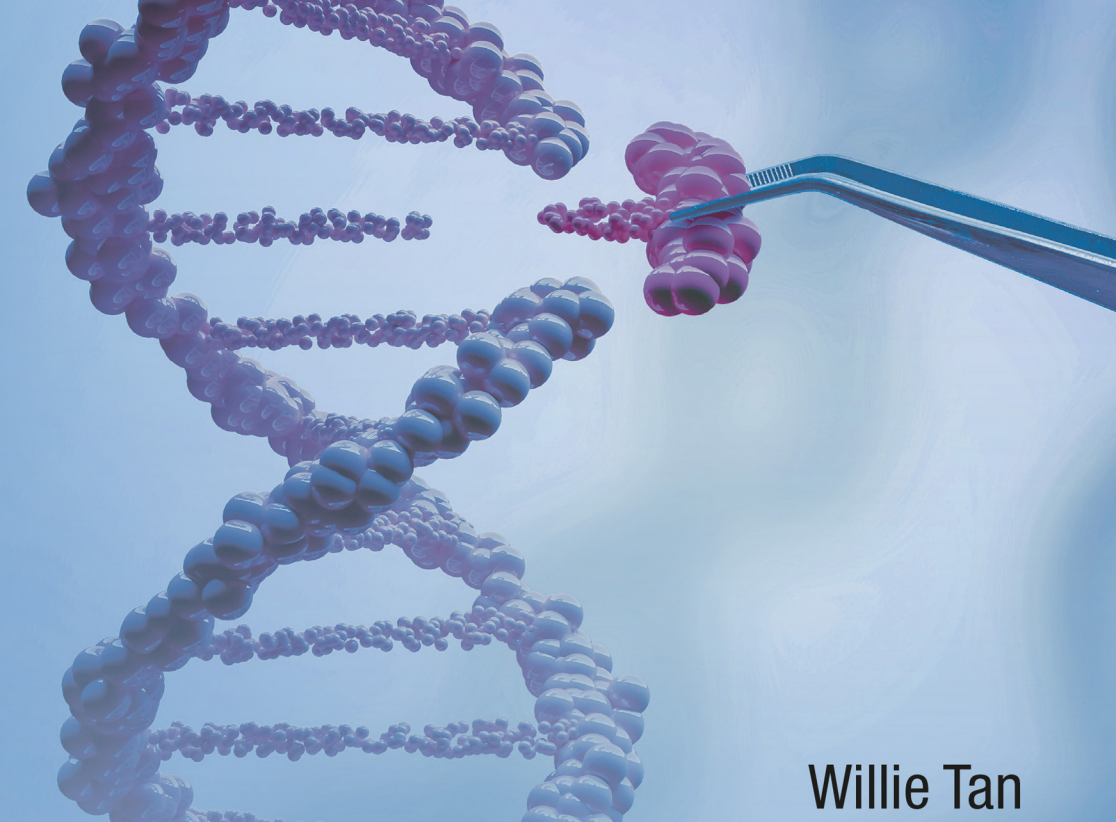

Willie Tan

RESEARCH METHODS

A Practical Guide for Students and Researchers

Second Edition

World Scientific

RESEARCH METHODS

A Practical Guide for
Students and Researchers

Second Edition

Other Titles on Research and Writing

Scientific Writing 3.0: A Reader and Writer's Guide
by Jean-Luc Lebrun and Justin Lebrun
ISBN: 978-981-122-883-4
ISBN: 978-981-122-953-4 (pbk)

Science Research Writing: for native and non-native speakers of English Second Edition
by Hilary Glasman-Deal
ISBN: 978-1-78634-783-1
ISBN: 978-1-78634-784-8 (pbk)

The 21st Century Guide to Writing Articles in the Biomedical Sciences
by Shiri Diskin
ISBN: 978-981-3231-86-3
ISBN: 978-981-3233-75-1 (pbk)

The Grant Writing and Crowdfunding Guide for Young Investigators in Science
by Jean-Luc Lebrun and Justin Lebrun
ISBN: 978-981-3223-23-3
ISBN: 978-981-3223-24-0 (pbk)

The Grant Writer's Handbook: How to Write a Research Proposal and Succeed
by Gerard M Crawley and Eoin O'Sullivan
ISBN: 978-1-78326-759-0
ISBN: 978-1-78326-414-8 (pbk)

Planning Your Research and How to Write It
edited by Aziz Nather
ISBN: 978-981-4651-03-5
ISBN: 978-981-4651-04-2 (pbk)

Directions for Mathematics Research Experience for Undergraduates
edited by Mark A Peterson and Yanir A Rubinstein
ISBN: 978-981-4630-31-3

RESEARCH METHODS

A Practical Guide for Students and Researchers

Second Edition

Willie Tan

National University of Singapore, Singapore

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI • TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Names: Tan, Willie, author.
Title: Research methods : a practical guide for students and researchers / Willie Tan, National University of Singapore, Singapore.
Description: Second edition. | New Jersey : World Scientific, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2022017936 | ISBN 9789811256936 (hardcover) | ISBN 9789811257957 (paperback) | ISBN 9789811256950 (ebook) | ISBN 9789811256943 (ebook other)
Subjects: LCSH: Research--Methodology.
Classification: LCC Q180.55.M4 T3596 2022 | DDC 001.4/2--dc23/eng20220722
LC record available at https://lccn.loc.gov/2022017936

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

For any available supplementary material, please visit
https://www.worldscientific.com/worldscibooks/10.1142/12863#t=suppl

Desk Editor: Amanda Yun

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore

Preface

The aim of this edition is the same as that of the first edition. It is to provide a concise and practical guide to researchers and students tackling increasingly complex problems, particularly in engineering and the built environment. Solving these problems requires both qualitative and quantitative approaches and, with expanding disciplinary and inter-disciplinary knowledge of what used to be seemingly unrelated fields, it becomes necessary to revise the first edition to broaden the toolkit. In essence, research methods have gone digital even though the basic principles of research remain.

The structure of this edition follows that of the first edition. It covers the entire research process of problem formulation, literature review and development of framework or hypothesis, research design, methods of data collection, data analysis, concluding the study, and publishing the findings. A feature of this book is that the data analyses are linked to the research designs. Hence, the chapters on data analyses are based on specific research designs. This link provides a better understanding of the research process as an integrated chain of reasoning.

The book uses many examples from the engineering and built environment sectors. There are also examples from other sectors for three main reasons. First, they may be classics in a particular field that demonstrate clearly many valuable lessons. Second, they are intended to show applications outside one's discipline and help students make the connections to their own disciplines and appreciate different ways of thinking about problems. Finally, these examples may explain certain concepts and research designs better, such as the frequent use of country analysis in political science to illustrate the comparative method.

Many chapters have been extensively rewritten to provide greater clarity, better organization, update existing content, and add new topics. In particular,

- Chapter 3 has been rewritten to provide greater clarity on theories, hypotheses, models, and frameworks;
- Chapter 6 provides a more comprehensive treatment on the various types of comparative designs;
- Chapter 9 includes digital forms of data collection;
- Chapter 11 has been expanded to include thematic analysis and interpretive phenomenological analysis;
- Chapter 12 includes basics of statistical tests, properties of estimators, circular data, and expanded treatment on index numbers;
- Chapter 15 includes Poisson regression and expanded treatment on nonlinear least squares;
- Chapter 16 covers new material on distributed lag models, panel regression, VAR model, the causality test, spectral analysis, spatial models, and rank-deficient models; and
- Chapters 17 and 18, on machine learning and meta-analysis and its variants, respectively, are new; the former includes the random forest classifier, naive Bayes classifier, support vector machine, cluster analysis, principal components analysis, factor analysis, and associative methods.

Machine learning ties in nicely with the newer methods of collecting big data using crowdsourcing, websites, the Internet of Things, and sensors. The latter are widely used in digital twin technology.

Although this book uses an intuitive approach and avoids most mathematical proofs, some basic background knowledge of calculus, probability, linear algebra, and statistics are still necessary. Readers who are unfamiliar with these topics may check them out from online sources and videos. That said, important basic topics such as the logic of statistical tests, eigenstructures, properties of estimators, and nonlinear optimization have been included to provide better support to readers.

I thank Alexander Lin (NUS), Calvin Yeung (Changi Airport Group), Cheng Zhuoyuan (Singapore University of Social Sciences), Dilini Thoradeniya (University of Moratuwa), Du Hongjian (NUS), Eric Tan

(SIFT Analytics), Ernawati Mustafa Kamal (Universiti Sains Malaysia), Ke Yongjian (University of Technology Sydney), Imriyas Kamardeen (Deakin University), Jonathan Lian (NUS), Liu Junying (Tianjin University), Neo Kim Han (Housing and Development Board), Ravi Shankar (IIT Madras), Richard Wu (University of NSW), Shankar Sankaran (University of Technology Sydney), Stephen Tay (NUS), Winston Hauw (NUS), and Zhang Yajian (Northeastern University).

Lastly, I thank Ms Amanda Yun, Senior Editor, World Scientific Publishing, for her kind assistance and patience while I seem to work endlessly on the draft because of my busy schedule.

Willie Tan

This page intentionally left blank

Contents

This page intentionally left blank

Chapter 1

Introduction to Research

What is Science?

Research is the bedrock of science, which is why universities require that students in the physical and social sciences learn how to conduct scientific research effectively. But what is science?

In terms of aims, all sciences seek to explore, describe, interpret, explain, predict, control, or evaluate phenomena or behaviors. But so do non-sciences such as religion, witchcraft, *feng shui*, and astrology (Carter, 2000; DK, 2000; Edington, 2020). For critical theorists, science also aims to empower people to overcome certain mindsets and social structures (Marcuse, 1964; Habermas, 1971; Parker, 2012).

Science is based on empirical, or observable facts. It does not deal with the spiritual, such as religion or witchcraft. Spirits are metaphysical, that is, beyond the physical, and hence beyond science.

Scientists often use facts to build theory (*induction*) (Fig. 1.1) and discover empirical regularities, which are called *scientific laws*. An example of an empirical regularity is the law of gravity, which states that objects near the surface of the earth must fall to the ground. It is a law of nature that all objects must obey. If most, but not all, objects obey, it is called a tendency or a rule. For example, the rank-size rule (Zipf, 1949) states that the population of the nth city is P/n, where P is the population of the largest city in the country. Hence, the population of the third-largest city is $P/3$, and that of the fourth-largest city is $P/4$. However, not all countries have this distribution (Rosen and Resnick, 1980), which makes it a rule rather than a scientific law.

Induction is also used in qualitative research to develop more general themes from the data. For example, if we look at what students write

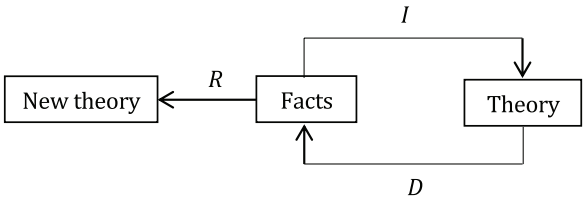

Fig. 1.1 Induction (*I*), deduction (*D*), and retroduction (*R*).

on a university's social media page, several themes may emerge, such as concerns with hostels, security, finances, public transport, parking, food, and academic matters. More recently, machine learning uses induction to predict events rather than build theory. For example, in recommender systems, the online store (e.g. Amazon) uses machine learning to recommend items. The algorithm uses correlation from data collected from millions of buyers to predict possible purchases by different categories of buyers based on certain characteristics and browsing history (see Chapter 17).

Induction is not the only route to generate scientific knowledge. Instead of starting from the facts, scientists use reason to build theories or hypotheses and test them against facts. This process is called *deduction*. For Popper (2002), science progresses through this endless interaction between conjectures and refutations, or trial and error. Symbolically, we may write it as $T \rightarrow F \rightarrow T'$. The theory ($T$) is tested against facts (F), which leads to a modified theory (T'). For Popper, astrology is not a science because its predictions are not falsifiable. He is also critical of induction because while no amount of observations can verify the statement "all swans are white," the appearance of a black swan will falsify it. As discussed later, Popper's falsification approach is not without problems.

The third route to knowledge is *retroduction* or *abduction* (Peirce, 1998). It begins with puzzling observations that cannot be explained by existing theories. It then seeks to discover new hypotheses or the best possible explanation. For Peirce, the "best" explanation is the simplest and most plausible.

Other than theory and evidence, another distinguishing feature of science is its rigor or *methodologies* (Poincare, 2007). A methodo*logy* is the logic of a method. It is the thoroughness that distinguishes science from non-scientific ways of acquiring knowledge through traditional habits of thought or practices, intuition, common sense, and experience. Scientific rigor ensures that the process is less prone to errors. Because of rigor, scientific fields are also called "disciplines."

In summary, scientists aim to understand, explain, or predict phenomena. Knowledge is generated from a rigorous testing process oscillating between theory and evidence, and puzzling observations lead to the refinement of theories.

Theories

Science deals with theories. A theory links *concepts* or constructs (ideas) such as poverty, demand, heat, cell, and love. Concepts, as ideas, exist only in our heads. They are the constituents of thought. Concepts may be abstract ideas that do not exist in time and space (e.g. numbers), or concrete ideas such as trees and rocks. Ideas may be highly abstract, such as energy and political power.

Scientists use *abstraction* to extract the critical features or properties of an entity to create theories and models. Using induction, they also abstract from particular instances to draw general conclusions. As a simple example, it is not possible to explain price fluctuations by just visiting a market and watching the haggling process. We need to abstract from our observations of the market and develop the concepts of demand and supply to explain price fluctuations.

To test our models or theories, we need to measure or *operationalize* the concepts as observable *variables*. For example, to operationalize "suicide" we may look at the number of deaths that are classified as suicides (s) in official statistics. We also need the population (p) to derive the suicide rate (s/p). The data for s and p may contain errors, making the correspondence between the concept and variable imperfect. For instance, for Catholics, death by suicide is a grave matter because God gives life. Thus, coroners in predominantly Catholic countries may be reluctant to classify

death cases as suicide. Moreover, death by jumping in front of a moving train is often classified as suicide, but not when one dies from being knocked by a speeding car, which is often classified as jaywalking or reckless driving.

As a second example, consider the claim that "John is an influential politician who gets things done." There are two concepts that require measurement if we are to test the claim, namely, "influential politician" and "getting things done." "Political influence" is not a simple concept. John may be influential inside and outside parliament. Due to time and resource constraints, the researcher may decide to measure part of the concept, such as influence in parliament, and leave out measuring John's social and political network. To measure "influence in parliament," we may look at the number of bills he has introduced and getting them passed. Alternatively, we may use analysts' ratings of political influence. A third possibility is to construct a weighted measure using both indicators. Finally, to measure "getting things done," we may consider the projects John has initiated and implemented in the community.

The scientific disciplines use different terminologies for "theory" (Table 1.1). For researchers looking for causes, the terms "theory," "hypothesis," "perspective," and "model" are often used interchangeably. A *paradigm* (Kuhn, 1962) is a school of thought comprising a cluster of related theories that share similar assumptions, concepts, and worldviews. These schools often clash, such as among Keynesians, Monetarists, Austrians, and Marxists in economics.

Interpretive researchers tend to use the term "framework" to refer to a structure that guides the study, rather than as a grand theory or less developed theory. In this book, we will use theory, hypotheses, perspectives, and models interchangeably and reserve "framework" as a structural guide for interpretive studies. We will not deal with mathematical conjectures and postulates. Lastly, the theories or research hypotheses in the table are not the same as statistical hypotheses. The latter are narrower claims about parameter values. On their own, statistical hypotheses do not explain or interpret phenomena. For example, if we postulate that $Y = \alpha + \beta X + \varepsilon$, where ε is a noise term, then a statistical finding that $\beta \neq 0$ implies that X and Y are related. The statistical test does not tell us why they are related. If $\beta = 0$, then X and Y are unrelated.

Table 1.1 Different terminologies for "theory."

	Meaning	Example
Theory	A general term that links causes and effects	Theory of natural selection, rational choice theory
Hypothesis	Testable theory	Tiebout hypothesis, economic base hypothesis
Model	Formal theory, often with equations	Simple Keynesian model: $Y = C + I + G$ $C = a + bY$; $I = I_0$; and $G = G_0$
Approach or perspective	A way of looking at something; sometimes used in place of theory	Functionalist perspective, liberal perspective
Paradigm	School of thought	Newtonian physics, neoclassical economics
Framework	Grand theory	Marxist theory, modernization theory
	Less developed theory	Leadership theories, theories of organization culture
	Interpretive theory	Labeling, interpretation of culture
	Structure to guide understanding	Liebow (1967), McKinsey 7S model
Equation	Theory expressed as an equation	$E = mc^2$
Conjecture	Mathematical proposition with supporting evidence but not conclusive proof	Frobenius conjecture, Borel conjecture
Postulate	A theory accepted as true as basis for further reasoning	Euclid's postulates, segment addition postulate
Law	Theory expressed as a regularity or scientific law	Ohm's Law, law of gravity
Statistical hypothesis	A claim about the value of a population parameter	H_0: $\mu = 0$, H_1: $\mu > 0$

Methodology

A research methodology is a series of logical steps from formulating a research problem to arriving at a conclusion. It provides the link between theory and evidence, including the use of agreed standards to maintain rigor. Roughly speaking,

Methodology = Philosophy + Research designs + Methods.

Some researchers use the terms designs and methods interchangeably (e.g. Fetterman, 2010; Richardson et al., 2011). We will follow the standard separation of research designs from methods of data collection.

These elements of methodology are briefly discussed below to provide an overview of the research process.

Philosophies of Science

Broadly, there are two main philosophies of science, namely,

- causal science; and
- interpretive science.

Within each philosophy, there are several variants (Ladyman, 2001). For example, in causal science, there are positivist, post-positivist, neo-positivist, realist, critical rationalist, critical realist, Marxist, and conventionalist approaches. In interpretive science, there are interpretivism, hermeneutics, constructivism, discourse analysis, grounded theory, critical theory, symbolic interactionism, ethnomethodology, gender, and phenomenological approaches. The main differences between causal and interpretive approaches are shown in Table 1.2.

Causal scientists assume there is an objective reality "out there" independent of our conceptions of it. For example, the housing market is "out there" for us to discover its workings, or causal *mechanisms* (Harre, 1970; Bhaskar, 1975). We use the mechanisms of supply and demand to explain the market price. The predictions are then tested against the evidence.

Table 1.2 Key differences between causal and interpretive science.

Feature	Causal science	Interpretive science
Reality	Objective; "Out there"	Subjective, in the actor's head
Purpose	Discover causal mechanisms	Discover understandings or meanings
Strategy	Test hypothesis	Use framework
Design	Experiment, regression, comparison, case study	Case study, survey
Data	Numeric, quantitative	Linguistic, symbolic, qualitative
Data analysis	Good statistics	Good narratives

For interpretivists, the world is not an objective external reality "out there" outside our heads to be discovered. Instead, the individual subjectively experiences and understands reality, resulting in multiple realities or different views of the same event (Collingwood, 1946; Taylor, 1971). These subjective understandings form the basis of human action. If we think people are generally honest, we act accordingly. Similarly, we sell our stocks if we think the equity market is going to crash. Every stock transaction has a buyer who is more optimistic than the seller.

Interpretivists use a conceptual *framework* as a structure or set of ideas (see Table 1.1) to *discover* different views, perspectives, or *understandings* on certain issues. An understanding refers to comprehension, and an object (e.g. a ring), activity (e.g. working), or event (e.g. death) may have different meanings for individuals or communities. For example, Willis (1977) used the following framework to describe the culture of a group of working-class schoolboys:

- opposition to authority;
- informal group;
- dossing and wagging;
- having a laugh;
- boredom and excitement;
- sexism; and
- racism.

Interpretivists use qualitative data in their narratives (stories). They are aware that parties may tell stories from *their* points of view by deliberately *constructing* the narrative to *persuade* the listener. In general, both quantitative and qualitative researchers use rhetoric or literary devices such as metaphors, analogies, and appeals to authority, data, or statistical tests to persuade readers (McCloskey, 1986). Finally, a narrative may be *contested* or resisted by other parties. After all, why should anyone believe *your* story?

Research Designs

A research design is a strategy to carry out the research to ensure that we test the hypothesis correctly or develop appropriate interpretations.

Hence, the aim of research designs is to rule out alternative explanations. Inappropriate or weak designs are likely to invite criticisms or alternative views.

The research design may consist of a

- case study;
- survey;
- comparison;
- experiment;
- regression; or
- mixed designs.

As shown in Table 1.2, this roadmap is underpinned by a particular philosophy of science. For example, a researcher who adopts a causal view of science will tend to use an experiment, survey, comparative study, or case study to determine causes and effects. In contrast, an interpretive researcher may use a case study design to probe more deeply to gather different views on an issue. A mixed design is a combination of quantitative and qualitative approaches (Tashakkori and Teddlie, 2003), such as the use of survey and case study designs. This may be done concurrently or sequentially. For example, a researcher may first conduct a broad survey of how hospitals are coping with the COVID-19 pandemic and then follow up with a case study to probe more deeply into the issues. It is also possible to begin with a qualitative and exploratory case study and use the findings to develop a questionnaire for the subsequent survey.

In an experiment, the researcher tries to vary one or more variables while holding other extraneous factors constant. For example, if the effect of a drug on cholesterol level depends on a subject's age, lifestyle, and gender, the researcher may fix the values of age and gender by selecting a sample comprising only males aged 40 to 50. These extraneous variables are not of interest to the experimenter. However, their effects on the outcomes of the experiment must be taken into account.

The social scientist is not so fortunate. She cannot fix variables such as interest rates or household income. Her "laboratory" is society, and many political, economic, and social factors are beyond her control. A possible solution is to use a *regression* design, which is a form of

statistical rather than experimental control. A further problem for social scientists is that humans, unlike physical objects, behave differently even under similar circumstances. Humans can exercise free will and their behaviors are shaped by values, beliefs, attitudes, incentives, constraints, and subjective interpretations of the world around us.

A comparative design seeks to uncover probable causes by examining a small number of different cases (Mill, 1884; Ragin, 2014; Mello, 2021). The number of cases has to be kept small because it is difficult to compare a large number of cases to draw certain conclusions. The cases may have similar or different features, and similar or different outcomes. For example, we may examine

- similar learning strategies of academically strong students;
- different learning strategies of academically strong students; or
- learning strategies of strong and weak students.

A comparative study that identifies the presence of common factors does not explain why these factors cause the outcome. At best, it identifies probable causes. Further, if there are too many factors and too few cases, it is not possible to identify common factors.

A case study probes in-depth into a unit (case) to trace a process or discover something new. It may be causal, such as a historical case study to trace the development of a unit or system over time. A case study may be interpretive. The researcher probes in-depth to understand the context and the actor's point of view. Here, context refers to the local situation or local knowledge or, more broadly, other factors. For example, Western management theories may not be applicable to firms in developing countries because of differing contexts. The politics, markets, and social norms are different.

Methods

Methods are ways to collect, process, and analyze data. They are the nuts and bolts of research that occupy much of this book. This process is often split into four separate steps, namely, planning, collection, processing, and analysis. In the planning phase, the researcher decides on the appropriate

methods of data collection. Data may then be collected using observations, interviews, questionnaires, simulation, or past records. They are then processed and organized in a form suitable for subsequent data analysis using qualitative or quantitative techniques.

In general, qualitative researchers tend to downplay the importance of numbers. They understand the need for numbers, and use numbers themselves. However, they are wary of the possible misuse of numbers by quantitative researchers. For qualitative researchers, what lies behind the numbers is a story or different stories as perceived or told, for one reason or another, by different actors. In other words, numbers do not provide the richness of data desired by qualitative researchers (Billups, 2020).

Research Process

The research process consists of the following steps:

- identify the research question (or problem);
- review the literature to develop a hypothesis, theory, model, or framework;
- determine an appropriate research design to test the hypothesis or apply the framework;
- devise appropriate methods to collect data;
- collect and process the data so that they are suitable for subsequent testing or interpretation;
- analyze the data; and
- conclude and publish.

This process underpins the structure of this book. For this reason, we will discuss these steps in subsequent chapters.

The research process is linked, that is, it is not possible to skip a step. For example, without a framework or hypothesis, it is not possible to design the study and decide on ways to collect and analyze the data. Even if a study uses machine learning on big data, it is necessary to start with a framework. We need to know what types of data to collect before applying machine learning algorithms.

Testing Theories

Earlier, I alluded to difficulties in applying Popper's falsification criterion when testing theories. Testing a causal theory involves looking at

- its assumptions;
- the causal mechanism; and
- the data.

For assumptions, there are three possibilities. Realistic assumptions approximate reality, such as the absence of friction in flight models or that people act rationally in maximizing utility by seeking the lowest price for any given product quality. Perhaps surprising, assumptions may be deliberately unrealistic (Friedman, 1953), such as a featureless plain in urban economics models or a closed economy to exclude external trade. For Friedman, the important issue is whether the theory predicts well, and not the realism of its assumptions. Finally, a theory may be constructed from axioms or self-evident truths, which is common in mathematical theories.

If a theory fails an empirical test, it is seldom rejected outright. It may be saved by

- stating that its assumptions are, after all, unrealistic;
- declaring that the purpose of the model is to provide insight and not an agreement with reality;
- reducing its applicable scope;
- adjusting or introducing additional (auxiliary) assumptions (Lakatos, 1978);
- blaming it on the measuring instrument;
- blaming it on inadequate data or sample size;
- questioning the statistical technique;
- stating that it omits certain variables as a limitation; and
- questioning the acceptance criterion of the test if the evidence falls in the gray area, and not black or white.

In summary, it is not easy to test a theory, which is why debates about the efficacies of vaccines for COVID-19 dominate the social media or

why disciplines in the social sciences such as economics, sociology, business, and political science tend to fragment into different schools of thought.

As an example, consider the Conservative claim that the public school system is "failing," and education should be privatized (Ravitch, 2011). How do we know the system is "failing"? Conservatives may argue that schools are inadequately funded, there is a shortage of good and qualified teachers, students did poorly in international tests and have disciplinary problems, and the curricula contain too little science and technology. The issue now shifts from testing whether governments have done too little or too much in education to testing each of these claims. If each claim can be tested, it is possible that only *some* claims are supported by the evidence. There is also much to dispute over imperfect measures of "teacher quality," "discipline problems," and so on.

Many-model Thinking Approach

Some researchers use many models to make sense of complex phenomena (Page, 2021). The core idea is that the use of different models allows us to look at a complex issue from different perspectives.

In the traditional approach, researchers use only one theory to explain an event. The main reason is that the different theories may be inconsistent. For example, theories of the business cycle are underpinned by different philosophies of science, assumptions, and causal mechanisms (Knoop, 2015). In other words, scientists belong to different paradigms that are incommensurable (Kuhn, 1962). For the free-market economist, the government is often the problem and main cause of economic crises. In contrast, for liberal and radical interventionists, the market is the problem, not the solution. It is not easy to reconcile such major differences (Heidlebaugh, 2001).

References

Bhaskar, R. (1975) *A realist theory of science*. London: Reed Books.
Billups, F. (2020) *Qualitative data collection tools*. Thousand Oaks: Sage.

Carter, K. (2000) *Move your stuff, change your life*. New York: Simon and Schuster.
Collingwood, R. (1946) *The idea of history*. London: Oxford University Press.
DK (2020) *A history of magic, witchcraft, and the occult*. London: DK.
Edington, L. (2020) *The complete guide to astrology*. London: Blackwell.
Fetterman, D. (2010) *Ethnography*. Thousand Oaks: Sage.
Friedman, M. (1953) *Essays in positive economics*. Chicago: Chicago University Press.
Habermas, J. (1971) *Knowledge and human interests*. New York: Beacon Press.
Harre, R. (1970) *The principles of scientific thinking*. London: MacMillan.
Heidlebaugh, N. (2001) *Judgment, rhetoric, and the problem of incommensurability*. Columbia: University of South Carolina Press.
Knoop, T. (2015) *Business cycle economics*. New York: Prager.
Kuhn, T. (1962) *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Ladyman, J. (2001) *Understanding philosophy of science*. London: Routledge.
Lakatos, I. (1978) *The methodology of scientific research programs*. London: Cambridge University Press.
Liebow, E. (1967) *Tally's corner*. Boston: Little, Brown, and Co.
Marcuse, H. (1964) *One-dimensional man*. New York: Beacon Press.
McCloskey, D. (1986) *The rhetoric of economics*. Wisconsin: University of Wisconsin Press.
Mello, P. (2021) *Qualitative comparative analysis*. Washington DC: Georgetown University Press.
Mill, J. (1884) *A system of logic*. London: Longman.
Page, S. (2021) *The model thinker*. New York: Basic Books.
Parker, R. (2012) *Critical theory: A reader for literary and cultural studies*. London: Oxford University Press.
Peirce, C. (1998) *The essential Peirce*. Indianapolis: Indiana University Press.
Poincare, H. (2007) *Science and method*. New York: Cosimo.
Popper, K. (2002) *Conjectures and refutations*. London: Routledge.
Ragin, C. (2014) *The comparative method*. Los Angeles: UCLA Press.
Ravitch, D. (2011) *The death and life of the great American school system*. New York: Basic Books.
Richardson, P., Goodwin, A., and Vine, E. (2011) *Research methods and design in psychology*. Thousand Oaks: Sage.
Rosen, K. and Resnick, M. (1980) The size distribution of cities: An explanation of Pareto Law and primacy. *Journal of Urban Economics*, **8**(2), 165–186.

Tashakkori, A. and Teddlie, C. (Eds.) (2003) *Handbook of mixed methods in social and behavioral research*. Thousand Oaks: Sage.

Taylor, C. (1971) Interpretation and the sciences of man. *Review of Metaphysics*, **25**, 3–51.

Willis, P. (1977) *Learning to labor*. New York: Columbia University Press.

Zipf, G. (1949) *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.

Chapter 2

The Research Problem

Introduction

The research problem is an issue or concern that warrants attention from the researcher, scientific community, government, or society. It should be well-articulated and stated as concisely as possible.

The formulation of the research problem is the most important step in the research process and probably the most difficult as well. This is because many novice researchers are unclear as to what constitutes a well-articulated and crystal-clear research problem. If the problem is unclear, the research quickly loses its focus.

Normally, there is only one research problem, although it may be broken down into a small set of sub-problems (Andrews, 2003). In many cases, it suffices to state the research problem and not the sub-problems. In some instances, there may be a cluster of research questions. For example, Blaikie and Priest (2017, p.44) provided the following research questions on the problem of child sexual abuse within the Catholic Church:

- What is the nature and extent of sexual abuse of children and young people in the Catholic Church?
- Why have the perpetrators engaged in these activities?
- Why have victims generally remained silent for so long?
- Why have church leaders and administrators not dealt with perpetrators in appropriate legal and responsible humanitarian ways?
- Why have church leaders and administrators not dealt with victims in appropriate morally and religiously responsible ways?
- What have been the consequences of this abuse for victims?

Clearly, it is difficult to answer such a large set of research questions in a single study. Hence, it is prudent to use a single or small set of questions.

As a second example, consider a study to establish the relation between Y and X_1,..., X_k, where Y is the dependent variable and the Xs are k independent variables. For example, Y may be the change in the share price of a listed Real Estate Investment Trust (REIT) and the Xs include the quality of management, portfolio of property assets, dividend policy, borrowing cost, debt to equity ratio, quality of sponsor, operational costs, and market conditions. The research problem is to determine how the Xs affect Y. However, there may be supplementary questions such as why X_i varies across firms, and whether the Xs interact.

Scanning for Topics

How does one find a research problem? Apart from a strong personal interest in the topic to sustain effort and its value in terms of a new idea, understanding, process, technique, product, or design, consider the likelihood of obtaining the data. Sometimes, there may also be new patterns in data that require explanation. All said, be reflective and creative at this stage and walk off the beaten track.

To find the research gap, read a book or journal article that summarizes the state of the art or, more likely, use a web search engine such as Google Scholar. Use broad concepts (e.g. homelessness or housing finance) as keywords for the search.

When scanning for topics, jot down the source and possible research gaps in the form of possible research questions. At this preliminary stage, do not immerse in the details of journal articles.

Scope

The next step is to select a suitable research question among the few shortlisted ones and start to narrow the scope so that the problem is logically solvable and it can be done within time and cost constraints. This establishes the boundaries of the study and ensures that the scope is not too broad or narrow.

It is possible to narrow a topic by concept, such as from "housing" to "rental housing" or "public rental housing." The topic may also be bounded by time and space, such as a study of 17th-century traditional houses, or a study of traditional houses in a particular town. In the child abuse case discussed earlier, it is possible to scope the study by focusing on one set of participants, such as victims.

Justification

The next step is to develop the justification for the study. It tries to answer the question "Why are you doing this?"

The justification may be based on theoretical or practical grounds, such as to contribute to knowledge or policy, or to solve practical problems. It should not be of a personal nature, such as to educate yourself in a new field. If you are contributing to knowledge, it is helpful to briefly state the existing gap in the literature. Importantly, this provides the context of the study within the current body of knowledge. It should not be isolated from what researchers in the same area are pursuing.

It may be useful to cite evidence to catch the attention of the reader. For example, if it contributes to policy, such as by discouraging teenagers from taking up smoking, how many people are or will be affected by it? Similarly, in a study of workplace accidents, you may cite recent data, such as to highlight a worsening trend or a sudden spike in accidents that warrants investigation.

Objectives

The objectives of the study are what you want to achieve by doing the research. There are often a handful of objectives, such as examining something and then recommending what needs to be done. What is required to carry out the research such as to review the literature or collect data are not research objectives.

In stating the research objectives, use verbs such as describe, explore, understand, determine, explain, evaluate, predict, or recommend. For example, one may wish to determine the impact of a particular policy.

Getting Feedback

The final step in formulating a research problem or question is to put it to the test by asking seasoned researchers or potential supervisors in the area. It is important to take feedback positively. Sometimes a simple remark that "this problem is not researchable" or that "it has been well researched" will set you thinking critically instead of blindly searching for a research problem. More often, the experienced researcher will hint to the student to refine the research question. This may take many iterations, and one should take it positively as a learning process.

Examples

Let us examine three examples on how to write the research question, scope, justification, and research objectives.

Tally's corner

The first example is Liebow's (1967) work, titled *Tally's corner*. It is a classic example of an ethnographic study. The research question is

> How do poor urban Afro-American men view themselves and the world around them, and how do they adapt?

In this case, it is possible to split the research problems into three parts: (i) how they view themselves, (ii) how they view the world around them, and (iii) how they adapt to it.

Liebow delimited the *scope* of the research to poor Afro-American men who occupied a street corner in the Second Precinct of Washington DC. He was clearly not studying all types of poor men from different backgrounds. He restricted the study to a particular race. This raises the issue of whether the study is generalizable, given that it is an American setting in the early 1960s, only poor Afro-American men were involved, and he chose only a precinct of Washington DC. Further, Liebow did not consciously select the men for representativeness because he did not intend to develop generalizations. The research strategy was to probe

intensively into the lives of these men. This argument, as we have learned in Chapter 1, is a classic defense of interpretive small-*N* case studies where *N* is the sample size.

Liebow provided three justifications for the study. First, the problems faced by and generated by low-income urban families are of major concerns to policymakers in America (Orshansky, 1965). These concerns are nothing new, and have been existing for many years. The charities have been trying hard to help resolve them, but they are unable to deal with the scale and complexity of the problems. Second, much of what we know about poor Afro-American families is biased towards women and children, with a corresponding neglect of adult men. This is because the men are harder to reach than women and children (Moynihan, 1965). Third, much of what we know about the problems were gathered through surveys using interviews and questionnaires, which tend to provide only superficial understandings of the problem (Rohrer and Edmonson, 1960). For Liebow, what was needed was an in-depth probe using participant observation of how these men viewed themselves and the world around them in their own terms, and how they adapted to it.

Finally, the objectives of the study are to

- explore in-depth on how the street corner men view themselves and the world around them, and how they adapt to it; and
- recommend ways to help these men to escape poverty.

Occupational injury in America: An analysis of risk factors using data from the general social survey

The second example is Smith and Dejoy's (2012) study on how psychological and organizational factors affect workers' experience of safety and health issues in their work environment. The research question is

> How do psychological and organizational factors affect workers' experience of safety and health in their work environment?

The *scope* is restricted to how psychological and organizational factors affect workers' experience of work injury. They deliberately did not

wish to consider other factors, either because we know much about their effects or because of resource constraints. It is also possible to restrict the study to workers in a particular industry.

