according regression by filling in the gaps. Use the variable `canum` as fixed effects.

```
#create regression `reg6`
reg6 <- feols(lpamzn ~ lrank + lreview + lR  + dnytpost1_3 + dnytpost6_3 +
                lrR + dnytpost10_3 + dothpost_3 + dothpost10_3 + dnytpostpre_1 +
                dnytpostpre_2 + dnytpostpre_3 + dothpostpre_1 + dothpostpre_2 +
                dothpostpre_3 + dnytpost1r_3 + dnytpost6r_3 + dnytpost10r_3 +
                dnytpostprer_1 + dnytpostprer_2 + dnytpostprer_3 + epos +
                epos2 + epos3 + eneg + eneg2 + eneg3 | canum, vcov = "hetero",
              data = dataUS)
#create regression `reg7`
reg7 <- feols(lpamzn ~ lrank + lreview + lR  + dnytpost1_3 + dnytpost6_3 +
                dnytpost10_3 + dothpost_3 + dothpost10_3 + dnytpostpre_1 +
                dnytpostpre_2 + dnytpostpre_3 + dothpostpre_1 + dothpostpre_2 +
                dothpostpre_3 + dnytpost1r_3 + dnytpost6r_3 + dnytpost10r_3 +
                dnytpostprer_1 + dnytpostprer_2 + dnytpostprer_3 + epos +
                epos2 + epos3 + eneg + eneg2 + eneg3 | canum, vcov = "hetero",
              data = dataUS)
```

**Task:** Run the following chunk to visualize the regressions using `modelsummary`.

```
#Create table 8
RegTable8 <- modelsummary(list(reg6, reg7), statistic = "({std.error})",
                          coef_omit = "DOTH|DNYT|epos|eneg|postpre|_1|_2",
                          coef_rename = c("lpamzn" = "Log Amazon price",
                                          "lreview" = "Number of ratings",
                                          "lR" = "log star rating",
                                          "lrR" = "log number of ratings x
                                          log stars", "dnytpost1_3" =
                                          "NYT: 0-5 days", "dnytpost6_3" =
                                          "NYT: 6-10 days", "dnytpost10_3" =
                                          "NYT: 11-20 days", "dnytpost1r_3" =
                                           "NYT Rec: 0-5 days", "dnytpost6r_3"
                                          = "NYT Rec: 6-10 days",
                                          "dnytpost10r_3" =
                                            "NYT Rec: 11-20 days", "dothpost_3"
                                          = "OTH: 1-10 days", "dothpost10_3" =
```

Table 8: Price Regressions

| | (1) | (2) |
|---|---|---|
| lrank | 0.010*** | 0.010*** |
| | (0.000) | (0.000) |
| Number of ratings | −0.068*** | −0.023*** |
| | (0.003) | (0.000) |
| log star rating | −0.007 | 0.045*** |
| | (0.004) | (0.005) |
| NYT: 0-5 days | 0.010*** | 0.010*** |
| | (0.003) | (0.003) |
| NYT: 6-10 days | 0.008*** | 0.008*** |
| | (0.003) | (0.003) |
| log number of ratings x log stars | 0.031*** | |
| | (0.002) | |
| NYT: 11-20 days | 0.009*** | 0.009*** |
| | (0.003) | (0.003) |
| OTH: 1-10 days | −0.036*** | −0.034*** |
| | (0.004) | (0.004) |
| OTH: 11-20 days | −0.009*** | −0.009*** |
| | (0.003) | (0.003) |
| NYT Rec: 0-5 days | 0.003 | 0.003 |
| | (0.004) | (0.004) |
| NYT Rec: 6-10 days | 0.005 | 0.004 |
| | (0.004) | (0.004) |
| NYT Rec: 11-20 days | 0.006* | 0.005 |
| | (0.003) | (0.003) |
| Num.Obs. | 1 795 265 | 1 795 265 |
| R2 | 0.914 | 0.914 |
| Std.Errors | Heteroskedasticity-robust | Heteroskedasticity-robust |
| FE: canum | X | X |

* p < 0.1, ** p < 0.05, *** p < 0.01

```
                                "OTH: 11-20 days"),
                    stars = c('*' = .1, '**' = 0.05 ,'***' = .01),
                    title = "Price Regressions",
                    gof_omit = "BIC|R2 Adj.|RMSE|AIC|R2 Within|Log.Lik.")
RegTable8
```

The occurrence of New York Times reviews within the first five days of publication is associated with a 0.01 percent price increase. This price increase remains relatively even over time looking at review occurrence within five days to 20 days of publication. Chapter 2.2 also assumed that Amazon star ratings tend to increase demand in long-term, which may lead to lower prices due to more competition. I added an interaction term for "reg7" to illustrate the dependence of Amazon star ratings on the number of reviews. When estimating the effect on the Amazon price, the addition of this interaction term results in a positive coefficient on the `log star rating` of 0.045 instead of a negative coefficient of -0.007.

#< quiz "Des_vs_Pred" question: What does this mean for the assumption of a long-term price increase? sc: - The assumption is incorrect because descriptive analyses cannot account for dependencies between different effects.* - The assumption is correct and the interaction term is inappropriate for this situation. success: Great, your answer is correct! As a rule, descriptive analyses do not take into account the dependencies of

the variables. In this case, the descriptive analysis leads to a false assumption.
failure: Try again. #>

**Summary**

After generating first estimation results in section 3.1, we have replicated the main regression from the underlying paper in table 7, and have found out that the coefficients differ from the original estimation results. These differences are due to the lagged sales rank, which was not reproducible. However, the regression from this problem set shew similar coefficients with higher standard errors. We found large short-term effects of New York Times reviews, while the overall effect of other professional reviews was less present. In addition, we learned how interaction terms work and how they are impacting the estimations results. Second, we examined the effects on the Amazon price. We found overall small but positive effects from New York Times reviews and negative effects from other journals. By using an interaction term, we discarded the assumption that an increase in Amazon star ratings is associated with a decrease in Amazon prices.

In the following chapter event studies are explained and implemented in R.

## 3.3. Introduction and Implementation of Event Studies

In their paper, Reimers and Waldfogel used a so-called **Event Study** to illustrate and estimate the impact of events (here: publication of professional reviews) that took place. Originally, event studies come from the financial sector from James Dolly in 1933. He examined the price effects of stock splits, studying the occurrence of price changes at the time of the split (MacKinley, 1997, ppt. 13). MacKinley listed several methods for applying event studies, including *Cross-Sectional Models* that use regressions (MacKinley, 1997, ppt. 33). Reimers and Waldfogel also used such an approach to visualize short-term effects on the Amazon sales rank.

**Task:** Check the following chunk to read in the main data set.

```
data <- readRDS("material/dataEst.RDS") %>%
  arrange(canum, ddate)
dataUS <- data %>%
  filter(cno == 3)
```

The first step of the authors procedure is to place all books chronologically on top of each other. For this purpose, the variables `NYT_elapse` and `OTH_elapse` are suitable. For both scenarios, the variable is filtered 20 days before and 40 days after the review is published. Before implementing this event study, the data structure need to get reviewed.

**Task:** Create the data set `dataESNYT` and `dataESOTH` to count the number of books with information about 20 days before and 40 days after the review was published.

```
#create aggregated data set for the NYT event study
dataESNYT <- data %>%
  filter(cno == 3 & NYT_elapse >= -20 & NYT_elapse <= 40) %>%
  group_by(titleno) %>%
  summarize(Sum_observations = n())
nrow(dataESNYT)
```

```
## [1] 1193
```

```
sum(dataESNYT$Sum_observations)
```

```
## [1] 46641
```

```
#create aggregated data set for the OTH event study
dataESOTH <- data %>%
  filter(cno == 3 & OTH_elapse >= -20 & OTH_elapse <= 40) %>%
  group_by(titleno) %>%
  summarize(Sum_observations = n())
nrow(dataESOTH)
```

```
## [1] 322
```

```
sum(dataESOTH$Sum_observations)
```

```
## [1] 11917
```

A total of 46641 observations are available for the New York Times and 11917 observations are available for other magazines in the U.S. during this period. This corresponds to a number of 1193 and 322 books. The next step is to create a regression as follows:

$$LogSalesRank = \beta_0 + \beta_1 LogLaggedSalesRank + \beta_2 LogPrice + \beta_3 LogNumberOfReviews + \beta_4 LogStarRating + \varepsilon$$

**Task:** Implement the regression above in R and save this regression under `reg7`. Use Robust Standard Errors and `canum` as fixed effects. Only observe U.S. data and press `check` to confirm.

```
#create regression `reg7`
reg8 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR | canum,
              vcov = "hetero", data = dataUS)
```

Now, this approach consists in aggregation the regression residuals on the filtered variable `NYT_elapse`. Before that, however, the residuals must be adjusted so that they have a value of zero for `NYT_elapse = -1`. This ensures that we can ideally estimate the effects after the event starts.

