CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

PART I. MODELS AND METHODS IN THE SOCIAL SCIENCES

Andrew Gelman

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

1. Introduction and Overview

WHAT THIS BOOK IS ABOUT

The seven parts of this book cover different examples of quantitative reasoning in social science applications, along with interspersed discussions on the role of quantitative theories and methods in understanding the social world. We are trying to show how researchers in different social sciences think about and use quantitative ideas. These approaches vary across the different disciplines, and we think that the best way for a student to get a broad overview is to see how quantitative ideas are applied in a variety of settings.

Hence, we focus on applications, including combat in the First World War, demographics in nineteenth-century Mexico, forecasts of economic conditions by the Federal Reserve, racial disturbances in the 1960s, voting in congressional committees, Pavlovian conditioning, and the effect of migration on solidarity. Each part of the book has several examples, with discussion of the theories used to explain each phenomenon under study, along with the data and models used to describe what is happening. We thus hope to give a general perspective of how these models and methods are used in the social sciences. It is not the usual straightforward story of the form, "Here are some data, here's the model, and look how well the method works." Rather, we explore the strengths but also the limitations of the models and methods in the context of the real problems being studied. Mathematical models and statistical methods can be powerful – you can fit a linear regression model to just about anything – but one needs some sense of what the models mean in context.

The style of the book is conversational and episodic. We do not intend to cover every mathematical model and every statistical method, but rather to give a specific sense of the diversity of methods being used. We believe that the reader of this book or the student in a course based on this book will gain a broader view and some specific tools that he or she can use in attacking applied problems with a social science dimension. This can be especially valuable for a student trained in a specific social science who has only a vague sense of what ideas are used in other fields. A psychological view can be important in understanding a problem that was thought of as pure economics, political scientists need to appreciate the strengths and weaknesses of their sources of historical data, social psychologists

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

need to comprehend the large-scale studies of sociologists, economists need to think seriously about the workings of governments, and so forth.

Just to be clear: This is not a statistics or a research methods book but rather a social science book with a quantitative focus. We certainly think that social science students need to learn some statistics, and we see the book's ideas as complementing that study. We do discuss linear regression (in Part III) and Poisson models (in Part IV), but we focus on how these answer (or fail to answer) the substantive questions in the particular applications considered. We also discuss some purely mathematical models (for example, game theory in Chapter 3 and decision theory in Chapter 20).

GOOD SOCIAL SCIENCE IS INTERDISCIPLINARY

We don't attempt to define social science except to say that whatever it is, it is studied from different perspectives by historians, economists, sociologists, political scientists, and psychologists. Everybody has his or her own turf, and that's fine, but as a student – or, for that matter, as a practitioner – it's good to know your way around the farm. To some extent, this already happens – quantitative political scientists learn economics, economists use results from cognitive psychology, and just about everybody should learn some history. In this book, we're trying to lay out some concepts in each of the social sciences that can be part of everybody's toolkit in understanding the social world.

We focus on quantitative methods because that's a common ground, an objective reality on which we can all agree – or, where we disagree, we can focus our disagreements on specific questions that, at least in theory, can be answered by hard data. This point arises in Part II of the book; for example, ideas about families that many of us take for granted are associated with demographic patterns in Europe hundreds of years ago. In Part III, there's a discussion of some methods used to assess how much information different economic actors have in predicting exchange rates, and Part IV describes a model for promotion of employees within a large corporation. Different methods can be used to attack similar problems. In political science there's a lot of research on the topic of bargaining, which can be studied quantitatively by looking, for example, at how legislators vote on different issues and to what extent these votes are consistent with political ideology. Part VI discusses several different kinds of theories used to explain decisions and behavior. Some of these theories are more quantitative than others, and it is important to understand under what circumstances such theories can be tested. This issue arises in all the social sciences.