On justification, Smith and Dejoy argued that occupational safety and health remain a significant problem because of the large number of injuries and the consequent economic costs (Weil, 2001; Schulte, 2005). Moreover, most studies of the distribution of occupation injury use various job and employment factors (Dembe et al., 2004; Simpson et al., 2005). Much less is known about other risk factors, particularly those pertaining to psychological and organizational factors.

For Smith and Dejoy, the objectives of their study are to:

- examine the impact of psychological and organizational factors on workers' experience of work injury; and
- recommend ways to reduce the rate of injuries.

Assessing creative writing

Our third example is Mozaffari's (2013) paper on creative writing. The research question is

> How do teachers assess creative writing?

For scope, she restricted it to creative writing in literature classes in an Iranian university. On justification, some researchers think creative writing is subjective and cannot be assessed while others think we should develop more objective criteria to assess it. Current techniques of assessing creativity include

- divergent thinking tests (Richards and Schmidt, 2002; Silvia, 2008), such as many uses of a physical object (e.g. a safety pin), which are inappropriate for creative writing;
- subjective expert panel assessments (Baer and Mckool, 2009); and
- vague attributes, such as "originality" or the ability to "move the reader" without specifying what these terms mean (Weigle, 2002).

Finally, the objective of her study was to develop and test a rubric (scoring guide) to assess creative writing. Note, in particular, the need to cite the literature to provide the context of the study.

Organization of Study

In addition to the research problem, scope, justification, and research objectives, the final section of the first chapter of a research report contains a short description of how the study is organized, such as,

> Chapter 2 provides the literature review and develops the hypothesis. The methodology is given in Chapter 3, where it consists of a regression model using a sample of 200 houses. Data on house prices and characteristics were collected using valuation reports. Chapter 4 provides the regression results, and Chapter 5 concludes the study.

This section should be brief, as the details will be discussed in each chapter. It merely provides a road map of what is in the research report. Write this section near the end of the study when you are finalizing the report.

References

Andrews, R. (2003) *Research questions*. London: Continuum.

Baer, J. and Mckool, S. (2009) Assessing creativity using consensual assessment techniques. In S. Schreiner (Ed.) *Handbook of research on assessment technologies, methods, and applications in higher education*. (pp. 65–77). Hershey, PA: Information Science Reference.

Blaikie, N. and Priest, J. (2017) *Social research: Paradigms in action*. Cambridge: Polity Press.

Dembe, A., Erickson, J., and Delbos, R. (2004) Predictors of work-related injuries and illnesses: National survey findings. *Journal of Occupational and Environmental Hygiene*, **1**(8), 542–550.

Liebow, E. (1967) *Tally's corner*. Boston: Little, Brown & Co.

Moynihan, D. (1965) *The Negro family: The case for national action*. Washington DC: US Department of Labor.

Mozaffari, H. (2013) An analytic rubric for assessing creativity in creative writing. *Theory and Practice in Language Studies*, **3**(12), 2214–2219.

Orshansky, M. (1965) Counting the poor: Another look at the poverty profile. *Social Security Bulletin*, January, 3–29.

Richards, J. and Schmidt, R. (2003) *Dictionary of language teaching and applied linguistics*. New York: Pearson.

Rohrer, J. and Edmonson, M. (1960) *The eighth generation: Cultures and personalities of New Orleans Negroes*. New York: Harper and Row.

Schulte, P. (2005) Characterizing the burden of occupational injury and disease. *Journal of Occupational and Environmental Medicine*, **47**(6), 607–622.

Silvia, P. (2008) Discerning and creativity: How ell can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts*, **2**, 139–146.

Simpson, S., Wadsworth, E., Moss, S., and Smith, A. (2005) Minor injuries, cognitive failures and accidents at work: Incidence and associated features. *Occupational Medicine*, **55**(2), 99–108.

Smith, T. and Dejoy, D. (2012) Occupational injury in America: An analysis of risk factors using data from the General Social Survey (GSS). *Journal of Safety Research*, **43**(1), 67–74.

Weigle, S. (2002) *Assessing writing*. New York: Cambridge University Press.

Weil, D. (2001) Valuing the economic consequences of work injury and illness: A comparison of methods and findings. *American Journal of Industrial Medicine*, **40**(4), 418–437.

CHAPTER 3

Theories, Hypotheses, Models, and Frameworks

Literature Review

Recall from Chapter 1 that the next stage of the research process is to review the literature from a clear research question. The "literature" refers mostly to journal articles and books, and the purposes of the review are to

- develop the theory, hypothesis, or model (in causal studies) or framework (in interpretive studies);
- discover productive ways to improve the methodology and data analysis; and
- be familiar with previous findings.

There are two stages in a literature review. The first stage is the *preliminary review* to *scan* the literature to identify the research problem. This was discussed in Chapter 2. The second stage involves a *thorough review* and consists of the following steps:

- identify good journals and books;
- develop keywords from your research question;
- decide if a journal article (or book) is relevant;
- understand the methodology and data analysis;
- note the findings;
- document the review; and
- structure the review.

These steps are discussed below.

Good journal articles and books

Leading researchers publish their works in good journals with high impact factors (IFs) or citation counts. IFs should be compared among peer journals, and not across disciplines. As before, use a search engine such as Google Scholar to find the literature. The journal articles are often extracted from online databases or other hosting sites such as those uploaded by the authors on their webpages. Similarly, the quality of a book depends on the quality of the researcher(s) and publisher. You should also read reviews of the book by reputable researchers.

Keywords

Develop about five to ten keywords from your research question using the following steps:

- cross out words that are not specific;
- develop synonyms;
- broaden or narrow concepts; and
- go beyond your discipline or industry.

For instance, if we are studying organization learning in the local construction industry, the possible keywords are:

> organization learning (different industries)
> learning by doing; learning curve (types of learning)
> construction: building (synonym)
> productivity, efficiency (broader concepts)

Relevance

Start by reading the abstract to see if the journal article is relevant. It is important to cast a wide net. Many disciplines look at the same problem in different ways. Next, look for clues on the quality of the article, such as the citation count, quality of the references, and affiliation of the author(s).

Depending on whether you are doing causal or interpretive studies, the next step is to look for concepts, themes, and mechanisms. Typically,

a few key recent articles will suffice because the major articles will cite previous research and are unlikely to miss the key points. Do not be distracted by peripheral or tangential works.

Methodology and data analysis

The literature review is also an opportunity to discover productive ways of doing your research. For example, if you are testing a hypothesis, you should review the research design, sampling, how the concepts are measured, the sources of data, and how they are processed and analyzed.

Findings

Familiarization with previous findings will allow you to compare your results with what other researchers have done. Pay attention to factors that may account for the differences so that the comparison makes sense.

Documentation

Summarize the contents of each book or journal paper on a card or a piece of A4 size paper. The cards or sheets will then form the bibliography.

Structuring the review

The final step is to structure your literature review using one of the following options:

- themes;
- key concepts;
- rival explanations;
- chronological order; or
- methodologies.

The thematic approach is common in interpretive studies, such as Liebow's classification of the men he studied, as discussed below.

Regression studies tend to focus on the key concepts or the dependent and independent variables. Where a theory is less developed, it is also possible to focus on the key concepts because we are less sure of the causal links among these concepts. Sometimes, one may find literature reviews that focus on authors instead of key concepts. Avoid this approach because the focus of the literature review is the development of concepts towards a theory, not the chronological citation of authors.

Rival explanations are often found in the social sciences, such as the different schools of thought in political science, sociology, and economics. If there is only one theory, use a chronological approach to trace its historical development. Finally, researchers in disciplines such as computer science and engineering may also be interested in improving methodologies, algorithms, and optimization techniques rather than just in the development of theory. It is often necessary to review the different models or techniques to justify a particular choice. If one is trying to improve a particular technique, it is advisable to review its historical development.

Examples

Tally's corner

I will use Liebow's (1967) study as the first example of how to develop an interpretive framework for the study. Recall from Chapter 2 that he wanted to understand how poor urban Afro-American men viewed themselves and the world around them, and how they adapted to it.

Liebow used *roles* as the key concept. The *framework* considered the roles of these men in their daily lives as workers, fathers, husbands, lovers, and friends. These structural positions in society come with certain responsibilities and expectations. If the men are unable to fulfill them, they may feel frustrated or even ashamed and may walk out of the family altogether. They may also have a poor *view of themselves*, and think society holds a similar opinion. But people do want to improve their lives, especially that of their children. What, then, are the *structural constraints* that prevent poor people and their children from making it in society?

Within each theme, it is necessary to generate sub-themes or sub-questions. For instance, as workers, why don't the men work? For Liebow, the job failed the men because

- it did not pay enough to support a family; and
- construction work was hard, irregular, or at remote sites.

Further, the men failed the job because of

- inadequate skills or qualifications; and
- poor job experience.

As one of the men put it on his unpleasant work experience, "I graduated from high school, but I don't know anything. I am dumb. Most of the time I don't even say I graduated, 'cause then somebody asks me a question and I can't answer it, and they think I was lying about graduating." (p.55).

7S framework

Peters and Waterman (1982) popularized the 7S framework of business performance, which consists of

- shared values;
- management style;
- strategy;
- organization structure;
- business systems;
- skills; and
- staff.

Even though it is not an interpretive study, they called it a "framework" in the sense of a less developed theory. Within each element of the framework are further concepts. For example, under "style," there are many theories of leadership focusing on personalities, tasks, transactions, transformation, teamwork, and so on. The framework is applied to analyze the performance of an organization, including how the different elements "fit" together to improve performance.

Housing bubble

Governments worry about housing bubbles because of possible speculation, inflation, erosion of the work ethic, and threats to the stability of the financial and economic systems.

A housing bubble is a rapid rise in house price, usually relative to benchmarks such as

- household income;
- rents;
- user cost; or
- its fundamental value.

If disposable household income is stable and house prices rise rapidly over the same period, there is evidence of excessive price appreciation. Usually, both house prices and household incomes rise over the same period. Hence, we can compare the ratio of median house price (P) to median disposable household income (Y). The median rather than the mean is used to eliminate the effects of outliers. The median is the middle score, and it is unaffected by extreme values. The question is: beyond which level in P/Y do we consider house prices to be in bubble territory? Here, opinions differ. According to the *Demographia International Housing Affordability Survey* (Demographia, 2021), the results are as follows:

US	4.2
Canada	5.4
Singapore	4.7
Australia	7.7
New Zealand	10.0
Hong Kong	20.7

Demographia defines a ratio greater than 5.1 as "severely unaffordable." By this measure, Canada, Australia, New Zealand, and Hong Kong fall into this category. The ratio for Hong Kong is clearly an outlier. The housing affordability index (HAI) is an alternate measure that incorporates the impact of changing credit conditions (i.e. loan terms and

mortgage interest rates) rather than just the P/Y ratio. It computes the monthly repayment of a median-income family using the price of a typical home, 20% down payment, and the prevailing mortgage interest rate.

Similarly, by viewing housing purchase as any other investment in an asset, the house price/rent ratio (P/R) or the rent/house price ratio (R/P) should remain relatively constant. The latter is also called the rental yield, and it should not move too far out of line as a market rate of return on the investment. This is similar to investing in the stock market, where the dividend/price ratio of a company's share in a mature industry should not deviate significantly over time. If it does, then the share price may be too cheap or too expensive, and the market will correct the discrepancy. If it is cheap, there will be fewer sellers and more buyers, and this will bid up the price. If it is expensive, there will be many sellers and few buyers, resulting in falling prices. Thus, according to this so-called asset pricing model, any sharp deviation in R/P or P/R indicates the presence of a housing bubble. Similar to the P/Y ratio, opinions differ on the value of the ratio to classify a housing market as containing a bubble.

The annual user cost of housing (U) adjusts the rent for mortgage loan interest (r), taxes (t), operating cost (c), depreciation (d), and possible house price changes (g) for capital gain or loss, that is,

$$U = (r + t + c + d - g)P.$$

The components are expressed as a percentage of house price (P). For example, if $r = 2\%$, $t = 1\%$, $c = 2\%$, $d = 1\%$, and $g = 2\%$, then $U = 4\%$ of P. Changes in the user cost are indicators of affordability and possible presence of bubbles.

Lastly, the fundamental value (V) of an asset is the present value of annual future net rents (R_t):

$$V = \sum_{t=1}^{n} \frac{R_t}{(1+k)^t}.$$

The net rent is gross rent less outgoings. Here k is the discount rate, and n is the remaining lease of the house in years. Any large deviation between house prices and fundamental values indicates the presence of a bubble (Garber, 2000). However, V depends on the choice of k and

uncertain future values R_t. Some researchers deny the possibility of bubbles; if prices rise substantially, then the fundamentals have changed, not because of the presence of a bubble. The basis for this belief is the efficient market hypothesis (Malkiel, 2016), where assets cannot be overvalued or undervalued.

Supply-side explanations of bubbles focus on constraints that prevent housing supply from adjusting rapidly to rising demand during a housing boom (Maisel, 1963; Quigley, 1997). These constraints include difficulties in obtaining land, securing financing, collecting housing market information, and hiring construction workers in a boom market.

Demand-side explanations focus on demand shifts, namely,

- fundamental factors such as demography, household income, and interest rates (Poterba, 1991; Hubbard and Mayer, 2009);
- psychological factors in terms of buyers' irrational expectations of future house prices (Case and Shiller, 2003); or
- the impact of purchases by out-of-town or foreign speculative buyers (Chinco and Mayer, 2014).

Finally, how does one test for the existence of housing bubbles? For those who stress fundamental demand and supply factors, a logical choice is to regress house prices on these factors to determine their explanatory power (Kaufmann and Muhleisen, 2003). If the regression fit is low, then fundamental factors do not explain much of the house price change, that is, there is evidence of a bubble.

Another approach is to test for structural change in the housing market, that is, we use two regressions — one before the rapid rise in house prices, and one during the boom period (Jones and Leishman, 2003). If there is a bubble, the regression coefficients will differ significantly between the two periods.

A third strategy is to test whether house prices move in tandem with fundamental factors or there are sharp deviations. This asset pricing approach uses co-integration analysis (Meese and Wallace, 1994; Arshanapalli and Nelson, 2008).

Lastly, researchers who stress the importance of psychological factors may use a survey to ask recent house buyers of their opinions on future

house prices and whether a housing bubble exists (Case and Shiller, 2003; Quigley, 2003).

Obviously, you have to decide which test to use by reviewing the strengths and weaknesses of each approach. In addition, you will need to improve the model or apply it in a different context as part of your research contribution. The improved model is your research hypothesis.

In summary, the literature review for a causal study may be organized as follows:

Definition of housing bubble
Explanations of housing bubbles
- Supply factors
- Demand factors
  - Fundamental factors
  - Psychological factors
  - Foreign buying

Methods of testing for bubbles
- Regression on fundamental factors
- Tests for structural change
- Asset pricing approach
- Survey of buyers' price expectations

Hypothesis
- Which testing method to use, and why
- Dependent variable
- Independent variables

Agrarian revolution

The next example of a hypothesis concerns peasant rebellion in many parts of the developing world. The classic study is Paige's (1975) *Agrarian revolution*. It is a complex piece of work and a simplified form of Paige's hypothesis is shown in Table 3.1.

Paige's review of the literature to develop the hypothesis above is too lengthy to be discussed here because it involves a detailed discussion of each mode of production and the respective outcomes in terms of reform

Table 3.1 Paige's hypothesis (Paige, 1975, p.11).

		Cultivators	
		Land	Wages
Non-cultivators	Land	Commercial hacienda (Agrarian revolt)	Sharecropping/Migrator labor (Agrarian revolution)
	Capital	Smallholding (Commodity reform)	Plantation (Labor reform)

or violent struggles. Instead, I will focus on Paige's hypothesis itself as an example of how it is stated.

The conflict arises over the different income sources of cultivators who derive their incomes from land or wages, and non-cultivators who extract the rural surplus through ownership of land and capital. The bottom row leads to reform but not revolt or revolution. For the top row, commercial hacienda production leads to agrarian revolt but not revolution. Hence, Paige's interest lies in the use of sharecropping or migratory labor, which leads to revolution.

Why is it a toxic combination? Paige argued that the peasants' reliance on wages entails greater risk. Although the plantation mode of production also involves wage labor, the non-cultivators who rely on an expanding capital income are more willing to compromise towards labor reform.

In summary, Paige hypothesized that agrarian revolution is more likely to occur in sharecropping or migratory labor agricultural production because of the land-wage nexus rather than in other modes of rural production. For an empirical test of this hypothesis, see (Anderson and Seligson, 1994).

Research Proposal

A research proposal is required to secure funding or approval. It normally consists of the following (Schimel, 2011):

- Title;
- Research problem;

- Justification;
- Preliminary literature review;
- Research design;
- Data collection method;
- Data collection plan;
- Data processing plan;
- Data analysis plan;
- Schedule of costs;
- Schedule of deadlines (see Table 3.2);
- Risk assessment;
- Background of the researcher(s) involved; and
- Bibliography.

The schedule of costs and background of the researcher(s) are not required in student research proposals. These proposals are used to facilitate the allocation of thesis or dissertation supervisors.

Many research evaluators pay close attention to costs that may include the salaries of researchers and assistants, stationery, postage, photocopying, computing, equipment, materials, transport, books, and so on. It is important to cater for contingencies because research is a risky endeavor. For instance, the response rate for the survey may be too low, and this may necessitate a change in research strategy, such as by interviewing a new group of respondents.

Table 3.2 Example of research schedule.

	Months					
Tasks	2	4	6	8	10	12
Problem formulation						
Literature review & hypothesis						
Design & method						
Data collection & processing						
Data analysis						
Writing of research report						

Research Ethics and Risk Assessment

Research ethics and risk assessment are related, which is why they are considered together (Israel, 2014).

The researcher or a competent external party is normally required to conduct a risk assessment of possible harm or discomfort to humans and ways to manage these risks. The risk assessment cycle comprises the stages of recognizing hazards, assessing risks, managing the hazards or risks, and preparing for emergencies. A formal assessment requires the computation of the risk/benefit ratio. The risk assessment should also cover possible financial loss to the organization as well as damage to property and equipment such as through negligent use, accidents, or fire.

Individuals can then decide whether they wish to participate in the project, through informed consent in the sense that they are aware of the research purpose, its procedures, risks, and benefits. They have the right to withdraw from the study without giving any reason. However, not everyone is capable of giving informed consent. For example, for studies on children or the elderly, the researcher will need to seek the consent of their parents or caregivers. They will also need to take special care to protect these subjects from possible harm or discomfort.

The following information, if any, should be provided to potential participants:

- expected time commitment;
- boring or repetitive tasks;
- considerable physical exertion;
- physical harm, such as exposure to dangerous chemicals;
- privacy issues, such as income or access to medical or employment records;
- confidentiality, such as a worker's "feedback" on management or that names will not be disclosed in the research publication;
- recall of distressing events;
- possible allergies;
- measures of performance, to avoid embarrassment to poor performers; and

- studies that may link performance or deficiencies to nationalities, race, ethnicity, religion, income, or gender.

The researcher should provide full disclosure. A debrief is also conducted soon after the study so that respondents can assess whether the objectives of the study have been achieved.

Avoid the use of deception to gather field data (Tourish, 2019). These techniques include misrepresenting the purpose of research or the level of risks, concealed observation or secret recording of behavior. For example, should a researcher conceal his identity when studying the behaviors and activities of a street gang (Fexia et al., 2020)? Some researchers do not reveal their identities until after the study has been completed because they believe it will affect their membership into the gang. Sound judgment is required; for example, in mass observation studies of pedestrian behavior, concealed observation from an overhead bridge is not an issue. However, if you are observing how a project manager conducts a site meeting, you should reveal your identity and obtain informed consent.

Plagiarism is the stealing of the work of other researchers, even if obscure (Tota and Hove, 2019), and presenting it as one's own work. It is a serious offence, and you should give proper credit when it is due. This does not mean that one should cite everything, which is overkill. For example, it is well known that Marxism originates from Karl Marx, and it is pointless to cite the originator of the usual matrix inversion. Self-plagiarism occurs if you recycle your old ideas in a new paper without referencing them. Another form of self-plagiarism is to try and publish two similar papers in different journals.

Do not falsify or fabricate the data. A more common problem is the deletion of outliers to make the results look better or "clean up the image," such as by obtaining a better regression fit with less messy data. This is a pity, because outliers, if they are not obvious mistakes or mismeasurements, play a major role in scientific discoveries. They provide a possible refutation of the theory.

Many universities and research institutions have Institutional Review Boards to tackle such issues on research ethics. The principal investigator should ensure that research assistants are properly selected, trained, and follow protocols.

References

Anderson, L. and Seligson, M. (1994) Reformism and radicalism among peasants: An empirical test of Paige's Agrarian Revolution. *American Journal of Political Science*, **38**(4), 944–972.

Arshanapalli, B. and Nelson, W. (2008) A cointegration test to verify the housing bubble. *International Journal of Business and Finance Research*, **2**(2), 35–43.

Case, K. and Shiller, R. (2003) Is there a bubble in the housing market? *Brookings Papers on Economic Activity*, **2**, 299–342.

Chinco, A. and Mayer, C. (2014) Misinformed speculators and mispricing in the housing market. *NBER Working Papers* 19817. Massachusetts: NBER.

Demographia (2021). *The International Housing Affordability Survey*. St Louis: Demographia.

Fexia, C., Sanchez-Garcia, J., and Brisley, A. (2020) Gangs, methodology, and ethical protocols. *Journal of Applied Youth Studies*, **3**, 5–21.

Garber, P. (2000) *Famous first bubbles: The fundamentals of early manias*. Cambridge: MIT Press.

Hubbard, R. and Mayer, C. (2009) The mortgage market meltdown and house prices. *B. E. Journal of Economic Analysis and Policy*, **9**(3), 1–45.

Israel, M. (2014) *Research ethics and integrity for social scientists*. Thousand Oaks: Sage.

Jones, C. and Leishman, C. (2003) Structural change in a local housing market. *Environment and Planning A*, **35**(7), 1315–1326.

Kaufmann, M. and Muhleisen, M. (2003) Are house prices overvalued? *IMF Country Report No. 3/245*. Washington DC: IMF.

Liebow, E. (1967) *Tally's corner*. Boston: Little, Brown, and Co.

Maisel, S. (1963) A theory of fluctuations in residential construction starts. *American Economic Review*, **53**, 359–383.

Malkiel, B. (2016) *A random walk down Wall Street*. London: W. W. Norton.

Meese, R. and Wallace, N. (1994) Testing the present value relation for housing prices: Should I leave my house in San Francisco? *Journal of Urban Economics*, **35**, 245–266.

Paige, J. (1975) *Agrarian revolution*. New York: Free Press.

Peters, T. and Waterman, R. (1982) *In search of excellence*. New York: HarperCollins.

Poterba, J. (1991) House price dynamics: The role of tax policy and demography. *Brookings Papers on Economic Activity*, **22**(2), 143–183.

Quigley, J. (1997) *The economics of housing*. Northampton: Edward Elgar.

Quigley, J. (2003) Comment on Case and Shiller: Is there a bubble in the housing market? *Brookings Papers on Economic Activity*, **34**(2), 343–362.

Schimel, J. (2011) *Writing science: How to write papers that get cited and proposals that get funded*. London: Oxford University Press.

Tota, A. and Hove, P. (2019) *Memoirs of a book thief*. London: SelfMadeHero.

Tourish, D. (2019) *Management studies in crisis: Fraud, deception, and meaningless research*. London: Cambridge University Press.

This page intentionally left blank

CHAPTER 4

Research Design I: Case Study

Features of Case Studies

Recall from Chapter 1 that the research process begins with a research question or problem, followed by a literature review to develop the theory, hypothesis, model, or framework. The next step is to develop an appropriate research design, and we begin the discussion of research designs with case studies.

A case study tells a big and in-depth story through the lens of a small case (Walton, 1992). It seeks to develop a complete or comprehensive understanding of the selected case using qualitative and quantitative data to make a point, describe a phenomenon, or determine causal relations.

A "case" is a unit of study. It may be a person, team, project, organization, province, country, process, activity, or situation. Examples of processes, activities, and situations include recruitment processes, class teaching, and project disputes. The unit should be a clearly defined and bounded system.

Case studies differ from *case methods* that are widely used in business schools to teach students how to apply the principles of management. In case methods, students study a written business case in advance of each class and debate the issues in class. They do not conduct research to solve a scientific problem.

Many case studies are interpretive, where the actor's perspective guides the research. Some case studies, such as historical ones, are primarily causal. However, unlike experimental studies, there is little control over extraneous variables in a historical case study. The historian cannot influence what has happened. This limitation, together with the historian's personal interests and perspective, makes it vulnerable to alternative interpretations (Carr, 1961). Nowadays, few historians defend Elton's (1967)

depiction of the practice of history as the search for objective truth. At the other extreme, history is also not merely linguistic fiction (White, 1973) but is largely accepted as studies that look at the past from different perspectives. "The" study of history has become "a" study of history (Toynbee, 1988) or simply "true stories" (Arnold, 2000). It is "true" by agreeing with the facts, and it is a "story," that is, an interpretation or argument about a messy, inaccessible, and complex past.

Many disciplines in the social sciences such as sociology, political science, business management, anthropology, psychology, and education regularly use case studies to study phenomena by "soaking and poking" with subjects rather than historical records over extensive field periods (Liebow, 1967; Merriam, 1998; Samuels, 2012). The aim is to provide "thick" holistic descriptions (Geertz, 1973) rather than the "thin" descriptions one finds in a survey questionnaire. In the latter, the questions often involve opinions and preferences with limited standardized options or boxes to tick.

Sampling

The researcher needs to justify why a particular case or a small number of cases are selected. There are several options (Stake, 1995; Ellet, 2018) (Table 4.1). Note that case studies seldom require representative samples. A related misapprehension is their generalizability; the purpose of a case study is to probe to achieve in-depth *understanding*, not to generalize. Logically, it is impossible to generalize from a small sample. Often, the researcher wishes to study the unique, the unusual, and test cases.

The "problematic case" is listed here because it is a common goal of case studies to solve problems. If the researcher is actively engaged during the problem-solving process, it is called action research (Stringer and Aragon, 2020).

Examples

Urban regimes

The first example of a causal case study is Stone's (1989) study of the urban regime in Atlanta in the United States from 1946 to the 1980s.

Table 4.1 Sampling case studies.

Type of sample	Purpose	Example
Typical case	Discover the usual pattern	What is the typical village hospital like?
Unique case	Discover the unique	Why does this organization excel?
Problematic case	To solve problems	What are the solutions to project defects?
Test case	To test theory	Does additional funding improve vocational school performance?
Multiple cases	Compare and contrast (Stake, 2005)	What explains the differing performance of contractors?
	Build theory (Glaser and Strauss, 1967)	How can the performance of local engineering firms be improved?
Embedded case	Probe deeper using sub-units (Yin, 2017)	Probe sub-units of an organization, e.g. a project team.

An urban regime is an informal but stable governing coalition comprising leading politicians (often the mayor), senior bureaucrats, business leaders, and possibly the labor union leaders and other interest groups. A regime may

- support economic growth (developmental regime);
- maintain the status quo (maintenance regime); or
- protect the environment (progressive regime).

There is no formal structure, but the developmental regime that Stone considered was interested in the redevelopment of the city in the face of global competition for capital investment.

The context matters. The federal urban policy began with the Housing Act of 1937 to eliminate unsafe and insanitary housing. After World War II, the Act was amended to include slum clearance and urban renewal, which resulted in the destruction of working-class communities and subsequent building of dense and high-rise public housing blocks in remote places (Anderson, 1967). In addition, the large concentration of poor and mostly unemployed people in these housing estates was a potential time bomb.

In response, the Johnson administration (1963–1969) declared its War on Poverty by implementing federally supported programs to help the poor, racial minorities, and elderly through training and education, food stamps, insurance, social security benefits, and better economic opportunities. There was a policy shift from building large-scale mass public housing towards smaller projects, social housing, preservation of heritage, and conservation. Despite these efforts, many urban riots took place in American cities during the Civil Rights Movement of the 1950s and 1960s.

In a major shift in urban policy, the conservative Reagan administration (1981–1989) no longer viewed the "city problem" as a federal problem. It then cut categorical grants sharply, leaving cities to fend for themselves (Stoesz, 1992). The other plank of Reagan's urban policy was deregulation and privatization, which encouraged local governments to collaborate with the private sector to revitalize ailing cities. An example of such collaboration was the establishment of urban enterprise zones in deprived communities. The local government would provide tax and other incentives to encourage firms to locate in these poor areas.

Consequently, during the 1980s, there was a substantial shift in the fortunes of American cities. Cities in the old "rust belt," such as St. Louis, Detroit, Cleveland, and Pittsburgh, lost about 22 to 31 percent of their population between 1970 and 1984, while cities in the South and West, such as Los Angeles, Houston, and Phoenix gained between 10 and 46 percent over the same period (Stoesz, 1992).

For Stone, nothing just happens for these declining or growing American cities. Neither the free market nor the tide of deindustrialization and globalism alone shapes the future of cities (Petersen, 1981). The visible hand of the urban regime is what makes urban redevelopment possible; conversely, it may also end in failure through poor leadership and management, corruption, and other reasons. Hence, the capacity of an urban regime to effectively organize and revitalize a city is what matters.

Stone subdivided the research question regarding the urban regime in Atlanta into three possible sub-questions:

- who formed the governing coalition;
- how did it accomplish its mission; and
- what were the consequences.

He answered these questions by examining the context, actors, ideas, structure, processes, and constraints in a causal way in terms of how they evolved historically. The narrative is an analysis of the long political struggles and conflicts before a bi-racial governing urban coalition that supported urban development emerged.

On the issue of sampling, why did Stone select Atlanta? He argued that Atlanta's urban regime excelled in getting strategically important people in the city to act together. Thus, he is using a unique case to discover how the urban regime was formed and how it managed to act coherently. Stone is not looking for a representative case or cases to compare and contrast.

Can we generalize the study of urban regimes beyond Atlanta? Critics argue that urban regimes may not exist outside the United States because of differing political contexts (Keating, 1991). For example, in Europe, Africa, and Asia, the central governments play a larger role in urban affairs despite the current neoliberal trend towards decentralization (Cheema and Rondinelli, 2007).

God's choice

Our example of an interpretive case study is Peshkin's (1986) study of the "total world" of a fundamentalist Christian school, the Bethany Baptist Academy (BBA, a pseudonym), in Illinois, America. The school immersed itself in a "total world" with its church and Christian families, with little contact with outsiders. Other examples of these "total institutions" include nursing homes, mental hospitals, military training camps, and prisons (Goffman, 1961). The film *One flew over the cuckoo's nest* (1975) is based on Goffman's idea of total institutions in public mental hospitals. These are places of order, discipline, isolation, and tension.

The objectives of Peshkin's study were to discover what such a school was like, what made it attractive to many American families during that period, and the strains and tensions within the school. The research framework comprised

- the school's doctrine of a single truth as outlined by its founding principal (pastor) and implemented by its headmaster;
- the Christian teachers;

- the structure of control;
- its socializing regime;
- the students' beliefs;
- the impact of Christian school on its graduates;
- the students who did not fit well;
- the school as a total institution; and
- the benefits and costs of such a system to American society.

Peshkin structured the chapters of the book based on this framework. Except for the last chapter on benefits and costs, he wrote the chapters from the actor's perspective. For example, in the chapter on the Christian teachers (Chapter 3), Peshkin first provided the background of the 12 full- and part-time teachers. They came from large middle-class families in different US states and backgrounds, and most were married. All of them attended college and believed that God "called" them to be teachers. The typical week began on Sunday, which was occupied with church-related activities. From Monday to Friday, other than the usual classes, the teachers performed various tasks such as collecting children before school and driving them home after school, driving the soccer team for competitions, community visitation to spread the Christian message, and afternoon sports duty. On Saturday, the men had prayer breakfast followed by bus visitation. Peshkin noted that the teachers engaged in these activities quite willingly because of their "calling."

What did the teacher do in their free time? Few teachers were serious readers, and only one belonged to any organization outside the school church. Why? Because "Church and church friends take all my time." (p.71) The breakdown was to spend about 65 percent of their waking time to school, 15 percent to the church, and the remaining 20 percent to family or friends. The teachers were contented with their busy schedules. Nonetheless, there were a number of strains, such as not having privacy in one's dating life, but most teachers were happy with the friends they have in church and school.

On the issue of pedagogy, teachers mined the Bible for pedagogic principles, believing in its richness. As Headmaster McGraw put it, "the Word of God should permeate every moment of the school." (p.75) For example, in the teaching of science, the goal was to develop a Christian

mind so that students see everything from God's perspective. Teachers also viewed students as "clay," and were challenged to shape their character. They looked for signs that identified the unsaved, such as their inability to pray or the use of inappropriate language. Once the teacher has identified a student, she would use a plan of salvation to engage the student in a conversation.

Peshkin then went on to probe how teachers viewed their relations with other stakeholders. Students were seen as disciples. Parents were viewed as co-workers, colleagues as friends, and administrators as leaders. They were loyal to administrators and avoided gripes that demoralized.

How did the teachers contrast their work with those in public schools? They acknowledged that public schools were better resourced, and teachers were better paid. However, there were clear disadvantages with teaching in public schools, such as poor discipline, social problems, less dedicated teachers, indifferent parents, and the inability to develop a student's character and teach the truth. For these teachers, the Christian school was just about right, and they were happy teaching there.

I hope this brief summary of Chapter 3 of Peshkin's book has provided a glimpse of the richness of his study and the importance of understanding the local context, particularly of the link between the church and school. The two entities are inseparable. It also illustrates how the basic idea of the Christian teacher in the research framework is further developed into related concepts such as how teachers saw the world around them, their goals as teachers, the activities they did on a weekly basis, their pedagogies, their relations with other stakeholders, and what motivated them to teach in a Christian school. While Peshkin had his own views, he let the actors speak for themselves.

Why did he choose the BBA? Peshkin, a Jewish professor of comparative education, was fascinated with religious schools. He chose the BBA because permission was granted after several rejections by other Christian schools. A refusing pastor likened Peshkin to "… a Russian who says he wants to attend meetings at the Pentagon just to learn … No matter how good a person you are, you will misrepresent my school because you don't have the Holy Spirit in you." (p.12)

This example raises a number of questions. It is beyond doubt that an intensive study of an American fundamentalist Christian school is fascinating, as there are few such intensive studies. However, is Peshkin's portray of the BBA as a total institution that served only God's Will accurate, given the cautionary misrepresentational remarks by the pastor? This leads us to the question of validity.

Reliability and Validity

Designs of case studies need to pay close attention to reliability and validity. A measurement is reliable if the results can be repeated with little error. For example, a thermometer is a reliable instrument. Our observations of human behavior are less reliable, which necessitates the use of different observers. Documentary data may not be reliable because of forgeries, biases, and silences or gaps.

Unlike quantitative research, we may not be primarily interested in reliability in interpretive studies because we are not measuring anything objectively when looking for different viewpoints. Hence the main issue concerns validity, the correctness or accuracy of our framework, sampling, fieldwork, thick description, and interpretation. This requires a thorough literature review to develop a good framework, the selection of appropriate case(s), and the application of a suitable methodology.

In the field, respondents may not tell the truth or they tell another story because of the presence of the researcher. Sometimes, the researcher's prior assumptions or knowledge may also bias the results. Hence, the researcher needs to build trust and be constantly aware of his presence and assumptions. This self-awareness is called *reflexivity*.

To ascertain whether the thick description is accurate, the researcher should obtain feedback from respondents on the transcript. Another technique is to triangulate the observations by using more than one observer.

Finally, it is good practice to subject the interpretation to expert review to keep it in check. It is also useful to examine negative cases similar to how quantitative researchers examine outliers as possible refutations of the theory.

References

Anderson, M. (1967) *The federal bulldozer*. New York: McGraw-Hill.

Arnold, J. (2000) *History: A very short introduction*. London: Oxford University Press.

Carr, E. (1961) *What is history?* London: Palgrave Macmillan.

Cheema, S. and Rondinelli, D. (Eds.) (2007) *Decentralizing governance*. Washington DC: Brookings Institution Press.

Ellet, W. (2018) *The case study handbook: A student's guide*. Cambridge: Harvard Business Publishing.

Elton, G. (1967) *The practice of history*. London: Fontana Books.

Geertz, C. (1973) *The interpretation of cultures*. New York: Basic Books.

Glaser, B. and Strauss, A. (1967) *The discovery of grounded theory*. Chicago: Aldine.

Goffman, E. (1961) *Asylums*. New York: Doubleday.

Keating, M. (1991) *Comparative urban politics*. Aldershot: Edward Elgar.