**Task:** `Check` the following Code to adjust the residuals from `reg7`.

```
#adjust the residuals from reg8
reg8$residuals <- reg8$residuals -
  mean(reg8$residuals[which(dataUS$NYT_elapse == -1)])
```

**Note:** The command `reg7$residuals` provides access to every single residual from `reg7`.

Further, we aggregate the residuals on `NYT_elapse`. As a result, for each expression of `NYT_elapse`, an average residual is generated.

#< info "How to use the function `aggregate()`"

The `aggregate()` function in R is used to perform an aggregate operation on a data set. The data set is grouped among selected variables while a function is applied to each part of this group. The structure is as follows:

```
# `Fun =` should apply the calculation method
aggregated_data <- aggregate(column/vector,
                             by = list(Variables_to_be_group_by),
                             Fun = mean/sum/...)
```

**Note:** Usually, `aggregate()` can be replaced by the combination of `group_by()` and `summarise()`, which is faster in computation. In this case I decided to use "aggregate" because of the different data sources.

#>

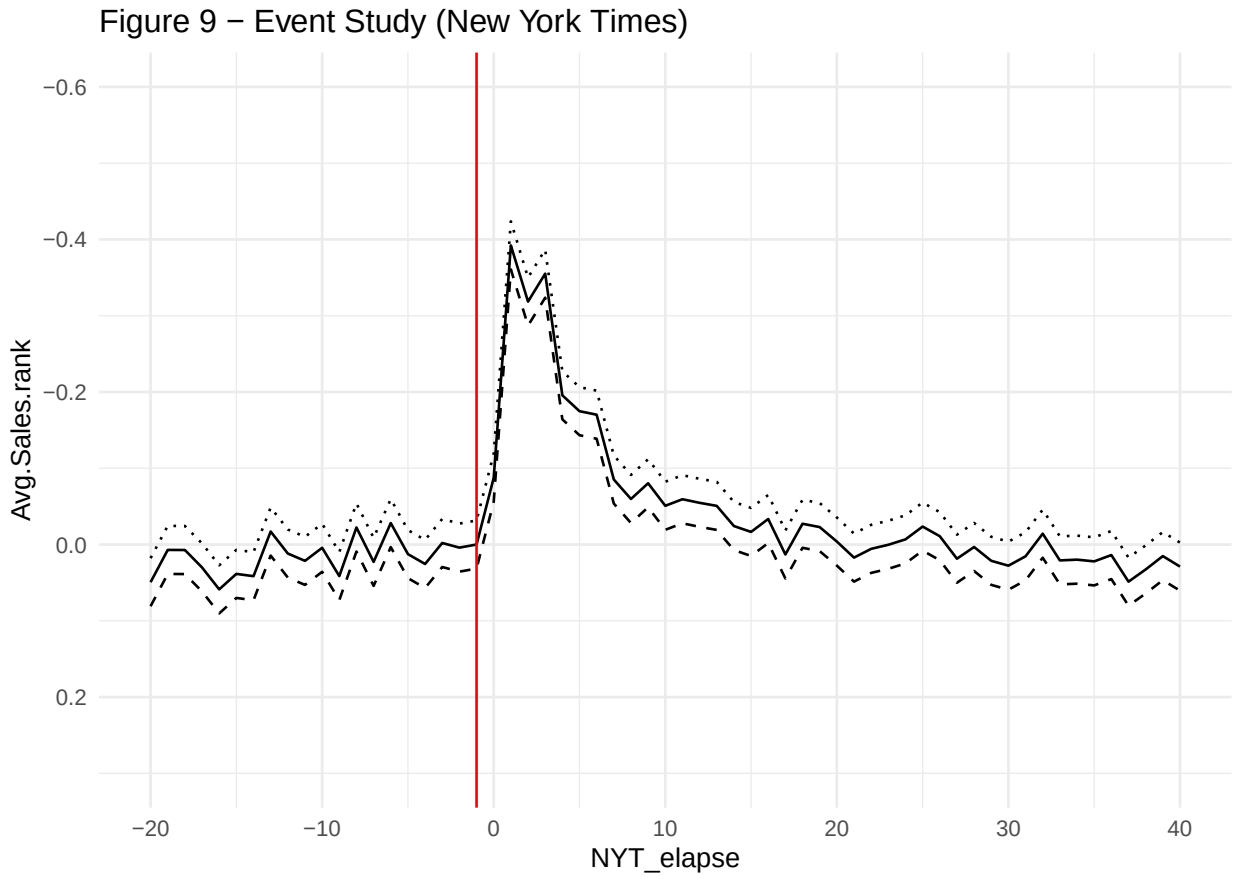
**Task:** `Check` the following chunk to aggregate the residuals on `NYT_elapse`.

```
#aggregate the mean residuals on `NYT_elapse`
reg8_agg <- aggregate(reg8$residuals, by = list(dataUS$NYT_elapse), FUN = mean)
  colnames(reg8_agg) <- c("NYT_elapse", "Avg.Sales.rank")
 reg8_aggr <- reg8_agg %>%
   mutate(Max95 = Avg.Sales.rank + (1.96*sd(Avg.Sales.rank)/
                                    sqrt(length(reg8))),
          Min95 = Avg.Sales.rank - (1.96*sd(Avg.Sales.rank) /
                                    sqrt(length(reg8))))
```

The variables `Max95` and `Min95` are indicating the 95 percent *confidence interval*. A confidence interval indicates to x percent (here: 95 percent) how probable it is that the actual value is in this interval or graphical range.

**Task:** Create a `ggplot`- graph to visualize the average residuals on the y-axis and `NYT_elapse` on the x-axis. Add a vertical line to illustrate the day before the review was published. Information can be found here.

```
#create figure 9
ESNYT <- ggplot(reg8_aggr, aes(x = NYT_elapse, y = Avg.Sales.rank)) +
   geom_line() +
   ggtitle("Figure 9 - Event Study (New York Times)") +
   scale_x_continuous(breaks = seq(-20, 40, by = 10), limits = c(-20, 40)) +
  geom_line(aes(y = Max95), linetype = "dashed") +
  geom_line(aes(y = Min95), linetype = "dotted") +
  geom_vline(xintercept = -1, color = "red", size = 0.5) +
  scale_y_reverse() +
  ylim(0.3, -0.6) +
  theme_minimal()
ESNYT
```




As figure 9 shows, a published review in the New York Times is associated with a huge increase in log sales rank of about 0.40 units. After almost two weeks, the sales ranks returns to its base. The proximity of the confidence interval line to the main line indicates a high level of significance. Analogous, we implement the same event study for all other journals.