Why do the social sciences use such different methods? One reason is that people study different phenomena in different fields; for example, a discrete Markov transition model[1] might be reasonable to describe changes in a person's

[1] In a "discrete Markov model," a person can be in one of several *states* (for example, 1, 2, or 3, corresponding to "unemployed," "underemployed," or "employed"). The person starts in a particular state, and then at each of a sequence of time points (for example, every month), he or she can stay in the same state or switch to a different state, with specified "transition probabilities" corresponding to each pair of states. Examples of Markov models appear in Parts IV and V of this book, and they can

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

employment status but inappropriate for modeling gradual changes in public opinion. Another reason for the differences might be personal tastes, traditions, and even political ideology. In any case, if you can understand and communicate in these different fields, you can move to integrate them in solving applied problems, which are never confined to a single academic discipline. Part VII, the last part of this book, provides a more comprehensive, although by no means exhaustive, discussion of causal inference in the social sciences.

CONTROVERSIES IN ESTIMATING THE DOLLAR VALUE OF A LIFE

Mathematical and statistical models can be extremely effective in studying social situations, especially when many methods are used and none is relied on entirely. To take just one case, the economist Peter Dorman wrote a book in 1996 on workers' compensation and valuing human life. The value of a life has long been a contentious topic. From one perspective, it seems immoral to put a dollar value on lives, but in practice this must be done all the time, in settings ranging from life insurance to building code regulations to the cost of air bags in automobiles. For workers' compensation (insurance payments for on-the-job injuries), it is sometimes suggested that dollar values be set based on workers' own valuations of job risks, as revealed by the "risk premium" – the additional amount of money a person will demand in order to take on a risky job. This risk premium can itself be estimated by statistical analysis – for example, running a regression of salaries of jobs, with the level of risk as a predictor variable. In his book Dorman discusses the history of economic analyses of this problem, along with problems with the current state of the art. (For one thing, jobs with higher risks pay *lower* salaries, so the simplest analysis of risks and salaries will "show" that workers actually prefer riskier jobs. An appropriate analysis must control for the kind of job, a perhaps impossible task in practice.)

But what really makes Dorman's book interesting is that he includes historical, political, and psychological perspectives. Historically, workers' compensation has been a government solution to a political struggle between unions and management, so nominal dollar values for risks have been set by political as well as economic processes. In addition, psychologists have long known that attitudes toward risk depend strongly on the sense of responsibility and control over the dangers, and it should be no surprise that people are much less willing to accept new risks than to tolerate existing hazards. Dorman puts all this together, along with some theoretical economic analysis, to suggest ways in which labor,

apply to entities other than persons (for example, a plot of land could be unused or used for residential, agricultural, industrial, commercial, or other purposes, and it can move between these states over time). Such models can be relevant for studying the progress of individual persons or chains of events, but they are generally less appropriate for studying gradual behavior of aggregates. For example, one might consider modeling public opinion in the United States as supporting the Democrats or the Republicans or as balanced between the parties. Such a characterization would probably not be as useful as a time-series model of the continuous proportion of support for each of the parties.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

management, and government could negotiate reasonable compromises in assignment of risks.

Dorman's particular interdisciplinary approach is controversial, however. In many different studies, economists have attempted to refine the regressions of wages on risk (these attempts are considered an example of "hedonic regression," so called because they estimate individuals' preferences or judgments about what will make them happiest) to adjust for differences among the sorts of people who take different jobs (see, for example, Costa and Kahn 2004; Viscusi 2000).

As Dorman points out in his book, these questions are politically loaded, and he and others are necessarily giving their own perspectives. But I still believe that the quantitative analyses are useful as long as we are able to evaluate their underlying assumptions – both mathematical and historical.

FORECASTING AS A SOCIAL SCIENCE PROBLEM AS WELL AS A STATISTICAL PROBLEM

I'd like to try to separate the details of a statistical forecasting model from what we learn simply from the existence of a successful forecast. This general idea – that quantifying your uncertainty can reveal the underlying structure of a system – is important in all the observational social sciences, most notably economics, and also is used in engineering in the area of quality control.

Forecasting is an important – in some ways fundamental – task in economics, as discussed in Part III. In addition to its importance for practical reasons (for example, to estimate costs and set prices), the limits of your ability to forecast tell you how much information you have about the underlying system that you are studying. More precisely, the predictive error of your forecast reflects some combination of your ignorance and inherently unpredictable factors (which will always occur as long as humans are involved). Economists sometimes talk about an ideal forecast, which is not a perfect, zero-error prediction but rather is the most accurate prediction possible given the inherent variation in a system between the current time T_0 and the time T_1 about which you are trying to forecast.