Liebow, E. (1967) *Tally's corner*. Boston: Little, Brown and Co.

Merriam, S. (1998) *Qualitative research and case study applications in education*. New York: Jossey-Bass.

Peshkin, A. (1986) *God's choice: The total world of a fundamentalist Christian school*. Chicago: University of Chicago Press.

Petersen, P. (1981) *The limits of the city*. Chicago: University of Chicago Press.

Samuels, D. (2012) *Case studies in comparative politics*. London: Pearson.

Stake, R. (1995) *The art of case study research*. Thousand Oaks: Sage.

Stake, R. (2005) *Multiple case study analysis*. New York: Guilford Press.

Stoesz, D. (1992) The fall of the industrial city: The Reagan legacy for urban policy. *Journal of Sociology and Social Welfare*, **19**(1), 149–167.

Stone, C. (1989) *Regime politics*. Kansas: University Press of Kansas.

Stringer, E. and Aragon, A. (2020) *Action research*. Thousand Oaks: Sage.

Toynbee, A. (1988) *A study of history*. London: Oxford University Press.

Walton, J. (1992) Making the theoretical case. In C. Ragin and H. Becker (Eds.), *What is a case?* (pp. 121–137). Cambridge: Cambridge University Press.

White, H. (1973) *Metahistory: The historical imagination of 19*th *century Europe*. Baltimore: Johns Hopkins University Press.

Yin, R. (2017) *Case study research and applications: Design and methods*. Thousand Oaks: Sage.

This page intentionally left blank

CHAPTER 5

Research Design II: Survey

Features of Surveys

A survey may be used in descriptive, interpretive, or causal studies. In descriptive studies, a survey uses a sample to obtain broad characteristics of the target population. Researchers may also ask respondents for their viewpoints, preferences, or reasons (causes) for their actions. However, surveys are less suitable for identifying causes through mechanisms. For example, if we ask people about the causes of recession, we tend to obtain opinions rather than the links from causes to effects.

Surveys are popular because they provide a quick and efficient way to obtain broad answers based on a sample before generalizing it to the population. The weaknesses of surveys include possible researcher, sampling, and response biases. They are also less appropriate if in-depth answers are required.

Types of Surveys

Surveys may be ad hoc or carried out at regular intervals. Most surveys are cross-sectional studies that gather information about a population using a sample at a point in time. For example, we may ask a group of firms at a point in time about their organizational leadership, marketing strategies, or human resource policy. Such surveys are ill-suited for understanding changes over time.

In a longitudinal study, we collect data over time to monitor changes. There are three types of longitudinal studies, namely,

- trend studies using different samples;
- cohort studies using different samples from the same cohort; and
- panel studies using the same sample.

For example, in a trend study, we may sample different consumers every five years to track changes in online shopping behavior over time.

A cohort is a group of individuals who share a common characteristic such as birth or class. In a cohort study, we take different samples from the same birth cohort (age group), such as those in their 20s, 30s, 40s and 50s.

Finally, a panel study traces the development of the same sample over time, such as starting from a sample of shoppers in their 20s and studying them every five years to track their shopping behavior. Another common use of a panel study is to increase the sample size to improve the precision of our sample estimates. For example, a researcher may not be able to access many firms in an industry. With only a few firms and a large number of variables, it will be difficult to carry out a proper regression analysis using cross-sectional data. One way of overcoming this problem is to use panel data (see Chapter 16), that is, data collected from these firms over time (Hsiao, 2003).

Longitudinal surveys tend to be more expensive than cross-sectional surveys because of the need to monitor samples over time. In addition, if the same sample is used, there is a risk of participants dropping out of the survey.

Sampling

Sampling is a key part of every research design and not just of surveys. In surveys, sampling consists of identifying the following:

- sampling unit or element;
- target population;
- sampling frame (if any);
- sampling method; and
- sample size.

A sampling unit or element is the smallest unit of observation that is of interest to the researcher in a sample (Fig. 5.1). In consumer surveys, the sampling unit may be individuals or the household as the basic decision-making unit on consumption expenditure. In business surveys, the sampling unit is often a manager, professional, or employee and not the firm itself.

The target population (*TP*) is the theoretical aggregate of all sampling units, such as all the contractors in the city or country. We do not sample from the population because it is just a concept.

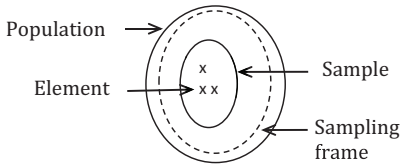

Fig. 5.1 Basic sampling concepts.

The sampling frame (*SF*) is the actual list of elements from which sampling takes place. For example, we may obtain the list of contractors from the local contractors' association. It is slightly smaller than *TP* because not all contractors, particularly the smaller or foreign ones, belong to the association. Similarly, in a school setting, we only have a list of students who are in school when sampling takes place; we cannot sample those who are absent. Obviously, if we are studying truancy, then it is the absent ones that are of interest. In this case, *TP* consists of students who are often absent from class without valid reasons. As shown in Fig. 5.1, the *SF* should be as close to *TP* as possible to reduce sampling bias. For example, in sampling university students, *SF* should comprise students living on and off-campus.

Many surveys do not have *SF*s. For instance, it is not possible or difficult to obtain a list of every fish in a lake or every student in a large university. If an *SF* is available, probability samples are used; otherwise, we use non-probability samples.

Probability Samples

Probability samples use random sampling to draw samples from an *SF*. If *SF* is close to *TP*, probability samples are more accurate than non-probability samples.

The main types of probability samples are

- simple random samples;
- systematic samples;
- stratified samples; and
- cluster samples.

Simple random sample

In a simple random sample, we draw the sampling units randomly. However, simple random samples are rarely used in actual surveys because of the need for a large *SF*. Hence, it is conducted only if *TP* is small and manageable, such as drawing randomly from a party hat. For large populations, we need to find other ways to draw the sample, such as using stratified or cluster samples.

Systematic sampling

In systematic sampling, we first divide *TP* into smaller lists and then draw the elements following a particular pattern after a random start. For instance, if a town has a phone directory comprising 500 pages and the desired sample size is 1,000, then we randomly draw two elements from each page, such as the 10^{th} and 50^{th} names. Hence, systematic sampling is similar to simple random sampling because of random selection.

The lists should not be ordered. Ordered lists are often used in the military, such as that of a platoon where the order is often the platoon commander, sergeant, corporals, and then privates.

Stratified sampling

Stratified sampling is often used in actual surveys. We first stratify *TP* according to some theoretical criteria such as gender, income, workplace, residence, or cohort. We then draw the elements *randomly* from each stratum. In proportional stratified sampling, the samples are drawn in equal proportion. For instance, in a class of 40 boys and 60 girls, we may draw a sample comprising 20 percent of the students from each category, that is, 8 boys and 12 girls.

In disproportionate stratified sampling, the sample proportions are unequal. If a stratum is fairly homogeneous, such as similar ball bearings, we only need to draw a small sample. In contrast, firms are heterogeneous and a larger sample is required for studying dissimilar units.

Subject characteristics are often used to stratify a sample. In Table 5.1, the students have been stratified by class level (Year 1 to 4), religion

Table 5.1 Sample of students stratified by class, religion, and gender.

	Population				Sample			
	Males		Females		Males		Females	
Religion	Yes	No	Yes	No	Yes	No	Yes	No
Year 1	30	30	30	30	3	3	3	3
Year 2	35	35	35	35	4	3	4	3
Year 3	40	40	40	40	4	4	4	4
Year 4	45	45	45	45	4	5	4	5
Total	150	150	150	150	15	15	15	15

(Yes or No), and gender. Proportional stratified sampling of about 10 percent is used.

Cluster sampling

In cluster sampling, selection is based on random clusters rather than individual elements. For example, the use of simple random sampling to select 200 students living on campus will require a large *SF*. A better strategy is to use multi-stage cluster sampling. For example, if there are 7 hostels on campus, we first select 4 hostels randomly. In the second stage, we select 50 students from each of these 4 hostels. Thus, instead of a large SF, only a short *SF* comprising the 7 hostels is required.

Cluster sampling is often used in city surveys to select households by first selecting suburbs, followed by neighborhoods, and then the residential blocks within selected neighborhoods (Blair and Blair, 2015).

Non-probability Samples

Non-probability samples are used if there is no *SF*. Since chance selection is not used, the probability of an element being selected is unknown. However, non-probability samples are easier to collect, which explains their popularity despite the higher probability of bias (Lohr, 2021).

Non-probability samples include

- convenience samples;
- purposive samples;
- quota samples;
- snowball samples; and
- adaptive samples.

We discuss these techniques below.

Convenience sampling

In convenience sampling, we select the elements out of our convenience or because we think they are likely to be good respondents. It is useful for exploratory work, for the pre-testing of questionnaires, or where a quick opinion is required.

Reporters often use this sampling technique for the evening news. It is also widely used in mall intercepts, street surveys, or email surveys. However, the sampling errors may be large, depending on who is selected.

Judgmental sampling

Judgmental sampling is purposive, that is, our judgment or the choice of experts is preferred to random sampling. For example, in the construction of the Consumer Price Index (*CPI*), we deliberately choose the sample basket of goods and services. The basket should represent a typical household consumption pattern.

The sampling errors depend on the quality of judgment. Different experts may not agree on what is representative.

Quota sampling

A quota sample is similar to a stratified sample except that chance selection is not used in each stratum. Hence, a quota sample is not a probability sample. In the hostel example, we may select the 50 students from each of the 4 hostels out of convenience rather than draw them randomly.

Quota samples are popular because stratification reduces the need to select large samples. The greater homogeneity of the sampling elements with each stratum also reduces the sampling error. Finally, since an *SF* is not required, it is cheaper to draw a quota sample.

Snowball sampling

A snowball sample begins with a few respondents who provide referrals for the researcher to contact additional respondents. This may happen if the initial sample is very small, such as if we are sampling people who suffer from a rare disease, gangsters, the homeless, or corporate leaders who are harder to access.

Adaptive sampling

In adaptive sampling, the sample size is not fixed at the initial stage. It changes as the survey progresses (Thompson and Seber, 1996). For example, if there is an outbreak of disease, the survey is extended to neighboring sites after the detection of an initial cluster. This technique has also been applied to surveys of rare animal species. Unlike snowball sampling, it does not use referrals; instead, it expands the spatial reach of the search from an initial discovery.

Sample Size

There is no simple guide on sample size because it depends on the concept of *statistical power* (Cohen, 1988) and other considerations.

We begin with test power. A statistical hypothesis is about the values of a population parameter, such as H_0: $\mu = 5$ kg and H_1: $\mu = 6$ kg for the population mean. As discussed in Chapter 1, it differs from a research hypothesis, which is about causal mechanisms. The two types of hypotheses are related because a causal mechanism may suggest certain values of a parameter. H_0 is called the null hypothesis, and H_1 is the alternative hypothesis. To decide whether to select H_0 or H_1, we compute the *test statistic* from our sample. In this case, the sample mean ($\overline{x}$) is a good test statistic.

In Fig. 5.2, the left curve is the probability distribution of μ under H_0 and the right curve is the distribution under H_1. For example, the left curve is the distribution of the weight of 3-month-old babies, and H_1 is that of babies of the same age given a special diet. The H_1 curve is to the right because the hypothesized μ is larger at 6 kg. For convenience, it is assumed that the distributions are normal. The logic of statistical testing applies to other distributions.

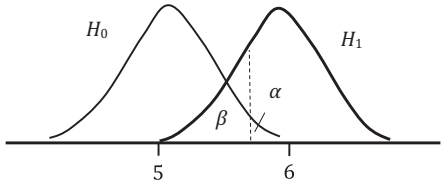

Fig. 5.2 Statistical power.

To conduct the statistical test, we fix the critical region α, which is often 2.5% or 5% of the area under the H_0 curve, depending on whether it is a one-sided or two-sided test. Here, it is a one-sided test because we think the special diet will increase a baby's weight. Hence, $\alpha = 0.05$. The value of α is called the significance level, risk, or Type I error. If our test statistic ($\overline{x}$) lies to the left of the critical region, we accept H_0 because $\overline{x}$ is closer to 5 kg than 6 kg. But it is still possible, with a 5% probability, that we have a random sample of heavier babies when H_0 is actually true. Hence α is an error if we reject H_0 when it is actually true.

However, in accepting H_0, we may commit a Type II error (β), that is, we accept H_0 when it is actually false. The size of β is the area to the left of the vertical line under the H_1 curve. In other words, we conclude that the special diet does not work when, in fact, it works. We just happen to have a sample of smaller babies who have been fed the special diet.

In summary, in deciding whether to select H_0 or H_1, we can make two errors:

- Type I (α): Reject H_0 when it is actually true; or
- Type II (β): Accept H_0 when it is actually false.

The power of a test is given by $1 - \beta$. It is the probability of rejecting H_0 when it is false. A test of high power implies that β should be as small as possible. As seen in Fig. 5.2, β will be reduced, if

- the spread of each curve is smaller (called spread size); and
- the curves are further apart (called effect size).

From Central Limit Theorem, the variance of the sample mean is given by

$$Var(\bar{x}) = \frac{\sigma^2}{n}.$$

Hence, the spread size, or standard deviation of the mean, depends on $\sigma/\sqrt{n}$ where σ is the standard deviation of x and n is the sample size. To reduce the spread size, we should reduce σ or increase n. Hence, we should select babies that are similar in weight for the experiment and, if possible, increase the sample size.

The effect size is the difference in population means. In Fig. 5.2, the effect size is 6 – 5 = 1 kg. To increase the effect size, the special diet must be effective; otherwise, it is more difficult to detect a difference, that is, the test has low power.

There are important caveats to the above rules. First, they only refer to a single variable, the weight of babies. If there are many variables in our survey, the rules are difficult to apply. Second, if we are seeking qualitative responses, we should select samples for varied, rather than similar, responses. Hence, for qualitative researchers, statistical precision may not be a useful criterion.

In regression designs, we also want to spread the data more evenly across the *X*-axis to obtain a better estimate of the slope of the regression line. Here, "getting more of the same" cluster of data points near the lower values of *X* will not improve our estimate of the slope (Fig. 5.3). Hence, regression samples should have large variations in the independent variables.

A final determinant of sample size is access to potential respondents or the cost of doing so. If they are hard or costly to reach, it will affect the sample size.

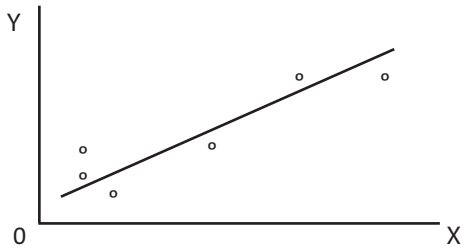

Fig. 5.3 Estimating a regression line.

In summary, there is no simple rule on sample size. In qualitative surveys, the aim is to gather as many differing views as possible. In quantitative surveys, the spread and effect sizes come into play, together with considerations of the cost and difficulties in accessing potential respondents. In regression designs, samples should have large variations in the independent variables.

It is tedious to compute the test power to determine the sample size. Hence, many researchers select sample sizes using reasonable guesses and do not report the power of their statistical tests. Nowadays, there are free statistical software (e.g. G*Power) that compute the sample size (n) based on inputs of the spread size, effect size, α, and β for different types of statistical tests. This is called *a priori* analysis. In post hoc analysis, the test power is computed based on actual sample values. However, as noted above, test power is just one of many considerations on sample size.

Pilot Survey

A pilot survey involving a small but similar sample is often conducted to determine if the survey design and methods of data collection (for example, a questionnaire) may be improved before the actual larger-scale survey. Logistical and other field issues may also surface during the pilot study.

A small sample can provide a surprisingly large amount of valuable feedback (Sampson, 2004). If the target group is hard to reach, try to select a pilot sample that is as close as possible. For example, in a study of how young mothers cope with their pregnancy, it is possible to use experienced mothers as part of the pilot group.

Examples

Perceptions of train service

The first example of a survey research is the design of a perception survey among commuters of the services provided by a mass rapid transit (MRT) system. Train operators often carry out such surveys to obtain feedback from commuters to improve their service (Jones and Stopher, 2003; BART, 2014).

For the research framework, it is hypothesized that commuter perception depends on rater characteristics (income, age, and gender), and the following services are important to commuters:

- ease of access to train stations;
- degree of integration with bus services;
- ease of buying tickets;
- fare structure;
- extent of overcrowding;
- waiting time;
- cleanliness of station;
- station shopping experience;
- helpfulness of service staff;
- safety;
- security;
- internal carriage temperature;
- cleanliness of trains;
- quality of wireless connections;
- response to breakdowns; and
- usefulness of commuter information.

A simple 5-point rating scale is used to rate each service.

There is no *SF* because of the difficulties in generating a list of names for commuters. Hence, non-probability sampling is used. If there are 40 stations in the network, we will select 60 commuters per station during the morning peak hours, resulting in a sample size of 2,400. The quota sample is stratified by gender and age (Table 5.2). If desired, income may

Table 5.2 Quota sample of commuters per station.

Age group	Males	Females
20–39	10	10
40–59	10	10
60 and over	10	10
Total	30	30

also be used for stratification. A pilot study involving two stations may be used to obtain feedback on the questionnaire.

Perceptions of shopping mall

The second example of a survey design is shoppers' perceptions of the attractiveness of a shopping mall. We will use a simple framework where perception depends on shopper characteristics and the following aspects are salient:

- mall location,
- pricing,
- shopping mix, and
- facilities.

For sampling, *TP* is potential shoppers to the mall. There is no *SF* because of the difficulties in obtaining a list of potential shoppers. Hence, a non-probability sample is used. As convenience sampling is unlikely to be representative, quota sampling stratified by type of housing, age, and gender is used (Table 5.3). Housing type (public or private housing) is a proxy for income.

The sample size is 240. It is possible to stratify further, such as using more varied age groups instead of just two simple categories.

The use of mall intercepts to interview respondents will miss those who do not like to shop there. Hence, the sample should be drawn from households within the market area of the mall. There is some subjectivity in drawing the market boundary because shoppers take different modes of transport (Vandell and Carter, 1993).

Table 5.3 Quota sample of potential shoppers.

Males				Females				
<30 years		≥30 years		<30 years		≥30 years		Total
Public	Private	Public	Private	Public	Private	Public	Private	
30	30	30	30	30	30	30	30	240

Reliability and Validity

Recall from Chapter 4 that reliability refers to the replicability, stability, or consistency of our measures. If a respondent answers "Yes" to a question, he should also answer "Yes" to the same question tomorrow. In the *test-retest* check on reliability for a questionnaire, the correlation coefficient should exceed 0.7 for it to be reliable. However, the responses may change because circumstances have changed. To minimize this problem, the time period between tests should be short.

Different interviewers may also affect the responses. To test *inter-rater reliability*, we may compute the kappa correlation coefficient to determine the extent of agreement among the responses. Again, it should exceed 0.7 with proper training and supervision of interviewers. However, the most common measure of reliability is *Cronbach's alpha* (Cronbach, 1951). In a questionnaire comprising 5(k) questions on attitudes towards the environment, we expect those with positive attitudes to have similar responses to the questions. Thus, if the responses are poorly correlated, the questionnaire is not reliable. Cronbach's alpha compares the sum of the variance of the scores from each question (s_i^2) to the variance of the total score from all questions (s^2). Specifically,

$$\alpha = \frac{k\,(s^2 - h)}{(k-1)s^2},$$

where $h = \sum s_i^2$. The data table is shown below:

Respondent	Q1	Q2	Q3	Q4	Q5	Total
1	3	4	5	5	2	19
2	3	3	2	3	5	16
3	4	2	2	3	4	15
...						
n	3	4	3	2	1	13
Variance	0.3	0.2	0.5	0.5	0.3	6.1

Thus $h = 0.3 + 0.2 + 0.5 + 0.5 + 0.5 + 0.3 = 1.8$ and

$$\alpha = \frac{5(6.1 - 1.8)}{4(6.1)} = 0.88,$$

which is above the acceptable reliability value of at least 0.7. For an extensive discussion on reliability in surveys, see (Alwin, 2007).

There are three types of validity in surveys. *Construct validity* is about correspondence between constructs and measures. For example, measures of "attitudes towards the environment" should include items like how much one values the environment, the trade-off between the environment and economic progress, beliefs on whether we have crossed the tipping point on climate change, and so on.

Internal validity is about the logical tracing of causes and effects, that is, causal mechanisms. Since surveys are usually not appropriate for finding causal mechanisms, the issue of internal validity may not arise.

Finally, *external validity* refers to the extent to which the results may be generalized. This requires careful sampling so that the sample is representative. The other requirement for generalizability is context. If the contexts are different, it is risky to generalize. For example, a teaching method that works well in a developed country may not work in a developing country. The contexts are different, such as high absenteeism among students and teachers in some countries.

There are other sources of errors in surveys (McNabb, 2013). Administrative errors arise from mistakes in data collection or processing, such as talking to the wrong person or keying in the wrong data. Respondent errors may occur if respondents cannot recall, misunderstand the question, respond only to strategic incentives such as the free provision of public goods, wish to hide sensitive information, are not in the mood to talk, or are put off by the questions or interviewer.

References

Alwin, D. (2007) *Margins of error: A study of reliability in survey measurement*. New York: Wiley.

BART (2014) *2014 Customer satisfaction study*. San Francisco: BART Marketing and Research Department.

Blair, E. and Blair, J. (2015) *Applied survey sampling*. Thousand Oaks: Sage.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences*. London: Routledge.

Cronbach, L. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Hsiao, C. (2003) *Analysis of panel data*. Cambridge: Cambridge University Press.

Jones, P. and Stopher, P. (Eds.) (2003) *Transport survey quality and innovation*. London: Emerald.

Lohr, S. (2021) *Sampling: Design and analysis*. London: CRC Press.

McNabb, D. (2013) *Non-sampling error in social surveys*. Thousand Oaks: Sage.

Sampson, H. (2004) Navigating the waves: The usefulness of a pilot in qualitative research. *Qualitative Research*, **4**, 383–402.

Thompson, S. and Seber, G. (1996) *Adaptive sampling*. New York: Wiley.

Vandell, K. and Carter, C. (1993) Retail store location and market analysis: A review of the research. *Journal of Real Estate Literature*, **1**(1), 13–45.

This page intentionally left blank

Chapter 6

Research Design III: Comparative Design

Features of Comparative Designs

A comparative study compares and contrasts several cases to draw *causal* inferences. Hence, it differs from multiple case studies that are interpretive and usually do not draw causal inferences.

A comparative design is a small-*N* design. It falls somewhere between a case study and a large-*N* survey. It is attractive to researchers who wish to explain phenomena where there are only a few cases and it is not possible to vary and control the variables to determine causes. Unlike experimental studies, variations in comparative cases occur naturally without deliberate interventions by the researcher. Thus, comparative studies are widely used in political science to study social revolutions in a few selected countries (Paige, 1975; Skocpol, 1979; Lange, 2013). The small sample also makes it difficult to use multivariate statistical analyses or broad large-sample surveys.

In cultural studies, Parker (2001) compared the police system in Japan and the United States (US) to account for the different crime rates. Brake (1985) compared youth culture in the US, Britain, and Canada. Zippel (2006) studied differences in sexual harassment in the US, European Union, and Germany. The comparative method is also used in education (Bray et al., 2014).

Many economists and sociologists used the comparative approach to explain the economic performance of the East Asian "dragon" economies (Vogel, 1991; Schuman, 2010; Perkins, 2013). Klein (1988) compared slavery in the two New World colonies of Cuba and Virginia. Both colonies were plantation economies with similar proportions of slaves in the

population. He used institutional differences to explain why slaves in Cuba achieved greater social integration and occupational mobility.

Comparison does not prove cause. There is no causal mechanism linking similarities or differences to the outcome. Hence, comparative designs can only help to identify possible causes.

Types of Comparative Designs

There are four possible comparative designs (Mill, 1884; Tilly, 1984; Pickvance, 2001) (Table 6.1). They are renamed in Table 6.2 as *SS*, *SD*, *DS*, and *DD* strategies, where *S* stands for similarity and *D* stands for difference.

Suppose we want to compare learning strategies (features or independent variables) among academically strong and weak students (outcomes or dependent variables). The six possible strategies, (a) to (f), are shown in Table 6.3. The choice of design depends on the objectives of the study; for example, knowing the learning strategies of weak students is less useful. Hence, we may only be interested in designs (a), (c), and (f). For example, in design (f), we wish to compare the different learning strategies of strong and weak students.

Table 6.1 Four comparative strategies.

		Outcomes	
		Similar	Different
Features	Similar	Method of Agreement	Differentiating comparative analysis
	Different	Universalizing comparative analysis	Method of Difference

Table 6.2 Simplified comparative strategies.

		Outcomes	
		Similar	Different
Features	Similar	*SS*	*SD*
	Different	*DS*	*DD*

Table 6.3 Comparative learning strategies and outcomes.

		Outcomes	
		Similar	Different
Learning strategies	Similar	(a) Similar strategies among strong students (b) Similar strategies among weak students	(e) Similar strategies among different students
	Different	(c) Different strategies among strong students (d) Different strategies among weak students	(f) Different strategies among different students

Comparative Sampling

Comparative sampling depends on the selected strategy. For example, if strategy (a) is selected, then the sample consists of only academically strong students. Similarly, if we select strategy (f), the sample will consist of strong and weak students. Clearly, strategy (f) is better than (a) because even if a common set of learning strategies is identified among strong students, we are never sure if weak students also use the same strategy.

Apart from the outcome criterion above, it is also necessary to consider the independent variables. Generally, diversity is desired, that is, the learning strategies should be as diverse as possible. For example, if $X_1, \ldots, X_k$ causes Y, it is difficult to persuade readers if we only sample cases with, say, X_1 and X_2. There are just too few independent variables to make a conclusion. For example, if Y is economic growth, X_1 is external trade, and X_2 is resource endowment, there are insufficient independent variables to explain economic growth. There is a long list of omitted variables such as policies, geography, culture, institutions, and politics.

In summary, in comparative sampling, we prefer to select different outcomes and a diverse set of independent variables.

Examples

Four little dragons

The first example of a comparative design is Vogel's (1991) study of the spread of industrialization in East Asia from the 1960s to the 1980s, in the

so-called "four little dragons" of South Korea, Taiwan, Singapore, and Hong Kong. It is one of many studies of the rapid growth of East Asian economies (e.g. (World Bank, 1993)).

Vogel identified nine factors that may explain the rise of East Asia, namely,

- massive external aid from the US and international agencies;
- destruction of the old order and its replacement by a "strong State" that is relatively insulated from the local elites;
- sense of political and economic urgency for industrial development on the part of its leaders and the community;
- eager and plentiful labor force;
- familiarity with the good performance of the Japanese model of economic development, beginning with labor-intensive industries, export growth, skills and technological upgrading, and the crucial role of the Developmental State in guiding these changes;
- culture, such as a respected and meritocratic bureaucracy, an entrance examination system that rewards learning and discipline, group cohesion, and self-cultivation and learning;
- consumerism, the passionate drive to acquire new goods as incomes rise;
- export orientation; and
- the success cycle, where successful industrialization develops the confidence to acquire new skills for another round of achievement.

To make comparisons easier, it is helpful to reduce the number of variables to a more manageable set. The "massive external aid" after World War II is a one-off event and may be removed from consideration. The "urge to industrialize" is also unnecessary because most post-World War II governments in the developing world prefer industrialization to agricultural development to create jobs and shift away from excessive reliance on cyclical crop production.

The cultural factors may be consolidated as having an effective and coherent bureaucracy in line with the literature on "good governance" (World Bank, 2017). There is nothing particularly East Asian about self-cultivation (Bruford, 2010), and groupism may result in collective myopia (Chikudate, 2015).

For the last three factors, consumerism appears to be an effect of rising income rather than a cause of growth. Similarly, the success cycle kicks in only *after* the economy has been set in motion. Hence, we are left with only export orientation as a possible cause of rapid growth in East Asia.

In summary, there are five common factors:

- destruction of the old order;
- its replacement by a Developmental State and effective bureaucracy that is relatively insulated from the local elites;
- disciplined and plentiful labor force;
- export orientation; and
- familiarity with the Japanese model of development.

These factors have also been identified in the Developmental State literature (Johnson, 1982; Amsden, 1989). Other factors may be added, such as market-conforming interventions, active consultations with the private sector, industrial targeting, becoming an entrepreneur as a last resort, having a pilot agency, and the ability to resist rent-seekers and discipline workers as well as poor corporate performers. The term "soft authoritarianism" is sometimes used to describe such Developmental States (Scalapino, 1989).

Are the above set of factors the recipe for economic success in East Asia? There are doubts. For neoclassical economists, the East Asian miracle is a myth (Krugman, 1994) because it is largely driven by "brute force" growth of capital and labor inputs rather than by total factor productivity growth. Simply put, much of the growth was due to perspiration rather than inspiration (innovation).

In recent years, intense international competition, rising land and labor costs as well as slower or falling growth of the labor force have reduced economic growth in these countries (Kozul-Wright and Rayment, 2007) even before the global subprime crisis of 2007–8. Exports have also fallen or grown far more slowly, in line with the slower growth in global trade (Hoekman, 2015). From this perspective, the high growth rates from the 1960s to the 1980s in East Asian economies is a one-time event that is unlikely to be repeated. The future is more uncertain, and it is difficult for the State or entrepreneurs to pick winners.

From a different perspective, the long post-World War II boom may have led to the development of powerful distributional coalitions (interest groups) that slow down a country's capacity to respond to structural changes and thereby reduce the rate of economic growth (Bates, 1981; Olson, 1982; Smith, 2017).

Finally, many other African States have imitated the industrialization strategies of the East Asian dragons, and yet failed miserably or achieved only modest economic growth (Mkandawire, 2001). The common factors also exist in failed cases, which is why we should compare both successful and unsuccessful cases.

Colonial origins of comparative development

The second example is the work of Acemoglu et al. (2001) on the colonial origins of comparative development.

The basic claim is that the current gap between rich and poor countries lies in their colonial origins. For poor countries, the European colonizers established "extractive" institutions to transfer as much wealth back to home. In contrast, in neo-European colonies such as Australia, Canada, and New Zealand, the imperialists replicated European institutions and emphasized local economic development, protection of property rights, and strong checks on State power. In terms of Table 6.2, it is a *DD* design comparing dissimilar cases (poor colonies and neo-European colonies) with dissimilar outcomes (poor and rich countries).

Why did the European power establish different colonial institutions? Acemoglu et al. argued that high mortality rates among the early settlers were a key consideration, which explains why they planted extractive institutions in the tropical countries of South America, Africa, and Asia.

Why did the extractive colonial institutions persist long after the colonies achieved political independence (Young, 1994)? For Acemoglu et al., the small group of new elites has the incentive to perpetuate these institutions, and changing them would be costly. However, this idea of persistent colonial extractive institutions is not persuasive. It seems strange to blame poverty on colonial institutions established more than a century ago. A better approach is to take into account the dynamic struggles among contending groups, which led to the development of failed, predatory, and developmental States (McSherry, 2005; Bates, 2015).

Reliability and Validity

The reliability issue in terms of consistency of measures is not unique to comparative studies. On validity, there are several issues such as

- the inability to prove cause;
- difficulties in generalizing from limited cases;
- problems in finding comparable or diverse cases;
- difficulties in sorting out rival explanations if there are many variables and only a few cases (Lijphart, 1971); and
- dealing with only binary outcomes.

It is possible to address some of the shortcomings. For example, in Paige's (1975) study of the effect of agricultural exports on the social movements of cultivators (see Chapter 3), he excluded the oil-producing countries and small city-states without monocrop economies. However, the theory becomes less general.

Comparative studies with binary or dichotomous outcomes involve the presence or absence of causes or effects. It is all or nothing. In practice, it is usually a matter of degree, not of kind. There have been attempts to overcome this limitation by using truth tables and fuzzy or Boolean logic, called Qualitative Comparative Analyses (Ragin, 1987). So far, there has been limited success.

References

Acemoglu, D., Johnson, S., and Robinson, J. (2001) The colonial origins of comparative development: An empirical investigation. *American Economic Review*, **91**(5), 1369–1401.

Amsden, A. (1989) *Asia's next giant: South Korea and late industrialization*. Oxford: Oxford University Press.

Bates, R. (1981) *Markets and States in tropical Africa*. Berkeley: University of California Press.

Bates, R. (2015) *When things fell apart: State failure in late-century Africa*. London: Cambridge University Press.

Brake, M. (1985) *Comparative youth culture: The sociology of youth cultures and youth subcultures in America, Britain, and Canada*. London: Routledge.

Bray, M., Adamson, B., and Mason, M. (Eds.) (2014) *Comparative education research*. New York: Springer.

Bruford, W. (2010) *The German tradition of self-cultivation*. Cambridge: Cambridge University Press.

Chikudate, N. (2015) *Collective myopia in Japanese organizations*. London: Palgrave MacMillan.

Hoekman, B. (Ed.) (2015) *The global trade slowdown: A new normal?* London: CEPR Press.

Johnson, C. (1982) *MITI and the Japanese miracle: The growth of Industrial policy, 1925–1975*. Stanford: Stanford University Press.

Klein, H. (1988) *Slavery in the Americas: A comparative study of Virginia and Cuba*. London: Oxford University Press.

Kozul-Wright, R. and Rayment, P. (2007) *The resistible rise of market fundamentalism*. London: Zed books.

Krugman, P. (1994) The myth of Asia's miracle. *Foreign Affairs*, **73**(6), 62–78.

Lange, M. (2013) *Comparative historical methods*. Thousand Oaks: Sage.

Lijphart, A. (1971) Comparative politics and comparative method. *American Political Science Review*, **65**, 682–693.

McSherry, P. (2005) *Predatory states*. Lanham, Maryland: Rowman and Littlefield.

Mill, J. (1884) *A system of logic*. London: Longman.

Mkandawire, T. (2001) Thinking about developmental States in Africa. *Cambridge Journal of Economics*, **25**(3), 289–313.

Olson, M. (1982) *The rise and decline of nations*. New Haven: Yale University Press.

Paige, J. (1975) *Agrarian revolution*. New York: Free Press.

Parker, C. (2001) *The Japanese police system today: A comparative study*. London: Routledge.

Perkins, D. (2013) *East Asian development*. Massachusetts: Harvard University Press.

Pickvance, C. (2001) Four varieties of comparative analysis. *Journal of Housing and the Built Environment*, **10**, 162–84.

Ragin, C. (1987) *The comparative method: Moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.

Scalapino, R. (1989) *The politics of development*. Cambridge: Harvard University Press.

Schuman, M. (2010) *The miracle: The epic story of Asia's quest for wealth*. New York: Harper Business.

Skocpol, T. (1979) *States and social revolutions: A comparative analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

Smith, M. (2017) *Power, norms, and inflation: A skeptical Treatment*. London: Routledge.

Tilly, C. (1984) *Big structures, large processes, huge comparisons*. New York: Sage.

Vogel, E. (1991) *The four little dragons*. Cambridge: Harvard University Press.

World Bank (1993) *The East Asian miracle*. Washington DC: World Bank.

World Bank (2017) *Governance and the law*. Washington DC: World Bank.

Young, C. (1994) *The African colonial state in comparative perspective*. New Haven: Yale University Press.

Zippel, K. (2006) *The politics of sexual harassment: A comparative study of the United States, the European Union, and Germany*. London: Cambridge University Press.

This page intentionally left blank

CHAPTER 7

Research Design IV: Experiment

Features of Experimental Design

Suppose the relation between causes (*X*s) and an effect or outcome (*Y*) is given by

$$Y = f(X_1,\ldots, X_k),$$

where Y is the dependent variable and $X_1,\ldots, X_k$ are k independent variables. The function $f(.)$ specifies how Y and Xs are related. It may be known, imprecisely known, or we simply do not know.

An experimental design is suitable if k is small and it is possible to vary some variables to ascertain their effects on Y by keeping other Xs constant. For example, if $k = 5$, it may be of interest to see how X_1 and X_2 affect Y, keeping X_3, X_4, and X_5 constant. Here, X_1 and X_2 are called *treatments*. It may be a vaccine, teaching method, type of fertilizer, an incentive, a new car financing scheme, a policy change, and so on (Orr, 1998; Abelson et al., 2003; Dunning, 2012).

Strictly speaking, experiments are based on correlations, not causes, for lack of an explicit mechanism linking causes to effects. The researcher can only observe how X and Y vary.

Social experiments are less common because k is often large and it is difficult to fix or vary the values of socioeconomic variables. For example, the economist cannot vary the interest rates to observe its effects on business investment. He can only observe how the two time series vary. Hence, the economist conducts observational rather than experimental studies. Through regression analysis, he uses *statistical* rather than experimental control to examine possible causes and effects.

