1

Motivation

The way in which we conduct empirical social science has changed tremendously in the last decades. Lewis Fry Richardson, for example, was one of the first researchers to study wars with scientific methods in the first half of the twentieth century. Among many other projects, he put together a dataset on violent conflicts between 1815 and 1945, which he used in his *Statistics of Deadly Quarrels* (Richardson, 1960). Richardson collected this information on paper, calculating all of the statistics used for his book manually. Today, fortunately, empirical social science leverages the power of modern digital technology for research, and data collection and analysis are typically done using computers.

Most of us are perfectly familiar with the benefits of digital technology for empirical social science research. Many social science curricula – for example, in political science, economics, or sociology – include courses on quantitative methods. Most of the readers of this book are trained to use software packages such as SPSS, Stata, or R for statistical analysis, which relieve us of most of the cumbersome mathematical operations required for this. However, according to my experience, there is little emphasis on how to prepare data for analysis. Many analyses require that data from different sources and in potentially different formats be imported, checked, and combined. In the age of "Big Data," this has become even more difficult due to the larger, and more complex, datasets we typically work with in the social sciences.

I wrote this book to close this gap in social science training, and to prepare my readers better for new challenges arising in empirical work in the social sciences. It is a course in data processing and data management, going through a series of tools and software packages that can

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

assist researchers getting their empirical data ready for analysis. Before we discuss what this book does and who should read it, let us start with a short description of the research cycle and where this book fits in.

1.1 DATA PROCESSING AND THE RESEARCH CYCLE

Most scientific fields aim to better understand the phenomena they study through the documentation, analysis, and explanation of empirical patterns. This is no different for the *social* sciences, which are the focus of this book. I fully acknowledge that there is considerable variation in the extent to which social scientists rely on empirical evidence – I certainly do not argue that they necessarily should. However, this book is written for those that routinely use empirical data in their work, and are looking for ways to improve the processing of these data.

How does the typical research workflow operate, and where does the processing of data fit in? We can distinguish three stages of an empirical research project in the social sciences:

1. Data collection
2. Data processing
3. Data analysis

The first stage, data collection, is the collection or acquisition of the data necessary to conduct an empirical analysis. In its simplest form, researchers can rely on data collected and published by someone else. For example, if you conduct a cross-national analysis of economic outcomes, you can obtain data from the comprehensive *World Development Indicators* database maintained by the World Bank (2021). Here, acquisition for the end users of this data is easy and just takes a few mouse clicks. Similarly, excellent survey data can be obtained from large survey projects such as the *Demographic and Health Surveys* (US Agency for International Development, 2021) or the *Afrobarometer* (2021). In other cases, data gathering for a research project is more difficult. Researchers oftentimes collect data themselves, for example by coding information from news reports or other sources, or by conducting surveys. In these cases, data collection is a fundamental part of the contribution a research project aims to make, and requires considerable resources.

The output of the first stage is typically a (set of) *raw* dataset(s). Before the raw data can be used in the analysis, it needs to be processed in different ways. This data processing can include different operations. For example, we may have to adjust text-based codings in our data, since our

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

statistical package can only deal with numbers. In many cases, we need to aggregate information in our dataset; for example, if our original raw data contains survey results at the level of households, but we conduct our analysis at the level of villages, we have to compute the sum or the average over all households in a village. In other cases, we have to combine our original dataset with others. For instance, if we study the relationship between the level of economic development and the level of democracy, we may have to combine information from the *World Development Indicators* database with data on regime type, for example from the *Varieties of Democracy* project (Coppedge et al., 2019).

The third stage in our simple research workflow is data analysis. "Analysis" refers to any kind of pattern we derive from the data prepared in the previous stage, such as a correlation table, a graphical visualization of the distribution of a particular variable, or the coefficients from a regression model estimated on the data. Data analysis – whether it is descriptive, graphical, or statistical – requires that our data be provided in a particular format, a format that is not necessarily the most convenient one for data collection or data storage. For example, if we analyze the relationship between development and regime type as mentioned earlier, it is necessary to combine data from different sources into a single dataset that is ultimately used for the analysis. Hence, separating data processing from data analysis – as we do in this book – is not simply a convenient choice in the research workflow, but rather a necessity.