**Task:** According to above, adjust the residuals on `OTH_elapse = 1` and call it `reg7.1$residuals`. Furthermore, create `reg7.1_agg` with `aggregate()` on the variable `OTH_elapse`. Finally, add the minimum

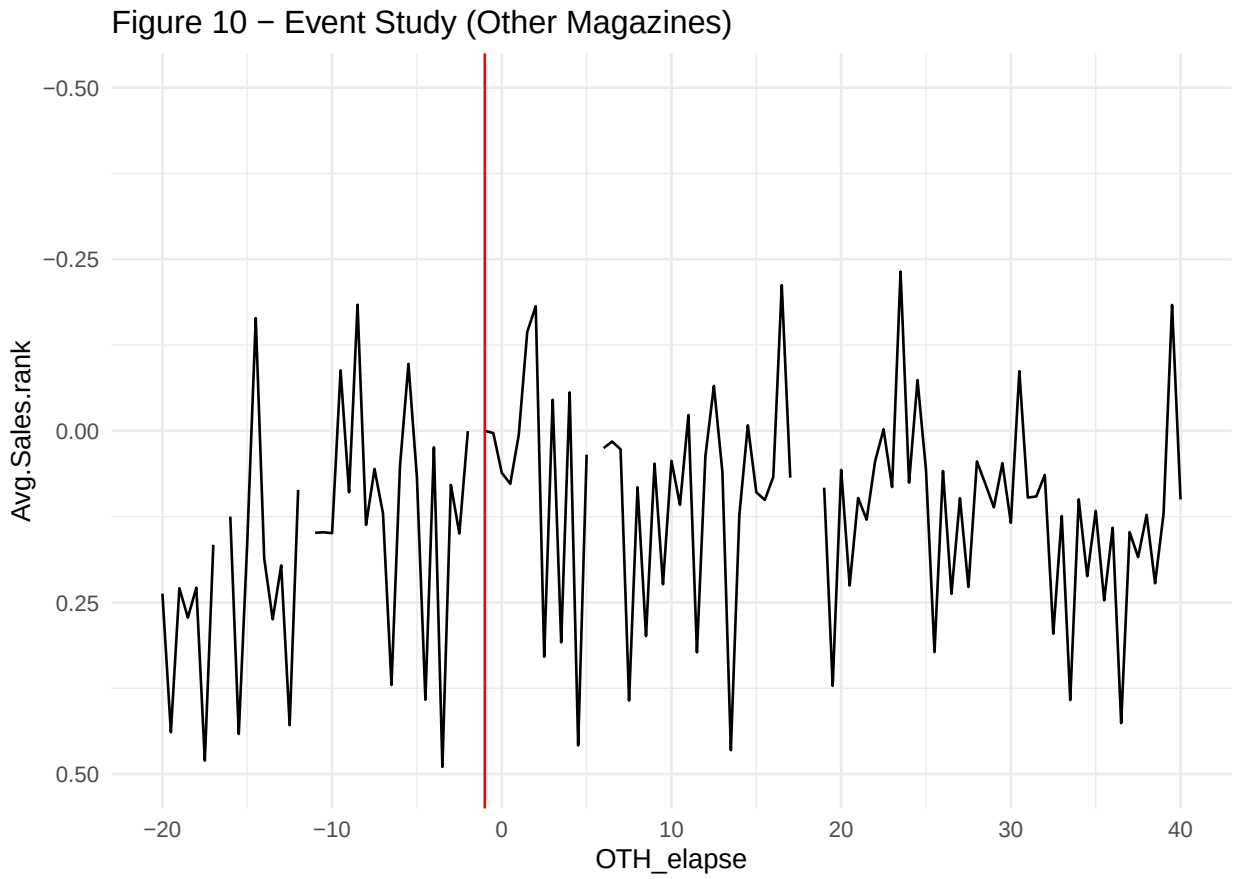
and maximum value of the 95 percent confidence interval.

**Note:** Use the name `reg7.1_aggr` for the calculation of `reg7.1_agg` at the second step, because RTutor cannot store the same variable names with two different values.

```
#aggregate the mean residuals on `OTH_elapse`
reg8$residuals <- reg8$residuals -
  mean(reg8$residuals[which(dataUS$OTH_elapse == -1)])
reg8.1_agg <- aggregate(reg8$residuals,
                        by = list(dataUS$OTH_elapse),
                        FUN = mean)
colnames(reg8.1_agg) <- c("OTH_elapse", "Avg.Sales.rank")
reg8.1_aggr <- reg8.1_agg %>%
   mutate(Max95 = Avg.Sales.rank + (1.96*sd(Avg.Sales.rank)/
                                      sqrt(length(reg8))),
          Min95 = Avg.Sales.rank - (1.96*sd(Avg.Sales.rank) /
                                      sqrt(length(reg8))))
```

**Task:** Press `check` to illustrate the differences between effects of New York Times and other magazines.

```
#create figure 10
ESOTH <- ggplot(reg8.1_aggr, aes(x = OTH_elapse, y = Avg.Sales.rank)) +
  geom_line() +
  ggtitle("Figure 10 - Event Study (Other Magazines)") +
  scale_x_continuous(breaks = seq(-20, 40, by = 10), limits = c(-20, 40)) +
  geom_vline(xintercept = -1, color = "red", size = 0.5) +
  scale_y_reverse() +
  ylim(0.5, -0.5) +
  theme_minimal()
ESOTH
```




While `ESNYT` delivers a similar course to the New York Times event study from Reimers and Waldfogel, there are major differences in the graph produced here. This could be due to `L1.lrank`, which is not reproducible and differs from the variable created by Reimers and Waldfogel. Since no added value is generated from the figure 10, the effects (based on event studies) of other magazines are not discussed further in the course of this problem set.

However, we estimate effects of reviews, which are assigned in an absolutely autonomous decision-making process. Furthermore, the occurrence of professional reviews from different magazines could overlap, resulting in contamination of the effects from these magazines. In the following, a intersection table is provided to display an absolute proportion of contamination. Contamination can potentially lead to higher standard errors and inconsistencies in our estimate and result in misinterpretations. Consequently, especially when we specifically examine professional reviews, which occur in only 12.6 percent of all books, we need to check these considered conditions to assess the validity of the estimated coefficients.

**Task:** First, read in the data set `DesIntersection`. Then, create a table based on the function `kbl()` to visualize the intersections between the single magazines. `Check` your results.

```
#read in the data set `DesIntersection`
DesIntersection <- readRDS("material/DesIntersection.RDS")
#change row names
row.names(DesIntersection) <- c("New York Times", "Los Angeles Times",
                                "Boston Globe", "Chicago Tribune",
                                "Washington Post", "Wall Street Journal",
                                "Without Intersections", "Sum")
#create table 8
DesIntersection  %>%
```

Table 9: Intersections Between The New York Times And Other Magazines

| | New York Times | Los Angeles Times | Boston Globe | Chicago Tribune | Washington Post | Wall Street Journal |
|---|---|---|---|---|---|---|
| New York Times | NA | 22 | 56 | 61 | 40 | 22 |
| Los Angeles Times | 22 | NA | 7 | 9 | 5 | 1 |
| Boston Globe | 56 | 7 | NA | 14 | 9 | 2 |
| Chicago Tribune | 61 | 9 | 14 | NA | 10 | 2 |
| Washington Post | 40 | 5 | 9 | 10 | NA | 3 |
| Wall Street Journal | 22 | 1 | 2 | 2 | 3 | NA |
| Without Intersections | 1162 | 16 | 21 | 68 | 39 | 46 |
| Sum | 1363 | 60 | 109 | 164 | 106 | 76 |

```
kbl(col.names = c("Intersection_NYT" = "New York Times",
                  "Intersection_LAT" = "Los Angeles Times",
                  "Intersection_BG" = "Boston Globe",
                  "  Intersection_CHI" = "Chicago Tribune",
                  "  Intersection_WAPO" = "Washington Post",
                  "Intersection_DWSJ" = "Wall Street Journal"),
    caption = "Intersections Between The New York Times
    And Other Magazines") %>%
  kable_paper(full_width = TRUE) %>%
  pack_rows("", 8, 8) %>%
  kable_styling(bootstrap_options = c("hover", "condensed",
                                      "responsive"))
```

By focusing on New York Times reviews, we found several 201 books were reviewed by at least one other magazine. Conversely, a larger proportion of the reviews from other magazines were also rated by the New York Times. To determine the exact proportion, we calculated that at least 30 percent (Wall Street Journal) and at most 67 percent (Boston Globe) of non-New York Times magazines were also reviewed by the New York Times. As a result, estimates in this regard may be contaminated. In this context, contamination are intersections resulting from overlaps, whereby respective effects can no longer be clearly assigned. The row or column sum does not reflect the total number of reviews in the journals, as we only consider individual overlays. In reality, there are also multiple overlays where three or four journals review one book.