There are many types of experimental designs (Dean and Voss, 1999; Box et al., 2005; Montgomery, 2009). The basic designs are discussed below.

Classical Experimental Design

In the classical experimental design (CED), an experimenter uses two comparable groups of subjects, called the *control* group (*C*) and the *experimental group* (*E*). The control group acts as the counterfactual or comparison group. She then administers a *treatment* (*T*) to *E* and compares the effects:

$$\begin{array}{llll} E & y_E & T & Y_E \\ C & y_C & & Y_C \end{array}$$

For simplicity, we assume that both groups are of equal size (*n*), although it is unnecessary as long as they are comparable. If there is some doubt, each subject may be given a pre-test (Table 7.1). The pre-test scores (*y*s) are then compared; if they are comparable, the pre-test scores may be disregarded. We only need to compare the post-test scores (*Y*s) of both groups to determine if *T* is effective. If pre-test scores are used, then

$$\text{Impact} = (Y_E - y_E) - (Y_C - y_C).$$

Table 7.1 Classical experimental design.

Subject	Pre-test score	Treatment	Post-test score
1	90		95
2	80		70
3	60	*T*	80
…			
n	70		75
A	60		65
B	70		75
C	85	*Nil*	80
…			
n	90		85

This is also called the difference-in-difference (*DD*) approach.

Quasi-experimental Designs

If it is not possible to implement the full CED, some adjustments are made, such as

- not using a control group;
- not conducting a pre-test; or
- not using comparable groups.

Campbell and Stanley (1963) called these departures quasi-experimental designs (QEDs) (Reichardt, 2019). Obviously, conclusions from QEDs are less persuasive than the CED. For example, without a control group, it is difficult to separate treatment effects from trends and other influences. If a firm advertises its product (the treatment) for a month and then checks the change in sales, advertising may not be the reason for the change in sales. It may be due to economic growth and seasonal effects.

Parallel Group Design

In a parallel group design, there are two treatments, *T* and *U* (Table 7.2), and the objective of the study is to determine which treatment is more

Table 7.2 Parallel group design.

Subject	Pre-test score	Treatment	Post-test score
1	90		95
2	80		70
3	60	*T*	80
…			
n	70		75
A	60		70
B	70		70
C	85	*U*	80
…			
n	90		95

effective. The two groups should be comparable and independent. It is a QED because there is no control group.

Repeated Measures Design

A repeated measures design may be used if it is difficult to find comparable groups. Only one group is used and each subject is given two treatments (*T* and *U*) over two periods. It is also called a "within subject" design because each subject acts as his own control, or a "crossover" design because the treatments are administered at sufficiently long periods apart. This design is common in clinical trials where the effects of the first treatment must not be present when the second treatment is administered.

Let *T* and *U* be two different types of wine. In a parallel group design, two groups of people are used, one for each wine, and we obtain the ratings for each wine on a scale of 1 to 10, for example. There is large variability in ratings because the two groups do not have similar tastes (Table 7.3).

In a repeated measures design (Table 7.4), there is only one sample and each participant is given *U* and *T* in random order to ensure that the order of treatment does not matter. Observe that some wine tasters are higher scorers (e.g. Participant 1), while others are low scorers (e.g.

Table 7.3 Parallel group design in wine tasting experiment.

Participant	Wine	Rating
1		9
2		7
3	*U*	4
…		
n		7
A		3
B		7
C	*T*	8
…		
n		5

Table 7.4 Repeated measures design in wine tasting experiment.

Participant	Wine *U*	Wine *T*	*d*
1	8	9	1
2	4	5	1
3	6	8	2
…			
n	7	7	0

Participant 2). This variability *between* participants does not matter because we are only interested in the difference in ratings (d) for each participant.

Randomized Block Design

In a randomized block design (RBD), there are two variables, namely, the variable of interest and the extraneous ("noise") variable that affects the result but is not of interest. For example, there is considerable evidence that boys tend to do better than girls in mathematics (Niederle and Vesterlund, 2010). Thus if we wish to test a new method of teaching (T), gender (G) is a possible blocking variable.

Suppose there are 100 students comprising 60 boys and 40 girls in a class. In a completely randomized design (CRD), we will randomly assign 50 students each to E or C. However, if we wish to block out the effects of gender, we randomly divide the boys into two groups of 30 each, and similarly, divide the girls into two groups of 20 each (Table 7.5).

Observe that subjects within each E or C are homogeneous, that is, they comprise all boys or all girls. The gender effect has been blocked out because we only compare the results within each block.

Latin Square Design

A Latin square design may be used if there is a variable of interest and two extraneous variables. For example, a firm wishes to test four types of machines (A, B, C, and D), as shown in Table 7.6. The performance of a

Table 7.5 Randomized block design.

Total	Gender split	Samples	Groups
100	60	30	*E*
		30	*C*
	40	20	*E*
		20	*C*

Table 7.6 Latin square design.

	I	II	III	IV
a	*A*	*B*	*C*	*D*
b	*B*	*C*	*D*	*A*
c	*C*	*D*	*A*	*B*
d	*D*	*A*	*B*	*C*

machine depends on two extraneous variables, namely, the operators (*a*, *b*, *c*, and *d*) and factory conditions (I, II, III, and IV). Each treatment (machine type) appears once in each row or column. Each operator will operate the different machines under varying factory conditions.

The dependent variable is the output from each type of machine (*Y*), that is,

$$Y = f(M, O, F).$$

Here *M* is the type of machine, *O* is the assigned operator, and *F* is the selected factory condition.

Reliability and Validity

Reliability in experiments refers to consistency in measuring variables, such as care in the use of calibrated instruments, and preventing cheating.

On validity, we have seen that experiments are not causal studies. They are used to detect certain effects from changes in the independent

variables but are by themselves not explanatory mechanisms. Validity issues also arise from possible experimenter and participant biases. Experimenter bias include

- omitting or failing to control for extraneous variables;
- using small sample sizes;
- using non-comparable groups;
- expecting certain results;
- change of experimenters; and
- sensitizing subjects.

The groups may be too small, in which case the remedy is to use a larger sample. To obtain comparable groups, we may use random allocation, repeated measures design, *matching*, or *discontinuity* design. Depending on the type of experiment, participants in different groups may be matched on a few significant characteristics such as age, race, health status, height, weight, income, and gender. A simple example is a tug of war game where participants are matched on weight as a proxy for strength so that the two sides are comparable.

In a discontinuity design, subjects are selected near a threshold or discontinuity. For example, to study the impact of scholarships on subsequent academic performance at university, we may select students just below and marginally above the cut-off score for scholarships. The assumption is that these two groups near the threshold are academically comparable.

Experimenters may be biased towards what they are looking for, such as whether the treatment is likely to work. A possible solution to this problem is to use a *double-blind experiment*, that is, experimenters and participants do not know whether the results come from *E* or *C*. Editors of scientific journals often use blind peer reviews, that is, the reviewer does not know the author(s) of the manuscript so that she can review it objectively. In a double-blind review, the editor does not know the reviewers. This ensures that she does not reject manuscripts just because the reviewer, who is a leading researcher in the field, rejects the paper.

A change of experimenters may bias the results because of differing skills, experience, and expectations. Finally, the presence of experimenters

may sensitize subjects. The *Hawthorne effect* refers to subjects behaving differently because they know they are under experimental observation (Mayo, 1933).

Participant bias in experiments include

- self-selection;
- maturity;
- testing effect;
- influence of external events;
- predisposition; and
- interaction.

Self-selection occurs if those who volunteer for the experiment are different from the target population. For example, those who volunteer for employment training programs may be more progressive and are likely to improve their performance after training. Similarly, graduates from good universities tend to earn higher salaries because the better students self-select to study there. If monetary incentives are given to participants, it may also attract those who are there for the money.

Subjects may mature and perform better at the post-test if the time period between tests is too long. Or they may become better after a pre-test because of the experience. Sometimes, external events may also bias the results. For example, a teaching experiment may be disrupted by home-based learning because of COVID-19.

Participants may be predisposed to expect the treatment to be effective. In a *blind experiment*, the subjects in each group do not know whether they have been given the treatment or a placebo (an inactive substance) because both pills look identical. This will reduce the possibility of subjects exaggerating the main or side effects if they know that they have been given the active drug.

If subjects in *E* and *C* interact, they may bias the outcome. For example, if students in both groups exchange notes or tutor each other in a teaching experiment, it will affect the pure treatment effect.

Lastly, in terms of external validity, can the results of the experiment be generalized? In general, it may be harder to generalize from social experiments because of different institutional and other contexts. For

physical experiments, generalization is possible if the above biases are small.

References

Abelson, R., Frey, K., and Gregg, A. (2003) *Experiments with people: Revelations from social psychology*. New York: Psychology Press.

Box, G., Hunter, W., and Hunter, J. (2005) *Statistics for experimenters*. New York: Wiley.

Campbell, D. and Stanley, J. (1963) *Experimental and quasi-experimental designs for researchers*. Belmont, California: Wadsworth.

Dean, A. and Voss, D. (1999) *Design and analysis of experiments*. New York: Springer.

Dunning, T. (2012) *Natural experiments in the social sciences*. London: Cambridge University Press.

Mayo, E. (1933) *The human problems of an industrial civilization*. New York: MacMillan.

Montgomery, D. (2009) *Design and analysis of experiments*. New York: Wiley.

Niederle, M. and Vesterlund, L. (2010) Explaining the gender gap in mathematics test scores: The role of competition. *Journal of Economic Perspectives*, **24**(2), 129–144.

Orr, L. (1998) *Social experiments*. London: Sage.

Reichardt, C. (2019) *Quasi-experimentation*. New York: Guilford Press.

This page intentionally left blank

CHAPTER 8

Research Design V: Regression

Features of Regression Design

A regression design examines the influence of independent variables (X_2, …, X_k) on a dependent variable (Y), that is,

$$Y = f(X_2, \ldots, X_k).$$

Here, $f(.)$ denotes a function. For brevity, we write it as $Y = f(X)$ if there is a single independent variable, and $Y = f(\mathbf{x})$ if there are many variables. The independent variables are *exogenous*, or determined outside the model. The dependent variable is *endogenous*, that is, determined by the model or, specifically, by $f(\mathbf{x})$. Hence, regression differs from correlation, where the relation is symmetric without any specification on whether Y depends on X or vice versa. For example, a person's height correlates with her weight; they do not cause each other.

If $f(.)$ is linear, the *population regression model* is

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon. \tag{8.1}$$

Here β_1 is the constant or intercept term, the other βs are parameters or slope coefficients, and ε is the error term.

Taking the expectation,

$$E[Y] = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

because $E[\varepsilon] = 0$ by assumption. $E[.]$ gives the mean value; for example. $E[Y]$ is the mean of Y. The partial derivative of $E[Y]$ with respect to X_i is

$$\frac{\partial E[Y]}{\partial X_i} = \beta_i .$$

It is clearer if we write it as

$$\Delta E[Y] = \beta_i \Delta X_i.$$

That is, β_i represents the effect of a unit change in X_i on $E[Y]$, holding other variables constant. Hence, regression is a form of *statistical control*, which accounts for its popularity in the social sciences where, as discussed in Chapter 7, experimental control is difficult, if not impossible.

A regression model is not just a correlation exercise. There must be good reasons why certain variables are included in the model. In many cases, we know that a few independent variables (e.g. X_2, ..., X_5) affect Y. They are what former US Secretary of Defense Donald Rumsfeld called the "known knowns" at a news briefing on 12 February 2002. Then there are variables (e.g. X_6, ..., X_9) that affect Y but are omitted from the model because we do not have data, the so-called "known unknowns" or things we know we do not know. Finally, there are "unknowns unknowns" (e.g. X_{10}, ..., X_{13}), the things we do not know we do not know. The model is

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_5 X_5 + \varepsilon.$$

The error term contains the "known unknowns" and the "unknown unknowns." It is a measure of our ignorance. The novice focuses on estimating the betas, whereas the expert worries over the error term.

Sampling

In Equation (8.1), there are k parameters, and so the number of data points should be much higher than k. To see this, consider the *population* regression model

$$Y = \alpha + \beta X + \varepsilon.$$

The parameters are unknown and must be estimated from a sample. The *sample* regression model is

$$Y = a + bX + e \qquad \text{or} \qquad Y_i = a + bX_i + e_i$$

where a is the estimated intercept, b is the estimated slope coefficient, and e is the estimated error or *residual*. The regression *line*, estimated using ordinary least squares (OLS), is $Y = a + bX$ (see Fig. 8.1). Clearly,

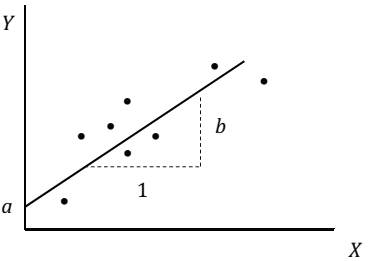

Fig. 8.1 Linear regression line.

different samples will produce different regression lines and hence different values of a and b. These different values are called the *sampling distributions* of a and b.

To estimate the sample regression line properly, a rule of thumb is at least 10 data points for each unknown. In this case, there are 2 unknowns (α and β) and we should use at least 20 data points.

It is also necessary to spread the points in the X-direction, as shown in Fig. 8.1. This ensures that the estimated regression line will not be sensitive to any outlying data point. If we use only the cluster of five points near the middle of the graph, the estimated regression line will be nearly horizontal, which incorrectly reflects the upward-sloping linear relation between Y and X.

Finally, the sample should be representative. The estimated regression line is only as good as our sample.

Examples

Hedonic price model

The price (P) of a multiple-attribute asset such as a house depends on

- the property attributes, such as plot size, design, structure, age, and orientation;
- the neighborhood, such as the extent of crime, school quality, and accessibility to employment, shopping, and recreational centers; and
- the environment, such as noise, air quality, and water quality.

A potential buyer will consider these attributes jointly, and the house price function is

Table 8.1 Data for hedonic price model.

House	Transacted price (P, in $)	Land area (A, in m^2)	Age (G, in years)	…	Noise (N, rating scale)
1	800,000	150	5	…	2
2	900,000	160	6	…	1
3	1,100,000	180	8	…	3
…	…	…	…	…	…
150	1,600,000	210	2	…	3

$$P = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon. \tag{8.2}$$

The βs are implicit prices that are not directly observable (Rosen, 1974). For each property, we can only observe the transacted house price (P) and about a dozen attributes (Xs). The implicit prices are estimated from a large sample of transacted prices (Table 8.1). Noise is measured on a 1 to 5 rating scale, from quiet to noisy.

Uses of the hedonic price model include

- mass appraisal for local property tax purposes (Benjamin et al., 2004);
- developer's selection of property attributes that are highly valued by house buyers, such as tenure, land area, house design and conditions, security, amenities, and proximity to good primary schools (Ridker and Henning, 1967); and
- valuation of environmental quality such as noise or air pollution in cost-benefit analysis (Colony, 1967; Nourse, 1967; Nelson, 1980; Tan, 2021).

The hedonic price model has weaknesses. We will discuss these issues of reliability and validity in the next section.

Learning curve

The learning curve shows how the cost or time per unit (Y) decreases as the cumulative number of units produced (X) increases (Fig. 8.2), as Wright (1936) discovered in the production of airplanes. The cumulative output is a proxy for the learning experience of the team.

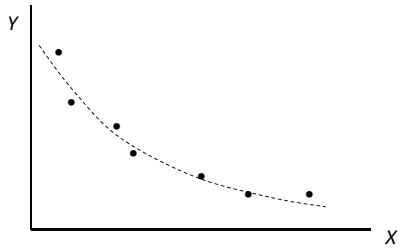

Fig. 8.2 The learning curve.

The sample regression model is

$$Y = aX^b e,$$

where a is the intercept, b is the learning coefficient, and e is the residual. To estimate b from the sample data, we take logs on both sides so that

$$\log Y = \log a + b\log X + \log e.$$

Letting $y = \log Y$, $c = \log a$, $x = \log X$, and $u = \log e$, we have a linear regression model

$$y = c + bx + u.$$

Since Wright's study, learning curves have been estimated for many industries, particularly in manufacturing. They have also been applied beyond production to include quality, inventory, and productivity (Jaber, 2011).

Reliability and Validity

On reliability, measurement errors can cause problems. The measurement error in the dependent variable is less serious because it will be incorporated into the error term. That is, if u is the measurement error in Y,

$$Y + u = \alpha + \beta X + \varepsilon.$$

Thus,

$$Y = \alpha + \beta X + (\varepsilon - u).$$

Letting $v = \varepsilon - u$, we obtain a similar regression equation. We can use OLS to estimate it if v is a random error term.

If X is measured with an error, then

$$X_m = X + w,$$

where X_m is the measured variable, X is the true value, and w represents measurement error. Then,

$$Y = \alpha + \beta(X_m - w) + \varepsilon = \alpha + \beta X_m + (\varepsilon - \beta w).$$

Hence, w is correlated with X_m because $X_m = X + w$, and it is also correlated with $(\varepsilon - \beta w)$. Thus, X_m and $(\varepsilon - \beta w)$ are correlated, which violates the OLS assumption that $Cov(X, \varepsilon) = 0$. We need to use other estimation techniques such as instrumental variables (IV) or 2-stage least squares (2SLS).

Another measurement issue is the use of crude dummy variables or rating scales in measuring housing attributes. A dummy variable takes the value of 0 or 1, such as whether a house is near a park. What is considered "near" may be subjective.

On validity, the local housing market may not be stable (King, 2010). If price fluctuations are large, it is difficult to estimate implicit prices with precision. Hedonic price models also tend to suffer from the following information problems:

- insufficient variables on housing attributes, resulting in omitted variable bias (Abbott and Klaiber, 2011);
- small sample size because of limited sales data;
- imperfect information about the environmental characteristics of a property, such as soil or water contamination;
- buyers have limited capacity to evaluate the possible impacts of externalities on house prices (Bartke, 2011); and
- buyers cannot find houses with a bundle of attributes that suit them, such as a large, new house with a garden in the inner city area.

If the relations between Y and the Xs are nonlinear, the estimated regression model may also contain large errors (Halvorsen and Pollackowski, 1981). Finally, the hedonic model has limited ability to generalize beyond the local housing market because of differing

institutions and rules. The preferences of consumers for housing features will also differ across space.

References

Abbott, J. and Klaiber, H. (2011) An embarrassment of riches: Confronting omitted variable bias and multiscale capitalization in hedonic price models. *Review of Economics and Statistics*, **93**(4), 1131–1142.

Bartke, S. (2011) Valuation of market uncertainties for contaminated land. *International Journal of Strategic Property Management*, **15**(4), 356–378.

Benjamin, J., Guttery, R., and Sirmans, C. (2004) Mass appraisal: An introduction to multiple regression analysis for real estate valuation. *Journal of Real Estate Practice and Education*, **7**(1), 65–78.

Colony, D. (1967) *Expressway traffic noise and residential properties*. Toledo, Ohio: University of Toledo Press.

Halvorsen, R. and Pollackowski, H. (1981) Choice of functional form for hedonic price equations. *Urban Economics*, **10**, 37–49.

Jaber, M. (Ed.) (2011) *Learning curves*. London: CRC Press.

King, P. (2010) *Housing boom and bust*. London: Routledge.

Nelson, J. (1980) Airports and property values. *Journal of Transport Economics and Policy*, **14**, 37–52.

Nourse, H. (1967) The effect of air pollution on house values. *Land Economics*, **43**, 181–189.

Ridker, G. and Henning, J. (1967) The determinants of residential property values with special reference to air pollution. *Review of Economics and Statistics*, **49**(2), 246–257.

Rosen, S. (1974) Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, **82**, 34–55.

Tan, W. (2021) *Managing infrastructure projects*. Singapore: World Scientific.

Wright, T. (1936) Factors affecting the cost of airplanes. *Journal of the Aeronautical Sciences*, **3**(4), 122–128.

This page intentionally left blank

CHAPTER 9

Methods of Data Collection

Data Collection Methods

After determining the research design, the next step in the research process is to collect data from

- observations;
- interviews;
- questionnaires;
- standardized tests;
- physical instruments;
- simulation;
- documents;
- crowd-sourcing;
- sensors; and
- websites.

Observations, interviews, and documents are widely used in qualitative research to collect contextual details and probe more deeply into a respondent's perspectives, interpretations, and experiences. Quantitative researchers tend to use questionnaires, standardized tests, physical instruments, simulation, crowd-sourcing, sensors, and websites.

Many research studies use multiple means to collect data. For example, we may use standardized tests to assess the performance of students, followed by qualitative data on why students differ substantially in their performance.

Scales

Scales are used to categorize, rank, and assess magnitudes (Table 9.1). There are two types of variables, namely,

Table 9.1 Types of variables and scales.

Type of variable	Scale	Use	Example
Discrete	Nominal	Classification	Gender
	Ordinal	Ranking	Rating
Continuous	Interval	Distance	2005–2010
	Ratio	Ratio	Weight

- discrete variables that take integer values, such as 20 girls; and
- continuous variables that take any real number, such as 2.3 kg.

A nominal scale counts group membership, such as the number of boys and girls in a class or the mode of transport a person takes to work. It categorizes the data using labels or numbers, such as 1 for Walk, 2 for Cycle, 3 for Public transport, and 4 for Private transport. The numbers are just labels to identify the mode of transport, and the order or magnitude does not matter.

An ordinal scale provides ranks or ratings. A rank places objects in some order, such as a teacher ranking her students by test scores or the ranking of universities based on aggregate scores on peer perceptions of research, teaching, facilities, student quality, graduate employment, and so on.

A rating is a personal preference or opinion on something, such as the cleanliness of a school, on a suitable scale. For rating scales, the number of points (e.g. 5 or 7) depends on the desired sensitivity of the responses. The difference between using odd and even number of points lies in whether "Neutral" is an acceptable answer. If the intervals are roughly equal, it is possible to find the mean rating using a *summative rating scale* comprising individual items that measure a construct. An example is the Likert scale where the response format is 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree. The individual scores from various items are then summed and averaged.

An interval scale consists of equal intervals that measure the relative distances (differences) between points. Examples of such a scale include IQ scores, temperature, or time. Ratios are meaningless, that is, a person with an IQ score of 160 is not twice as intelligent as one with a score of

80. Similarly, the period 2005 – 2010 makes sense in calendar time, but not the ratio 2005/2010.

In a ratio scale, the intervals are regular and ratios are meaningful. For example, 10 kg is twice as heavy as 5 kg.

In general, precision depends on purpose and cost. For instance, we should measure the room temperature using a thermometer if this level of accuracy is important, rather than merely "hot" or "cold." The scale will also affect the statistical tests that may be used. For instance, in the nominal scale, the mean and standard deviation are meaningless. If there are 20 boys and 10 girls in a class, the average (15) does not make sense.

Observations

We may collect data by observing

- physical arrangements;
- people and their behavioral patterns;
- what they are trying to accomplish;
- actions;
- activities;
- use of equipment or facilities;
- flows of people or traffic;
- processes;
- occurrences, such as an accident;
- sequence of events; and
- feelings or emotions.

Such observations may be independent or participatory, that is, whether the researcher as an observer participates in the activities to "soak and poke" beyond appearances (Fenno, 1978). In contrived observations, the observer may purposely lodge a "complaint" to see the reaction.

It is common to use a checklist to guide the observations. Notes and sketches should be made as soon as possible and, if possible, supplemented with recording media. Be sure to obtain permission to record or film conversations and activities.

To reduce observer bias as well as improve coverage, triangulate by using two or more observers and compare notes.

Interviews

Most qualitative researchers use interviews to ask probing questions through conversations and open discussions.

In an unstructured interview, the interviewer does not wish to impose any prior framework. The questions cover awareness of the issue (e.g. onset of a major illness), the adaptive responses, damage control, the consequences, and so on. The aim is to understand the respondent's perspective on the issues.

In a semi-structured interview, the interviewer has a list of questions based on the research framework but the answers can be open-ended to enable the researcher to probe more deeply into an issue.

A structured interview is highly standardized. Respondents answer the same set of questions in the same order. These structured questions contain a limited set of possible answers to ensure that aggregation and comparisons are possible.

A focus group is a group interview comprising about five to ten respondents. The researcher facilitates and moderates the collective discussion to explore ideas, share views, or make recommendations on an issue (Hall, 2020). Its effectiveness will depend on the composition and quality of the participants, the skills of the moderator, the ground rules, and the issues raised.

The Delphi method (Adler and Ziglio, 1996) uses an expert panel to generate business forecasts or scenarios over several rounds. The experts use the results from each earlier round to revise their forecasts. This method assumes that group forecasts are more reliable than individual ones; that is, they rely on the wisdom of the panel.

Researchers may use public forums and hearings to gather ideas from a wider range of stakeholders. Such meetings should include all sections of the community. They should also be participatory and not be dominated by certain stakeholders. Finally, do not assume that all stakeholders will be present for a public meeting.

Researchers may carry out the interviews face to face, over the telephone, or through video conferencing. Start by putting the interviewees at ease in physical or online interviews so that they are ready to provide information. Other considerations in interviews include authorizations,

informant access to key respondents, timings, venues, seating arrangements, use of visual aids, use of recorders if permission is given, note-taking, and training of interviewers (Seidman, 2019).

Questionnaires

A questionnaire is a list of questions intended to collect information by asking people directly. It is often used in conjunction with a physical, telephone, or online interview. However, a postal, email or web-based questionnaire does not involve an interview.

Most questionnaires contain highly structured questions together with a limited set of answers. They usually contain factual questions and ratings, and occasionally, opinions and reasons. An example is given below.

ABC Organization
Address
Contact details

Date

Dear Sir/Madam,

Study of Tenant Satisfaction

I am conducting a survey on tenant satisfaction to serve you better. Your response will be beneficial in helping us understand the areas we have done well and the areas that need improvement. It would be appreciated if you could convey your views by answering the questions below.

Thank you for sparing your valuable time.

Sincerely,
Name and position

Note: If a covering letter is not used, a brief introduction outlining the purpose of the survey should be given in the questionnaire.

Please fill in the blanks or tick the appropriate box:

A Tenant information

1 Mall: _____ Unit number: _____
2 Name (optional): _____ E-mail address: _____
Contact number: _____ Job title: _____

B Tenant satisfaction

Please circle your *level of satisfaction* (1 = Low; 7 = High) for each listed item.

1 Lease management

a. Lease period	1	2	3	4	5	6	7
b. Rent collection	1	2	3	4	5	6	7
c. Etc							

2 Marketing

a. Marketing expenditure	1	2	3	4	5	6	7
b. Marketing activities	1	2	3	4	5	6	7
c. Etc							

3 Layout of shopping center

a. Size of shops	1	2	3	4	5	6	7
b. Shopper circulation	1	2	3	4	5	6	7
c. Etc							

4 Etc

Do you have any suggestions or comments?

.....
.....
.....
.....
.....

End of Questionnaire. Thank you once again.

A pre-test using a small sample of likely respondents is often conducted to obtain feedback on the length, structure, sequencing, and

content. The length should not exceed five pages, and the structure should avoid disruptive jumps from one section to another. Check the content for construct, internal, and external validity. Recall that construct validity is concerned with whether we are measuring the right thing. Internal validity is about the causal mechanism that links causes to effects. Finally, external validity is about whether the results can be generalized to other contexts.

To improve the questionnaire (Bradburn et al., 2004; Brace, 2013; Harris, 2014),

- use simple words; if a technical word is necessary, provide a short explanation;
- use fixed-alternative questions that are theoretically sound and not artificially imposed, and state clearly if multiple answers to a question are possible;
- use frequency counts instead of vague words such as "often" or "seldom";
- use open-ended questions where many answers are possible;
- avoid leading questions, for example, "Should *unproductive* speculators be taxed?";
- avoid double-barreled questions, for example, "Is your work easy *and* challenging?" poses a dilemma if it is easy but not challenging;
- state clearly the units of measurement, for example, gross or net monthly income;
- ask in units that people remember, for example, monthly take-home pay rather than annual income;
- use ranges for sensitive issues, for example, income;
- de-sensitize phrases, for example, "Drivers sometimes park their car illegally; have you done this before?" is more acceptable than "Have you ever parked your car illegally?";
- avoid hypothetical questions that are poor predictors, for example, "Do you intend to buy this product?";
- avoid questions on competency, for example, "How do you rate yourself as a computer user?" is prone to over-rating;
- avoid social desirability bias, for example, "Do you support this project to help unfortunate children?"; and
- be aware that the way questions are worded or asked may not reflect how respondents view them.

Recall from Chapter 5 that it is advisable to conduct reliability tests such as the test-retest, inter-rater, or Cronbach's alpha.

Standardized Tests

Standardized tests are commonly used in psychological and educational research to collect data, such as in an experiment to test mental ability or the effectiveness of a teaching method. Such tests are also used for recruitment. Performance may be compared among participants or against some criteria.

The challenges in designing standardized tests include ensuring that the content is appropriate, there is sufficient time to complete the test, and the test is not too easy or difficult. Standardization makes such tests less suitable for answering complex questions where there are no simple solutions.

In education testing, it is possible to teach to the test, that is, by helping students learn to ace the test rather than understand the fundamental principles. What is not tested will be devalued, and learning can become superficial. Further, such tests may not reflect the quality of instruction or learning and are harmful to the self-confidence and morale of low-performing students (Kohn, 2000; Harris et al., 2011).

Physical Instruments

Physical instruments are widely used in the natural sciences to measure velocity, acceleration, temperature, distance, mass, pressure, weight, volume, and so on. The decision will depend on factors such as cost, availability, accuracy, precision, ease of use, calibration requirements, and reliability.

Simulation

A simulation model is a mathematical representation of a real-world system. By experimenting with the input variables and parameters in the model, it is possible to compare the system performance and make design changes. For example, we may simulate the energy performance of a

building using the design drawings and use the results to improve the design before construction. Simulation models are widely used to study ecological processes (Acevedo, 2019), business processes (Laguna and Marklund, 2005), and the behavior of economic systems (Fontana, 2006).

Digital twins are virtual models or algorithms that use sensors to collect real-time operational data on an asset or system. These data and operating environmental conditions data are then fed into the virtual twin or replica to check, understand, optimize, or predict its performance before applying adjustments to the parameters to optimize asset or system performance. For instance, Tesla builds a digital twin of each car it sells. The sensors send operational data for analysis to determine if the car works properly. It then fixes the problem by sending software updates. In this way, Tesla knows how the cars work under different environmental conditions (Coors-Blankenship, 2020).

As an example of simulation, let

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

where Y_i is the monthly consumption expenditure of the ith household, X_i is the monthly household income, α is the intercept, β is the slope coefficient, and ε_i is the error term. In regression analysis, β is estimated using a sample of households. Alternatively, to see how β varies across different samples, we may conduct a Monte Caro simulation as follows:

- fix the values of α and β, for example, 500 and 0.4, respectively;
- select, say, 100 values of X_i;
- generate 100 values of ε_i using a random number generator;
- compute 100 values of Y_i using the regression equation; and
- use the method of least squares to estimate β.

The process is repeated many times (e.g. 3,000), and the different estimates of β are plotted as a histogram.

In agent-based models, we simulate how agents interact with their environment (Railsback and Grimm, 2011). For example, Schelling (1978) used agent-based modeling to show how households tend to segregate into racial clusters. He first divided the "urban area" into a grid of cells, one for each household. From an initial random allocation, a

household in any cell will survey its eight neighbors to check if they are racially the same or different. If the proportion of different households exceeds a certain tolerance (for example, 0.3), the household is "unhappy" and will relocate to another vacant cell. Schelling showed that racial clusters appear after many rounds of shifting.

Although simulation can handle models with many interacting variables, it may degenerate into black box modeling where the results are less certain. The user may not be able to check the computations.

Review of Documents

We may also collect data from published documents such as

- internal organizational accounting records, sales data, maps, commercial reports, and other miscellaneous records;
- academic journals, directories, magazines, newspapers, commercial reports, reference books, dissertations, theses, encyclopedias, websites, and books; and
- private diaries, letters, speeches, memos, photographs, and films.

Permission is required to access these sources. Accuracy is important; if possible, consult the original source that often contains the actual words, intent, methodological details, warnings, and standard errors that are not reported by subsequent users. It is also necessary to verify the authenticity and credibility of the source. Even "official" sources may suppress statistics on worksite accidents or use different methodologies or words to make the numbers or organization look good.

Crowdsourced Data

Crowdsourced data refers to data collected from public users such as from smartphones through mobile applications or sensors. For example, a local municipality may collect data on air quality, temperature, flooding, and the condition of facilities. Similarly, the transit operator may collect data from riders through a mobile application on operational issues. With digital payments, it is also possible to collect household consumption patterns as well as food and other prices.

Sampling is important in crowdsourced data. To tap the wisdom of crowds (Surowiecki, 2004), it has to be a large and unbiased sample if representativeness is desired. However, in cases such as the reporting of operational issues, the goal is to rely on the eyes and ears of users rather than obtaining an unbiased sample.

Crowdsourcing has also been used to gather opinions, generate funds or ideas, test software, and so on.

Sensors

The Internet of Things (IoT) are devices connected to the Internet. They consist of not just computers but include smartphones, air-conditioners, water heaters, security alarms, and smoke detectors. Sensors in these devices are connected through Wi-Fi, Bluetooth, or other interfaces to a gateway for initial processing or aggregation before the transfer of data to the cloud for storage and analysis. This makes it possible to remotely control the operation of the device, such as switching off the home air-conditioner from work from a mobile application on the smartphone.

IoT has been applied in different fields (Kranz, 2017; Veneri and Capasso, 2018), and will continue to find new applications.

Website Data

Many websites provide textual and other data on narratives, prices, logistics (e.g. shipping information), and product reviews (Costa and Condie, 2018). These quantitative and qualitative data may be mined for analysis. However, they are not in standard formats, which makes the manual "copy and paste" process rather tedious. There are commercial software that automate the process by "scraping" websites for data using machine learning algorithms (see Chapter 17).

References

Adler, M. and Ziglio, E. (Eds.) (1996) *Gazing into the oracle*. Boston: Addison-Wesley.

Acevedo, M. (2019) *Simulation of ecological and environmental models*. London: CRC Press.

Brace, I. (2013) *Questionnaire design*. London: Kogan Page.

Bradburn, N., Sudman, S., and Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design*. New York: Jossey-Bass.

Coors-Blankenship, J. (2020) Taking digital twins for a test drive with Tesla, Apple. *Industry Week*, 29 April.

Costa, C. and Condie, J. (Eds.) (2018) *Doing research in and on the digital*. London: Routledge.

Fenno, R. (1978) *Home style*. Boston: Little, Brown, and Co.

Fontana, M. (2006) Simulation in economics: Evidence on diffusion and communication. *Journal of Artificial Societies and Social Simulation*, **9**, 1–15.

Hall, J. (2020) *Focus groups*. Gorham: Myers Education Press.

Harris, D. (2014) *The complete guide to writing questionnaires*. New York: I & M Press.

Harris, P., Smith, B., and Harris, J. (2011) *The myth of standardized tests*. Washington DC: Rowman and Littlefield.

Kohn, A. (2000) *The case against standardized testing*. Portsmouth: Heinemann.

Kranz, M. (2017) *Building the Internet of Things*. New York: Wiley.

Laguna, M. and Marklund, J. (2005) *Business process modeling, simulation, and design*. New Jersey: Prentice Hall.

Railsback, S. and Grimm, V. (2011) *Agent-based and individual-based modeling: A practical introduction*. New Jersey: Princeton University Press.

Schelling, T. (1978) *Micromotives and macrobehavior*. New York: W. W. Norton.

Seidman, I. (2019) *Interviewing as qualitative research*. New York: Teachers College Press.

Surowiecki, J. (2004) *The wisdom of crowds*. New York: Anchor.

Veneri, G. and Capasso, A. (2018) *Hands-on industrial Internet of Things*. Birmingham: Packt Publishing.

CHAPTER 10

Collection and Processing of Data

Introduction

After developing the research question(s), hypothesis (or framework), research design, and methods of data collection, the next step in the research process is the actual collection and processing of data.

The issues during data collection include

- access to respondents;
- management of research assistants;
- care and use of equipment;
- protocol for review of documents;
- note-taking;
- establishing the chain of evidence;
- enhancing reliability; and
- tracking of progress.

These issues are similar for all research designs, with minor variations between interpretive and causal studies. These variations will be highlighted below.

Access

Researchers often need to gain access to individuals and organizations to gather information. It opens the door to field research. Yet, this process of gaining access is not well theorized or documented. The literature is sparse (Feldman et al., 2003), and a researcher may spend considerable time and effort trying to gain access. Not all organizations value academic research or welcome outsiders to probe into their activities.