1.2 WHAT WE DO (AND DON'T DO) IN THIS BOOK

The focus on the three stages of an empirical analysis suggests the following workflow, depicted in Figure 1.1. The first stage, data collection, produces one or more raw datasets that are input to the second stage, data management. At this stage, the data are processed in various ways: they are cleaned, recoded, and combined in order to yield one or more datasets for analysis, which are used during the data analysis stage. The gray box shows what is covered in this book: how to get from the raw dataset(s) to those used for analysis.

This depiction of the research workflow does not mean that these three stages are always carried out in strict sequence. Rather, in reality researchers will likely go back and forth between them. This is necessary when adding an additional variable to the analysis, for example, as a new control variable: In this case, we have to adjust the data processing step, such that the variable is part of the analysis dataset. In some instances,

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

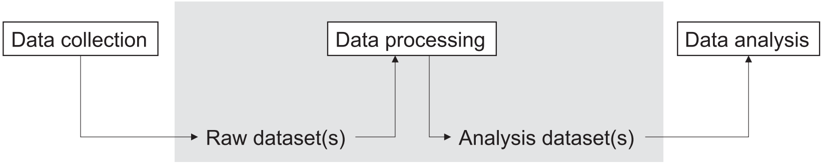

FIGURE 1.1. Research workflow in an empirical social science project. The gray area is what we cover in this book.

this may even require us to go back to the first stage (data collection), since we may not yet have the necessary information in our raw data. Similarly, we may have to go back to the data processing stage for much simpler operations, for example, when the format of a variable in our analysis cannot be processed by our statistical software or our visualization toolkit.

While useful for the purpose of illustration, the three stages are clearly a simplification of the typical research workflow and omit a number of additional steps. To name just one example, the "Data Lifecycle" by the US Geological Survey (Faundeen et al., 2014) is a more complete illustration of the different phases involved in an empirical research project. It includes six phases: (1) Plan, (2) Acquire, (3) Process, (4) Analyze, (5) Preserve, and (6) Publish/Share. The planning of a project in Stage 1 is obviously of key importance, and a social science research proposal normally includes all the necessary details for how the empirical analysis should be carried out. At this stage, one would also pre-register a study, for example, within the *Center for Open Science's Open Science Framework*. Stage 2 in their model corresponds to what we call "data collection," and Stages 3 and 4 correspond to our second and third stages. Stage 5 covers the documentation and physical storage of the data, such that it can later be accessed and used again. This stage is closely related to Stage 6, since documentation, anonymization, and technical description are required for both.

This book focuses on the practical aspects of data processing, and thus covers primarily the second step in our three-stage model. Research design and, in particular, data analysis are part of most social science programs that focus on empirical work, while basic questions of data management and processing are typically left out. This means that most researchers will be perfectly able to design their study and carry out an empirical analysis, while struggling with the processing of data. In this book, we will see that some of the standard practices that have evolved in the community can lead to inefficient workflows that make life difficult

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

for researchers, or even introduce errors in the data. For example, in contrast to conventional wisdom, spreadsheets are in most cases not a good choice for managing your data, even if they appear to be simple and intuitive. Also, managing and analyzing data in the same software can be challenging or even impossible, since data processing oftentimes requires specialized functionality that data analysis software simply lacks. Hence, I recommend to treat data processing as a step that is different from – but of course closely connected to – data collection (which comes before) and data analysis (which is what we do after the processing of our data).