Similarly, an uneven distribution of professional ratings in terms of author name recognition could lead to possible bias. For instance, if professional reviewers at the New York Times prefer certain well-known authors because of personal relations or preferences, those reviewers might tend to review or even recommend just those books rather than other books. The same applies for preferences to particular genres of the reviewers. Below, we use a density function to examine whether the professional reviewers at the New York Times and other magazines had any preference in selecting the books to be evaluated.

**Task:** Create a data set `DesGenre` in which you calculate the proportion of professional reviews per book title. Differentiate according to the genre of the books (`genre`). Do not forget to check your solution.

```
#create the required data set `DesGenrefinal` including information for the
#validity check
DesGenre  <- data %>%
```

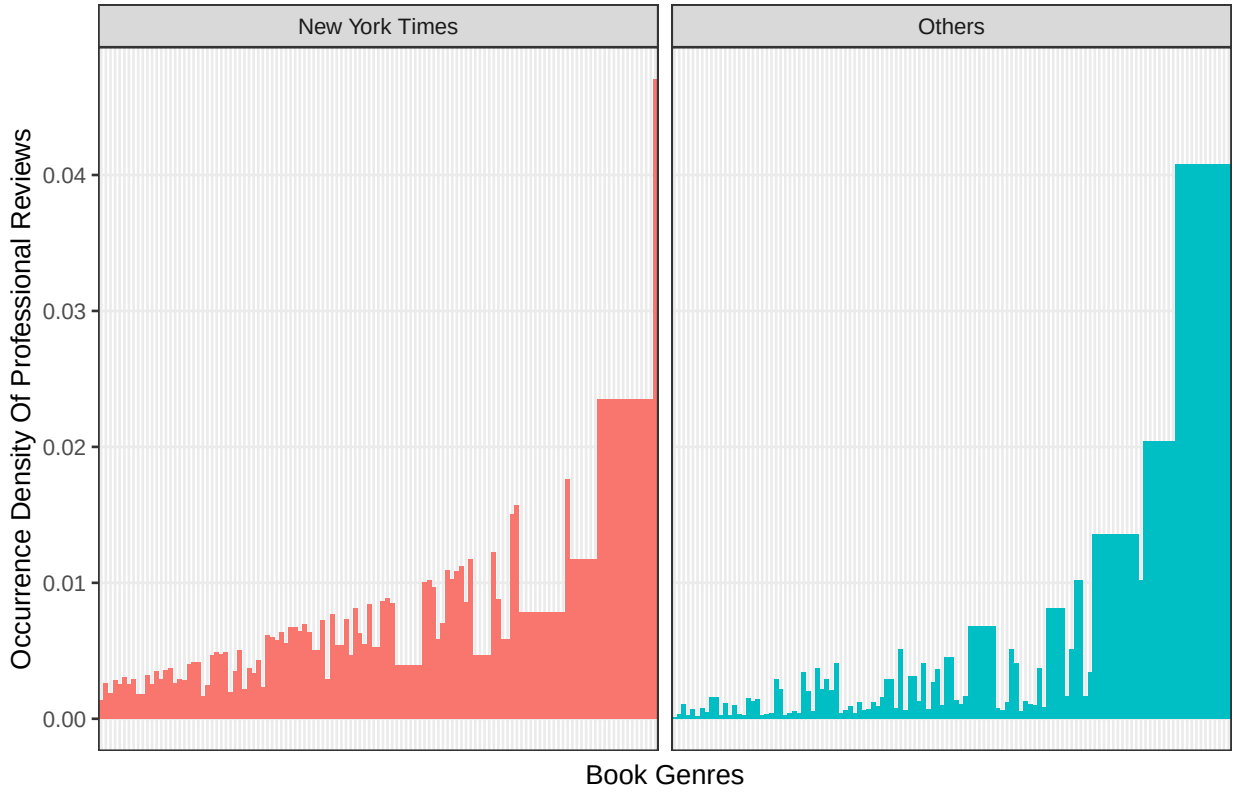
```
  group_by(genre) %>%
  summarise(Number_Ratings = n_distinct(titleno),
            Number_NYT = n_distinct(titleno[which(data$DNYT == 1)])
            / n_distinct(titleno),
            Number_OTH = n_distinct(titleno[which(data$DOTH == 1)])
            / n_distinct(titleno)) %>%
  arrange(Number_NYT)

y1 <- DesGenre [,c(1, 3)] %>%
  mutate(Group = "New York Times",
         Number_Proportion =  Number_NYT/sum(Number_NYT))
y2 <- DesGenre [,c(1, 4)] %>%
  mutate(Group = "Others",
         Number_Proportion =  Number_OTH/sum(Number_OTH))

colnames(y1) <- c("Genre", "Number", "Group", "Number_Proportion")
colnames(y2) <- c("Genre", "Number", "Group", "Number_Proportion")

DesGenrefinal  <- rbind(y1, y2)
```

After we creating the data set, we can proceed with the creation of the chart.

**Task:** Create the graph representing the density of professional reviews among the number of published books. Save this graph as `f1` and use `facet_grid()` to split this graph by the variable `Group`.

```
#create figure 11
 f11 <- ggplot(data = DesGenrefinal) +
  facet_grid(~Group) +
  geom_bar(aes(x = fct_reorder(Genre, Number_Proportion, .fun = sum,
                               .asc = TRUE,
                               .subset = Group == "New York Times"),
               y = Number_Proportion, fill = Group), stat = "identity") +
  theme_bw() +
  labs(title = "Figure 11 - Distribution Of Professional Reviews
       Among Book Genres",
       x = "Book Genres",
       y = "Occurrence Density Of Professional Reviews") +
  guides(fill=guide_legend(title="")) +
  theme(legend.position = "",
  axis.text.x=element_blank(),
  axis.ticks.x = element_blank(),
  panel.grid.minor=element_blank(),plot.background=element_blank())
f11
```

Figure 11 − Distribution Of Professional Reviews Among Book Genres




Every underlying bar in figure 11 represents the density of professional reviews on the individual book genre. In total, the Amazon data set provides information on 121 different book genres, which are arranged equally in the diagram to show similarities and differences. The aim is to identify possible distributional skews that might reject the assumption of an even selection process by professional reviewers. Apparently, the New York Times and other magazines prefer to rate some of the genre categories more frequently than others. However, with the exception of about one-fourth of the book genres, the probability density for other books does not appear uniformly.

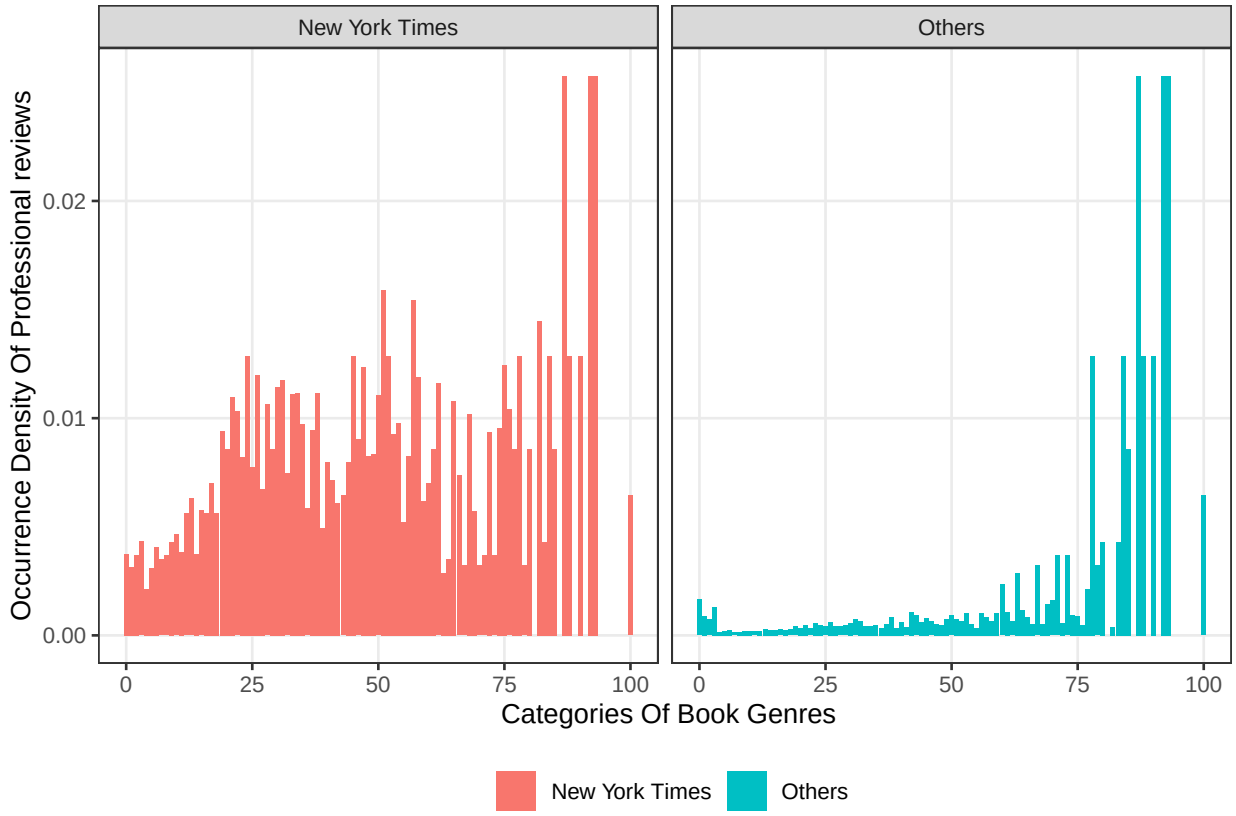
#< quiz "Genre_Guess" question: What could this probably be due to? sc: - This could be due to the fact that book genres vary in their demand on reviews. For instance, the need for professional reviews is likely larger for novels than for travel guides.* - The assumption of an even selection process by professional reviewers can generally be rejected. - Novels and other few categories generate higher sales than many other genres, making them stand out. success: Great, your answer is correct! failure: Try again. #>