The gatekeeper controls access to the organization. For example, the gatekeeper of a school is the principal, who will decide whether you can observe and interview the school administrators, teachers, and students. You will need to address her concerns to gain access, such as

- whether you are trustworthy;
- the purpose of the study;
- whether it is a sensitive topic;
- why the school has been selected;
- what are you going to do with the results;
- will it disrupt classes;
- how will other stakeholders (e.g. parents and alumni) react; and
- how the school can benefit from participating.

As an inducement, you may want to share your research findings with the school. For example, you may want to share whether a new technique for teaching mathematics is effective.

Other ways of gaining access include

- formal introductions by someone influential such as the chairperson of the school advisory council;
- informal introductions and networks, such as through a teacher or a friend; and
- past links to the school, such as if you are an alumnus.

In a commercial setting, gaining access to workers presents a different set of problems. One issue is the lack of trust between the researcher and workers. For some workers, the researcher may be a "spy" from management. Further, there may be little incentive for workers to share sensitive information or views with outsiders. The firm will also be concerned about the loss of productive time.

In the case of questionnaire surveys, access to potential respondents is often through an impersonal mailing list. The researcher has to consider the cost (for example, postal, email or web-based), geographical spread, timeliness, ease of obtaining permission, and the likely response rate. The response rate depends on issues such as the topic, quality of questionnaire, currency of mailing list, ease of reply, possibility of rewards, and so on.

All forms of access have time limits. They cannot go on forever. For example, if you are interviewing a senior manager, be clear upfront on the expected number of interviews and the period. Do not assume that subsequent access to a respondent to clear up some issues is automatic.

Research Assistants

The management of research assistants includes issues such as planning and budgeting, hiring, training, supervising, scheduling the fieldwork, logistics, safety and health, quality control, and mentoring (Stouthamer-Loeber and van Kammen, 1995).

In the selection of interviewers, honesty, humility, possession of the relevant skills, a good voice, a pleasant personality, good attitude, energetic, and a liking for field conditions are pre-requisites. Good command of certain languages is mandatory if not all respondents (for example, tourists) speak the same language. Interviews can be a little frustrating at times, and people with short fuses are unsuitable. It is uncertain whether biases may occur between paid and unpaid field staff, as well as between internal and external field assistants. Motivation and effort are also affected by perceptions of the adequacy of payment.

Training should be provided. Someone familiar with the entire research, such as the principal investigator, should conduct the training. The briefing includes the nature and purpose of the study as well as data collection procedures. They should be trained on specific procedures to be followed when contingencies arise, such as (Lavrakas, 1993)

- the respondent is not at home;
- no one is answering the phone;
- the call is directed to an answering machine;
- the call is answered by a non-resident;
- the line has been disconnected;
- the selected respondent is unable to answer because of physical disability;
- there is a language barrier;
- the interview is incomplete; and
- the selected respondent refuses to be interviewed.

A demonstration interview, followed by a trial interview by each field staff member, is recommended. Interviewers may use probes to prompt respondents for their "best guess" answers but they must be mindful of possible biases in the responses. Training should also cover ethical issues and expectations of integrity, such as truthful reporting, processing, and analyses of field data, data protection, avoiding plagiarism, and not infringing copyright.

The administration of field data collection includes scheduling, workload distribution, logistics, safety, and health. Finally, use the opportunity to supervise and mentor research assistants. They are not just "a pair of hands" for collecting data. Mentored research assistants will be more productive and find the experience more satisfying.

Equipment

Check that equipment such as tape recorders, cameras, and other measuring devices are calibrated and in good working order. Field equipment should be looked after to prevent damage and, for safety reasons, ensure only qualified people handle the equipment. Leaving them unattended invites theft and gives participants the impression of professional irresponsibility.

For interviews, notebooks, instruction manuals, survey forms, and maps should be properly handled.

Documents

Most research designs require the review of physical and digital documents for qualitative and quantitative data. There should be a protocol or checklist for the review of such documents to extract the data meaningfully and effectively.

The process begins with an assessment of the types of information required, and hence the types of documents to review. Some of the information may have been published elsewhere, or there are alternate sources of information. For example, information on construction statistics may be published by different government agencies. As far as possible, the researcher should triangulate the data from these sources to minimize errors.

Where possible, use original sources when collecting documentary data to minimize transcription and interpretation errors. The original data may have been reorganized by subsequent users and important footnotes on how these data have been collected may have been omitted.

Note-taking

In many research designs, there will be large quantities of notes that need to be managed as a database.

Note-taking requires skill. Observers must be trained on what to observe, how to fill in observation checklists, forms, and logbooks. Observations should be unobtrusive because subjects may react differently when an observer is busy tracking every step. If the situation is fluid, leave nothing to memory by taking notes as soon as possible. Use color pens to insert footnotes, personal opinions, and follow-up action. It is a good idea to take notes in stages starting with sketchy notes to keep abreast of what is happening and filling in the details when spare time is available.

As an observer, you may not be sure what is important. Hence, you should jot down the events whenever possible. The perspective may change and what is previously thought to be unimportant may become important.

Tracking of Progress

For interpretive studies, the tracking of research progress is less of a problem once access has been secured and respondents continue to co-operate. This is not so for survey research, where the response rate may be uncertain.

The tracking of research progress also involves field supervision to ensure that research assistants follow field procedures and workloads are reasonable. It is not unusual for supervisors to verify a small portion of the interviews or questionnaires by re-interviewing or asking respondents whether they have been interviewed.

Supervisors should collect survey forms on a regular basis and edit them in the field for legibility and completeness. Where problems occur,

these issues are communicated to field assistants, and additional training may be necessary. A reminder may be sent, and follow-ups are made soon after the cut-off date.

Data Processing

After data have been collected, the next step is to process them into information suitable for presentation, visualization, and analysis. The processing of qualitative data is part of data analysis rather than a separate process. For quantitative data, processing is necessary to ensure the integrity of the data. Data processing goes through the stages of editing, transforming, coding, and development of themes.

The data is edited for errors, contradictions, inconsistencies, and omissions that have escaped preliminary field editing. Falsified data are usually rejected. If errors or missing data have been spotted, then a decision has to be made to discard the information, re-contact the respondent, use an average value among similar respondents, interpolate from other values, or use subject matter knowledge to guess an appropriate value. Care must be exercised in handling outliers just because they do not fit the theory. They may provide a refutation of the theory. Data obtained from published documents may also require editing. They may contain biases such as arbitrary accounting conventions and failure to take into account quality change, price discounts, and reporting errors. They may also contain misprints.

The next stage is to transform the data through conversion, adjustment, or reconstruction. This may involve converting one currency to another, converting monthly to annual income, deriving net from gross values, or rebasing a time series to a new base year.

The third stage is to code the data by labeling, classifying, and organizing them for subsequent analysis. Coding is generally straightforward for quantitative data, such as in developing the data table or matrix for regression or multivariate analyses using statistical software. The researcher should avoid heaping, where too much data fall into a particular category. If it occurs, reclassification is necessary.

For qualitative data, a group of related codes is used to develop more general themes. The development of the storyline is fundamental in

qualitative studies, so it is not just a matter of labeling and classifying data (Auerbach and Silverstein, 2003). This will be covered in the next chapter.

Spatial data may be organized into thematic layers such as transport networks, utility networks, drainage, buildings, water bodies, terrain, land use, and vegetation. These data must be edited for

- positioning errors, such as incorrect coordinates;
- typological errors, such as the failure of road intersections to meet at a single point; and
- feature errors, such as the incorrect naming of a building.

Spatial data from different providers may also be based on different projection or coordinate systems. Hence, geometric data transformations are required to transform the data into a common coordinate system (Comber and Brunsdon, 2021).

Big Data

Big data are data sets that are so large that the traditional data processing techniques and software outlined in the previous sections are inadequate in dealing with their vastness. The common sources of big data are transactional data and sensor data from smart devices (Internet of Things). These include data on financial transactions, transit trips, utility consumption, manufacturing, smart cities, disaster tracking and management, healthcare, cyber security, and mobile communications.

Big data have the following characteristics, namely,

- big volume that is based on observations of what happened, rather than merely sampled;
- high velocity, where data are obtained dynamically, such as every second or minute; and
- large variety in terms of text, images, audio, and video, which means the data are also less structured.

The storage, maintenance, aggregation, and querying of the database is a major challenge. Specialized software and hardware, called big data

platforms, will have to be used (Grover et al., 2015). The connected servers may play different roles through parallel processing, such as to generate or process different parts of the data.

References

Auerbach, C. and Silverstein, L. (2003) *Qualitative data: An introduction to coding and analysis*. New York: NYU Press.

Comber, L. and Brunsdon, C. (2021) *Geographical data science and spatial data analysis*. Thousand Oaks: Sage.

Feldman, M., Bell, G., and Berger, M. (2003) *Gaining access: A practical and theoretical guide for qualitative researchers*. Walnut Creek, California: Altamira Press.

Grover, M., Malaska, T., Seidman, J., and Shapira, G. (2015) *Hadoop application architectures: Designing real-world big data applications*. Sebastopol: O'Reilly Media.

Lavrakas, P. (1993) *Telephone survey methods*. London: Sage.

Stouthamer-Loeber, M. and van Kammen, W. (1995). *Data collection and management*. London: Sage.

Chapter 11

Qualitative Data Analysis

Types of Qualitative Data

Qualitative data, which are used extensively in interpretive and constructivist frameworks (see Chapter 1), comprise texts and visual images. They may be primary data collected from direct observations and interviews in the form of field notes or secondary data drawn from books, diaries, political scripts, speeches, articles, newspapers, advertisements, paintings, symbols, artifacts, photographs, films, audio and video recordings, websites, open-ended responses to interviews, recollections, and so on.

Reflexivity

To be reflexive is to be aware of the researcher's own biases and preconceptions as well as how your presence may affect the study (Hibbert, 2021). In quantitative studies, the researcher may not be present in the field for long periods, and subjects (or respondents) act independently of the researcher.

There are ways to reduce researcher biases in qualitative research. Start by providing your background in the final report so that readers are aware of where you are coming from. For example, in Peshkin's (1986) study of a fundamentalist Christian school in the US, he provided his background, that of a male Jewish professor of education. It is also useful for respondents to know your background, just as you are also trying to know their backgrounds to understand their points of view. It is difficult to put across a point of view to a total stranger.

Provide the context or local situation. This requires "thick" (holistic) rather than "thin" descriptions (see Chapter 4). The context better allows readers to understand the situation or circumstances for certain behaviors

or perspectives. The core idea is to provide contextual details to allow the reader to understand or interpret cultural meanings and intentions. It is not about merely recording the details; it is about the circumstances that lead to a particular action or perspective. These circumstances may include luck, respondent sensing of opportunities, resource constraints, social pressure, norms, regulations, institutional obstacles, and so on.

Be aware that your presence may affect the study. Participants may react differently for many reasons, such as feeling awkward because of your personality or mannerism, receiving your feedback on their behavior, the way you ask probing questions, their perceptions of your expectation of their behavior, and uncertainty over your motive for conducting the study. Participants want to know which side of the fence you are sitting on even if you declare or come across as neutral.

A researcher's preconceptions, prior theories, experiences, and feelings can have similar effects. You will need to consciously set aside or "bracket out" these thoughts and focus on the participant's point of view and experiences (Moustakas, 1994). If they know of your preconceptions, they may change their responses to suit your needs or expectations and confirm your hypotheses.

It may be helpful to discuss with fellow researchers everyone's personal biases and jot them down in personal journals to create awareness and act as reminders. It is also useful to report such biases in the research report so that readers are aware of these possible slants and can decide for themselves. Group biases cut both ways; researchers working together tend to agree with each other, but so do respondents in group interviews.

There are two sides to a coin. The parties will tell the story in particular ways, such as blaming project delays on other parties or external factors rather than their own mistakes and failings. One party may impose its views of "reality" on another, particularly if the power structure is excessively unbalanced. For example, the school's discipline master may impose her view of what constitutes unacceptable behavior on a student. The student may accept, challenge, or negotiate the master's "definition of the situation" (Thomas, 1923).

If possible, triangulate observations by cross-checking them with fellow observers or other participants. This will help to ensure that observations are factually correct even though opinions may differ.

Finally, words are not neutral; speakers deliberately choose certain words, or discourses, to persuade. As discussed later in the chapter, discourse analysis is a technique of qualitative data analysis.

In summary, reflexivity is about reducing researcher bias through continuous self-questioning and not taking anything for granted — not even language. There is awareness of the presence of self-bias and recognition of multiple views, reasons, and explanations. It is a key principle of qualitative research.

Codes

As discussed in the previous chapter, the raw data require some processing before the actual data analysis. These may include making back-ups of original copies, indexing sources for easy reference and retrieval, and transcribing audio recordings into texts to facilitate analysis. Transcribing does not mean copying verbatim from the recording. The aim is to identify the key points or concepts for further analysis.

The coding of texts consists of assigning words, phrases, symbols, or numbers to each category. The number of codes is a matter of preference, but using too many codes will result in a loss of focus. The researcher then annotates the page margin with informal notes (Table 11.1).

Codes may be preset or emergent ideas that crop up during coding. These new ideas may change in the initial storyline. Codes may also be redefined along the way, such as by merging preset codes in situations where there is no data.

Thematic Analysis

We begin our analyses of qualitative data with thematic analysis. It tries to discover themes or patterns from qualitative data (Braun and Clarke, 2021). The first step is to be familiar with the data, reading it over and over again to identify open codes. These codes are then grouped together to form categories or themes. For example, Codes 4, 14, and 18 in Table 11.1 relate to transport issues, which may be a category (sub-theme) or theme. As the research progresses, the researcher may refine this initial theme.

Table 11.1 Example of a coded interview transcript.

Line			Code	Notes
30	Interviewer	How can the park design be improved?		
31	John	We need more cycling paths for cyclists.	04	Cycling
32	Jim	The children's playground is too close to the canal. I have safety concerns.	08	Playground safety
33	Kay	I have safety concerns as well, but for women. The park is poorly lit, with many hiding places.	08	Safety for women
34	Joe	We need for shade in our tropical climate.	09	Shade
35	Jane	As our population ages, we need to cater to their needs in terms of access and facilities.	12 13	Access Facilities for older people
36	Joe	Speaking of facilities, there should be sufficient parking spaces for private vehicles.	14	Parking
37	Jim	If you look at East Coast Park, public transport connectivity is poor.	18	Public transport
38	John	Parks should be connected to other green spaces, so that I can jog or cycle for longer distances.	20	Park connectivity

A major issue in thematic analysis concerns the development and agreement of codes among a team of researchers. A more flexible approach is to allow codes to evolve, either from the data (emergent codes) or theory (deductive codes). Even here, there is room for disagreements over what constitutes a theme. Researchers may also disagree on the significance of various themes.

Narrative Analysis

A narrative is a storyline (Riessman, 2007; Clandinin, 2013). There are three types of narratives, namely,

- participant's narrative;
- researcher's narrative; and
- document narrative.

Participants tell stories, and these stories, such as life experiences or personal justifications for certain actions (e.g. selling a company), are your data. As noted earlier, they may tell stories to put forward certain points of view.

In contrast, the researcher's narrative is a method of data analysis. The research framework provides the guideposts for the story. The researcher then proceeds to code the data and develop themes to re-tell the individual stories, knowing that the narrative may change as the research proceeds to discover new ideas and evidence. The term "analytic narrative" refers to the attempt to construct general theories out of individual narratives or case studies and then subject the explanation to empirical tests (Bates et al., 1998). The researcher adopts a particular style or language to tell the story. Metaphors and analogies are widely used to persuade readers. For example, McCloskey (1986) argued that four master tropes are widely used in economics, namely metaphor, metonymy, synecdoche, and irony. The production *function* $f(A_t, K, L)$ is a metaphor for making things; the capital (K) and labor (L) inputs are metonymies (symbols for the real thing); technical change A_t is a synecdoche, or a part representing the whole or host of other factors such as education, culture, politics, and institutions; and, finally, the use of irony. Here, economists are fond of saying that perfect markets plainly do not exist in reality; we both know it, but let us proceed with the analysis, nonetheless.

Narratives are also found in documents, images, and videos, such as history books, biographies, or films. Both cinematic films and non-fiction documentaries need to tell stories, as well as earlier physical theatrics and stage plays (Nagler, 1959; Morrison, 1997). Even cinematographers tell their own stories, albeit visually, by manipulating the camera, color, lighting, sound, and images (Brown, 2002).

Stories may have structures or plots, such as a simple start, trigger, mid-story crisis or disequilibrium, turning point (epiphany), and resolution or new equilibrium. These plots contain conflicts or struggles in particular settings or environments, such as the struggles of a start-up entrepreneur. Stories of nation building in historical case studies may be of this genre (White, 1990), and they typically include a "hero" (protagonist) who organized the struggle against colonial or other forms of oppression (antagonist).

Sometimes, the structure is reversed: a happy start, a mid-story crisis, a turning point, the current unhappy state, everything is in a mess, and perhaps the way forward. The struggles may be internal to the actor rather than against an external protagonist. In postmodern narratives, there may not be the traditional (modernist) structure or chronological time, such as the film *Star Wars* (Gibson, 1996).

Discourse Analysis

A discourse uses a set of statements to construct an object (Parker, 1992). For example, one could construct "the poor" as those with certain personal deficiencies, such as health, race, intelligence, character, or culture (Lewis, 1959; Herrnstein and Murray, 1994; Harrison, 2000). These social constructions appeal to reason, statistics, partial evidence, tradition, values, expertise, analogies, metaphors, and anecdotes to persuade. An older term is "ideology" or certain worldviews, such as the German ideology or conservative ideology (Marx and Engels, 1998). However, because of its association with the ideas of Marxist classes, discourse has replaced ideology in the study of the use of language to create an object.

Deconstruction seeks to counter this discourse to reveal the ways they create certain views of "reality" and sustain ways of life through such cognition and power relations (Derrida, 1976; Foucault, 1991). For example, poverty is often constructed as something personal, such as a person who is deficient in certain traits (e.g. intelligence), a characteristic of a particular race or group at the bottom of society, or characteristic of a (colonial) society that is deficient in its culture. For the latter, its beliefs, values, norms, institutions, and practices are "traditional" as opposed to "modern" ones (Levy, 1967). These traditional cultural elements, such as the unwillingness to take risks or exploit new market opportunities, are obstacles that hold them back. In other words, you are poor because of your poor cultural capital. In deconstruction, poverty is not just personal; there are many structural constraints on a person's life chances. The poor kid is likely to start with fewer resources, limited social networks, parents who may not value education highly, and poorly-funded neighborhood schools. Similarly, groups in society struggle over the generation and distribution of resources. Hence, it is not just

a simple matter of cultural deficiency; clearly, politics, institutions, and economics matter (Beckford, 1999).

Content Analysis

Content analysis involves quantifying the contents of physical and online texts by looking for the occurrences of particular words or images. It analyzes contents rather than people. For example, McClelland (1962) tried to measure the achievement motive (AM) by examining children's books and found that countries that stressed successes in life tended to experience higher economic growth compared to those that did not. A more general framework for such a view is the modernization theory of the 1950s and 1960s, where cultural ideas, views, beliefs, practices, and norms in "traditional" societies were thought to have held back economic development (Rostow, 1960; Levy, 1967).

However, critics argued that his economic growth data, from the 1920s to the late 1950s, were not reliable. Several studies using updated data found little correlation (Mazur and Rosa, 1977; Gilleard, 1989). Further, the values in children's stories may reflect those of the authors rather than those taught to children at home or in school. Finally, to focus on a single variable, AM, as the cause of economic growth is naïve. Current accounts of economic growth, as mentioned in the previous section, stress political, institutional, and economic factors (Easterly, 2001).

Grounded Theory

The aim of grounded theory is to generate theory from a systematic analysis of qualitative data (Glaser and Strauss, 1967). The approach is inductive, from evidence to the discovery of theory. It is a method of qualitative data analysis and not a theory.

The steps are as follows:

- review the literature to develop a preliminary framework to guide data collection;
- interview a small sample to identify ideas to build the theory;
- take another small sample to identify more ideas and, if necessary, refine the theory;

- continue until you reach theoretical saturation, that is, the samples provide fewer and fewer insights; and
- link the fragmented concepts into a coherent theory.

These steps are called *theoretical sampling* where, instead of sampling units, we sample for new ideas. The most difficult part of the process is the final step, the conceptual leap from concepts to theory. In most cases, it will be difficult to develop a new theory. A more likely approach is to modify an existing theory from the new ideas generated from theoretical sampling.

Beware of the inductive trap of generating ideas from the ground without a careful literature review. It is possible that the ideas generated are not new and have been discussed in the literature. It is just that they have not been found because of a faulty literature review. For instance, it is possible that hotels adopt different strategies to cope with labor problems during the COVID-19 pandemic. Hotels that employ primarily local workers tend to have excess labor, and those that hire foreign workers tend to face shortages because of the closure of international borders to prevent the spread of the virus. To understand why hotels adopt different labor strategies, it is tempting to interview hotel operators directly instead of starting from a thorough literature review to uncover reasons such as differing competitive strategies, economies of scale, relative capital-labor-land costs, the extent of government support for different types of hotels, corporate culture, the extent of worker resistance, availability of different types of technologies, current skill sets, differing processes, and so on.

Interpretive Phenomenological Analysis

Interpretive phenomenological analysis (IPA) is the study of lived experiences, such as life in a refugee camp, the struggles of an entrepreneur, or coping with a terminal illness (Smith and Nizza, 2021). In studying these experiences, we look for

- major transitions;
- unexpected events;
- proactive actions;
- reflections;

- lessons learned;
- positive experiences;
- negative experiences; and
- significant events.

For example, IPAs on the struggles of entrepreneurs often consider the initial frustration with working life, visioning of an untested but interesting business idea, the exploration and the transition, unexpected twists, the proactive actions of developing a business model, selling the idea to lenders and investors to raise capital, and the nuts and bolts of the business. This is followed by the initial setbacks, positive learning experiences, dealing with uncertainties, and the road to recovery or success. Finally, the qualities of the entrepreneur, such as hard work, persistence, and determination, often feature in such accounts.

There is no shortage of business leaders sharing their lived experience on how they lead or manage (Iacocca and Novak, 1986; Welch and Welch, 2005), turn things around (Gerstner, 2003) and fix problems (Harvey-Jones, 1993), what it takes to transform companies (Schwarzman, 2019), and the lessons learned (Iger, 2019).

References

Bates, R., Grief, A., Levi, M., Rosenthal, J., and Weingast, B. (1998) *Analytic narratives*. New Jersey: Princeton University Press.

Beckford, G. (1999) *Persistent poverty: Underdevelopment in plantation economies of the Third World*. Barbados: University of West Indies Press.

Braun, V. and Clarke, V. (2021) *Thematic analysis*. Los Angeles: Sage.

Brown, B. (2002) *Cinematography: Theory and practice*. London: Routledge.

Clandinin, J. (2013) *Engaging in narrative inquiry*. London: Routledge.

Derrida, J. (1976) *Of grammatology*. Baltimore: Johns Hopkins University Press.

Easterly, W. (2001) *The elusive quest for growth*. Massachusetts: MIT Press.

Foucault, M. (1991) *Discipline and punish: The birth of a prison*. London: Penguin.

Gerstner, L. (2003) *Who says elephants can't dance?* New York: Harper Business.

Gibson, A. (1996) *Towards a postmodern theory of narrative*. Edinburgh: Edinburgh University Press.

Gilleard, C. (1989) The achieving society revisited. *Journal of Economic Psychology*, **10**(1), 21–34.

Glaser, B. and Strauss, A. (1967) *The discovery of grounded theory*. Chicago: Aldine.

Harrison, L. (2000) *Underdevelopment is a state of mind*. Lanham: Madison Books.

Harvey-Jones, J. (1993) *Managing to survive*. London: Heinemann.

Herrnstein, R. and Murray, C. (1994) *The bell curve*. New York: Free Press.

Hibbert, P. (2021) *How to be a reflexive researcher*. Cheltenham: Edward Elgar.

Iacocca, L. and Novak, W. (1986) *Icocca*. New York: Bantam.

Iger, R. (2019) *The ride of a lifetime: Lessons learned from 15 years as CEO of the Walt Disney Company*. New York: Random House.

Levy, M. (1967) *Social patterns and problems of modernization*. New Jersey: Prentice-Hall.

Lewis, O. (1959) *Five families: Mexican case studies in the culture of poverty*. New York: Basic Books.

Marx, K. and Engels, F. (1998) *The German ideology*. New York: Prometheus.

Mazur, A. and Rosa, E. (1977) An empirical test of McClelland's "achieving society" theory. *Social Forces*, **55**(3), 769–774.

McClelland, D. (1962) Business drive and national development. *Harvard Business Review*, **40**(4), 99–113.

McCloskey, D. (1986) *The rhetoric of economics*. Brighton: Wheatsheaf.

Morrison, M. (1997) *The tragedies of G.B. Giraldi Cinthio: The transformation of narrative source into stage play*. New York: Edwin Mellen.

Moustakas, C. (1994) *Phenomenological research methods*. London: Sage Publications.

Nagler, A. (1959) *A source book in theatrical history*. New York: Dover.

Parker, I. (1992) *Discourse dynamics*. London: Routledge.

Peshkin, A. (1986) *God's choice*. Chicago: University of Chicago Press.

Riessman, C. (2007) *Narrative methods for the human sciences*. Los Angeles: Sage.

Rostow, W. (1960) *The stages of economic growth*. Cambridge: Cambridge University Press.

Schwarzman, S. (2019) *What it takes: Lessons in the pursuit of excellence*. New York: Simon and Schuster.

Smith, J. and Nizza, I. (2021) *Essentials of interpretive phenomenological analysis*. Washington DC: APA.

Thomas, W. (1923) *The unadjusted girl*. Boston: Little, Brown, & Co.

Welch, J. and Welch, S. (2005) *Winning*. New York: Harper Business.

White, H. (1990) *The content of the form: Narrative discourse and historical representation*. Baltimore: Johns Hopkins University Press.

CHAPTER 12

Quantitative Data Analysis I: Survey Data

Nature of Survey Data

Survey data usually contain large amounts of information on respondent characteristics and their views. The data may be spatial, cross-sectional, or temporal, and include frequency counts, ratings, ranks, and continuous variables.

The measurement errors from survey data vary. These errors may be large if we are gathering sensitive information such as income, wealth, religion, race, political opinions, monetary contribution, and taxation. For example, respondents may not be truthful in revealing their demand for a public good if they are required to contribute towards its provision based on their responses (Clarke, 1971). There is an incentive to "free ride" by not paying one's fair share and letting other people foot the bill.

In this chapter, we will consider the use of simple data analytic techniques involving means, variances, ranks, indexes, frequencies, and possible relations among the variables.

Exploratory Data Analysis

The purpose of exploratory data analysis (EDA) is to examine simple data patterns such as

- relations among variables;
- the presence of outliers;
- trends and turning points; and
- distributional assumptions (Tukey, 1977).

The display or presentation of data may take the form of simple tables, texts, plots, graphs, and charts. EDA is usually the first step in

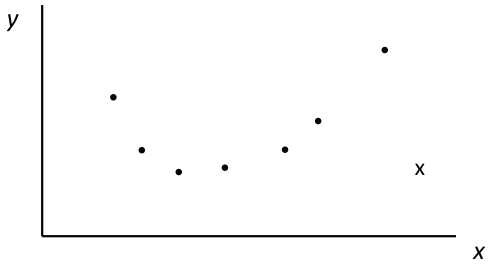

Fig. 12.1 Nonlinear correlation.

quantitative data analysis, where we explore patterns in data prior to more rigorous statistical analyses.

The correlation between any two continuous variables may be visually inspected by plotting the data in a scatter diagram. If it is linear, the data pattern looks like a line. If it is nonlinear, the data pattern resembles a curve (Fig. 12.1).

For the linear case, the correlation coefficient is given by

$$r = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}.$$

The formula is a simplification from the definition of correlation, which is

$$r = \frac{Cov(x, y)}{s_x s_y}.$$

Here, *Cov*(.) is the covariance and *s* is the standard deviation. For the sample of $n = 6$ data points in Table 12.1,

$$r = \frac{6(40) - 11(15)}{\sqrt{[6(31) - 11^2][6(55) - 15^2]}} = 0.91.$$

Since *r* lies between 0 and 1, a value of 0.91 is considered strong. In practice, *r* is easily computed using statistical software such as EXCEL. After keying in the input data (first two columns), select "Data" from the top menu, followed by "Data analysis." Then select "correlation" from the dialog box.

Table 12.1 Sample data for computation of correlation coefficient.

x	y	xy	x^2	y^2
2	3	6	4	9
4	5	20	16	25
0	1	0	0	1
1	0	0	1	0
1	2	2	1	4
3	4	12	9	16
$\sum x = 11$	$\sum y = 15$	$\sum xy = 40$	$\sum x^2 = 31$	$\sum y^2 = 55$

To test whether r is significantly different from zero, the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

is distributed as Student t distribution with $n - 2$ degrees of freedom. In this case, $t = (0.91\sqrt{4})/\sqrt{(1 - 0.91^2)} = 4.39$, which is greater than the two-tailed 0.05 critical value of 2.776 for 4 degrees of freedom. The computed value of r is significant.

Outliers are "far" from the main cluster of points or data pattern, such as the point "x" in Fig. 12.1. If they are not measurement errors, outliers may provide a refutation of the theory. A statistical distribution with more outliers has a heavy or fat tail, unlike the thin tail of the normal distribution.

Time series data observed at regular time intervals may be plotted to identify trends and correlations (Fig. 12.2). For each series, it may be possible to determine the amplitudes and turning points, and hence the period of the cycle, that is, the mean length of time measured from peak to peak, or from trough to trough. For economic data, these cycles may be short (seasonal), mid-length (5 to 10-year business cycles), or long (50-year waves) (Berry, 1991). Spectral analysis may be used to detect cycles and less regular waves (Stoica and Moses, 1997), as discussed in Chapter 16.

For frequency or count data, histograms are used to check the distribution, such as its shape, mean, and variance. Visual inspections are useful but insufficient. Formal tests of statistical hypotheses are required, as discussed below.

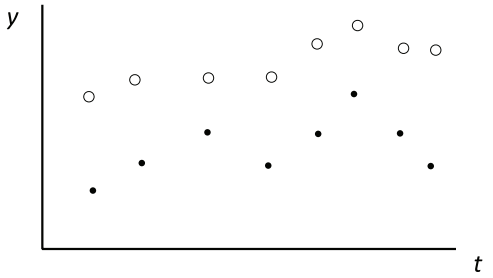

Fig. 12.2 Time series data.

Basics of Statistical Tests

Statistical tests help us to make certain decisions in a formal way. For instance, in the previous section, it may be difficult to decide if the data follow a normal distribution just by looking at the histogram. A formal test for normality is the Jarque–Bera (*JB*) test where, under the null hypothesis (H_0) of normality, the test statistic is

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right].$$

Here, n is the sample size and S is the skewness where

$$S = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})^3}{[\frac{1}{n}\Sigma(x_i - \bar{x})^2]^{3/2}}$$

and K is the kurtosis, that is,

$$K = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})^4}{[\frac{1}{n}\Sigma(x_i - \bar{x})^2]^2}$$

In both expressions, $\bar{x}$ is the sample mean. If the peak is to the left, such as income distribution, it is positively skewed (see also Fig. 12.3). If it is to the right, such as test scores where most students obtain between 60 to 90 marks, it is negatively skewed. Kurtosis is a measure of the "peakness" of the curve, which also affects the thickness of the tails of the distribution. Both measures are relative to the normal distribution, which

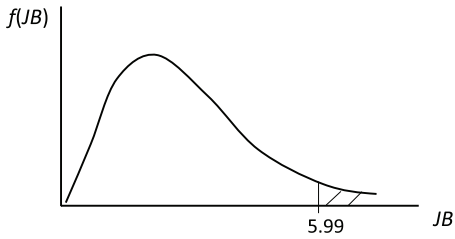

Fig. 12.3 Test of significance.

is symmetrical and has relatively thinner tails (i.e. fewer outliers), that is, $S = K = 0$.

Under H_0, *JB* is distributed as chi-square with 2 degrees of freedom. We reject H_0 if the computed value of *JB* exceeds the critical value of 5.99 at 0.05 level of significance. Figure 12.3 shows the distribution of *JB* (i.e. chi-square with 2 degrees of freedom) and the shaded critical region. The function *f*(*JB*) is the probability density function.

In summary, the steps in conducting a statistical test are:

- postulate the null hypothesis (e.g. H_0: $\mu = 0$);
- postulate an alternative hypothesis (e.g. H_1: $\mu > 0$);
- compute the test statistic under H_0;
- identify the significance level and the critical value; and
- reject H_0 if the computed value of the test statistic lies in the critical region; otherwise, do not reject H_0.

The alternative hypothesis may be one-sided, as shown above, or two-sided ($\mu \neq 0$). It depends on whether μ can take only positive values or it can be positive or negative. If it is two-sided, there are two critical regions. In the *JB* example, it is one-sided because *JB* can never be negative.

The derivation of the distribution of the test statistic often requires advanced statistics. Hence, we shall often take the distribution as given, such as chi-square distribution for the *JB* statistic. For one-sided tests, it is usual to use the 0.05 significance level, which gives a 1-in-20 chance of wrongly rejecting H_0 (see Chapter 5). The critical value is then taken from the relevant distribution in statistical tables. Finally, rejecting H_0 means

our sample does not support H_0; that is, the computed value of the test statistic is near the tail of its distribution under H_0.

Instead of stating the significance level of the test, some researchers provide the probability or *p*-value if the value of the test statistic is close to the critical value. For instance, a *p*-value of 0.03 suggests that the null hypothesis may be rejected at the 0.05 level but not at the 0.01 level.

Confidence Interval

A confidence interval shows the lower and upper bounds within which the value of an unknown parameter will lie. For example, the sample mean $\bar{x}$ will vary from sample to sample, and this variation is called its *sampling distribution*. The central limit theorem (CLT) states that, even if the distribution of x is not normal, $\bar{x}$ will be normally distributed with mean μ and variance σ^2/n, where σ^2 is the variance of x and n is the sample size. In short, CLT states that if $x \sim (\mu, \sigma^2)$, then $\bar{x} \sim N(\mu, \sigma^2/n)$ for sufficiently large n. The notation $N(.,.)$ stands for a distribution with mean and variance. Here, N stands for normal distribution. The Poisson distribution with mean λ and variance λ is written as $P(\lambda, \lambda)$.

Given a sample of size n, we often compute the sample mean $\bar{x}$ and sample variance s^2. The sample standard deviation is s, the square root of the variance. A 95% confidence interval for μ is found by computing

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}.$$

The value of 1.96 is taken from the standard normal distribution ($N(0, 1)$) table found in the Appendix. The area under the curve from −1.96 to 1.96 is 0.95, that is,

$$p\left(-1.96 < \frac{\bar{x} - \mu}{s/\sqrt{n}} < 1.96\right) = 0.95.$$

For example, if $\bar{x} = 4$, $s = 1$, and $n = 16$, then a 95% confidence interval for μ is

$$4 - 1.96 \frac{1}{\sqrt{16}} < \mu < 4 + 1.96 \frac{1}{\sqrt{16}}$$

i.e.,

$$3.51 < \mu < 4.49.$$

Properties of Estimators

If μ is the population mean, the sample mean $\bar{x}$ is called an *estimator* or formula for estimating μ. There are many ways of estimating μ, such as by finding the average of any two data points, rather than the average of all points in the sample mean. We prefer $\bar{x}$ because it is *unbiased*, that is, its expected value is equal to μ. It is written as

$$E(\bar{x}) = \mu + b$$

where b, the bias, is zero. If we use only the first two data points as an alternative estimator, then

$$E\left(\frac{x_1 + x_2}{2}\right) = \frac{E(x_1) + E(x_2)}{2} = \frac{\mu + \mu}{2} = \mu.$$

Hence, this estimator is also unbiased. However, $\bar{x}$ is more *efficient* in the sense of having a smaller variance if n is large:

$$Var(\bar{x}) = Var\left(\frac{x_1 + \cdots + x_n}{n}\right) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n};$$

$$Var\left(\frac{x_1 + x_2}{2}\right) = \frac{1}{2^2}(2\sigma^2) = \frac{\sigma^2}{2}.$$

Sometimes, it is difficult to prove unbiasedness or efficiency in small samples because the estimator is a complicated formula. In such situations, we need to prove *consistency*, that is, the probability of an estimator θ^* being close to the parameter of interest (θ) increases with the sample size:

$$\lim_{n \to \infty} P[|\theta^* - \theta| > c] = 0,$$

where c is an arbitrary small number. The shorthand is plim(θ^*) = θ. Since θ^* is centered at θ as n tends towards infinity, a simpler way to show

consistency *for an unbiased estimator* is to show that its variance vanishes as n gets larger. For example, $\bar{x}$ is unbiased and

$$Var(\bar{x}) = \frac{\sigma^2}{n} \to 0 \, as \; n \to \infty.$$

Hence, $\bar{x}$ is a consistent estimator.