To put it another way – this may be a familiar idea if you've studied linear regression – the residual error or unexplained variance represents the variation in the outcome that is not explained by the predictors. For example, a regression of income on job category, years of experience, and an interaction between those predictors will not be perfect – that is, its R-squared will be less than 100% – which makes sense since other factors affect your income (in most jobs). From a social science perspective, the very fact that the regression predictions are accurate only to a certain level is informative, because it tells us how much other factors affect income in this particular society. Statistically, this is one of the ideas motivating the analysis of variance, which involves studying how important different factors are in explaining the variation we see in the world. In addition, the coefficients in a regression model can tell a story by revealing which predictors carry information, which is discussed in detail at the end of Chapter 9.

To return to time-series forecasting, if a variable can be accurately predicted two years ahead of time, then intervening factors during those two years cannot be important in the sense of affecting the outcome. What is more common is that a

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

forecast several years in advance will have a large error variance, but as the time lag of the forecast decreases, it will become increasingly accurate – but with some limit on the accuracy, so that a very short term forecast (for example, of tomorrow's prices) will still not be perfect.

From a social science point of view, the situations and time lags where a forecast has large errors provide an opportunity to try to understand the intervening variables that make a difference. (A similar point comes up in Part IV, where data on racial disturbances do *not* follow a Poisson model, motivating a search for further understanding.) We consider this in more detail for election forecasting.

There are many ways to predict the outcomes of elections. In the following discussion we focus on U.S. congressional and presidential elections, but the same ideas apply to other contests in the United States and other places where elections are held regularly. You can forecast two years ahead of time, or three months before the election, or one week before, or on election night. Each of these forecasts can be useful for different purposes, and each uses different information. For example, to forecast an election two years early, you'll use past election results, forecasts of the economy in two years, and general political knowledge such as that the president's party often does poorly in off-year elections. Three months before the election the candidates are known, so you can include information on candidate quality. You also have some information from polls, such as presidential approval ratings and answers to the question "Who would you vote for?" as well as more reliable economic data. These forecasts have received a lot of attention in the economics and political science literature (studies by Ray Fair [1978], Steven Rosenstone [1983], Bob Erikson and Chris Wlezien [1996], James Campbell [1996], Michael Lewis-Beck and Tom Rice [1992], and others) because of their inherent interest and also because of the following paradox: The final months of election campaigns sometimes feature dramatic reversals, but if elections can be accurately forecast three months ahead of time, this seems to allow little room for the unexpected. As Gelman and King (1993) put it, why do preelection polls vary so much when elections are so predictable? In this setting, the very existence of a forecasting model has political implications.

To continue briefly with the election forecasting example, the next step is prediction on election night itself. Partial information – the election results from the first precincts and states to report – can be used to forecast the whole. Time-series forecasts are used in an interesting way here: From a previously fitted model, you can get a forecast for each precinct and state. As actual election returns come in, you can see where each candidate is performing better or worse than predicted and then make inferences about the forecast error – and, from there, the actual election results – in the rest of the country. The important fact, from a modeling standpoint, is that these cross-sectional errors are in fact correlated and that correlation is part of the model used to make the forecast.

Although interesting as a statistical problem and important to news organizations, election night forecasting is not particularly interesting from a political or economic perspective. It is certainly no surprise that elections can be forecast from partial information, and, in any case, you can always just wait for the morning to find out who won. An exception was the 2000 election, where

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

forecasting models had a sort of afterlife, being used, among other things, to show the implausibility of Patrick Buchanan's vote total in Palm Beach County, Florida (see, for example, Adams 2001).