#< quiz "Distribution_Guess" question: Focusing the **New York Times reviews**, What is your suggestion? sc: - The more books an author publishes, the more likely these books will be professionally reviewed. - The more books an author publishes, the less likely these books will be professionally reviewed. - The overall distribution is even.* success: Great, your answer is correct! failure: Try again. #>

**Task:** `Check` the following chunk to create a density graph for the distribution on published books.

```
DesAuth <- readRDS("material/DesAuth.RDS")
#create figure 12
f12 <- readRDS("material/f12.RDS")
f12
```

Figure 12 − Professional Reviews Among the Number Of Publications




The variable `numbooks` appropriates as an indication of prior author awareness. We focus on the distribution of professional reviews among the number of published books to review the assumption of an even selection process according to figure 11. For the New York Times, we overall note an even distribution across all popularity levels. In contrast, other magazines visibly preferred to specifically review books by already well-known authors. As a result, the assumption of an even selection process for other magazines can be rejected which calls into question all estimation results regarding the effects of other magazines in the further course of this examination.

**Task:** Create a data set `dataAdj` from which you filter out contaminated data where more than the New York Times has reviewed a book.

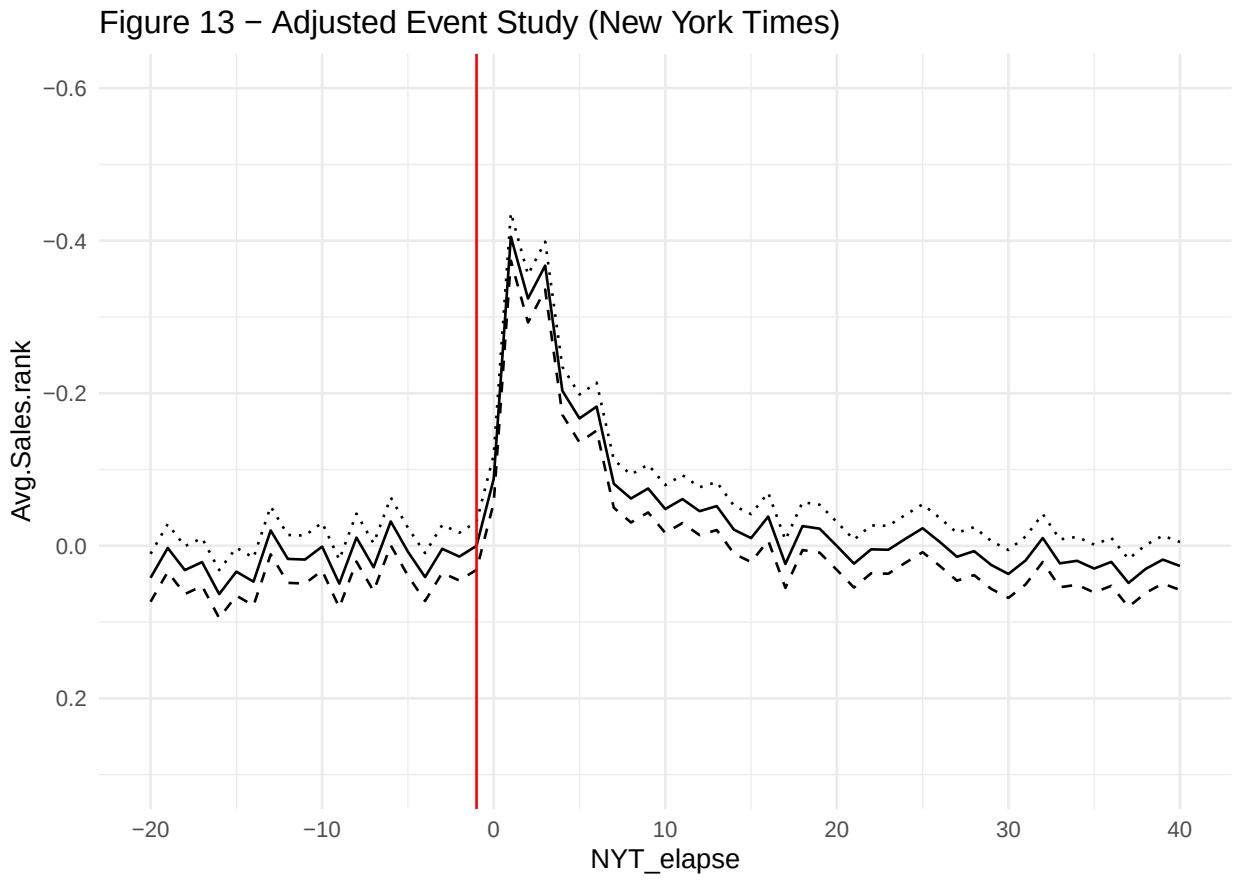
```
#filter out contaminated data
dataAdj <- data %>%
  filter(((data$DBG == 0 & data$DCHI == 0 & data$DLAT == 0 & data$DWAPO == 0 &
             data$DWSJ == 0 & data$DNYT == 1) | data$DNYT == 0) & cno == 3)
```

We now count about 50000 observations less and filtered out the contaminated data. Let us repeat creating an event study of New York Times reviews.

**Task:** `Check` the following code to create an additional event study.

```
# 1. Create a regression based on the adjusted data
reg9 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR | canum,
              vcov = "hetero", data = dataAdj)
# 2. Mean-adjustment on NYT_elapse = -1.
reg9$residuals <- reg9$residuals -
```

```
  mean(reg9$residuals[which(dataAdj$NYT_elapse == -1)])
# 3. Aggregation
reg9_agg <- aggregate(reg9$residuals,
                      by = list(dataAdj$NYT_elapse),
                      FUN = mean)
  colnames(reg9_agg) <- c("NYT_elapse", "Avg.Sales.rank")
 reg9_aggr <- reg9_agg %>%
   mutate(Max95 = Avg.Sales.rank + (1.96*sd(Avg.Sales.rank)/
                                      sqrt(length(reg9))),
          Min95 = Avg.Sales.rank - (1.96*sd(Avg.Sales.rank) /
                                      sqrt(length(reg9))))
# 4. Visualizing
ESNYTNew <- ggplot(reg9_aggr, aes(x = NYT_elapse, y = Avg.Sales.rank)) +
   geom_line() +
   ggtitle("Figure 13 - Adjusted Event Study (New York Times)") +
   scale_x_continuous(breaks = seq(-20, 40, by = 10), limits = c(-20, 40)) +
  geom_line(aes(y = Max95), linetype = "dashed") +
  geom_line(aes(y = Min95), linetype = "dotted") +
  geom_vline(xintercept = -1, color = "red", size = 0.5) +
  scale_y_reverse() +
  ylim(0.3, -0.6) +
  theme_minimal()
ESNYTNew
```




As figure 13 shows, there are almost no deviations from figure 9. However, event studies also entail some

disadvantages. First, the mean-adjusted method does not work that well. This is due to the fact that if many of these events occur simultaneously, possible seasonal fluctuation or other endogenous effects cannot be ruled out (V. Henderson Jr, 1990, pp. 288) As a result, the time period must be well chosen. Furthermore, event studies cannot provide causal evidence. Finally, other effects also influence the estimation, especially when the regression consists of only four explanatory variables.