Another property of an estimator is *sufficiency*, that is, it uses all the sample information. Clearly, $\bar{x}$ uses all the sample information, whereas the alternative estimator that uses only two data points is not a sufficient estimator.

Finally, it is possible for an unbiased estimator to have a larger variance than a biased estimator. In such cases, we compare the *mean square error*

$$MSE(\theta^*) = E(\theta^* - \theta)^2 = Var(\theta) + bias^2(\theta^*).$$

An estimator with a lower *MSE* is preferred. For example, the ridge estimator may be used if multicollinearity exists because, under certain conditions, it has lower *MSE* than the ordinary least squares estimator (see Chapter 15).

Continuous Data

For data measured using the ratio scale from a single population, there are a number of simple tests. The t test for a population mean for a *small* sample of size n uses the following test statistic:

$$t_{n-1} = \frac{\bar{x} - \mu}{s / \sqrt{n}}.$$

It follows the Student t distribution with $n - 1$ degrees of freedom. For large samples ($n > 30$), we replace it with the Z statistic, which follows the standard normal distribution (i.e. $N(0, 1)$). For example, to test H_0: $\mu = 5$ against H_1: $\mu > 5$ from a sample of size 16 with $\bar{x} = 7$ and $s = 2$,

$$t_{15} = \frac{7 - 5}{2/\sqrt{16}} = 4.$$

From the Student t distribution table, the one-tailed 0.05 critical value is 1.753 and we reject H_0.

To test the population variance, the test statistic

$$Q = \frac{(n-1)s^2}{\sigma^2}$$

is distributed as chi-square with $n - 1$ degrees of freedom. Using the same sample above, to test H_0: $\sigma^2 = 3$ against H_1: $\sigma^2 > 3$,

$$Q = \frac{(16-1)4}{3} = 20.$$

From the chi-square distribution table, the 0.05 critical value for 15 degrees of freedom is 25.00. We do not reject H_0. For tests on two populations, such as the unpaired and paired t tests, see Chapter 13.

Count Data

It is possible to test frequency or count data. If counts satisfy a Poisson distribution, that is, they are independent discrete events in a given time interval, distance, or area, then the test statistic for the null hypothesis H_0: $\mu_1 = \mu_2$ is

$$G = n_1/(n_2 + 1) \sim F(2(n_2 + 1), 2n_1).$$

The notation $F(.)$ stands for the F distribution with its numerator and denominator degrees of freedom. For example, if the number of potholes per km for two highways are 2 and 5, respectively, we may test H_0 using

$$G = 2/(5 + 1) \sim F(2(5 + 1), 2(2)) = F(12, 4).$$

Thus $G = 2/6 = 0.33$, and the 0.05 critical value for $F(12, 4)$ is 5.91. Hence, we do not reject H_0. If both highways are funded from the same source, the road operation and maintenance policies are likely to be similar and the events are not independent. The test will then not be valid.

The test may be generalized into a one-way contingency table. For example, does the evidence below support the claim that the number of potholes per km is equally distributed among four highways?

Highway	A	B	C	D	Total
Number of potholes per km	2	3	7	8	20

The expected count is $E = 20/4 = 5$. Let O_i be the observed count. The test statistic is

$$Q = \sum \frac{(O_i - E)^2}{E}.$$

Thus,

$$Q = (2 - 5)^2/5 + (3 - 5)^2/5 + (7 - 5)^2/5 + (8 - 5)^2/5 = 5.2.$$

It is distributed as chi-square with $k - 1$ degrees of freedom where k is the number of cells (= 4 here). From the chi-square distribution table, the 0.05 critical value is 7.81, and we do not reject the null hypothesis of equal distribution.

A two-way contingency table may be used to test for independence between two variables. For example, from a survey of 200 construction firms, we wish to test if firm size (classified as Small, Medium, or Large) is related to profitability (Low, Medium, or High). Intuitively, larger firms, with better access to markets and inputs, are likely to be more profitable. However, they may suffer from diseconomies of scale.

The observed frequencies are shown in Table 12.2 together with the row total (R_i) and column total (C_j). Obviously, good definitions of firm

Table 12.2 A 3 × 3 contingency table.

	Profitability			
Firm size	*L*	*M*	*H*	Row total
S	60	20	10	$R_1 = 90$
M	20	30	10	$R_2 = 60$
L	10	20	10	$R_3 = 40$
Column total	$C_1 = 90$	$C_2 = 70$	$C_3 = 30$	$N = 190$

size and profitability matter; otherwise, the frequency counts will not be meaningful. There should also be at least five observations in each cell.

An $r \times c$ contingency table has r rows and c columns. Here $r = c = 3$. If firm size and profitability are independent (H_0), the expected frequency for cell (1, 1) is N times the joint probabilities:

$$E = N \frac{R_1}{N} \frac{C_1}{N} = \frac{R_1 C_1}{N} = \frac{90(90)}{190} = 42.6.$$

In general, the expected frequency for cell (i, j) is $R_i C_j / N$ and these frequencies are shown below:

$$\begin{matrix} 42.6 & 33.2 & 14.2 \\ 28.4 & 22.1 & 9.5 \\ 18.9 & 14.7 & 6.3 \end{matrix}$$

The test statistic is the sum of the normalized squares of the difference between observed and expected frequencies:

$$Q = (60 - 42.6)^2/42.6 + (20 - 33.2)^2/33.2 + \cdots + (10 - 6.3)^2/6.3 = 27.2.$$

It can be shown that Q is distributed as chi-square with $(r - 1)$ x $(c - 1)$ degrees of freedom. From the chi-square distribution table, the 0.05 critical value for 4 degrees of freedom is 9.49. We reject H_0 and conclude that firm size does matter when it comes to profitability. For details on contingency tables, see (Kateri, 2016).

Spatial Data

Analyses of spatial data include

- network patterns, e.g., roads;
- spatial point patterns, e.g. the outbreak of diseases at different locations;
- spatial correlations among neighboring points or areas, e.g. rainfall patterns; and

Table 12.3 Data for interpolation example.

Station	x	y	z
A	1	1	0
B	2	4	5
C	5	3	10
D	2	3	?

- interpolation of the value of a variable at an unknown point, e.g. its height.

Readers who are interested in these techniques may consult (Cressie, 1993; Lloyd, 2010; and Chang, 2016).

As an example, given the rainfall data (z) for weather stations *A*, *B*, and *C* and their coordinates (x, y), we wish to estimate the value of z at point *D* (Table 12.3).

We can interpolate the data in many ways, such as by using the inverse distance weighted (IDW) method, kriging, or trend surface analysis (Stein, 1999). For example, in the IDW method, the predicted rainfall value at *D* is the weighted sum of the known values at other stations, that is,

$$z_D = \sum w_i z_i.$$

The weights are computed from

$$w_i = \frac{1/d_i^2}{\Sigma 1/d_i^2}.$$

The *i*th subscript refers to the station or point. The first step is to find the distances (d_i) between each station and *D*:

From	To	Distance, d_i
A	*D*	$\sqrt{5}$
B	*D*	1
C	*D*	3

Next, compute

$$\sum \frac{1}{d_i^2} = \frac{1}{5} + \frac{1}{1^2} + \frac{1}{3^2} = 1.31.$$

Hence, the weights are

$$w_A = (1/5)/1.31 = 0.15;$$
$$w_B = (1/1)/1.31 = 0.76; \text{ and}$$
$$w_C = (1/9)/1.31 = 0.08.$$

The predicted rainfall value for D is

$$z_D = 0.15(0) + 0.76(5) + 0.08(10) = 4.60.$$

Circular Data

Circular data, such as wind direction, are measured on a circle. Because 0^0 and 360^0 point in the same direction, we cannot simply add up the numbers to compute the arithmetic mean. In Fig. 12.4, two angles are measured from the north, and the mean direction θ^* is based on the dotted line formed by completing the resultant parallelogram.

The resultant length R is the length of the dotted line, and the mean resultant length is $R^* = R/n$, where n is the number of vectors (2 here). The circular variance is $1 - R^*$ and the circular standard deviation is $\sqrt{(-2 \log R^*)}$. The statistical tests for circular data are more complex and bear little resemblance to those for linear data (Fisher, 1995).

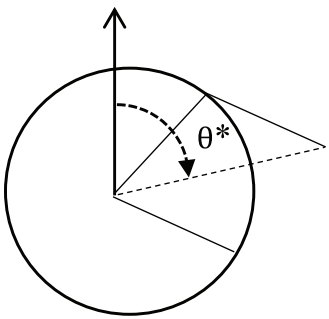

Fig. 12.4 Circular data.

Index Numbers

Index numbers are used to track temporal changes in prices, costs, water quality, and so on. We shall use the term "price" to obviate the need to repeat "cost" or "water quality" because the principles are similar.

A *nominal* price index uses market prices. To convert a nominal price index into a *real* price index, we deflate it by a suitable inflation index such as the consumer price index (*CPI*). There are other types of deflators, such as the Gross Domestic Product (*GDP*) deflator and producer price deflator. For example, if the nominal house price index is 105 and the *CPI* is 1.03, then the real house price index is 105/1.03 = 101.9. In the discussion below, we will be constructing nominal price indexes. They may be converted to the real price indexes using a suitable deflator.

Price index for homogeneous product

To construct the price index of a relatively homogeneous product such as an apple, we first decide on the "standard" apple (for example, rose apple) to remove quality differences. Then, if P_0, P_1, …, P_n are annual prices, each price is divided by its base period price to obtain the following index:

$$1, P_1/P_0, \ldots, P_n/P_0.$$

The index is 1 in the base period because $P_0/P_0 = 1$. We often multiply the index by 100 to shift the decimal point (e.g. $1.031 \times 100 = 103.1$). Basically, a price index is a normalized series.

If the annual apple prices are 50c, 55c, 60c, and 80c, respectively, then the price index is

$$1, 55/50, 60/50, 80/50,$$

that is,

$$100, 110, 120, 160.$$

Price index for heterogeneous product

To construct a fruit price index, we compare current year prices (P_t) with base year prices (P_0) using suitable base year weights (W_0) (Table 12.4). The weights, in quantity or dollar terms, may be based on imports, annual

Table 12.4 Construction of fruit price index.

	P_0	W_0	P_0W_0	P_t	P_tW_0
Apple	0.50	0.5	0.25	0.60	0.30
Orange	0.60	0.3	0.18	0.70	0.21
Durian	10	0.2	2	12	2.4
Total			2.43		2.91

Table 12.5 Price index and sales volume.

Property type	Price index	Sales volume
Public flats	120	2,000
Private apartments	130	3,000
Landed houses	150	1,000

national consumption, or annual consumption from a survey of households. Only three types of fruits are used for illustration. The index for the current year (t) is 100(2.91/2.43) = 119.75.

Similarly, we may construct a house price index for a city using sales volume as base year weights (Table 12.5).

This method of constructing a price index using base year weights produces a Laspeyres Index. Over time, the base year weights will change and it is necessary to use a new set of weights after every 5 or 10 years. This is called rebasing an index.

Human Development Index (HDI)

The *HDI* under the United Nations Development Program uses three indicators, namely, life expectancy index (*LEI*), education index (*EI*), and income index (*II*). *LEI* and *II* are normalized using

$$Index = \frac{(x - a)}{(b - a)}.$$

Here, a is the minimum value, b is the maximum value, and x is a country's value. This produces a value between 0 and 1. For example, if

the life expectancy of a country is 65 years, then the index value is (65 – 20)/(85 – 20) = 0.69.

For *II*, the formula is

$$Index = \frac{(\log y - \log 100)}{(\log 75,000 - \log 100)}.$$

Here, y is the gross national income per capita in purchasing power parity terms. The education index is computed differently using

$$EI = (MYSI + EYSI)/2.$$

The mean years of schooling index (*MYSI*) is *MYS*/15, and the expected years of schooling index (*EYSI*) is *EYS*/18. The number 18 refers to the number of years of education to obtain a master's degree. Finally, *HDI* is the geometric mean of the three sub-indexes:

$$HDI = (LEI.EI.II)^{1/3}.$$

In summary, the Laspeyres index uses base year weights to compare changes, whereas the *HDI* uses normalized values.

Productivity index

The productivity index uses factor shares as weights. Consider a production function $f(.)$ where the output (Q) is a function of capital input (K), including land, and labor input (L). Then

$$Q = f(K,L).$$

For estimating purposes, it is usual to specify a Cobb–Douglas production function

$$Q = AK^{\alpha}L^{\beta}.$$

Here, A represents the efficiency in which inputs are used to produce the output. It is called the Total Factor Productivity (TFP). It is "Total" because it estimates productivity from capital and labor inputs. A partial measure such as labor productivity holds capital input constant. TFP reflects factors other than inputs, such as better management and

organization, learning economies, scale economies, worker incentives and effort, research and development, and the regulatory environment. This long list of variables is the main reason why it is hard to pin down the causes of productivity changes over time.

Differentiating the Cobb–Douglas function with respect to time (t) and using the relation

$$\frac{d\log(x)}{dt} = \frac{dx}{dt}\frac{1}{x},$$

the so-called Divisia index is given by

$$A^* = Y^* - \alpha K^* - \beta L^*,$$

where A^* is the rate of productivity growth, Y^* is the rate of output growth, K^* is the capital growth rate, and L^* is the labor growth rate.

The parameter α represents capital's share of the output, which is about 0.35 for many economies (Weil, 2009). If constant returns to scale is assumed, $\beta = 1 - \alpha$. As an example, if for a particular year $Y^* = 3\%$, $K^* = 2\%$, and $L^* = 1\%$, then

$$A^* = 3 - 0.35(2) - 0.65(1) = 1.65\%.$$

For more capital-intensive industries, the value for α will be greater than 0.35 to reflect the greater share of capital's contribution to output.

The Törnqvist (1936) index is a discrete approximation to the Divisia index and is given by

$$\log\left(\frac{A_t}{A_{t-1}}\right) = \log\left(\frac{Q_t}{Q_{t-1}}\right) - \alpha\log\left(\frac{K_t}{K_{t-1}}\right) - \beta\log\left(\frac{L_t}{L_{t-1}}\right).$$

The reason is that, for small changes,

$$\log\frac{x + \Delta x}{x} \approx \frac{\Delta x}{x}.$$

Further, the Törnqvist index uses changing weights, that is,

$$\alpha = (\alpha_t + \alpha_{t-1})/2; \text{ and}$$
$$\beta = (\beta_t + \beta_{t-1})/2.$$

In other words, the factor shares are computed (chained) using the average of the shares of two successive periods.

Ratings

Unweighted ratings

Many surveys ask respondents to rate k factors that affect some event, process, or activity in terms of their frequency of occurrence (for example, 1 = Rare, 5 = Often) and criticality (for example, 1 = Not critical, 5 = Critical).

In Table 12.6, the "Count" column refers to the number of responses for each question. For instance, 100 respondents answered the question on Factor 1, but only 90 respondents replied to the question on Factor 2. The mean frequency and criticality ratings are based on the average rating for frequency of occurrence and criticality, respectively. An event may occur frequently (for example, rain) but its effects are not critical. On the other hand, an event may occur infrequently (for example, site accident) but its effects are critical. The impact of each factor is then computed as the product. Factors with high impacts require close attention and monitoring.

Weighted ratings

We may apply judgmental weights to the ratings. For instance, bidders for a project may be selected using suitable weights on items such as track record and expertise (A), financial strength (B), workload (C), bid price (D),

Table 12.6 A rating table.

Factor	Count	Mean frequency rating (a)	Mean criticality rating (b)	Impact [(a) × (b)]
1	100	3.4	4.5	15.3
2	90	2.0	3.2	6.4
…		…	…	…
k	70	3.0	2.5	7.5

Table 12.7 Selection of contractor using judgmental weights.

		Bidder			Bidder		
		X	*Y*	*Z*	*X*	*Y*	*Z*
Criteria	Weight	Original score			Weighted score		
A	0.2	6	7	8	1.2	1.4	1.6
B	0.1	5	6	9	0.5	0.6	0.9
C	0.1	9	6	5	0.9	0.6	0.5
D	0.4	6	8	7	2.4	3.2	2.8
E	0.2	6	8	5	1.2	1.6	1.0
Total	1.0				6.2	7.4	6.8
Bid ($m)					100	105	108
Value ratio					0.062	0.070	0.063

and schedule (*E*) (Table 12.7). We multiply the original scores by the weights to obtain the weighted scores. The total weighted score for each bidder is then divided by the bid price to obtain the value ratios. In this case, bidder *Y* provides the best value.

Alternatively, the weights may be derived using pairwise comparisons. In Table 12.8, a judging panel makes pairwise comparisons between pairs of criteria. In the first row, criterion *A* is considered to be more important than *B*, less important than *C* and *D*, and more important than *E*. We then transpose the first row result to the first column by symmetry (shown as boldface letters).

In the second row, *B* is considered to be more important than *C* but less important than *D* or *E*. Again, we transpose the results to the second column (shown as underlined letters). For each row, we note the frequency of the criterion from which weights are derived. For instance, in the first row, criterion *A* appears twice. In the second row, *B* appears once, and so on. A criterion with zero frequency is dropped from consideration.

Ranks

If data involve ranks, there are many ways of dealing with them beyond simple description. Some of these techniques are discussed below.

Table 12.8 Derivation of weights.

Criteria	*A*	*B*	*C*	*D*	*E*	Frequency	Weight
A		*A*	*C*	*D*	*A*	2	0.2
B	***A***		*B*	*D*	*E*	1	0.1
C	***C***	*B*		*D*	*E*	1	0.1
D	***D***	*D*	*D*		*D*	4	0.4
E	***A***	*E*	*E*	*D*		2	0.2
Total						10	1.0

Generally, these tests are non-parametric, that is, they do not assume that the data come from any particular parent distribution (Corder and Foreman, 2014). For example, if the samples do not come from a normal distribution, it may be possible to convert the continuous ratio scale data to ranks and use a suitable non-parametric test.

Rank correlation

The rank correlation coefficient computes the correlation between two sets of ranks. For example, we may wish to determine if two judges (or two groups of people) rank various brands of wine consistently (Table 12.9).

The Spearman rank correlation coefficient is given by

$$r = 1 - \frac{6R}{n(n^2 - 1)} = 1 - \frac{6(8)}{5(25 - 1)} = 0.6.$$

Since r lies between 0 and 1, the rankings are broadly in agreement. Wines *A*, *B*, and *C* form the top group, while *D* and *E* are in the bottom group. For large samples, the test statistic is

$$Z = \frac{6R - n(n^2 - 1)}{n(n + 1)\sqrt{(n - 1)}}.$$

It follows the standard normal distribution and can be used to test the null hypothesis of no correlation.

Table 12.9 Rankings of wine.

	Judge			
Wine	x	y	$d\ (= y - x)$	d^2
A	1	2	1	1
B	3	1	−2	4
C	2	3	1	1
D	4	5	1	1
E	5	4	−1	1
Total			0	$8 = R$

Table 12.10 Ranking of male and female singers.

M	1	2	5	8	9	11	13	$T_M = 49$	$n_2 = 7$
F	3	4	6	7	10	12		$T_F = 42$	$n_1 = 6$

Mann–Whitney test

Instead of computing just the rank correlation coefficient, we can formally test ranks for two groups of people similar to the parametric *t* test for two independent groups. Suppose 13 singers, comprising 7 males (*M*) and 6 females (*F*), are ranked by a panel of judges, as shown in Table 12.10. The penultimate column shows the sum of the ranks, and the last column shows the sample sizes with n_1 designated for the smaller sample. Is there evidence of a difference in the ranking?

In the Mann–Whitney test, the null hypothesis is no difference in ranks for both groups. The test statistic T is the rank sum of the smaller sample, that is, $T = 42$. This is checked against the critical values of the Mann–Whitney table. For large samples,

$$Z = \frac{T - E(T)}{S}$$

follows the standard normal distribution. Here $E(T)$ is the expected value of T and it can be shown to be given by

$$E(T) = \frac{n_1(n_1 + n_2 + 1)}{2}.$$

The value of S is computed from

$$S = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}.$$

In our example, $E(T) = 6(14)/2 = 42$ and

$$S = \sqrt{\frac{42(14)}{12}} = 7.$$

Thus,

$$Z = (42 - 42)/7 = 0.$$

Since the 0.05 critical value for the standard normal distribution is 1.645, we do not reject H_0. There is no evidence that the ranks differ substantially. This is not unexpected; the average rank for males is 49/7 = 7 and, for females, it is 42/6 = 7. There is no difference in average rank for both groups.

Friedman test

The Friedman test is used if there are more than two groups. In Table 12.11, five respondents have been asked to rank their preferences for three projects *A*, *B*, and *C*. For example, Fox prefers *C* to *A* to *B*. The column total (*R*) is the sum of ranks. Under the null hypothesis of no difference in total or mean rank, the Friedman test statistic for large n is given by

$$F_r = \frac{12S}{nk(k+1)} - 3n(k+1).$$

Here n = number of respondents (5), k = number of categories (3, that is, *A*, *B*, and *C*), and S = sum of squares of column ranks = $8^2 + 11^2 + 11^2 = 306$.

Table 12.11 Rankings of three projects.

Respondent	*A*	*B*	*C*
Fox	2	3	1
Joe	1	2	3
John	1	3	2
Ken	3	1	2
Ben	1	2	3
Total	$R_1 = 8$	$R_2 = 11$	$R_3 = 11$
Mean rank	1.60	1.83	1.83

It can be shown that F_r is distributed as chi-square with $k - 1$ degrees of freedom (Marden, 1996). Hence,

$$F_r = 0.2(306) - 60 = 1.2.$$

The 0.05 critical value for a chi-square variable with 2 degrees of freedom is 5.99. We do not reject the null hypothesis. Intuitively, we can see from the table that the mean ranks are fairly close.

Wilcoxon signed-rank test

This test is the non-parametric counterpart of the paired *t* test. Table 12.12 shows the test scores of 8 students over two semesters. The difference is computed, absolute values are taken and then ranked from smallest to largest value. If there is a tie, the ranks are split. If T^+ is the positive rank sum and T^- is the negative rank sum, then

$$T^+ = 5 + 8 + 1 + 3.5 + 6 + 7 = 30.5; \text{ and}$$
$$T^- = 3.5 + 2 = 5.5.$$

The null hypothesis is $T^+ = T^-$, that is, no difference in test scores. The test statistic, W, is the minimum of T^+ or T^-, which is 5.5. Further,

$$Z = \frac{W - E(W)}{S(W)}$$

Table 12.12 Data for Wilcoxon signed-rank test.

ID	Test 1	Test 2	Difference	|Difference|	Rank	Sign
1	50	60	10	10	5	
2	60	55	–5	5	3.5	(–)
3	65	80	15	15	8	
4	70	70	0	0	1	
5	75	71	–4	4	2	(–)
6	80	85	5	5	3.5	
7	85	96	11	11	6	
8	90	94	14	14	7	

follows the standard normal distribution for $n > 25$. For smaller sample sizes, we need to look up the distribution table for W. For illustration, we will use the normal approximation where

$$E(W) = \frac{n(n+1)}{4} = \frac{8(9)}{4} = 18,$$

and

$$S(W) = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{8(9)(17)}{24}} = 7.14.$$

Thus,

$$Z = \frac{5.5 - 18}{7.14} = -1.75.$$

From the standard normal distribution table, the 0.025 one-tailed critical value is −1.96 and we do not reject H_0

References

Berry, B. (1991) *Long-wave rhythms in economic development and political behavior*. Baltimore: Johns Hopkins Press.

Chang, K. T. (2016) *Introduction to geographic information systems*. New York: McGraw-Hill.

Clarke, E. (1971) Multipart pricing of public goods. *Public Choice*, **11**, 17–33.
Corder, G. and Foreman, D. (2014) *Non-parametric statistics*. New York: Wiley.
Cressie, N. (1993) *Statistics for spatial data*. New York: Wiley.
Fisher, N. (1995) *Statistical analysis of circular data*. London: Cambridge University Press.
Kateri, M. (2016) *Contingency table analysis*. New York: Birkhauser.
Lloyd, C. (2010) *Spatial data analysis*. London: Oxford University Press.
Marden, J. (1996) *Analyzing and modeling rank data*. London: Chapman and Hall.
Stein, M. (1999) *Interpolation of spatial data*. Berlin: Springer.
Stoica, P. and Moses, R. (1997) *Introduction to spectral analysis*. New Jersey: Prentice Hall.
Törnqvist, L. (1936) The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin*, **10**, 1–8.
Tukey, J. (1977) *Exploratory data analysis*. London: Pearson.
Weil, D. (2009) *Economic growth*. New York: Pearson.

This page intentionally left blank

CHAPTER 13

Quantitative Data Analysis II: Experimental Data

Unpaired *t* Test

The unpaired *t* test for two independent samples may be used to analyze data for the classical experimental design and its variants, such as the parallel group design and quasi-experimental designs.

The groups are assumed to be statistically independent, that is, they do not affect each other. It is further assumed that the samples come from a normal population with mean μ and variance σ^2. The samples should be of similar sizes (n_1 and n_2, respectively) and not be too small so that sample variances can be reasonably estimated. For sample sizes greater than 30, the *t* and normal distributions are identical.

Under the null hypothesis (H_0), where there is no difference in mean scores between the two groups, the test statistic is

$$t = D/S.$$

It follows the *t* distribution with $n_1 + n_2 - 2$ degrees of freedom. Here, *D* is the difference in sample means and *S*, the standard deviation of *D*, is computed from

$$S^2 = \frac{V}{n_1} + \frac{V}{n_2},$$

where *V* is the pooled variance, our estimate of σ^2. It is the weighted average of the sample variances, that is,

$$V = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Table 13.1 Data for unpaired *t* test.

Group	Mean	Variance	Sample size
1	7	1	10
2	6	2	10

The weights are based on sample sizes.

Example

A project company recruited 20 trainees and put one group of 10 trainees (Group 1) through a new training program. The other group followed the usual training program. The trainees are then assessed on their performance on a scale of 1 to 10, from poor to excellent (Table 13.1).

Here $D = 7 - 6 = 1$ and

$$V = \frac{9(1.0) + 9(2.0)}{18} = 1.5.$$

Thus,

$$S^2 = \frac{1.5}{10} + \frac{1.5}{10} = 0.3$$

and

$$t = \frac{D}{S} = \frac{1}{\sqrt{0.3}} = 1.825.$$

From the Student *t* distribution table, the 0.025 critical value for t_{18} is 2.101. We do not reject H_0. Statistically, there is no difference in group performance.

Paired *t* Test

Recall from Chapter 7 that, in a repeated measures design, each experimental unit is measured twice. As before, we assume that the population is normally distributed. Under the null hypothesis of no difference in test scores, the test statistic is

$$t_{n-1} = \frac{\bar{d}}{s/\sqrt{n}}.$$

Here, $\bar{d}$ is the mean difference in scores, s is the standard deviation of d, the difference in scores, and n is the sample size.

Example

In Table 13.2, 25 students took a mathematics test before they were exposed to a new teaching method, and then took a second test. The test scores are based on 0–100 marks. We then compute the differences in scores (d), its mean ($\bar{d}$), and standard deviation (s).

Under the null hypothesis of no difference in scores, the test statistic is

$$t_{24} = \frac{4}{5/\sqrt{25}} = 4.0.$$

From the Student t distribution table, the 0.05 critical value (2-tailed) for t_{24} is 2.064. We reject H_0 and conclude that there is a statistical difference in scores.

Linear Model Approach

Recall from Chapter 7 that we may wish to block out certain effects that are not of experimental interest, such as gender. In Table 13.3, a sample

Table 13.2 Data for teaching experiment.

	Test scores		
Individual	Before	After	Difference d
1	90	60	–30
2	70	70	0
3	60	80	20
…	…	…	…
25 = n	80	90	10
			$\bar{d}$ = 4; s = 5

Table 13.3 Randomized block design.

Total	Gender split	Samples	Groups
100	60	30	*E*
		30	*C*
	40	20	*E*
		20	*C*

of 100 students comprising 60 boys and 40 girls were split into experimental (*E*) and control (*C*) groups. The treatment is a new teaching method (*T*).

The traditional way of analyzing such data is to use analysis of variance (ANOVA) (Milton and Arnold, 1995). A simpler and more modern approach is to use the linear model (Searle, 1971; Rawlings et al., 1998; Weber and Skillings, 2000) and define

$$Y_i = \alpha + \beta G_i + \lambda T_i + \varphi G_i T_i + \varepsilon_i. \tag{13.1}$$

Here, Y_i is the math score for the *i*th student, T is a dummy variable representing the new teaching method, with $T = 1$ if it is administered and 0 otherwise (Table 13.4). G is the gender dummy variable with $G = 1$ for boy and 0 for girl. The interacting term GT is the product of G and T. Finally, α is a constant, β, λ, and φ are parameters, and ε is the error term.

The interacting term captures the possibility that the new teaching method may affect boys and girls differently. For boys ($G = 1$), the equation becomes

$$Y_i = \alpha + \beta + \lambda T_i + \varphi T_i + \varepsilon_i = (\alpha + \beta) + (\lambda + \varphi) T_i + \varepsilon_i.$$

For girls ($G = 0$), the estimating equation is

$$Y_i = \alpha + \lambda T_i + \varepsilon_i.$$

The interacting and slope coefficients differ between the two models. If the interacting term is not used, only the intercepts differ.

It is not necessary to estimate the equations separately for boys and girls. We just need to estimate Equation (13.1) using regression analysis,

Table 13.4 Data for randomized block design.

Student	Score (Y_i)	Teaching method (T_i)	Gender (G_i)
1	90	1	1
2	82	1	1
…	…	…	…
30	56	1	1
31	80	0	1
32	70	0	1
…	…	…	…
60	54	0	1
61	67	1	0
62	93	1	0
…	…	…	…
80	56	1	0
81	78	0	0
82	80	0	0
…	…	…	…
100	58	0	0

as discussed in the next chapter. Interestingly, the data analyses are similar for both experimental and regression designs. The advances in the analysis of linear models have made it possible to apply regression analysis flexibly to many types of experimental data. Finally, observe that data for the independent variables in Table 13.4 consist of only ones and zeros.

Example

For the data in Table 13.2, the estimating equation is

$$Y_i = \alpha + \beta T_i + \varepsilon_i. \tag{13.2}$$

Here T_i is 0 before the experiment and 1 thereafter. The data are reorganized and presented in Table 13.5. The 50 observations are stacked. The first 25 observations are test scores before the experiment, and the last 25 observations (in boldface) are test scores after the experiment. If the

Table 13.5 Data for repeat measures design.

Y_i	T_i
90	0
70	0
60	0
...	...
80	0
60	1
70	1
80	1
...	...
90	1

treatment (new teaching method) is effective, the estimated value of β will be significant.

As discussed in subsequent chapters, it is easy to compute the regression equation using statistical software such as EXCEL.

References

Milton, J. and Arnold, J. (1995) *Introduction to probability and statistics*. New York: McGraw-Hill.

Rawlings, J., Pantula, S., and Dickey, D. (1998) *Applied regression analysis*. New York: Springer.

Searle, S. (1971) *Linear models*. New York: Wiley.

Weber, D. and Skillings, J. (2000) *A first course in the design of experiments*. New York: CRC Press.

CHAPTER 14

Quantitative Data Analysis III: Regression Data (Part I)

Linear Regression

The simple *population* regression model postulates that a dependent variable Y is a linear function of an independent variable X, that is,

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1, 2, ..., n. \tag{14.1}$$

Here, α is the constant term or intercept, β is the regression coefficient or parameter, ε_i is the ith error or disturbance term, and n is the number of data points. For example, we may postulate that monthly household expenditure (Y) depends on disposal income (X). For each household, we observe Y_i and X_i, giving a total of n data points in an (X, Y) scatter diagram.

In many cases, there are other variables such as household size, government policy, stage of a life cycle, and tastes that affect household expenditure. Hence, the simple regression model is used here for expositional purposes. In most cases, we will deal with the multiple regression model.

For brevity, the subscripts are sometimes dropped so that we can write the model as $Y = \alpha + \beta X + \varepsilon$. If the *population* refers to households in a large city or town, there will be many data points. In practice, we take a much smaller *sample* of households and estimate

$$Y = a + bX + e. \tag{14.2}$$

Here, a and b are the estimated intercept and coefficient, respectively, and e is the *residual*, the sample estimate of ε (Fig. 14.1).

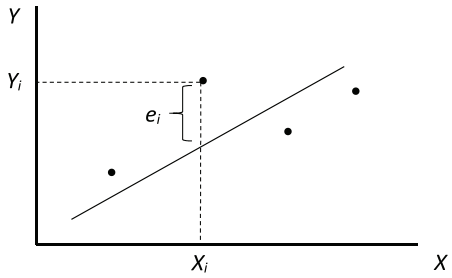

Fig. 14.1 Sample regression line.

The key coefficient is b, the estimated slope of the sample regression line. The hope is that b will be numerically "close" to β, the slope of the population regression line. Different samples will give different sample regression lines and hence different estimates of b. This variation, as discussed in Chapter 12, is called the sampling distribution of b.

Model Assumptions

In estimating the sample regression line, we have implicitly used various assumptions. These assumptions are discussed below. Departures from these assumptions are common and require careful analysis. We will discuss these issues in Chapters 15 and 16.

A1: Linearity

The model is assumed to be linear in the parameters but not in the independent variables. For example, $Y = \alpha + \beta X^2 + \varepsilon$ is linear in the parameter but not $Y = \alpha + \beta^2 X + \varepsilon$ or $Y = \alpha + \log(\beta^2 X) + \varepsilon$. There must not be other powers of β.

A2: $E(\varepsilon_i) = 0$

The errors are zero on average. This ensures that the regression line goes through the middle of sample points.

A3: $Var(\varepsilon_i) = \sigma^2$

The errors are distributed evenly (homoscedastic), as shown in Fig. 14.1. In contrast, the errors in Fig. 14.2 are heteroscedastic, that is, they tend to rise with increasing values of X.

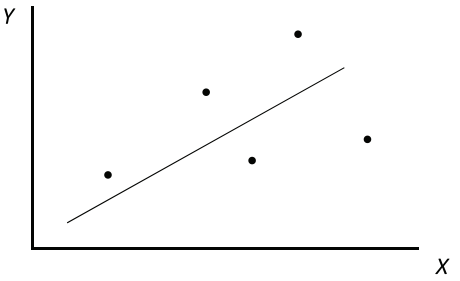

Fig. 14.2 Example of heteroscedasticity.

A4: X is fixed in repeated sampling

We assume that the X values are non-stochastic or fixed from sample to sample, which simplifies the properties of the model. For example, if we take the expectation,

$$E[Y] = \alpha + \beta E[X] + E[\varepsilon] = \alpha + \beta X$$

because $E[\varepsilon] = 0$ by A2. Observe that $E[X] = X$ by assumption. This assumption is reasonable in an experimental setting where the experimenter selects particular values of X before observing Y. However, it is unrealistic in many non-experimental situations where if we take another sample, the X values are likely to be different. In this case, we take the conditional expectation, that is, $E[Y|X] = \alpha + \beta X$. The conditional expectation $E[Y|X]$ means the expectation of Y given X. In summary, the linear regression model is applicable to both experimental and non-experimental data.

A5: $Cov(\varepsilon_i, \varepsilon_j) = 0$

The errors are independent, which implies that they are not correlated. This means that Y_i and Y_j are independent because X is fixed by A4.

A6: ε is normally distributed

This assumption is required to facilitate the derivation of statistical tests. It also implies that Y is normally distributed because X is fixed in repeated sampling. Although many variables are normally distributed, there are also variables that do not follow this distribution. Together, A2,

A3, and A6 imply ε_i is distributed as $NID(0, \sigma^2)$, where *NID* stands for normally and independently distributed. The error term is unobservable and represents

- the influences of omitted variables, presumably because of our ignorance;
- measurement errors; and
- random variations, such as in human behavior.

For these reasons, households with the same income level will vary in their spending on goods and services.

