Reserve Bank of India Occasional Papers
Vol. 42, No. 1: 2021

Quantitative Social Sciences: An Introduction by Kosuke Imai, 432 pp., Princeton University Press, USA (2018), US$95.00

Over the last few decades, quantitative social sciences have flourished in various areas at an exponential rate. Quantitative social sciences are an interdisciplinary field encompassing subjects such as economics, education, public policy, political science, psychology, and sociology. Data and computational revolution have contributed to the rapid growth in quantitative social sciences. Due to these technological changes, the sheer volume of data available to quantitative social scientists has increased drastically. With a wide variety of data availability, the boundaries of quantitative social sciences research have expanded. The book *Quantitative Social Sciences: An Introduction* written by Kosuke Imai, an eminent Professor in the Department of Government and the Department of Statistics at Harvard University, follows this modern trend. The author raises the concern that the traditional statistical approaches cannot meet the current demands of a society. The author believes that one can contribute to the society through data-driven discoveries requiring greater emphasis on learning to analyze data, interpret the results and communicate the findings. This book introduces three elements of data analysis for social science research: research context, programing techniques, and statistical methods. The book extensively uses R, which is a powerful programing language for data analysis. As an open-source statistical programing language it is available for free download and runs on any Macintosh, Windows, and Linux computer.

The book consists of eight chapters. Chapter 1 explains how vital quantitative social science research is to a modern society. This chapter also provides a brief introduction to R. Chapter 2 introduces causality, which plays an essential role in quantitative social science research. It helps us to understand whether a particular policy or program (intervention) changes the outcome of interest. In this chapter, the author discusses the fundamental issue of causal inferences. An estimation of causal effect involves a comparison

between factual and counterfactual outcomes. The chapter begins with the study of racial discrimination in the labor market, trying to understand whether a black American candidate who did not receive a job offer would have otherwise received it if the candidate was white. In this study, researchers sent fictitious job applicants' resumes to potential employers after randomly choosing the applicant's name as either black American or Caucasian. This kind of study is known as experimental. From this study, one could understand how randomization of treatment assignment (intervention) enables researchers to identify the average causal effect of treatment. The author also discusses causal inference in observational studies where researchers do not control the treatment assignment (no intervention). It is demonstrated by understanding the impact of increasing the minimum wage on employment. Many economists argue that an increase in minimum wage can reduce work because employers must pay higher wages to their workers, and as a result, employers hire a few workers. But the decision to increase the minimum wage is not random, and may depend on many factors, which could influence how firms find themselves in the treatment group—A simple comparison between those who received treatment and those who did not may lead to biased inferences. The author introduces various research design strategies to reduce bias, such as cross-sectional (comparing treatment units with control units after a treatment), before-and-after (comparing the same unit before and after a treatment), and difference-in-difference (evaluating pre-treatment and post treatment measurements obtained for both treatment and control groups). The results reported in the book suggest that an increase in minimum wage had no negative impact on employment.

Chapter 3 investigates measurements. Accurate measurement is essential for any data-driven discovery because the bias in measurement can lead to incorrect conclusions and faulty managerial decisions. The author explains how to measure public opinion through sample surveys and illustrates this by showing how researchers attempted to measure the degree of support among Afghan citizens for international forces and the Taliban during the Afghanistan War. The results show that civilian attitudes are asymmetric. The harm inflicted by international forces is met with reduced support for international forces and increased support for the Taliban. But Taliban inflicted harm does not translate

into greater support for international forces. This chapter shows the ability of randomization in survey sampling. Here the author discusses potential biases in survey sampling and how nonresponses can undermine the representativeness of a sample. Also, in survey sampling, misreporting poses a severe threat to statistical inferences. Further, the author discusses the measurement of latent or unobserved concepts essential for social science research. In order to explain this notion, the author discusses political ideology. The author demonstrates by showing how overall ideological orientation in the US Congress has changed over time by using the voting records of each member of congress. The result shows that political polarization has increased in recent years among major political parties. For data analysis, the author introduces a basic clustering algorithm called K means algorithm, an unsupervised learning technique. (In unsupervised learning, there is no outcome variable. Instead, the goal is to discover hidden structures in data).

Chapter 4 dwells on prediction. Prediction is the first step towards understanding the complex causal relationships that underlie human behavior. Prediction is an essential component of the policy and decision-making processes. The author illustrates this by predicting the US presidential election results through prelection polls. The exercise makes an accurate prediction by combining multiple polls. The author also analyses data from a psychological experiment in which subjects are shown the facial pictures of unknown political candidates and asked to rate their competence. The analysis indicates that quick facial impressions can predict election outcomes. The author uses the regression technique to forecast the values. The chapter also introduces regression discontinuity design for making causal inferences. It is demonstrated by investigating how British politicians accumulated wealth when they were holding political office.

Chapter 5 is about the discovery of patterns from various types of data. This chapter analyses textual data to discover topics and predict the authorship of documents based on the frequency of words used. The application illustrated here is the authorship prediction of The Federalist Papers, which forms the basis of the US constitution. Some of the papers have known authors, while

others do not. By analysing the frequency of certain words in papers with known authorship, the author can predict the unknown authorship of papers. Further, the author uses network data to record the relationship among various entities. It is illustrated by exploring the marriage network in Renaissance Florence and a modern example of the twitter-following network among politicians. Here, the author discusses various measures of centrality (this is done to identify influential entities in a given network). Finally, the author discusses spatial and spatial-temporal data. The author shows how maps can visualize spatial patterns effectively using John Snow's famous study of the Cholera outbreak in London during the 19th century. John Snow utilized a natural experiment to uncover the primary cause of the outbreak in London. John Snow showed that sewage-contaminated water caused the outbreak. Finally, the author illustrates using maps the expansion of Walmart stores in the US over the last several decades. To understand how the spatial pattern changes over time, the author uses animation that sequentially displays a series of maps.

Chapter 6 shifts the focus from data analysis to probability. It discusses the two dominant perspectives on probability: Frequentist and Bayesians. Further, it discusses Baye's rule and uses it in predicting individual ethnicity using surnames and resident location when no survey data is available. Further, it discusses the essential concepts of random variables and probability distribution. The author uses these concepts to quantify the sampling uncertainty regarding polling prediction of election results. This chapter explains two fundamental theorems in probability: the law of large numbers and the central limit theorem. Chapter 7 introduces a framework for methods of statistical inference. It enables quantifying the degree of uncertainty regarding the estimates. This chapter shows us how to distinguish signals from noise through the computation of standard errors and confidence intervals, and hypothesis testing. This chapter also discusses the pitfalls of multiple testing and publication bias. It concludes by describing ways to make inferences from linear regression models with measures of uncertainty. The last chapter offers a brief discussion of what readers may need to pursue after finishing this book. To become a practicing methodologist, the authors convey the importance of a solid foundation in multivariable calculus, linear algebra

and probability theory, after which, individuals need to learn about statistical theories and various modeling strategies in a rigorous fashion.

I wholeheartedly recommend this book to beginners of social science research and those having interest in designing and implementing quantitative research techniques. All data sets used in this book are available online. An added advantage is the online review questions, enabling one to learn the basics first before attempting the exercises. While the datasets and most of the examples in the book relate to the USA, that should not work as a constraint to effective learning. Importantly, the book requires no prior programming experience and exposure to only elementary algebra. Through this book, the researchers can discover the true power of data analysis.

Mayank Gupta*

* Mayank Gupta is Consultant (Applied Econometrics) in Department of Economic and Policy Research at the Reserve Bank of India, Mumbai.