The focus on hands-on data processing also distinguishes this book from what is typically referred to as "data management" without the focus on practical questions. This alternative definition of "data management" includes strategic questions of organizations about how to acquire, store, document, and disseminate research data (Henderson, 2017). These are typically issues that are addressed by dedicated organizational units, for example, university libraries. Of course, data management also needs to solve practical questions about processing and storage such as the ones we discuss here, and therefore overlaps partly with the content of this book. In other words, data management for large organizations requires much of the knowledge and skills I try to convey in this book, but also entails a number of other challenges we do not cover. Rather, the book caters to the needs of individual researchers or small research groups, who are oftentimes responsible for designing most of their data processing procedures themselves – something that, according to my own experience, is probably what most researchers in the social sciences do.

1.3 WHY FOCUS ON DATA PROCESSING?

Readers may wonder why we need an entire book on data processing. There are several reasons why researchers should devote more attention to working with data. In particular, I believe that there are several major advantages to treating data processing as a separate step in the research workflow, which requires particular skills and (potentially) specialized software.

DOCUMENTATION. One of the most important goals of this book is to show you how to properly document the processing of your data. That is, every operation you apply to the raw data you start from until you end up with a dataset for analysis must be written down, such that you – or someone else – can later return to it. For many of us, this type of documentation is standard practice for data analysis – that is, we produce

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

code in R, Stata, or another statistical package that executes the different steps required to produce a plot or to run a statistical model. Manually merging or aggregating data in a spreadsheet, for example, is very different; here, you essentially point-and-click to achieve the desired result, and these operations are difficult, if not impossible, to understand and repeat later. In contrast, (almost) all the different methods I present in this book are automated; they allow you to prepare and process your data using a set of instructions to a data management software. As a result, your research workflow improves in several ways.

CONVENIENCE. One of the advantages of fully documented data processing is simply the added convenience for you as the researcher. Automated data processing is more powerful and much faster, since you can let the computer process many entries in your dataset at once, rather than manually fixing them. Also, you can later modify your processing instructions in case you change your mind, or if you discover mistakes. By adjusting the data processing code, it is possible to change the coding of individual variables, introduce different types of aggregation, or derive datasets for different types of analysis from your raw data. All of this is extremely cumbersome if you resort to manual data management, for example, by using spreadsheet software.

REPLICABILITY AND TRANSPARENCY. Another major advantage of a fully documented data workflow from the raw data to the dataset used for analysis is the transparency resulting from this process. This documentation is not just an advantage to you as a researcher, but it allows you to share your data processing code in the research community, thus making your work perfectly replicable by others. The replication of empirical research has been at the forefront of current attempts towards increased research transparency (see, e.g. the DA-RT Initiative, 2015). Almost all major journals in the social sciences now require that the data and code used for the analysis be published along with an article. While this is a move in the right direction, increased transparency should also apply to the data processing stage. Thus, with the techniques presented in this book, it is possible to create fully documented data workflow, which can make data preparation transparent and replicable.

SCALABILITY. One of the benefits of the digital transformation is the increasing amount of data that becomes available to social scientists. Rather than analyzing a few dozen observations, as Richardson (1960) did in his empirical analysis of wars, researchers now possess datasets that are several orders of magnitude larger. For example, recent work

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

with data collected from social media can easily include millions of observations. This requires data processing techniques that can deal with large datasets, in other words, that scale with increasing amounts of data. This is where conventional tools quickly come to their limits. Spreadsheets are not suitable for these amounts of data, not just because manual editing of data is no longer possible, but also because they have an upper limit on the number of entries in a dataset they can process (for MS Excel, for example, this limit is about 1 million). Also, many statistical tools are not suitable, since they too have difficulties processing large datasets (although there are extensions that make this possible). In contrast, some of the more advanced tools I present in this book are perfectly scalable; they are designed to store and process large datasets while hiding most of the complexity of these operations from the user. Again, it is sometimes useful to use specialized, but different, software tools for data processing and data analysis, since each of them have different strengths and weaknesses.