A7: $Cov(X, \varepsilon) = 0$

This important assumption implies that X and ε are independent or uncorrelated. It will not be true if there are measurement errors in X, Y affects X (i.e. feedback), or if X is a lagged dependent variable. We will discuss these issues in Chapter 15.

A8: Linear independence

Consider the model

$$Y = \alpha + \beta X + \lambda Z + \varepsilon.$$

If Z is *collinear* or perfectly correlated with X, then Z is a linear function of X, such as $Z = \theta X$ for some constant θ, and

$$Y = \alpha + \beta X + \lambda\theta X + \varepsilon \ = \alpha + (\beta + \lambda\theta)X + \varepsilon.$$

The inclusion of Z does not add any information because the model can be represented as a simple regression equation. Technically, the model $Y = \alpha + \beta X + \lambda Z + \varepsilon$ cannot be uniquely estimated, and this is explained in the next chapter. If Z and X are collinear, we say that they are *linearly dependent*. Otherwise, they are linearly independent.

When Z is highly correlated with X, we say they are *multicollinear* rather than perfectly collinear. For example, if Y is the house price, X is the land area, and Z is the number of rooms, then X and Z are highly correlated because a large house is likely to have more rooms. In this case,

the inclusion of Z does not add much information. The methods of detecting and dealing with multicollinearity are discussed in Chapter 15.

$$\text{A9: } n > k$$

Finally, there must be more equations than unknowns for us to estimate the k parameters uniquely. To understand this concept, consider solving $x + y = 2$. There are two unknowns and only one equation. There is no unique solution. For example, (1, 1), (0, 2), (2, 0), and (3, −1) are possible solutions.

Least Squares Estimation

The regression model in Equation (14.1) is usually estimated using ordinary least squares (OLS). Other approaches, such as the maximum likelihood method, Bayesian method, robust regression or method of moments (Birkes and Dodge, 1993; Draper and Smith, 1998) are used if some of the model assumptions are violated. Since these techniques are more complicated, we will often use the OLS estimator.

The population regression model may be written as

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon. \tag{14.3}$$

The estimated model is

$$Y = b_1 + b_2 X_2 + \ldots + b_k X_k + e. \tag{14.4}$$

In matrix form, the corresponding equations are

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{14.5}$$

and

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}. \tag{14.6}$$

Here, $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times k$ design matrix, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors, $\mathbf{b}$ is a $k \times 1$ vector of estimated coefficients, and $\mathbf{e}$ is an $n \times 1$ vector of residuals. Note that lower case boldface letters represent vectors, and upper case boldface letters represent matrices.

For the data in Table 14.1, Equation (14.6) may be written as

$$\begin{bmatrix} 3 \\ 4 \\ 6 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 4 \\ 1 & 3 & 5 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

The column of 1s in the first column of **X** is necessary because of the intercept term. For example, if we multiply out the first row, we get

$$3 = b_1 + 0b_2 + 1b_3 + e_1.$$

The least squares solution vector **b** is found from the orthogonal (perpendicular) projection of **y** onto the space spanned by the column vectors of **X** (Fig. 14.3).

Table 14.1 Sample data.

[*i*]	Y	X_2	X_3
1	3	0	1
2	4	1	1
3	6	2	2
4	7	3	4
5	9	3	5

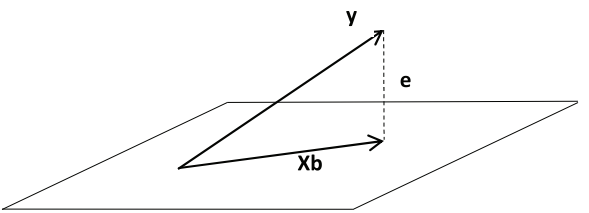

Fig. 14.3 Least squares solution.

To see this, first pre-multiply Equation (14.6) by $\mathbf{X}^T$, the transpose matrix of $\mathbf{X}$, so that

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{b} + \mathbf{X}^T\mathbf{e}. \tag{14.7}$$

The transpose matrix is formed by rewriting the *i*th column of $\mathbf{X}$ as the *i*th row of $\mathbf{X}^T$. For example, if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},$$

then

$$\mathbf{A}^T = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}.$$

The least squares solution implies $\mathbf{X}^T\mathbf{e} = \mathbf{0}$, that is the columns of $\mathbf{X}$ are orthogonal to $\mathbf{e}$. Thus, from Equation (14.7), the least squares solution is found by solving the *normal equations*

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{14.8}$$

This estimator is unbiased because, substituting Equation (14.5) for $\mathbf{y}$,

$$E[\mathbf{b}] = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = \boldsymbol{\beta}$$

because $\mathbf{X}$ is fixed and $E[\boldsymbol{\varepsilon}] = 0$. Further,

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}$$

so that

$$\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}.$$

Thus,

$$\begin{aligned} Var(\mathbf{b}) &= E(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})^T \\ &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}][(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}]^T \\ &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \end{aligned} \tag{14.9}$$

Table 14.2 Sample data for the project cost overrun model.

[*i*]	*Y*	*S*	*P*
1	0	10	0
2	0	15	0
3	2	20	0
4	4	25	0
5	5	30	0
6	5	35	1
7	7	40	1
8	8	50	1
9	12	60	1
10	15	70	1

It can be shown that an unbiased estimator for σ^2 is

$$s^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-k}. \tag{14.10}$$

In summary, we use Equation (14.8) to compute the least squares estimates and Equations (14.9) and (14.10) to compute the standard errors.

Example

Consider the model

$$Y = b_1 + b_2S + b_3P + e.$$

Here Y is the project cost overrun as a percentage of the contract value, S is the project size by contract value in million dollars, and P is the procurement strategy with $P = 0$ for a lump sum contract and $P = 1$ for design-build and other types of contract. Although many other variables affect cost overrun, such as the experience of the contractor, composition of the project team, method of financing, experience of the project owner, and complexity of design, we will use only two independent variables for illustration.

The model may be estimated using Microsoft EXCEL, Minitab, Statistical Analysis System (SAS), or Statistical Package for the Social

Sciences (SPSS). For EXCEL, key in the table and select the "Data" menu at the top and then "Data analysis." Then select "Regression" from the dialog box. The output is given below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.991404
R Square	0.982883
Adjusted R Square	0.977992
Standard Error	0.724914
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	211.2215	105.6107	200.9717
Residual	7	3.678505	0.525501	
Total	9	214.9		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	–3.07103	0.549135	–5.59249	0.000822
X Variable 1	0.263551	0.022161	11.89243	6.75E-06
X Variable 2	–0.77009	0.825935	–0.93239	0.382169

The estimated regression model is

$$Y = -3.07 + 0.263S - 0.770P + e \qquad R^2 = 0.983.$$
$$(11.89) \quad (-0.93)$$

The numbers in brackets are the *t* values. Alternatively, you may cite the standard errors, provided this is explained to the readers. Before interpreting the model, it is necessary to conduct diagnostic tests to check if the assumptions of the model hold. There are many diagnostic tests; in this chapter, we will deal with the standard ones, that is, the coefficient of

determination and tests of significance. For more complicated tests, see Chapters 15 and 16.

Coefficient of Determination

How well does the sample regression line fit the data? To answer this, we use the coefficient of determination, which is given by

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Here *RSS* is the residual sum of squares, that is,

$$RSS = \sum e_i^2.$$

TSS is the total sum of squares, that is,

$$TSS = \sum (Y_i - \bar{Y})^2,$$

where $\bar{Y}$ is the sample mean. If the sample regression line fits the data well, the residuals will be small. Thus, *RSS*/*TSS* will be close to zero, and R^2 will be close to one. Hence, R^2 provides a measure of goodness of fit. A value of R^2 close to 1 indicates a good fit. For our model,

$$R^2 = 1 - \frac{3.68}{214.9} = 0.983.$$

In practice, it is not necessary to compute it by hand; it is generated by the EXCEL output. Note that, in some textbooks, *RSS* refers to the regression sum of squares, not the residual sum of squares. This can be confusing and lead to errors. Hence, do not assume that the symbols mean the same thing.

We cannot compare R^2 between two models if

- the dependent variables are different, such as *Y* and log *Y*;
- one model has no intercept term; or
- sample sizes are not identical.

In both cases, the ratio *RSS*/*TSS* will not be the same. Further, it is possible to improve the R^2 by discarding points that do not fit the line well.

Thus, R^2 is a property of the sample, not of the population. For this reason, researchers should not take R^2 too seriously. Finally, because we can improve R^2 by adding more independent variables, a better measure of fit is the adjusted R^2, which is given by

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k}.$$

For our model,

$$R_a^2 = 1 - (1 - 0.98)\frac{9}{7} = 0.978,$$

which is close to the value of R^2.

Test of Overall Significance

Students often ask how high R^2 must be for the model to be acceptable. Since R^2 is subjective, it may be preferable to formally test for the overall significance of the model in Equation (14.3), that is,

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon.$$

The null hypothesis is H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$, and the alternative hypothesis is that H_0 is not true. The test statistic is

$$F = \frac{R^2/(\mathrm{k}-1)}{(1-R^2)/(n-k)}.$$

It is distributed as $F(k-1, n-k)$. The derivation is based on Equation (15.4) and will not be discussed here. Using the same project example,

$$F = \frac{0.983^2/(3-1)}{(1-0.983^2)/(10-3)} = 201.$$

This F value is shown in the regression output as 200.97. From Table A.3 in the Appendix, the 0.05 critical value for $F(2, 7)$ is 4.74. Hence the F value of 201 is highly significant, and we reject H_0. This is expected

because R^2 is very high. Therefore, the F test is necessary only if R^2 is not high.

Tests of Individual Significance

Recall that our estimated model is

$$Y = -3.07 + 0.263S - 0.770P + e \qquad R^2 = 0.983.$$
$$(11.89) \quad (-0.93)$$

We first need to test H_0: $\beta_2 = 0$ against H_1: $\beta_2 \neq 0$. The test statistic is

$$t_{n-k} = \frac{b_2}{s(b_2)} = \frac{0.263}{0.022} = 11.89.$$

Here, $s(b_2)$ is the standard error of b_2. There are $n - k = 10 - 3 = 7$ degrees of freedom. The 0.05 critical value for t_7 is 2.365 (two-tailed), and we reject H_0. This means that S has a significant effect Y.

Similarly, to test H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$, the corresponding t value is

$$t_7 = -0.770/0.826 = -0.93.$$

We do reject H_0 and conclude that P does not affect Y. In this case, we will drop P from the model and re-estimate it, as shown below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.990332
R Square	0.980757
Adjusted R Square	0.978351
Standard Error	0.718971
Observations	10

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	210.7647	210.7647	407.7327
Residual	8	4.135349	0.516919	
Total	9	214.9		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	–2.84593	0.489176	–5.8178	0.000397
X Variable 1	0.246364	0.012201	20.19239	3.78E-08

The model is

$$Y = -2.84 + 0.246S + e \qquad R^2 = 0.981.$$
$$(20.19)$$

The estimated value of b_2 differs only slightly from the previous model, and the t statistic is highly significant. The value of R^2 is also similar.

Forecasting

How do we interpret the results? On average, the cost over-run is given by $0.246S - 2.84$ because the average value of e is zero. It is possible to forecast the expected percentage cost overrun for a project of size \$55 m, that is

$$Y_F = -2.84 + 0.246(55) = 10.69\%.$$

Finally, note that \$55 m is within the range of project size values. Out-of-sample forecasts may not be reliable; for example, if the sample consists of small projects, it is unwise to use it to project the cost overrun for large projects.

If desired, it is possible to estimate the confidence interval of the forecast. The formula may be found in standard regression texts (e.g. Montgomery et al., 2001).

References

Birkes, D. and Dodge, Y. (1993) *Alternative methods of regression*. New York: Wiley.

Draper, N. and Smith, H. (1998) *Applied regression analysis*. New York: Wiley.

Montgomery, D., Peck, E., and Vining, G. (2001) *Introduction to linear regression analysis*. New York: Wiley.

CHAPTER 15

Quantitative Data Analysis III: Regression Data (Part II)

Dummy Variables

In Chapter 13, the dummy variable G is used to capture the effects of gender on test scores. If there are c categories, we need to use $c - 1$ dummy variables. For example, a house may face north, south, east, or west. Since there are 4 categories, we use 3 dummy variables — North, South, and East (Table 15.1). Observe that the second house faces West, the default direction.

The model is

$$P = \alpha + \beta\text{North} + \lambda\text{South} + \theta\text{East} + \Upsilon\text{Area} + \cdots + \varepsilon.$$

We use $c - 1$ dummy variables to represent c categories to avoid collinearity. Suppose a student's test score (S) depends on gender and monthly household income (X). If we use two variables, M and F, to represent gender (Table 15.2), the model is

$$S = \alpha + \beta M + \lambda F + \theta X + \varepsilon.$$

Table 15.1 Use of dummy variables in a regression.

House price ($m)	North	South	East	Area (m^2)	Other variables
1.56	1	0	0	150	…
1.60	0	0	0	160	…
2.00	0	1	0	200	…
…	…	…	…	…	…
1.40	0	0	1	140	…

Table 15.2 Illustration of collinearity.

S	*M*	*F*	*X* ($)
90	1	0	5,000
80	0	1	4,000
70	0	1	3,500
…	…	…	…
85	0	1	5,200

Observe that the sum of the values for *M* and *F* equals 1. The columns are collinear, which violates assumption A8 (see Chapter 14). To avoid this problem, we should replace *M* and *F* with a single variable *G* for Gender. Males are then coded 1, and females are coded 0, as shown in the table below.

S	*G*	*X* ($)
90	1	5,000
80	0	4,000
70	0	3,500
…	…	…
85	0	5,200

Interacting Variables

Let *Y* be a dichotomous variable where $Y = 1$ if a person has lung cancer and 0 otherwise. Let

$$Y = \alpha + \beta X + \lambda G + \theta XG + \varepsilon,$$

where *X* is the number of cigarettes smoked per day, *G* is a gender dummy variable ($G = 0$ for female and 1 for male), *XG* is *X* multiplied by *G* (an interacting term), and β, λ, and θ are the usual parameters. Then

$$E[Y] = \alpha + \beta X + \lambda G + \theta XG$$

since $E[\varepsilon] = 0$ by assumption. If $E(Y)$ is interpreted as the probability of having lung cancer, then it depends on *X*, *G*, and *XG*. To understand

the interacting term, we differentiate $E(Y)$ partially with respect to X so that

$$\frac{\partial E(Y)}{\partial X} = \beta + \theta G .$$

The change in $E(Y)$ resulting from the number of cigarettes smoked affects men and women differently. For men, it is given by $\beta + \theta$. For women, it is just β because $G = 0$. It does not make a difference only if $\theta = 0$.

Transformation of Nonlinear Functions

Recall from the previous chapter that the regression model is linear in the parameters. Often, it is possible to transform a nonlinear function into a linear one. For example, we can linearize the cost function

$$C = \alpha + \beta Q + \lambda Q^2$$

by letting $T = Q^2$ so that

$$C = \alpha + \beta Q + \lambda T.$$

Here, C is the cost of production, and Q is the output.

Next, consider a nonlinear multiplicative model

$$Z = AS^{\lambda}E^{\beta},$$

where Z is the wage rate, A is a constant, S is the years of schooling, and E is the years of working experience. To linearize it, we take natural logs on both sides so that

$$\log(Z) = \log(A) + \lambda\log(S) + \beta\log(E).$$

By adding an error term (ε), we can use ordinary least squares (OLS) to regress $\log(Z)$ against $\log(S)$ and $\log(E)$ by letting $z = \log(Z)$, $\alpha = \log(\mathrm{A})$, $s = \log(S)$, and $e = \log(E)$, that is,

$$z = \alpha + \lambda s + \beta e + \varepsilon.$$

To obtain A, we use the antilog function, that is, $A = \text{antilog}(\alpha)$. The raw data Z, S, and E must be converted into logs before estimating the model.

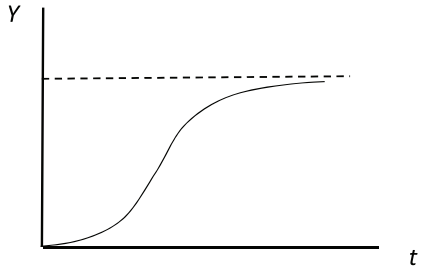

Figure 15.1 A logistic curve.

The logit function or "*S* curve" (Fig. 15.1) is given by

$$Y_t = (1 + e^{-(\alpha + \beta t)})^{-1}.$$

For example, it may represent how project cost (Y) varies with the duration of the project (t). To carry out the linear transformation, we first write it as

$$\frac{1}{Y_t} = 1 + e^{-(\alpha + \beta t)}$$

so that

$$\log\left(\frac{1}{Y_t} - 1\right) = -(\alpha + \beta t) = \lambda_1 + \lambda_2 t.$$

By letting the left-hand side be Z, we have a linear model

$$Z = \lambda_1 + \lambda_2 t.$$

Maximum Likelihood Estimation

In the previous section, we have seen how certain nonlinear functions may be transformed into linear equations. There are *intrinsically nonlinear equations* such as

$$Y = f(X, \alpha, \beta) = \alpha X^{\beta} + \varepsilon \tag{15.1}$$

that cannot be transformed into a linear model. There are two possible methods of estimating such models. We will discuss the maximum

likelihood estimation (MLE) here and nonlinear least squares (NLS) in the next section.

If the error terms are normally distributed, the probability density function (pdf) is

$$f(\varepsilon_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right),$$

where exp(.) is the exponential function, that is, $\exp(x) = e^x$, where e is the base of the natural logarithm. The errors are unobservable, so it is necessary to find the pdf of the observable Y values. From Equation (15.1),

$$\varepsilon_i = Y_i - \alpha X_i^{\beta}$$

so that

$$\frac{d\varepsilon_i}{dY_i} = 1.$$

The pdf of Y_i is

$$g(Y_i) = f(\varepsilon_i)\left|\frac{d\varepsilon_i}{dY_i}\right| = (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(Y_i - \alpha X_i^{\beta})^2}{2\sigma^2}\right).$$

Here, $|d\varepsilon_i/dY_i|$ is the *Jacobian* of the transformation. It scales the two density functions so that they are compatible.

For a sample of n observations $Y_1,\ldots, Y_n$, the *likelihood function* L is the product of the individual density functions so that

$$L = g(Y_1)\ldots g(Y_n).$$

This function is difficult to optimize because each $g(Y_i)$ contains a complicated exponential function. It is easier to maximize the log likelihood function

$$Q = \log L = -\left(\frac{n}{2}\right)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum(Y_i - \alpha X_i^{\beta})^2.$$

Since $\log(L)$ is a monotonically increasing function of L, maximizing $\log(L)$ is the same as maximizing L. To maximize Q, we set the partial derivatives to 0 and solve the resulting equations for α, β, and σ:

$$\frac{\partial Q}{\partial \alpha} = -\frac{1}{2\sigma^2}\sum 2(Y_i - \alpha X_i^{\beta})(-X_i^{\beta}) = 0;$$
$$\frac{\partial Q}{\partial \beta} = -\frac{1}{2\sigma^2}\sum 2(Y_i - \alpha X_i^{\beta})(-\alpha X_i^{\beta} \log X_i) = 0;$$
$$\frac{\partial Q}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (Y_i - \alpha X_i^{\beta})^2 = 0.$$

Simplifying, we get

$$\sum (Y_i - \alpha X_i^{\beta})(X_i^{\beta}) = 0;$$
$$\sum (Y_i - \alpha X_i^{\beta})(-\alpha X_i^{\beta} \log X_i) = 0;$$
$$\sum (Y_i - \alpha X_i^{\beta})^2 = n\sigma^2.$$

It is not easy to solve this set of equations. If $f(x)$ is a function of a single variable, Newton's method of computing the roots of the equation $f(x) = 0$ uses the following iteration:

$$\delta x_k = -\frac{f(x)}{f'(x)}.$$

For a set of equations $f_i(\mathbf{x}) = \mathbf{0}$, the iterative formula is similar except that the derivative is replaced by $\mathbf{J}_k$ and shifted to the left-hand side to avoid extensive computation of its matrix inverse:

$$\mathbf{J}_k \delta \mathbf{x}_k = -f(\mathbf{x}_k).$$

Here $\mathbf{J}_k$ is the Jacobian matrix evaluated at the kth iteration, and $\delta \mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$. The Jacobian matrix is

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}.$$

To illustrate how it works, consider solving the following set of equations:

$$f_1(\mathbf{x}) = x_1^2 + x_2 - 2 = 0;$$
$$f_2(\mathbf{x}) = x_1 + 2x_2^2 - 3 = 0.$$

The Jacobian matrix is

$$\begin{bmatrix} 2x_1 & 1 \\ 1 & 4x_2 \end{bmatrix}.$$

Let the initial guess be

$$\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Then, $\mathbf{J}_0 \delta\mathbf{x}_0 = -f(\mathbf{x}_0)$ is given by

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \delta\mathbf{x}_0 = -\begin{bmatrix} -2 \\ -3 \end{bmatrix}.$$

By Gauss elimination, the solution is

$$\delta\mathbf{x}_0 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

The new estimate is

$$\mathbf{x}_1 = \delta\mathbf{x}_0 + \mathbf{x}_0 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

Then, $\mathbf{J}_1 \delta\mathbf{x}_1 = -f(\mathbf{x}_1)$ is given by

$$\begin{bmatrix} 4 & 1 \\ 1 & 12 \end{bmatrix} \delta\mathbf{x}_1 = -\begin{bmatrix} 5 \\ 17 \end{bmatrix}.$$

By Gauss elimination,

$$\delta\mathbf{x}_1 = \begin{bmatrix} -0.91 \\ -1.34 \end{bmatrix}.$$

The iteration continues until convergence to the solution $(1, 1)^{T}$.

MLE estimators have desirable large sample properties if they satisfy certain technical regularity conditions (Rohatgi, 1976). For large samples, they are

- consistent (unbiased);
- efficient (i.e. have minimum variance); and
- normally distributed.

We can summarize these properties by writing $\mathbf{b} \approx N(\boldsymbol{\beta}, \mathbf{I}(\mathbf{b})^{-1})$ where $\mathbf{b}$ is the MLE estimator, $\boldsymbol{\beta}$ is the vector of parameters, and $\mathbf{I}(\mathbf{b})$ is the *Fisher information matrix*. Each element of $\mathbf{I}(\mathbf{b})$ is given by

$$I_{ij} = -\frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j}.$$

Note that $Var(\mathbf{b})$ is the inverse of the Fisher information matrix. The latter is a measure of the curvature of the log likelihood function.

Where possible, we avoid using maximum likelihood estimators because of the need to find the first-order optimal conditions as well as the necessity of solving a set of nonlinear equations. However, in instances where there are significant departures from OLS assumptions, researchers may turn to maximum likelihood estimators. Further, there are many statistical software that takes care of the tedious computations.

Nonlinear Least Squares

The nonlinear least squares (NLS) method linearizes a nonlinear model using Taylor series approximation about an initial set of parameter values and then applies OLS to the linear model. The estimated parameters are updated by applying OLS iteratively until the solution converges based on some criteria.

The Taylor series approximation for small h in the neighborhood of x is given by (Atkinson, 1989)

$$f(x + h) - f(x) = hf'(x).$$

The linearized multivariate equation is

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = \mathbf{h}^{T}\nabla f,$$

where ∇f is the gradient vector of partial derivatives, that is,

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_m} \end{bmatrix}.$$

We rewrite Equation (15.1) as

$$Y_i = \alpha X_i^{\beta} + \varepsilon_i$$

and the linearized model is

$$\begin{bmatrix} Y_1 - \alpha X_1^{\beta} \\ \vdots \\ Y_n - \alpha X_n^{\beta} \end{bmatrix} = \begin{bmatrix} \frac{\partial Y_1}{\partial \alpha} & \frac{\partial Y_1}{\partial \beta} \\ \vdots & \vdots \\ \frac{\partial Y_n}{\partial \alpha} & \frac{\partial Y_n}{\partial \beta} \end{bmatrix} \begin{bmatrix} \delta\alpha \\ \delta\beta \end{bmatrix} + \varepsilon.$$

We evaluate the model at initial values α_0 and β_0 and find OLS estimates for $\delta\alpha$ and $\delta\beta$. These are then used to update the initial guess using $\alpha = \delta\alpha + \alpha_0$ and $\beta = \delta\beta + \beta_0$. The process is then iterated until convergence. It may not converge to a solution if

- the initial guess value is far from the optimal point;
- the surface is "flat" near the optimal point; or
- the surface is too "sharp" near the optimal point, resulting in oscillations between iterations.

The last two cases are shown in Fig. 15.2. Generally, NLS is easier to implement than MLE, and convergence tends to be relatively fast (Seber and Wild, 2003).

Non-normal Errors

For the linear regression model $Y = \alpha + \beta X + \varepsilon$, it is assumed that $\varepsilon_i \sim N(0, \sigma^2)$ so that diagnostic tests based on normal distribution, such as *t*, chi-square, and *F* tests, may be constructed. If *Y* is not normally distributed, then ε is also not normally distributed because *X* is fixed in repeated sampling. In such cases, it may be possible to transform *Y* so that it is approximately normal.

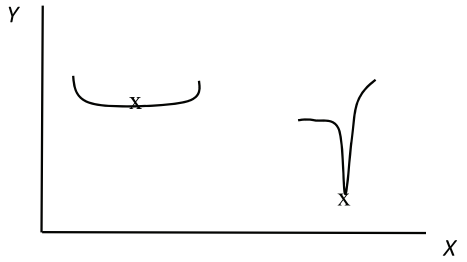

Fig. 15.2 Non-convergence for flat or sharp surfaces.

Box and Cox (1964) proposed a family of transformations:

$$\begin{aligned} Y^{(\lambda)} &= Y^{\lambda} && \text{if } \lambda \neq 0; \\ &= \log Y && \text{if } \lambda = 0. \end{aligned}$$

The second line is required because the first transformation is trivial if $\lambda = 0$, that is, $Y^0 = 1$. Some texts use

$$Y^{(\lambda)} = (Y^{\lambda}-1)/\lambda \qquad \text{if } \lambda \neq 0$$

instead of simply $Y^{(\lambda)} = Y^{\lambda}$ if $\lambda \neq 0$ because $(Y^{\lambda}-1)/\lambda$ tends towards $\log Y$ as λ tends towards 0. We will not use this version here.

The problem is to find λ. For example, if $\lambda = 2$, the model is

$$Y^2 = \alpha + \beta X + \varepsilon.$$

If $\lambda = 0$, the model is

$$\log Y = \alpha + \beta X + \varepsilon.$$

A search procedure may be used to select the value of λ that minimizes the sum of squares of residuals (*RSS*). For instance, we may start with $\lambda = -1$, regress

$$Y^{-1} = \alpha + \beta X + \varepsilon,$$

and compute $RSS(\lambda = -1)$. The next value may be $\lambda = 0$ so that we regress

$$\log Y = \alpha + \beta X + \varepsilon$$

and compute $RSS(\lambda = 0)$, and so on. By plotting *RSS* against λ, we can find the value of λ that gives the minimum *RSS*.

Outliers

The regression line is sensitive to outliers because the OLS estimator minimizes the sum of squares of residuals (*RSS*). This problem is particularly severe in small samples. In Fig. 15.3, points *A* and *B* are *outliers* because they are far from the main data cluster. However, *A* does not significantly affect the linear regression line if we exclude the bottom data point from the regression. In contrast, including *B* will significantly alter the slope of the regression line. It is both an outlier and an *influential point*. However, if the true relation is a nonlinear curve, then *A* is the influential point. In summary, with only a few data points, it is not easy to identify outliers and influential points.

Some outliers are gross errors, such as if we key in 18 instead of 81. However, there may be outliers that are not gross errors, and it is inappropriate to remove them just because they do not fit the theory.

If there are many data points and variables, it is difficult to identify outliers by inspection. We have to examine the residuals for outliers formally. Recall from Chapter 14 that

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = [\mathbf{I} - \mathbf{H}]\mathbf{y}. \tag{15.2}$$

In the above expression, we have used the normal equations $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, $\mathbf{I}$ is the identity matrix, and $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the *hat matrix*. Hence,

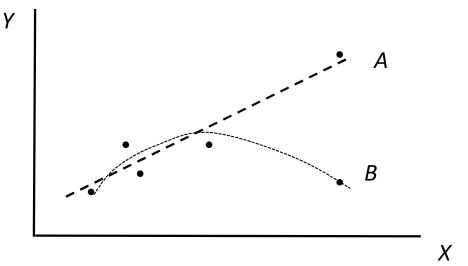

Fig. 15.3 Outliers and influential points.

$$Var(\mathbf{e}) = Var([\mathbf{I} - \mathbf{H}]\mathbf{y}) = [\mathbf{I} - \mathbf{H}]Var(\mathbf{y}|\mathbf{X})[\mathbf{I} - \mathbf{H}]^{\mathrm{T}} = \sigma^2[\mathbf{I} - \mathbf{H}]. \quad (15.3)$$

The derivation uses the relation

$$Var(\mathbf{By}) = \mathbf{B}Var(\mathbf{y})\mathbf{B}^{\mathrm{T}},$$

where **B** is a matrix of constants and $Var(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{I}$. Finally, $\mathbf{I} - \mathbf{H}$ is an *idempotent matrix*. A matrix is idempotent if $\mathbf{A}^2 = \mathbf{A}$. Thus, $[\mathbf{I} - \mathbf{H}][\mathbf{I} - \mathbf{H}]^{\mathrm{T}} = [\mathbf{I} - \mathbf{H}]$ because $[\mathbf{I} - \mathbf{H}]$ is a symmetric matrix, that is, $[\mathbf{I} - \mathbf{H}] = [\mathbf{I} - \mathbf{H}]^{\mathrm{T}}$.

Since $Var(\mathbf{e}) = \sigma^2[\mathbf{I} - \mathbf{H}] \neq \sigma^2\mathbf{I}$, we conclude that $Cov(e_i, e_j) \neq 0$, that is, the residuals are correlated. For each residual, we can write Equation (15.3) as

$$Var(e_i) = \sigma^2[1 - h_{ii}],$$

where h_{ii} is the ith diagonal element of **H**, called the *leverage*. Thus

$$e_i^* = \frac{e_i}{s\sqrt{(1 - h_{ii})}}$$

is called the *standardized residual* and a data point may be an outlier if e_i^* exceeds 2.5. Note that e_i^* does not follow the Student t distribution because e_i and s are not independent (see Equation (14.10)).

An alternative way of identifying outliers is to use the *Studentized residual*. Here, we replace s with $s_{(i)}$, the standard deviation of the residuals from a regression without the ith outlier. It provides a better measure of σ than s because an outlier will inflate the value of s. We can observe this in Fig. 15.2, where the inclusion of the influential point in the regression will inflate the value of *RSS*.

A simpler approach to detect outliers is to identify an observation as influential if

$$h_{ii} > 3k/n,$$

where, as before, k is the number of parameters and n is the number of observations. The logic for this choice is that the average value of h_{ii} is k/n. To see this, note that

$$\begin{aligned} Trace(\mathbf{H}) = Trace(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}) &= Trace(\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}) \\ &\quad \text{using } Trace(\mathbf{AB}) = Trace(\mathbf{BA}) \\ &= Trace(\mathbf{I}) = k. \end{aligned}$$

The *Trace*(.) operator gives the sum of the diagonals of a square matrix. If the sum of the diagonals of **H** is k, then its average value is k/n. For a more detailed discussion on outlier detection, see (Belsley et al., 1980) and (Barnett and Lewis, 1994).

Testing Restrictions on Parameters

In Fig. 15.4, we show two possible regression lines

$$Y_1 = \alpha_1 + \beta_1 X + \varepsilon_1; \text{ and}$$
$$Y_2 = \alpha_2 + \beta_2 X + \varepsilon_2.$$

There appears to be a break in the data, such as whether sales have fallen or whether a new safety policy has reduced workplace accidents. The above set of equations is called the *unrestricted* model.

The *restricted* model is the single dashed regression line

$$Y = \alpha + \beta X + \varepsilon.$$

The null hypothesis is that there is no structural change, that is,

$$H_0: \alpha_1 = \alpha_2, \text{ and } \beta_1 = \beta_2.$$

There are $d = 2$ restrictions in the null hypothesis. If the restrictions are valid, both models should fit the data well. Hence, $RSS_R - RSS_U$ should be "small," where RSS_U is the residual sum of squares from the unrestricted model, and RSS_R is the residual sum of squares from the restricted

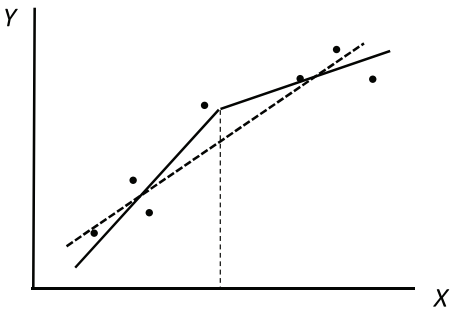

Fig. 15.4 Possible structural change.

model. To remove its dependence on the units of measurement, we can divide the difference by RSS_U, together with their appropriate degrees of freedom. Hence, it is possible to show that (Johnston and DiNardo, 1997)

$$\frac{(RSS_R - RSS_U)/d}{RSS_U/(n-k)} \sim F(d, n-k). \tag{15.4}$$

As before, n is the total number of observations, and k is the number of parameters in the unrestricted model. $F(.)$ is the F distribution with d and $n - k$ degrees of freedom, respectively. A large F value greater than the 0.05 critical value implies that the difference in RSS is "large," and we should reject H_0.

For example, if the results for a sample of 30 observations are

$$\begin{aligned} Y_1 &= a_1 + b_1X + e_1 \qquad & RSS &= 10 \text{ (observations 1 to 15)} \\ Y_2 &= a_2 + b_2X + e_2 \qquad & RSS &= 16 \text{ (observations 16 to 30)} \\ Y &= a + bX + e \qquad & RSS_R &= 40 \text{ (30 observations)} \end{aligned}$$

Then $RSS_U = 10 + 16 = 26$ and, from Equation (15.4),

$$\frac{(40-26)/2}{26/(30-4)} = 7.$$

The 0.05 critical value for $F(2, 26)$ is 3.39, and we reject H_0. There is sufficient evidence of a structural change.

We will use Equation (15.4) in subsequent chapters to test other restrictions. For other examples of such restrictions on functional forms, scale economies, and so on, see (Berndt, 1991).

Heteroscedasticity

Recall from Chapter 14 that the assumption of homoscedasticity, that is, $\varepsilon_i \sim N(0, \sigma^2)$ or, in matrix form, $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, may not hold in practice. Consider the model

$$Y = \alpha + \beta X + \lambda L + \varepsilon, \tag{15.5}$$

where Y is the rate of profit, X is the firm size, and L is the labor input. Because larger firms tend to have greater variability in the rate of profit

compared to smaller ones, the errors may be bigger for larger firms (Fig. 14.2), that is, they are heteroscedastic.

There are other ways to detect heteroscedasticity besides plotting Y against each independent variable (Maddala, 1986). In the Lagrange Multiplier or *LM* test, we regress the square of the residuals against the independent variables:

$$e_i^2 = \alpha_0 + \alpha_1 X_i + \alpha_2 L_i + u_i.$$

The null hypothesis is that the $\alpha_1 = \alpha_2 = 0$, and the alternative hypothesis is that at least one of them is non-zero. The test statistic is

$$LM = nR^2 \sim \chi^2(k).$$

For example, if $n = 20$ observations and $R^2 = 0.7$ from the residual regression, then $nR^2 = 14$. The 0.05 critical value for $\chi^2(3)$ is 7.81, so we reject H_0 and conclude that heteroscedasticity is present in the data.

To deal with heteroscedasticity, we may deflate (divide) Y by L, and Equation (15.5) becomes

$$Y/L = \alpha + \beta X + \varepsilon.$$

The dependent variable is now Y/L, or profit per worker. This transformation "compresses" the errors to reduce the heteroscedasticity. Since L is no longer an independent variable, λ is no longer relevant.

Alternatively, we may take logs on both sides of Equation (15.5) and regress

$$\log(Y) = \alpha + \beta \log(X) + \lambda \log(L) + \varepsilon$$

and obtain a similar compressive effect on the errors by virtue of the compressive property of the log function.

We may also model the heteroscedasticity explicitly. If the errors are heteroscedastic, then

$$Var(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}] = \sigma^2 \boldsymbol{\Omega} = \boldsymbol{\Sigma},$$

where $\boldsymbol{\Omega}$ (or $\boldsymbol{\Sigma}$), the covariance matrix, is no longer an identity matrix. The matrices $\sigma^2\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ are equivalent, and which one to use is a matter of convenience. The diagonal elements of $\boldsymbol{\Omega}$ (or $\boldsymbol{\Sigma}$) are no longer ones but

individual variances (i.e. σ_i^2). These variances are unequal because of the presence of heteroscedasticity.