**Summary**

Event Studies are useful methods to illustrate the effect of events occurring at different times. We have received information about the origin of event studies and how they work in general. We then focused on the implementation of event studies in R using so-called cross functional models, which can estimate the effect on a chronological sequence using aggregated residuals. We replicated these event studies from Reimers and Waldfogel and differentiated between New York Times reviews and reviews from other journals. While New York Times reviews show large effects on log sales rank, the estimate of the effect for other journals differs from the authors' estimate, making it unusable. In the course of a validity check, we focused on the distribution of the selection process of books of the professional reviewers and possible contamination of the effects. We measured a high degree of contamination for non-New York Times magazines and found partially unequal distributions regarding the selection process of books. Finally, we conducted a repeat event study of the New York Times considering contamination and found similar results as before.

In chapter 4., we link sales ranks to sales quantities to determine and examine price elasticities.

# 4. Translating Sales Ranks in Quantities and Price Elasticities

The research results of the sales ranks on Amazon in regressions are limited to relative statements about the sales level of a book as opposed to other books. In order to make economic statements regarding to price elasticity and welfare, absolute sales volumes are needed. Reimers and Waldfogel collected data from the Top 100 Weekly Bestsellers, where Reimers and Waldfogel claim they matched 876 books (hits in U.S. asin). The collected data was originally produced by the Nielsen Company, which is a market research company, and therefore the data cannot be replicated. Due to the fact that the Amazon data set is based on daily observations of sales ranks while the Nielsen data lists only weekly observations, the authors created a formula based on the general assumption that sales and ranks are exponentially related (Chevalier and Goolsbee, 2003) to reconcile daily sales ranks with weekly sales data:

$$q_{jw} = \sum_{t \in w} A_{rjt} - B + v_{jw}$$

where $t$ and $j$ stand for the day and the book index, $w$ denotes the week, $v_{jw}$ represents an error term and A and B are to be estimated by using least squares (Reimers and Waldfogel, 2021). By applying a 500-fold bootstrapping procedure (re-sampling method in which the coefficients are estimated using 500 different samples), estimators A and B were associated with average values of 10167 and 0.45.

#< quiz "Sampling_Methods" question: Which of these two statements could potentially have an adverse effect at an bootstrapping procedure? sc: - Re-sampling methods often require intensive computer power, especially if the data set is large.* - To rely on the solution of this method, the underlying data must be normally distributed. success: Great, your answer is correct!
failure: Try again. #>

In the following, not only price elasticities are calculated, but also elasticities from professional valuations and Amazon star ratings. Calculating these elasticities, Reimers and Waldfogel used the following formula:

$$\epsilon = B \cdot \text{Coefficient of Elasticity}/(1 - \text{Coefficient of Lagged Log Sales Rank})$$

The underlying formula draws a long-term estimated and dynamic elasticity model. According to Kranz (2023), this model implies a geometric effect under ceteris paribus, where the effect from the first year $\beta_1$ also influences the following coefficients $\beta_2$ and those of the further years. The formula looks as follows:

$$\beta_1 + \beta_1 \cdot \beta_2 + \beta_1 \cdot \beta_2^2 + ... = \beta_1 \sum_{t=0}^{\infty} \beta_2^t = \beta_1/(1 - \beta_2)$$

Hence, the extent to which these variables can change over time is to be stored in the variable `L1.lrank`. Due to the fact, that at least the coefficient for Amazon star ratings depends on exogenous time trends we use a static short-term model that assumes that these variables respond immediately to changes (here: professional and non-professional ratings). The formula is as follows:

$$\epsilon = B * Coefficient\ of\ Elasticity$$

**Task:** Check the following chunk to read in the main data set and to filter for U.S. observations.

```
data <- readRDS("material/dataEst.RDS") %>%
  arrange(canum, ddate)

dataUS <- data %>%
  filter(cno == 3)
```

To implement this formula in R, we use the coefficients estimated in `reg5`.

**Task:** `Check` this chunk to read in `reg5`.

```
#create regression `reg5`
reg5 <- feols(lrank ~ L1.lrank + lpamzn + lreview + lR + lrR + dnytpost1_1 +
                dnytpost1_2 + dnytpost1_3 + dnytpost6_1 + dnytpost6_2 +
                dnytpost6_3 + dnytpost10_1 + dnytpost10_2 +  dnytpost10_3 +
                dothpost_1 + dothpost_2 + dothpost_3 + dothpost10_1 +
                dothpost10_2 + dothpost10_3 + dnytpostpre_1 + dnytpostpre_2 +
                dnytpostpre_3 + dothpostpre_1 + dothpostpre_2 + dothpostpre_3 +
                dnytpost1r_1 + dnytpost1r_2 + dnytpost1r_3 + dnytpost6r_1 +
                dnytpost6r_2 + dnytpost6r_3 + dnytpost10r_1 + dnytpost10r_2 +
                dnytpost10r_3 + dnytpostprer_1 + dnytpostprer_2 +
                dnytpostprer_3 + epos + epos2 + epos3 + eneg + eneg2 +
                eneg3 | canum, vcov = "hetero", data = data)
```

Now, let us start to calculate the elasticities.

**Task:** `Check` the following Code to see how the 25th, the 50th and the 75th quantile of Amazon star rating elasticities get implemented in R.

```
#calculate different quantiles of star rating elasticities
star_elas_25 <- dataUS$B * (reg5$coefficients["lR"] +
                            reg5$coefficients["lrR"] *
                            quantile(dataUS$lreview,
                                     probs = 0.25, na.rm = TRUE))

star_elas_50 <- dataUS$B * ( reg5$coefficients["lR"] +
                               reg5$coefficients["lrR"] *
                               quantile(dataUS$lreview,
                                        probs = 0.5, na.rm = TRUE))

star_elas_75 <- dataUS$B * ( reg5$coefficients["lR"] +
                               reg5$coefficients["lrR"] *
                               quantile(dataUS$lreview,
                                        probs = 0.75, na.rm = TRUE))
#add them as a column to `dataUS`
dataUS$star_elas_25 <- star_elas_25
dataUS$star_elas_50 <- star_elas_50
dataUS$star_elas_75 <- star_elas_75
```

After this instruction the further elasticities are calculated by the reader.

**Task:** Determine the elasticities for `lpamzn`, `lreview`, `dothpost_3` and `dothpost10_3`. According to the chunk above, save your results as column in `dataUS` under the names `price_elas_mean`, `star_elas_mean`, `oth_1_10` and `oth_11_20`. Press `Check` to confirm.

**Note:** The coefficient for `lrR` was added to account for interaction effects. For the following chunks `lrR` is only needed to calculate `star_elas_mean`.

```
#calculate the price elasticity, star mean elasticity and the elasticity of
#professional reviews of non-NYT magazines
price_elas_mean <- dataUS$B*reg5$coefficients["lpamzn"]
dataUS$price_elas_mean <- price_elas_mean
```

```
star_elas_mean <- dataUS$B * (reg5$coefficients["lR"] +
                                reg5$coefficients["lrR"] *
                                mean(dataUS$lreview))
dataUS$star_elas_mean <- star_elas_mean

oth_1_10 <- dataUS$B *  reg5$coefficients["dothpost_3"]
dataUS$oth_1_10 <- oth_1_10

oth_11_20 <- dataUS$B *  reg5$coefficients["dothpost10_3"]
dataUS$oth_11_20 <- oth_11_20
```

Elasticities for the New York Times elasticity effects are calculated below.