The example of election forecasting illustrates how similar statistical models can have different social science implications. U.S. presidential elections can be forecast fairly accurately months ahead of time. This implies that campaigns, in aggregate, have little effect and in some ways seems to imply that our vote choices are less the product of free will than we might imagine. Social scientists found similar results decades ago (Converse 1964): From a series of opinion polls they found that voters' preferences were very stable, even across generations; people inherit their political views much as they inherit their religion (or, for that matter, their height and body type). In contrast, the accuracy of election night forecasts, based on a decomposition of vote swings at the national, regional, state, and local levels, is somewhat reassuring in that these forecasts reinforce the idea that the United States is a single political entity but with local characteristics. Once again, similar results have been found in public opinion polls: Page and Shapiro (1992) use the term "parallel publics" to refer to the phenomenon that national changes in opinion tend to occur in parallel throughout different segments of the population. Similar models are used for tracking information in other social sciences but to different ends: for example, the study of individual judgment and decision making in psychology and the study of information flow and communication in economics.

OVERVIEW OF THE BOOK

This book is about mathematical models; ways of understanding quantitative data; concepts of probability, uncertainty, and variation; and sources of historical and current data – all applied to the understanding and solution of social problems. With parts on each of five social sciences – history, economics, sociology, political science, and psychology – we teach by example some of the different ways in which researchers analyze social phenomena using mathematical ideas and quantitative data.

Although this is not a statistics textbook, we refer throughout to the use of statistical methods. Our focus is on the social problems being modeled. Also, we are not attempting to present a unified quantitative view of the social sciences; rather, we favor a more pluralistic approach in which different methods are useful for different problems. This book is intended to make the reader comfortable with the kinds of quantitative thinking used in the different social sciences.

In the next two chapters, I'll present some specific examples in which quantitative modeling or analysis has had an impact. My focus will be on how the mathematical or statistical methods link to the underlying social science. The rest of the book, as noted previously, is divided into separate parts for each of the major social sciences. The parts have little overlap, which is intentional: We want you to see the different perspectives of the different social sciences.

We continue our tour of the social sciences with history. Much of our understanding of history (including the recent and immediate past) is based on numbers, ranging from national statistics on populations, trade balances, public

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

health records, and voting records to data that are important for particular studies such as those on dietary consumption, television ratings, and sports attendance. (In between are data sources such as national opinion polls, which are privately collected but used for public policymaking.) In addition, nonnumerical data can be summarized quantitatively; for example, in the study of a historical debate, contemporary newspaper articles for and against the issue can be counted. Part II of this book begins with a discussion of the historical uses of statistics and then considers various primary sources of quantitative historical data. The lecture material is best supplemented with two kinds of exercises: first, those in which students must collate some historical data (for example, from city property records or old newspapers) and, second, those in which they must read secondary sources and summarize a scholarly debate about some quantitative question with historical implications (for example, the number of people living in the Americas in 1492).

We then move to economics, which is increasingly important in a variety of settings, from private-sector decision making (for example, deciding how seriously a consumer products company should take its sales forecasts) to cost-benefit analyses such as evaluating how much the population is willing to pay for public parkland. Part III begins with a general discussion of the difficulties of understanding and using economic data and then presents three examples of analysis of time series of interest and exchange rates. The focus is not on the particular statistical models being used, but rather on the way in which economic concepts are encoded into the models. All the analysis is done using simple least squares or two-stage least squares, and the key concerns are the predictor variables in the equations and the interpretation of the coefficients. One thing all these examples have in common is that they use macroeconomic data to study the flow of information, which is a major theme in modern economics. Doing the exercises for Part III is a good way to learn data analysis on the computer, starting with simple downloading and plotting of time-series data and then fitting some simple time series and regression models. For the model-fitting exercises, it is important to interpret the coefficients in the model in terms of the motivating economic questions in order to have a good understanding of the theoretical implications of the model

Part IV covers sociology, which in many ways is the most flexible and interdisciplinary of the social sciences. At the theoretical level, sociologists consider a variety of models for social interactions and the roles of individuals in groups, with no single underlying set of principles (unlike in economics). In addition, sociologists often invent new methods of analysis to study unusually structured data (for example, on social networks). Part IV begins with a general discussion of social science theories and the ways in which a single phenomenon can be understood from different theoretical perspectives. It then presents two examples of models for discrete events in which quantitative data are used to understand violent incidents in race relations. For these examples, we discuss the challenges of modeling as well as accurately summarizing historical data. Finally, there is a discussion of models for promotion of employees within large organizations. This sort of analysis can be useful in a wide variety of contexts when studying the fates of individuals within a large social structure. The mathematical ideas used in sociology, such as Markov chains and the Poisson distribution, are

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

not extremely complex, but they are new to many students (who are more likely to be trained in linear regression models). Hence, the exercises for the chapters in Part IV are focused on these models. For example, in one assignment you must come up with a set of count data to which you must fit the Poisson distribution and then comment on any lack of fit of the model.