VERSATILITY. Along with the increase in the amount of data that social scientists use for their work, we also witness an increase in the complexity of the data formats used. Rather than relying exclusively on single tables of data where observations are nicely arranged in rows and columns, there is now a variety of different types of data, each stored in a specific data format. For example, many different social science projects now use observations with geographic coordinates, where each observation in a dataset is tagged with a reference to a particular location on the globe: The *Demographic and Health Surveys*, for example, distribute geographic coordinates for their more recent survey waves, which makes it possible to locate each group of households that participated in the survey on a map. One potential use of these geo-coordinates is to combine the survey results with other information based on their location, for example, with night light emissions. Spatial data is just one example for new types of information requiring adjustments to the standard tabular data model; in this book, we present others, such that researchers can make an informed choice about their specific requirements and the software tools they should use for their work.

1.4 DATA IN FILES VS. DATA IN DATABASES

Most of the data we use in the social science comes in electronic files. That is, in a quantitative research project we rely on files to store the data we

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

have collected, and we use files to archive the data and to pass them on to other users. There are many different types of files for data: You are probably familiar with Microsoft's Excel files for storing spreadsheets, or Stata and SPSS files for tabular datasets. However, while files will continue to be the primary way by which we distribute data, they can be tricky to work with. For example, if the data is spread out across different files, you need to merge them before you can run your analysis. Also, you need to manually check for errors in your data, for example, whether a numeric variable mistakenly contains text. File-based data storage also means that multiple users can cause issues when accessing the same file, for example, if one user overwrites changes made by another user. Finally, file-based data storage can quickly get to its limits when we deal with large datasets. In most cases, the entire data contained in these files needs to be imported into your statistics package, where it stays as long as you work with it. This is not a problem if your dataset is not particularly large, but can be a real issue once you need to process a large amount of data. Even simple operations such as ordering or filtering your data can become extremely slow.

This is why for many applications, it is beneficial to store your data not in files, but in specialized data management software. We refer to these systems as "database management systems" (DBMS). DBMS have existed for a long time, and there are many different flavors. What they have in common is the ability to manage a set of databases for you. A database is a repository for all data required for a specific project. For example, if you intend to use a DBMS for a research paper, you would create a database for your project, which then contains a set of tables with all the data for this project. DBMS optimize the processing of the data contained in their databases. For example, some types of databases are designed for tabular data. They make sure that a table does not contain basic errors (e.g., non-numeric text in a numeric column), and they support the quick merging of tables that depend on each other. DBMS also facilitate efficient filtering and ordering of your data, and they can be accessed by different users concurrently. All this happens behind the scenes – users do not have to worry about where the data is physically stored, or how to enable fast and efficient access to them. Connecting to a database is possible from almost any type of statistical software or programming language – in this book, we use R to do this. Using R's database interface, you can send requests to the database server, for example, for changing or updating the data on the server, and for fetching data directly into R for further analysis.

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

1.5 TARGET AUDIENCE, REQUIREMENTS AND SOFTWARE

This book will be useful for any social scientist working with empirical data. Here, I define "social sciences" broadly, and include fields such as economics, sociology, psychology, anthropology, linguistics, communication studies and of course political science, my home discipline. I fully realize that there are significant differences across these fields when it comes to the predominant statistical methods they use; however, I also believe that these differences typically manifest themselves at the data analysis level. For example, while regression analysis is one of the main types of statistical approach used in my field (political science), psychologists may be more used to factor analysis. Importantly, however, there are few differences in the way the data needs to be prepared for these different types of analysis. In other words, the analysis dataset will likely be the same, regardless of whether it is later used in a regression model or a factor analysis. For that reason, the concepts and tools we cover in this book can likely be used across many different fields in the social sciences.