**Task:** Press `check` to calculate the remaining elasticities.

```
#calculate the elasticity of (recommended and not recommended) professional
#reviews of NYT magazines
nyt_1_5_not_rec <- dataUS$B * reg5$coefficients["dnytpost1_3"]
dataUS$nyt_1_5_not_rec <- nyt_1_5_not_rec
nyt_6_10_not_rec <- dataUS$B * reg5$coefficients["dnytpost6_3"]
dataUS$nyt_6_10_not_rec <- nyt_6_10_not_rec
nyt_11_20_not_rec <- dataUS$B * reg5$coefficients["dnytpost10_3"]
dataUS$nyt_11_20_not_rec <- nyt_11_20_not_rec

nyt_1_5_rec <- dataUS$B * reg5$coefficients["dnytpost1r_3"]
dataUS$nyt_1_5_rec <- nyt_1_5_rec
nyt_6_10_rec <- dataUS$B * reg5$coefficients["dnytpost6r_3"]
dataUS$nyt_6_10_rec <- nyt_6_10_rec
nyt_11_20_rec <- dataUS$B * reg5$coefficients["dnytpost10r_3"]
dataUS$nyt_11_20_rec <- nyt_11_20_rec
```

After collecting all elasticity values, we create a table to visualize them. The authors also created standard errors for the estimated elasticity values using a bootstrapping procedure. Basically, in addition to a variance-covariance matrix, they created normally distributed random variables to calculate within a bootstrapping procedure standard errors the elasticities that are present for us. For reasons of performance, these calculations will have no relevance in the further course.

Let us illustrate these elasticities in a table.

**Task:** First, create a data set with `data.frame` to list all elasticity values. Then, use `kbl()` to create a table containing the elasticity values. Fill in the _____ gaps and press `check` to confirm your input.

```
#aggregate the elasticities to create the data set `summary_data`
summary_data <- data.frame(Effects = c(mean(price_elas_mean),
                                       mean(star_elas_25),
                                       mean(star_elas_50),
                                       mean(star_elas_75),
                                       mean(star_elas_mean),
                                       mean(nyt_1_5_not_rec),
                                       mean(nyt_6_10_not_rec),
                                       mean(nyt_11_20_not_rec),
                                       mean(nyt_1_5_rec), mean(nyt_6_10_rec),
                                       mean(nyt_11_20_rec), mean(oth_1_10),
                                       mean(oth_11_20)))
```

Table 10: Elasticities

| | Effects |
|---|---|
| Price Elasticity | -0.0428129 |
| Star Elasticity 25% | 0.1002103 |
| Star Elasticity 50% | 0.1504656 |
| Star Elasticity 75% | 0.2022037 |
| Star Elasticity Overall | 0.1516704 |
| NYT 1-5 Days | 0.1342148 |
| NYT 6-11 Days | 0.0411287 |
| NYT 11-20 Days | 0.0119518 |
| NYT 1-5 Days rec | 0.1616906 |
| NYT 6-11 Days rec | 0.0996271 |
| NYT 11-20 Days rec | 0.0414070 |
| OTH 1-10 Days | 0.0073180 |
| OTH 11-20 Days | -0.0203794 |

```
#change row names of `summary_data`
row.names(summary_data) = c("Price Elasticity", "Star Elasticity 25%",
                            "Star Elasticity 50%", "Star Elasticity 75%",
                            "Star Elasticity Overall", "NYT 1-5 Days",
                            "NYT 6-11 Days", "NYT 11-20 Days",
                            "NYT 1-5 Days rec", "NYT 6-11 Days rec",
                            "NYT 11-20 Days rec", "OTH 1-10 Days",
                            "OTH 11-20 Days")
#create table 10
t10 <- summary_data %>%
kbl(caption = "Elasticities") %>%
  kable_paper("striped", full_width = TRUE) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed",
                                      "responsive"))
t10
```

The values differ from the elasticity values determined by Reimers and Waldfogel. However, this large differences are due to the fact that a static model has been used for the calculation and the values would otherwise be almost identical. In all likelihood, remaining differences are due to the regression coefficient of `L1.lrank` differing from the value calculated in the underlying paper. Nevertheless, the table provides valid values without considering long-term effects.

#< quiz "General_Elasticity_Question" question: Which of the following answers is **not** correct? sc: - Table 9 shows elasticities of demand. - The addition of the interaction term `lrR` can be related to the differences in the quantile effects of the star ratings. - Book prices on Amazon are generally elastic, so price increases are accompanied by disproportionately high changes in demand.* success: Great, your answer is correct!
failure: Try again. #>

The coefficient for `Price Elasticity` accounts approximately -0.04.

#< quiz "Coeff_Elasticity_Price" question: Which of the following answers is correct? sc: - Book prices on Amazon are elastic. - A price increase of one percent is associated with a decrease in sales of 0.04 percent.* - A price increase of one percent is associated with a increase in sales of 0.04 percent. success: Great, your answer is correct!
failure: Try again. #>

The coefficient for `NYT 1-5 days` is about 0.134, while the coefficient for `NYT 1-5 days rec` accounts ap-

proximately 0.162.

#< quiz "Coeff_Elasticity_NYT" question: Which of the following answers is **not** correct? sc: - The coefficient for `NYT 1-5 days` indicates that the occurrence of a New York Times review is associated with an increase in sales volume of 13.4 percent with the first five days of publication.* - The coefficient for `NYT 1-5 days` indicates that a one percent book price increase reviewed by The New York Times is associated with a 14.3 percent increase in sales volume within the first five days after publication. - New York Times recommendations have larger impacts on the sales volume than New York Times reviews without recommendation. success: Great, your answer is correct!
failure: Try again. #>

**Note:** $e^{0,134} - 1 = 0.143$

We now focus on the explanatory power of this model. The calculations are based on the estimated regression coefficients for determining the relative changes in the elasticity variables and the sales ranks (as a relative index for demand).

#< quiz "Expl_Elasticity_Power" question: Which of the following answers is **not** correct? sc: - These elasticities are based on estimated coefficients, so significance levels and other coefficients of determination should be included for interpretation purposes - Book prices on the U.S. Market are inelastic despite the fact that the prices are not statutory fixed. - Since the cost of a book is only a small part of the total consumption cost, the actual price elasticity tends to be larger than 1.* success: Great, your answer is correct!
failure: Try again. #>

The coefficient for `Star Elasticity Overall` is nearly 0.15, meaning that a one percent increase in Amazon star ratings is associated with a 0.15 percent increase in demand. In the boot-strapping process, the authors estimated an annual average effect of 2.8 percent of New York Times reviews on sales, which I did not include in the problem set for performance reasons.

**Summary**

In order to be able to make economic statements about price elasticity and actual demand with the help of the relative investigations of Amazon Sales Ranks, absolute sales figures are required. To generate these, the authors created an exponential formula based on sales numbers from the Nielsen Company and earlier research from Chevalier and Goolsbee (2003). As a result, it has become possible to estimate sales figures and thus determine price elasticities accordingly. In addition to the price elasticities, we examined the static elasticity effects of professional reviews and crowd ratings and estimated an effect of Amazon star ratings of about 1.5 percent and of New York Times reviews of about 14.3 percent. However, these elasticities are based on a simple static model that significantly underestimates the effect on sales.

Once you have worked through this chapter, you will have completed the content portion of this problem set and will move on to the conclusion.

# 5. Conclusion

### Recapitulation

In the recapitulation part, all tasks covered in this problem set are compactly summarized and evaluated.

The motivation chapter provided important information for understanding the following tasks. First, it classified the book as an economic good, explained the current situation in the book market, and discussed why the book market is particularly suitable for this study. Second, basics were explained about how a market with many market participants behaves and how pre-purchase information can influence it.

The second exercise begins with an explanation of the data set used for the research. Important variables are explained to understand the following exercises and to provide information on how Amazon classifies and identifies its products. Descriptive research is also conducted to first analyse crowd ratings and prices and sales ranks change due to different levels of star ratings. In addition, prices and sales ranks are analysed at different levels of expression of professional and non-professional reviews in order to make initial assumptions about their impact. Finally, prices were observed within one year to determine seasonal fluctuations.