We continue with Part V on political science, which encompasses both the particular study of individual governmental processes and the general study of political behavior. At a practical level, anyone trying to solve a problem – inside or outside of government – should be aware of the presence of governments at local, state, and national levels, whether as means for solving the problem or as arenas for competing possibilities. More generally, political science – the study of ways to work around conflict – is also relevant in nongovernmental settings. Along with economics, political science as a field can be highly quantitative due to the availability of vast amounts of numerical data (particularly in more industrialized countries) on taxes and spending, voting, and opinion polls. At the same time, political theory can be highly mathematical. Part V contains detailed discussions about how social theories can be constructed so that they can be testable using empirical data. This is followed by specific examples of linear regressions and more sophisticated methods using game theory and Markov chains to model political decision making and voting in committees.

Part VI focuses on psychology, which is both a cognitive and a social science – that is, it studies people's internal processes as well as their interactions with others. Psychology is clearly necessary, at some level, for understanding all social phenomena, since they all are ultimately based on human actions. For example, the laws of micro and macroeconomics and the observed regularities of politics require some sort of consistent behavior, either of individuals or in the aggregate. Part VI starts with a general discussion of theories of human behavior and how these theories can be tested or disproved. It continues with descriptions of some mathematical models of decision making, which in turn suggest where classical economics might be relevant to describing individual behavior and where it might fail. A useful exercise in Chapter 18 is to compare the predictions that several different psychological theories would each make about an actual experiement. Chapter 20 presents some exercises on decision making and the combination of information.

The book concludes with Part VII, a discussion of causal inference in the social sciences. It focuses on the ever-present issue of causal versus spurious relations and discusses the example of estimating the effects of migration on solidarity in migrant-sending communities.

Our chapters go back and forth between general discussions of social science modeling and applied examples that are designed to allow class discussion. The material can be used as a starting point to explore the variety of mathematical and statistical models used in social science, with the ultimate goal of enriching your understanding of the complexities of the social world.

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-86198-4 - A Quantitative Tour of the Social Sciences
Edited by Andrew Gelman and Jeronimo Cortina
Excerpt
More information

2. What's in a Number? Definitions of Fairness and Political Representation

A key starting point of quantitative social science is *measurement* – which encompasses direct observations from personal interviews and field observations, large-scale data collection efforts such as surveys and censuses, structured observations in designed experiments, and summary measures such as the Consumer Price Index. All these form the raw material for larger social science studies, and it is easy to get lost in these analyses and forget where the numbers came from and, even more importantly, what they represent.

We illustrate the choices involved in numerical measurements in the context of a subject of general interest – how people are represented in a political system – that can be studied both at a theoretical and an empirical level. I want to make the case that quantitative summaries can be helpful, as long as we are aware of the choices that must be made in summarizing a complex situation by a single set of numbers.

We want our political system to represent the voters and treat them fairly. At the simplest procedural level, this means giving a vote to each citizen and deciding elections based on majority or plurality rule. In practice, however, we are not all represented equally by the government, and as long as there is political disagreement, there will be some dissatisfaction. It would be appealing to have a mathematical definition of the amount of citizen "representation." Unfortunately, different measures of representation can interfere with each other, as we discuss with examples from national elections in the United States.

WHAT IS THE MEANING OF POLITICAL REPRESENTATION?

The United States is a representative democracy, and we vote for people who represent us: congressmembers, the president, state legislators and governors, and local officials. Indirectly, through our elected representatives, we vote for the justices of the Supreme Court and other persons in appointed positions.

What does it mean for us to be represented in this political system? For one thing, everyone's vote counts equally, and as a consequence of two Supreme Court rulings in the 1960s, most legislatures are set up so that the number of

© Cambridge University Press www.cambridge.org