This is an applied book, and I try to illustrate our discussion with real examples wherever possible. Owing to my own background, these examples are largely drawn from my own discipline. At the same time, however, the book requires no substantive background in political science, and each example will be briefly introduced so that all readers, regardless of their background, can understand what research question we deal with, and what data we use in the example. Similarly, the book is designed to address social scientists of different generations. I believe that a deeper engagement with practical questions of data management will be useful for social science students as part of their training in empirical methods, or while working on their first research project. Still, the book also speaks to more advanced researchers at universities, governmental and non-governmental organizations, and private companies that have an interest in improving their quantitative research workflow.

I fully realize that the book's readership will differ strongly in their technical experience, partly due to the variation in quantitative training that people have gone through. As regards the latter, it would be very difficult to custom-tailor this book to different statistical packages that readers have experience with. For this reason, I mainly use the R statistical toolkit in this book. R is free and open source, which means that there are no expensive licences to be purchased to work with the book. Also, R is one of the most flexible and powerful programming packages for statistical analysis out there, and can be used not just for estimating statistical

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

models, but also for advanced data management. Still, despite our focus on R in the code examples, many of the basic concepts and procedures for data management we cover are not tied to R, and therefore apply irrespective of which statistical software you use. While I provide many step-by-step examples in R, the book does not include a basic introduction to this software. It is therefore required that readers have some experience in R, either as part of their training in quantitative methods, or from one of the numerous R introductory books and courses that are available.

1.6 PLAN OF THE BOOK

The book consists of five parts. This and the following two chapters together constitute the introduction. So far, we discussed the role of data preparation and management within the research cycle, and defined the scope of the book. In Chapter 2, we go through the setup of the software used in this book. We rely mostly on R, but the advanced chapters on database systems require us to install a DBMS locally. Chapter 3 provides a conceptual introduction to data as a combination of informational content with a particular structure. We discuss the most important data structure in the social sciences (tables), but also talk about their design. In this first part of the book, we do not use any real examples for readers to practice – this starts only in the second part.

The second part of the book covers the processing of data stored in files. This is by far the most frequently used type of workflow in the social sciences, which is why we start with an overview of file-based data storage and different file formats in Chapter 4. In Chapter 5, we focus on data management with spreadsheet software such as MS Excel. This software is easy to use and most readers will be familiar with it. Still, it encourages certain bad practices for data management that we discuss (and caution against) in this chapter. We then turn to data management using R. The first chapter on this topic (Chapter 6) introduces R's basic features for reading data tables, as well updating and merging them. Chapter 7 presents a powerful extension of R's basic functionality: the `tidyverse` environment.

The third part of the book deals with data stored and processed in specialized systems, so-called databases. Here, data no longer reside in files, but are contained in database systems that users interact with via the network. This has a lot of advantages when it comes to avoiding errors and inconsistencies in the data, but also for handling large and complex

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press

datasets accessed by several users. Relational databases for tabular data constitute probably the most frequent type of database, and we cover them in two chapters. We start with a single table in Chapter 8, and later extend this to several tables in Chapter 9. In Chapter 10, we address important technical features of relational databases, such as the ability to efficiently work with large datasets, or to allow collaborative access by different users.

The fourth part of the book addresses more specialized types of data that do not neatly fit into the standard tabular model. For each of these data types, we discuss (i) file-based data processing using R and the corresponding extension libraries and (ii) data processing using a database system. We start with a discussion of spatial data, that is, observations that have geographic coordinates attached to them (Chapter 11). The subsequent chapters cover text as data (Chapter 12) and network data (Chapter 13).

In the fifth part of the book, we conclude our introduction to data management with some recommendations for collaborative data projects, as well as for the publication and dissemination of research data.

The aim of this book is *not* to give readers detailed, in-depth introductions to the different tools and techniques we discuss. Rather, my goal is to convey a good intuition of the key features, but also the strengths and weaknesses, of the different tools and approaches for data management. This way, readers get a good overview of the available techniques, and can later choose a software and workflow that best matches their research needs.

https://doi.org/10.1017/9781108990424.003 Published online by Cambridge University Press