In exercise three, empirical research is conducted to estimate the specific effects on Amazon sales and prices. First, a naive regression is introduced to provide basic understanding of how crowd ratings behave after step by step supplementing multiple methods that are also used by the authors. These instruments include control variables, fixed effects, robust standard errors, and logarithmic estimators. To convey this, samples are drawn from the main data set. Before continuing with the replication of the main regression, an introduction has been provided how calculate long-term effects by using the lagged sales rank. In addition, an event study is replicated and explained to graphically illustrate the impact of professional reviews. To review the estimate for validity, distributions and intersections of targeted attributes from the data set were checked.

The fourth exercise shows how to convert relative Amazon sales ranks into quantities to determine price elasticities and provide a transition to welfare analysis. First, it is explained how the authors have developed a specific function based on book data to which they had access and a similar formula have been used by other researchers. This formula estimates conversion factors based on existing data to estimate a specific sales volume per book. As a result, an elasticity table is created that, unlike the authors, uses a static model to calculate elasticities.

Moreover, the underlying paper examines the welfare effects of those two considered types of pre-purchase information. The authors have found out that satisfaction is easier to track using Amazon star ratings than professional reviews. Two different welfare analyses are performed for each type of pre-purchase information (with and without information) in order to determine an ex-post consumer surplus using an up-scaled utility function. The results are reported in the form of a delta difference.

### Results

The paper, which forms the basis of the underlying problem set, examines the influence of professional reviews from magazines such as the New York Times and non-professional crowd ratings on a five-point scale based on Amazon sales ranks. In examining the impact of professional reviews, the authors distinguish between short- and long-term effects and found that a New York Times review without a recommendation increased a book's estimated sales by 55 percent within the first five days of publication. With a recommendation, the effect should account approximately 80 percent. The long-term effect of New York Times reviews on the sales volume should account 2.8 percent. The effect from other magazines should account 0.12 percent within the first ten days of publication. The value added by Amazon star ratings is likely to be ten times higher than that of professional reviews, according to Reimers and Waldfogel.

The results of this problem sets state that New York Times reviews increase short-term sales by an average of 14.3 percent, while the short-term effect of New York Times reviews with recommendation within the first five days of publication averages 17.6 percent. Due to the changes in calculating the lagged sales rank, the estimated long-term effect would account 70 percent for New York Times reviews and 89 percent for recommendations from the New York Times (calculations attached in the material). The estimated short-term Amazon star rating elasticity accounts 0.15 on average. Likewise, the authors claim that their

formulations imply a causal relationship. This thesis illustrates multiple intersections between different magazines exist, so that about 15 percent of the entire New York Times reviews are contaminated, while other magazines are between 30 percent and 67 percent contaminated by the New York Times. As for the methodology, the use of fixed effects does not exclude but only reduces endogeneity effects. To ensure this, the instrumental variable approach or the difference-in-difference approach would have been more appropriate. Finally, New York Times reviewers are completely autonomous in the selection of their books, which is why, despite my research, it cannot be assumed that the study design is completely randomized. Still, the New York Times has large short-term effects on sales, while Reimers and Waldfogel contend that crowd ratings tend to have large impacts on the economic welfare.

**Related Literature**

In addition to the paper of Reimers and Waldfogel, similar studies have also been published. For instance, Chevalier and Goolsbee published their study back in 2003. In their study, they examined the price elasticity from Amazon and BN.com and found out, that the price elasticity for books sold on Amazon was about -0.6. Regarding the discussion of the extent to which authors with higher name recognition benefit from "positive" or "negative" reviews, Berger, Sorensen, and Rasmussen (2010) found that authors with higher prior awareness (more than ten published books) are expected to benefit almost twice as much from positive journal reviews. In contrast, less prior awareness authors (less than two published books) should benefit particularly well from bad reviews, while well-known authors should receive a negative impact from bad reviews on the number of books they sell. Chevalier and Mayzlin (2006) explained the phenomenon by noting that books listed on Amazon are generally better known than books that are not listed, so these listed books are more likely to suffer (due to a bad review) than unlisted books. A similar study was published by Reinstein and Snyder (2005), where the authors estimated the impact of professional reviews on movies using a difference-in-difference approach. Using two different levels of ratings (from positive ratings), the authors found that a single thumbs-up from the rater should result in an 11 percent increase in sales, while two thumbs-ups should have a 25 percent effect on sales.

# 6. Literature

## Bibliography


Berger, J., Rasmussen, Scott J., Sorensen, Alan T. (2010). Positive Effects of Negative Publicity: When Negative Reviews Increase Sales, 29(5), 815-827.

Beurich, A., Götz, G., Sprang, C., Fuchs, A. (2023). Professor Götz presents results of the book prize binding project at press conference in Berlin.

Chevalier, J., Goolsbee, A. (2003). The Effect of Word of Mouth on Sales: Online Book Reviews. Kluwer Academic Publishers, (1), 203-222.

Curcic, D. (2023). Impact of Amazon On The Publishing Industry.

Davis, Andrew P., Hill, Terrence D., Roos, M., French, Michael T. (2020). Limitations of Fixed-Effects Models for Panel Data, Sage Journals, 63(3), 357-369.

Henderson Jr, Glenn V. (1990). Problems and Solutions in Conducting Event Studies. American Risk and Insurance Association, 57(2), 282-306.

Kranz, S. (2022). 1 Predicting Prices for Bordeaux Red Wines.

Kranz, S. (2023). Better understand models with lagged dependent variable as regressor.

Kwan, J. (2013). Why books cost more in Canada.

MacKinley, Craig A. (1997). Event Studies in Economics and Finance. Journal of Economic Literature, 35(1), 13-39

Mayzlin, D., Chevalier, J. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. Kluwer Academic Publishers, 43(3), 345-354.

Mcloughlin, D. (2022). Amazon Print Book Sales Statistics.

Nakayama, M. (2015). For Whats It's Worth: Fixed Book Prices in Foreign Book Markets. Publishing Trends.

Rachev, S. (2007). Robust Estimation.

Reimers, I., Waldfogel, J. (2021). Digitization and Pre-purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings. American Economic Association, 111(6), 1944-1971

Reinstein, David A., Snyder, Christopher M. (2005). The Influence of Expert Reviews on Consumer Demand For Experience Goods: A Case Study of Movie Critics. The Journal of Industrial Economics, 53(1), 27-51.

Sorensen, Alan T. (2007). Besteller Lists and Product Variety, 55(4), 715-738.

Stieb, M. (2022). Amazon's War on Fake Reviews. New York Intelligencer.

Train, K. (2015). Welfare calculations in discrete choice models when anticipated and experienced attributes differ: A guide with examples, Journal of Choice Modelling, 16, 15-22.

Wieneke, D. (2019). Customer strategy foundation: Search, Experience and Credence (SEC) analysis determines how consumers buy.

Wisniach, B. (2023). Amazon Sales Rank: Explained.

## R Packages

Arel-Bundock, V. (2022). modelsummary: Data and Model Summaries in R. Journal of Statistical Software, 103(1), 1-23. URL https://www.jstatsoft.org/article/view/v103i01

Berge L., McDermott, G. (2023). Fast Fixed-Effects Estimation: Short Introduction. URL https://cran.r-project.org/web/packages/fixest/vignettes/fixest_walkthrough.html

Gotelli, N. (2018). Multiple plots in ggplot2 with patchwork. URL https://gotellilab.github.io/GotelliLabMeetingHacks/NickGotelli/ggplotPatchwork.html

R Documentation, ggplot2 (version 0.9.1). geom_vline: Line, vertical. URL https://www.rdocumentation.org/packages/ggplot2/versions/0.9.1/topics/geom_vline

Zhu, H. (2021). Create Awesome HTML Table with knit::kable and KableExtra. URL https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html

Zezula, P. (2021). Grafiken mit ggplot aus dem package library(ggplot2). URL https://md.psych.bio.uni-goettingen.de/mv/unit/ggplot2/ggplot2.html