In practice, we may not know the elements of $\boldsymbol{\Sigma}$ *a priori* and have to estimate it from the residuals of an OLS regression. This is the generalized least squares (GLS) model.

In weighted least squares (WLS), we assume that the off-diagonal elements of $\boldsymbol{\Sigma}$ are zero. Consider a non-singular transformation matrix $\mathbf{T}$ such that

$$\mathbf{Ty} = \mathbf{TX}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\varepsilon}.$$

That is,

$$\mathbf{z} = \mathbf{L}\boldsymbol{\beta} + \mathbf{u}, \tag{15.6}$$

where $\mathbf{z} = \mathbf{Ty}$, $\mathbf{L} = \mathbf{TX}$, and $\mathbf{u} = \mathbf{T}\boldsymbol{\varepsilon}$. Then,

$$E[\mathbf{u}\mathbf{u}^{\mathrm{T}}] = E[\mathbf{T}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}] = \sigma^2\mathbf{T}\boldsymbol{\Omega}\mathbf{T}^{\mathrm{T}}.$$

We choose $\mathbf{T}$ such that $\mathbf{T}\boldsymbol{\Omega}\mathbf{T}^{\mathrm{T}} = \mathbf{I}$ so that OLS may be applied to Equation (15.6). The normal equations are

$$\mathbf{b}_{\mathrm{GLS}} = (\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{z} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{y} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{15.7}$$

If $\boldsymbol{\Omega}$ is a diagonal matrix, then $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ is the weight matrix in the weighted least squares procedure. Finally,

$$Var(\mathbf{b}_{\mathrm{GLS}}) = \sigma^2(\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1} = \sigma^2(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}.$$

In summary, if the errors are heteroscedastic, we should use the GLS estimator in Equation (15.7). However, $\boldsymbol{\Omega}$ or $\boldsymbol{\Sigma}$ is often unknown, and we may estimate it from the residuals of an OLS regression. A simpler procedure is to use WLS.

Multicollinearity

Recall that it is not possible to solve the normal equations in Equation (14.8) if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible. Its determinant is zero or, equivalently, the matrix is singular.

Multicollinearity refers to cases where at least two columns of $\mathbf{X}$, and hence $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, are highly correlated so that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is nearly singular or ill-conditioned. For instance, if we regress the house price on the land area (*LA*) and the number of rooms (*NR*), then *LA* and *NR* are likely to be highly correlated because they provide roughly the same amount of

information. If $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill-conditioned, the estimated parameters will have large standard errors because of Equation (14.9) despite the high R^2. This is analogous to computing $1/y = y^{-1}$ when y is a very small number.

Generally, we suspect multicollinearity if the absolute values of the correlations are higher than 0.8 or, as alluded earlier, $Var(\mathbf{b})$ is large despite the high R^2. An alternative way is to compute the variance inflation factor

$$VIF\,(b_j) = \frac{1}{1 - R_j^{2'}}$$

where R_j^2 is R^2 obtained by regressing X_j on all other independent variables in the model. If R_j^2 is close to 1, *VIF* will be large. A *VIF* value greater than 10 indicates that multicollinearity may be present. Finally, we may inspect the condition index

$$\sqrt{\frac{\lambda_{Max}}{\lambda_{Min}}},$$

the square root of the ratio of the largest to smallest eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. If $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill-conditioned, the smallest eigenvalue will be close to zero.

Since eigenvalues and eigenvectors play a fundamental role in linear algebra, a brief exposition will be given here. If

$$\mathbf{Ax} = \lambda\mathbf{x},$$

we call λ an eigenvalue of a square matrix $\mathbf{A}$ and $\mathbf{x}$ is the corresponding eigenvector. Thus $\mathbf{A}$ transforms $\mathbf{x}$ to $\lambda\mathbf{x}$, that is, it stretches the vector in the same direction by factor λ. To find the eigenvalues, we rewrite it as

$$\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0}.$$

That is,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0},$$

where $\mathbf{I}$ is the identity matrix consisting of ones along the diagonal and zeroes elsewhere. The equation has non-trivial solutions if $(\mathbf{A} - \lambda\mathbf{I})$ is not invertible. Hence, the determinant is zero:

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

For example, if

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix},$$

then $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ gives

$$\det \begin{bmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{bmatrix} = 0.$$

Thus $\lambda = 2, -1$, and -1. Note there is a repeat root, -1. To find the corresponding eigenvector for $\lambda = 2$, we solve

$$\mathbf{Ax} = 2\mathbf{x}$$

for $\mathbf{x}$. This gives

$$t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},$$

where t is a parameter as an eigenvector. Because eigenvectors are not unique, it is usual to normalize them as unit vectors.

Similarly, the eigenvector corresponding to the repeat root $\lambda = -1$ is obtained by solving

$$\mathbf{Ax} = -\mathbf{x}.$$

Not surprising, this gives two eigenvectors, namely,

$$t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \text{ and } s \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix},$$

where t and s are parameters.

Once we have detected the presence of multicollinearity, there are several remedies such as

- leaving things alone because there is no point in trying to estimate more precisely than reality allows;
- redesigning the model to remove highly correlated variables;
- dropping one of the correlated variables; and
- using ridge regression (Hoerl and Kennard, 1970).

The first approach assumes that multicollinearity is inherent in the sample. Redesigning the model may improve matters. For instance, if the two independent variables are similarly affected by inflation, they tend to be highly correlated. A simple redesign of the model that may remove multicollinearity is to deflate both variables by an appropriate inflation index. If there are many independent variables, dropping one of the highly correlated variables is a possible option because it does not add much information.

Ridge regression adds a small number c to the diagonals of $\mathbf{X}^T\mathbf{X}$ before inverting it to solve the normal equations, that is,

$$\mathbf{b}_R = (\mathbf{X}^T\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{15.8}$$

where $\mathbf{b}_R$ is the ridge estimator and $\mathbf{I}$ is the identity matrix. It is obtained by minimizing the Lagrangean

$$L = \mathbf{e}^T\mathbf{e} + c(\mathbf{b}^T\mathbf{b} - \varphi).$$

That is, we minimize $\mathbf{e}^T\mathbf{e}$ subject to $\mathbf{b}^T\mathbf{b} = \varphi$, the square of the Euclidean norm of the solution vector. Thus,

$$\partial L/\partial \mathbf{b} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} + 2c\mathbf{b} = \mathbf{0}$$

and solving it gives Equation (15.8).

The value of c should be as small as possible, that is, it should be just sufficient to stabilize the values of the estimated regression coefficients (Fig. 15.5). This is because the ridge estimator is biased, that is,

$$E[\mathbf{b}_R] = E[(\mathbf{X}^T\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}] = E[(\mathbf{X}^T\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] \neq \boldsymbol{\beta}$$

unless $c = 0$, which returns us to the OLS estimator $\mathbf{b}$. The greater the value of c, the larger is the bias. Another disadvantage of the ridge estimator is that, unlike OLS, it is not easy to evaluate

$$Var[\mathbf{b}_R] = Var[(\mathbf{X}^T\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}].$$

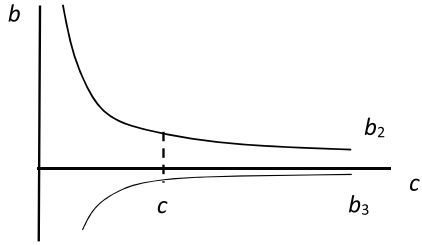

Fig. 15.5 Selection of c.

Observe that

$$\mathbf{b}_R = (\mathbf{X}^T\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{Z}\mathbf{b},$$

where **b** is the OLS estimator and

$$\mathbf{Z} = (\mathbf{X}^T\mathbf{X} + c\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}.$$

Since it can be shown that the Euclidean norm of **Zb** is smaller than that of **b** (Gruber, 1998), the ridge estimator is also called the *shrinkage estimator*. It shrinks the length of **b**, the OLS estimator.

Logistic Regression

The dependent variable (Y) may be a binary variable, such as whether a person has lung cancer, owns a car, or is a homeowner. In the cancer case, we can model it as

$$Y = \alpha + \beta X + \lambda G + \theta A + \varepsilon,$$

where $Y = 1$ if a person has lung cancer and 0 otherwise. The variable X is the number of cigarettes smoked per day, G is gender, and A is age. For brevity, we will drop the gender and age variables and consider the simpler model

$$Y = \alpha + \beta X + \varepsilon. \tag{15.9}$$

The principles are the same if there are more independent variables such as lifestyle and stress level.

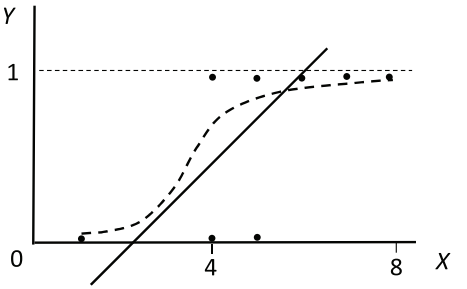

Fig. 15.6 Data on incidence of lung cancer.

In Fig. 15.6, we plot the data for a sample of eight persons. Five have lung cancer (that is, $Y = 1$), and three do not have lung cancer. The solid regression line fits the data poorly, resulting in low R^2. The residuals are heteroscedastic; for example, the two persons who smoke four cigarettes have large residuals compared to the rest of the group. Finally, the regression line exceeds the boundaries of 0 and 1, which is difficult to interpret.

A better option is to fit a logistic curve

$$Y = \pi = \frac{1}{1 + e^{-(\alpha + \beta X)}}.$$

For any X value, we interpret the corresponding Y value on the curve as the probability (π) of having lung cancer. To transform it into a linear model, we use simple algebra and rewrite it as

$$\frac{\pi}{1 - \pi} = e^{(\alpha + \beta X)}.$$

Taking logs on both sides gives

$$\log\left(\frac{\pi}{1 - \pi}\right) = P = \alpha + \beta X .$$

By adding an error term, we can regress P against X using OLS if we have data on P which, in turn, depends on π.

Table 15.3 Data for logistic regression.

X	No. of persons	No. with lung cancer	π	P
1	100	0	0	Undefined
2	100	0	0	Undefined
3	100	1	0.01	–1.9956
4	100	2	0.02	–1.6902
5	100	2	0.02	–1.6902
6	100	4	0.04	–1.3802
7	100	6	0.06	–1.1950
8	100	6	0.06	–1.1950

To estimate π, we use grouped data (Table 15.3). The first two data points are undefined because

$$P = \log\left(\frac{\pi}{1-\pi}\right) = \log\frac{0}{(1-0)}.$$

Hence, we can only use data from $X = 3$ onwards for the regression. A large sample is required to obtain reliable estimates of π for different categories of smokers.

If the sample size is small, we may use the maximum likelihood estimation (MLE) method, which uses individual rather than grouped data. The likelihood function is

$$L = Prob(Y_1,\ldots, Y_n) = f(Y_1)f(Y_2)\ldots f(Y_n).$$

For the logit model,

$$f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}.$$

This implies

$$f(Y_i = 1) = \pi_i(1-\pi_i)^0 = \pi_i$$
$$f(Y_i = 0) = \pi_i^0(1-\pi_i)^1 = 1 - \pi_i$$

as desired. That is, π_i represents the probability that the ith person who smokes X_i cigarettes a day has lung cancer. Substituting $f(Y_i)$ into L, the log likelihood function is

$$Q = \sum Y_i \log(\pi_i) + \sum (1 - Y_i) \log(1 - \pi_i).$$

Substituting for

$$1 - \pi_i = \frac{1}{1 + e^{(\alpha + \beta X_i)}},$$

we have

$$Q = \sum Y_i(\alpha + \beta X_i) + \sum \log\left(\frac{1}{1 + e^{\alpha + \beta X_i}}\right)$$

Differentiating Q with respect to the parameters and setting the results to zero gives

$$\frac{\partial Q}{\partial \alpha} = \sum Y_i - \sum \frac{Z_i}{1 + Z_i} = 0;$$
$$\frac{\partial Q}{\partial \beta} = \sum X_i Y_i - \sum \frac{X_i Z_i}{1 + Z_i} = 0.$$

For notational simplicity, I have used the expression

$$Z_i = e^{\alpha + \beta X_i}.$$

The solutions to these nonlinear equations are the maximum likelihood estimates. There are many statistical software that provides maximum likelihood estimates for logistic regression.

Since Y_i takes the value of 1 or 0, the coefficient of determination R^2 is lower in logistic regression. Some researchers have proposed alternative measures such as

$$R^2 = 1 - \left[\frac{L(0)}{L(\boldsymbol{b}^*)}\right]^{\frac{2}{n}}.$$

Here, $L(0)$ is the value of the likelihood function for the null model without independent variables, and $L(\mathbf{b}^*)$ is the value of the likelihood function using the estimated coefficients (Cox and Snell, 1989).

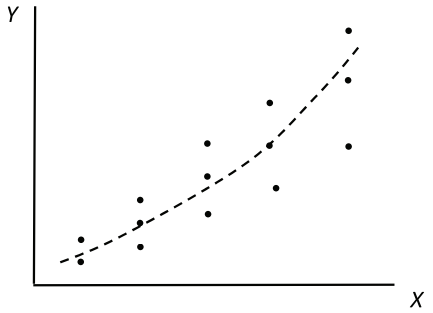

Fig. 15.7 Daily patient admissions to hospitals.

Poisson Regression

Recall that in logistic regression, the dependent variable Y is binary and takes the value of 0 or 1. In Poisson regression, Y is count data and takes only non-negative integer values. For example, Y may be the daily patient admissions to different hospitals (Fig. 15.7). The model is

$$Y = \alpha + \beta X + \varepsilon. \tag{15.10}$$

Here, X is the hospital size, and it is possible to add other independent variables such as the location of hospitals. If we assume that

$$Y = \exp(\alpha + \beta X + \varepsilon),$$

then

$$\log(Y) = \alpha + \beta X + \varepsilon.$$

As shown in Fig. 15.7, the errors are heteroscedastic for the count data. For the bigger hospitals, the variations in daily admissions are much larger than for smaller hospitals. This means that OLS will not be efficient, that is, $Var(b)$ will be large. One alternative is to use the Poisson regression.

The Poisson distribution or probability mass function is given by

$$P(Y = r) = \frac{\lambda^r e^{-\lambda}}{r!}. \tag{15.11}$$

Here, Y takes values 0, 1, 2, and so on, and λ is the mean and variance of Y. For example, if the mean is 9 for a small hospital, the variance is also 9, and the standard deviation is 3. If the mean is 64 for a larger hospital, the variance is also 64 and the standard deviation is 8. Observe that the standard deviation gets larger as the mean increases, which is what we desire (see Fig. 15.7). The expression $r!$ is the product of $r(r-1)...1$. For example, $3! = 3(2)(1) = 6$. Euler's number e is approximately 2.71828. As an example, if $\lambda = 9$, the probability of four admissions on a given day is

$$P(Y=4) = \frac{9^4 e^{-9}}{4!} = 0.033.$$

We are now ready to use the method of maximum likelihood to estimate the parameters in Equation (15.10). The likelihood function is

$$L = Prob(Y_1,\ldots, Y_n) = f(Y_1)f(Y_2)\ldots f(Y_n).$$

Each probability function $f(.)$ follows the Poisson distribution. The log likelihood function is

$$Q = \log L = \sum Y_i(\alpha + \beta X_i) - \sum(\alpha + \beta X_i) - \sum \log(Y_i!).$$

As before, we maximize Q by equating $\partial Q/\partial\alpha = 0$ and $\partial Q/\partial\beta = 0$ and solving the set of equations using Newton's iterative method. In practice, it is easier to use a statistical software package with Poisson regression.

Measurement Errors

Consider the model

$$Y = \alpha + \beta X + \varepsilon$$

and let

$$X_m = X + u,$$

where X_m is the measured variable, X is the true variable, and u is another error term. Thus

$$Y = \alpha + \beta(X_m - u) + \varepsilon = \alpha + \beta X_m + (\varepsilon - \beta u).$$

Since $X_m = X + u$, X_m depends on u. Hence X_m and $(\varepsilon - \beta u)$ are correlated and

$$E[\mathbf{b}] = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} \neq \boldsymbol{\beta}.$$

The OLS estimator is biased if X is correlated with the error term. One solution to this problem is to replace X with an *instrumental variable* Z that is correlated with X but uncorrelated with ε. In practice, it is difficult to find Z.

If there is another independent variable W, we can use the *2-stage least squares* estimator. In the first stage, we use OLS to estimate

$$\widehat{X} = a + bZ + cW.$$

In the second stage, we regress

$$Y = \alpha + \beta\widehat{X} + \varepsilon.$$

Again, the main difficulty is finding an appropriate instrumental variable Z.

So far, we have discussed the measurement error in X. If there is a measurement error in Y, it is not a major issue because the error ends up as part of the model error term. That is, if $Y_m = Y + v$, then

$$Y_m - v = \alpha + \beta X + \varepsilon$$

and

$$Y_m = \alpha + \beta X + (\varepsilon + v).$$

In summary, a measurement error in Y is not an issue. For measurement error in X, if there is only one independent variable, we may replace it with an instrument variable Z. If there is more than one independent variable, we use 2-stage least squares estimation. In both cases, it is, unfortunately, difficult to find an instrument variable (Bowden and Turkington, 1985).

Omitted Variable

There are two types of omitted variables. The first type involves an independent variable (X_j) and is straightforward. It occurs because we do not

have data on X_j, or we are ignorant and unconsciously omit X_j from the regression. For example, we may not have included the orientation of a house or the house number in a hedonic house price model.

The second type of omission is more serious. For the model

$$Y = \alpha + \beta X + \varepsilon,$$

there may be a variable H that affects both Y and X. If H affects Y, then it also affects ε through the equation above. Since H also affects X, it follows that X and ε are correlated. We then have the same problem as the measurement error in X. Again, it is difficult to find an appropriate instrumental variable.

As an example, if Y is the test score and X is the class size, we expect bigger classes to have lower test scores because it is presumably harder to teach a bigger class, particularly in elementary schools (Mishel and Rothstein, 2002; Chingos, 2013). Hence, reducing the class size is a popular education policy among politicians, educators, and parents, but unfortunately, it is costly because of the substantial increase in staffing.

Is there a variable that affects both the test score and class size? One such variable is family income. Higher-income families will tend to send their children to better schools with smaller class sizes. Such students will also tend to do better academically. The lesson here is, by looking at the equation above, we need to ask if there is a variable that affects both Y and X. If it exists, there is an omitted variable bias. Consequently, X and ε will be correlated.

Reverse Causality

In the linear regression model $Y = f(X)$, we assume that X affects Y and not the other way round. We say that X is *exogenous* or determined outside the system. However, if Y also affects X, there is reverse causality, and OLS will yield inconsistent estimates. Again, X and ε will be correlated.

Suppose we use OLS to estimate the model

$$Y = a + bX + cG. \tag{1}$$

Let us assume that we are interested in the coefficient of G. If Y also affects X, then there is a second equation

$$X = d + eY. \tag{2}$$

If we substitute X into the first equation, we have

$$Y = a + b(d + eY) + cG.$$

After some rearranging, we have,

$$Y = f + gG, \tag{3}$$

where

$$f = \frac{a}{1 - be};$$
$$g = \frac{c}{1 - be}.$$

There are two inconsistent estimates for the coefficient of G, namely c (in Equation (1)) and g (in Equation (3)).

As an example of reverse causality, if Y is a person's state of mental health, X is a dummy variable on whether a person is currently working, and G is gender, then X and Y may affect each other. Unemployment can affect a person's mental state, and this, in turn, may affect the decision on whether to seek employment.

If reverse causality is present, we should estimate the two equations as a system (Johnston and NiNardo, 1997), that is, we write the *structural equations* as

$$Y = a + bX + cG + u \tag{1}$$
$$X = d + eY + v. \tag{2}$$

We can rewrite it as

$$\begin{bmatrix} 1 & -b \\ e & 1 \end{bmatrix} \begin{bmatrix} Y \\ X \end{bmatrix} + \begin{bmatrix} -c & 0 \\ 0 & -d \end{bmatrix} \begin{bmatrix} G \\ 1 \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}.$$

In matrix form, we have

$$\mathbf{By} + \mathbf{Tx} = \mathbf{u}.$$

This arrangement is elegant and easier to manipulate because it puts variables in **y**, **x**, and **u** as separate vectors containing the endogenous variables, exogenous variables, and error terms, respectively.

Solving for **y**, the *reduced form* is

$$\mathbf{y} = -\mathbf{B}^{-1}\mathbf{T}\mathbf{x} + \mathbf{B}^{-1}\mathbf{u} = \boldsymbol{\pi}\mathbf{x} + \mathbf{v},$$

where $\boldsymbol{\pi} = -\mathbf{B}^{-1}\mathbf{T}$ and $\mathbf{v} = \mathbf{B}^{-1}\mathbf{u}$. Thus,

$$Y = \pi_{11}G + \pi_{12} + v_1$$
$$X = \pi_{21}G + \pi_{22} + v_2.$$

The *identification problem* refers to whether we can recover **B** and **T** from $\boldsymbol{\pi}$ by using the order conditions (Table 15.4). H is the number of equations in the system (= 2 here), and F is the number of excluded pre-determined variables in each equation. A structural equation is

not identified if $F < H - 1$,
identified if $F = H$, and
over-identified if $F > H - 1$.

Recall that the structural equations are:

$$Y = a + bX + cG + u \quad (1)$$
$$X = d + eY + v. \quad (2)$$

Y and X are endogenous variables, and G is an exogenous (pre-determined) variable.

In Equation (1), there is one pre-determined variable (G), no pre-determined variables are excluded, and $F < H - 1$. In Equation (2), there are no pre-determined variables, one pre-determined variable is excluded (G), and hence it is identified. Thus, we can use the reduced form to estimate the πs, but we cannot recover the parameters in Equation (1) because it is under-identified.

Reverse causality is common in economic, engineering, and social systems. It is called by different names and variants, such as the

Table 15.4 Order conditions for identification.

Equation	Pre-determined	Excluded	F	$H - 1$	Conclude
1	1	0	0	1	Not identified
2	0	1	1	1	Identified

simultaneous equations model, feedback models, systems dynamics model, path analysis, LISREL, and structural equations model.

References

Atkinson, K. (1989) *An introduction to numerical analysis*. New York: Wiley.

Barnett, V. and Lewis, T. (1994) *Outliers in statistical data*. New York: Wiley.

Belsley, D., Kuh, E., and Welsch, R. (1980) *Regression diagnostics*. New York: Wiley.

Berndt, E. (1991) *The practice of econometrics*. New York: Addison-Wesley.

Bowden, R. and Turkington, D. (1985) *Instrumental variables*. Cambridge: Cambridge University Press.

Box, G. and Cox, D. (1964) An analysis of transformations. *Journal of the Royal Statistical Society*, B, **26**(2), 211–243.

Chingos, M. (2013) Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management*, **32**(2), 411 – 438.

Cox, D. and Snell, E. (1989) *Analysis of binary data*. London: Chapman and Hall.

Gruber, M. (1998) *Improving efficiency by shrinkage*. New York: Marcel Dekker.

Hoerl, A. and Kennard, R. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Johnston, J. and DiNardo, J. (1997) *Econometric methods*. New York: McGraw-Hill.

Maddala, G. (1986) *Econometrics*. New York: McGraw-Hill.

Mishel, L. and Rothstein, R. (Eds.) (2002) *The class size debate.* Washington DC: Economic Policy Institute.

Rohatgi, V. (1976) *An introduction to probability theory and mathematical statistics*. New York: Wiley.

Seber, G. and Wild, C. (2003) *Nonlinear regression*. New York: Wiley.

CHAPTER 16

Quantitative Data Analysis III: Regression Data (Part III)

Time Series

So far, our discussion has primarily been on cross-sectional data. In this section, we consider linear regressions using time series such as

$$Y_t = \alpha + \beta X_t + \varepsilon_t.$$

The subscript has been changed from i to t for time series. The observations are ordered by time, such as monthly observations of rainfall and interest rates. As before, we use the simple regression model to avoid carrying too many independent variables. The principles may be extended to the multiple regression case. Where it creates special problems, these issues will be highlighted.

There are two recurring problems with time series data, namely,

- autocorrelation, where the errors ε_t and ε_{t-h} are correlated for some lag h, which violates the OLS assumption of independence among the error terms (see Chapter 14); and
- non-stationarity, where a time series is trending upwards or downwards and can result in spurious regressions.

These issues are discussed below.

Autocorrelation

We will consider the autocorrelation problem first. If there is a single lag, ε_t may be autocorrelated with ε_{t-1}, as shown in Fig. 16.1. Equivalently, the residual e_t is correlated with e_{t-1}. This means that if the residual is

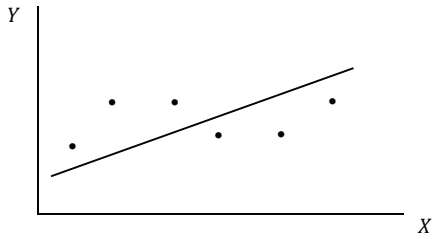

Fig. 16.1 Time series with autocorrelated errors.

positive in one period, it tends to stay positive for the next period. The first three residuals are positive, and the last three residuals are negative. Hence, the residuals are not randomly distributed; they are autocorrelated.

Autocorrelation is common in economic time series because the impact of an external shock such as a spike in oil prices or interest rates tends to persist for more than one period. Governments, firms, and consumers will take time to adjust to the new conditions. At first, higher oil prices will affect industries that use large quantities of oil, such as the oil and transport sectors. Subsequently, the impact spreads to other industries through higher prices for the outputs of these sectors. These higher prices may depress demand in other sectors and trigger a general recession. The process then works in reverse as unemployment rises, investment falls, and so on.

To see the implications of autocorrelation, consider the model

$$Y_t = \alpha + \beta X_t + u_t; \text{ and} \tag{16.1}$$

$$u_t = \rho u_{t-1} + \varepsilon_t. \tag{16.2}$$

The error term u_t is not random. It correlates with u_{t-1}, and ρ is the coefficient of correlation. If ρ is close to 1, the error terms u_t and u_{t-1} are highly correlated. If it is close to zero, they are uncorrelated. The error term ε_t is random (white noise) and follows the usual OLS assumptions, that is, it has a normal distribution with zero mean and constant variance σ^2. We further assume that u_t and ε_t are independent.

Taking the expectation and variance,

$$E(u_t) = \rho E(u_{t-1}) + E(\varepsilon_t) = 0; \text{ and}$$

$$Var(u_t) = \rho^2 Var(u_{t-1}) + Var(\varepsilon_t) = \rho^2 Var(u_{t-1}) + \sigma^2 > \sigma^2.$$

The expectation equation shows that the OLS estimator is still unbiased if autocorrelation is present, but $Var(u_t)$ is larger than $Var(\varepsilon_t) = \sigma^2$. Consequently, the usual diagnostic t and F tests will no longer be valid. Hence, we need to test for autocorrelation before regressing the time series, that is, we need to test whether $\rho = 0$.

The traditional procedure is to use the Durbin–Watson test, but this test may be inconclusive (Kmenta, 1986). The modern approach is to use the Lagrange Multiplier (LM) test. The unrestricted model is

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t.$$

The restricted model is

$$Y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t.$$

Then under H_0: $\rho_1 = \rho_2 = 0$,

$$\frac{(RSS_R - RSS_U)/d}{RSS_U/(n-k)} \sim F(d, n-k). \tag{16.3}$$

Here, $d = 2$ because there are two restrictions under H_0.

If autocorrelation exists, a popular way of estimating the model is to multiply Equation (16.1) by ρ and lag it by one period to obtain

$$\rho Y_{t-1} = \alpha\rho + \rho\beta X_{t-1} + \rho u_{t-1}.$$

We then subtract this equation from Equation (16.1) and use Equation (16.2) so that

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \varepsilon_t.$$

This equation may be written as

$$y_t = \alpha(1-\rho) + \beta x_t + \varepsilon_t, \tag{16.5}$$

where

$$y_t = Y_t - \rho Y_{t-1}; \text{ and}$$
$$x_t = X_t - \rho X_{t-1}.$$

We require an estimate of ρ to compute y_t and x_t in Equation (16.5). Durbin's (1960) method estimates ρ directly by shifting ρY_{t-1} in Equation (16.4) to the right-hand side and regressing

$$Y_t = \alpha(1-\rho) + \rho Y_{t-1} + \beta X_t - \beta\rho X_{t-1} + \varepsilon_t$$

using OLS. The coefficient of Y_{t-1} provides an estimate of ρ for use in Equation (16.5) to find y_t and x_t before applying OLS.

In summary, autocorrelation is likely to occur for economic time series data because economic activities require considerable periods of time to adjust to external shocks. We use the LM test to determine if autocorrelation is present. If it is absent, we can use OLS if the series is stationary, as discussed in the next section. If autocorrelation is present, we may use Equation (16.5).

Non-stationarity

A time series may trend upwards or downwards, that is, it is not stationary. For example, the series in Fig. 16.1 is trending upwards. Many economic time series, such as prices, trade volumes, housing starts, and interest rates, are not stationary. They exhibit cyclical or wavelike behavior and tend to trend upwards because of inflation, rising population, globalization, and expanding economic production.

In contrast, a (weakly) stationary series does not trend. Its mean and variance remain relatively constant over time, that is,

$$E(X_t) = \mu \text{ and}$$
$$Var(X_t) = \sigma^2 \text{ for all } t.$$

Further,

$$Cov(X_t, X_{t+h}) = \gamma_h$$

is also constant for all t. The covariance depends only on the lag (h) and not on the observation period. For example, if we have a long series of monthly interest rates, the covariance at, say, lag 2 is the same irrespective of whether we compute it using data from last year or this year.

A major problem with regressing Y on X if both series are not stationary is that the trend will dominate the relation. Thus, even though the series are unrelated, we will still obtain a high R^2, which is undesirable. The regression is said to be spurious (Granger and Newbold, 1974). It is

a problem because we can never be sure whether the relation between X and Y is real or spurious. Our guide is a theory on how they may be related, but it is ultimately only a theory. An example of a spurious regression is data on ice cream sales (S) and the number of drownings (D). If we regress

$$S_t = \alpha + \beta D_t + \varepsilon_t,$$

it is likely that β is significant. But it is incorrect to infer that drownings cause ice cream sales. Both series are highly correlated because of a third variable, such as a hot summer or heat wave.

The other problem with non-stationarity is that the error variance may be infinite, and this violates the OLS assumption that $Var(\varepsilon_t) = \sigma^2$. To see this, consider the model

$$Y_t = \rho Y_{t-1} + \varepsilon_t. \tag{16.6}$$

If $\rho = 0$, then $Y_t = \varepsilon_t$ and the series is stationary because the random error series is stationary. If $\rho = 1$, we have

$$Y_t = Y_{t-1} + \varepsilon_t.$$

Then, through repeat substitution,

$$\begin{gathered} Y_2 = Y_1 + \varepsilon_2; \\ Y_3 = Y_2 + \varepsilon_3 = Y_1 + \varepsilon_2 + \varepsilon_3; \\ Y_4 = Y_3 + \varepsilon_4 = Y_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4; \\ \ldots \\ Y_t = Y_1 + \varepsilon_2 + \ldots + \varepsilon_t. \end{gathered}$$

Thus,

$$Var(Y_t) = Var(Y_1) + Var(\varepsilon_2) + \ldots + Var(\varepsilon_t) = t\sigma^2.$$

In general, if we repeat the substitution, t will be large and the variance becomes infinite.

In summary, if $\rho < 1$, the series in Equation (16.6) is stationary. If $\rho = 1$, the series is non-stationary. If $\rho > 1$, the series is explosive because each Y_t is greater than Y_{t-1}. Hence, the boundary case is when $\rho = 1$, called the unit root. A test for stationarity is, therefore, a test for a unit root.

Unit Root Test

Consider the model

$$Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t.$$

Based on the discussion above, we want to test

H_0: $\rho = 1$ (series is non-stationary) against
H_1: $\rho < 1$ (series is stationary).

Under H_0, both Y_t and Y_{t-1} are non-stationary and, because it violates the OLS assumptions, we cannot use OLS to estimate ρ directly from the equation. Instead, we use the first-order differenced form by subtracting Y_{t-1} from both sides so that

$$Y_t - Y_{t-1} = \alpha + \rho Y_{t-1} - Y_{t-1} + \varepsilon_t.$$

In general, first-order differencing converts most non-stationary series to stationary ones (Box and Jenkins, 1976). We call Δ the difference operator, that is,

$$\Delta Y_t = Y_t - Y_{t-1}.$$

Thus,

$$\Delta Y_t = \alpha + (\rho - 1) Y_{t-1} + \varepsilon_t = \alpha + \delta Y_{t-1} + \varepsilon_t.$$

Hence, a test of H_0: $\rho = 1$ (series is non-stationary) is the same as testing H_0: $\delta = 0$. This is the Dickey–Fuller (DF) test for unit root. The Augmented Dickey–Fuller (ADF) test for unit root has a more generous lag structure:

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \Sigma \beta_i \Delta Y_{t-i} + \varepsilon_t.$$

The additional lags (ΔY_{t-i}) are included so that the error term becomes random. The asymptotic critical values depend on whether a constant term (α), trend (λt), or both are used (see Table 16.1).

In summary, if we use the simpler DF test, then H_0: $\delta = 0$ (non-stationary) and we use OLS to regress

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \varepsilon_t.$$

Table 16.1 Critical values for unit root test.

	Significance level		
	1%	5%	10%
No constant, no trend	–2.58	–1.95	–1.62
Constant, no trend	–3.43	–2.86	–2.57
Constant and trend	–3.96	–3.41	–3.12

Source: Fuller (1976).

We then conduct the usual test of significance using the critical values from Table 16.1 instead of the Student *t* distribution table. If we reject H_0, the series is stationary. The ADF test uses the following equation:

$$\Delta Y_t = \alpha + \lambda t + \delta Y_{t-1} + \Sigma \beta_i \Delta Y_{t-i} + \varepsilon_t.$$

Note the trend term (λt) is included for completeness.

Cointegration and Error Correction

If *Y* and *X* are non-stationary series, the usual approach is to regress in differenced form because, as mentioned in the previous section, most series are stationary after first-order differencing. Thus, we can apply OLS to

$$\Delta Y_t = \alpha + \beta \Delta X_t + \varepsilon_t.$$

To see why a first-order differenced series is often stationary, let

$$Y_t = \{1, 3, 4, 5, 7, 6\},$$

which is trending up (i.e. non-stationary). Then

$$\Delta Y_t = \{3-1, 4-3, 5-4, 7-5, 6-7\} = \{2, 1, 1, 2, -1\}$$

has no obvious trend, that is, it is stationary.

However, if we regress in differenced form, we are estimating how *Y* changes in response to changes in *X* in the *short* run. This process is called a dynamic adjustment in economics because of its association with the market adjustment of demand and supply towards the equilibrium price. The *long*-term relation between *Y* and *X* is given by

$$Y_t = \alpha + \beta X_t + \varepsilon_t.$$

Although X and Y are non-stationary, it is possible that ε_t is stationary. If this is the case, the two series are said to be *cointegrated*. We may then regress the above long-term relation without worrying about spurious relations because the stationarity of ε_t ensures that both series cannot drift "far apart" from each other. Hence, the first step in cointegration analysis is to regress the above equation and test whether the residual variable e_t is stationary using the DF or ADF test.

In the second step, we regress

$$\Delta Y_t = \alpha + \beta \Delta X_t + \lambda e_{t-1} + \varepsilon_t.$$

The term λe_{t-1} is an *error correction mechanism* (ECM). While X and Y have a long-term relation, there may be short-term deviations, and the ECM adjusts these deviations towards the long-term equilibrium relation (Engle and Granger, 1987). If there are more than two variables, there may be more than one cointegrating relation, and more complex procedures are required (Johansen, 1995).

Distributed Lag Model

A distributed lag or dynamic regression model contains lagged values of the independent variable, for example,

$$Y_t = \alpha + \beta X_{t-1} + \lambda X_{t-2} + \varepsilon_t.$$

This model may be estimated using OLS. Lags occur frequently in economic activities, such as the lagged impact of a rise in mortgage interest rates on monthly house prices. It may take a few months before the impact is felt because of procedural delays such as in approving new mortgage loans, or because potential house buyers think the change in interest rates is uncertain or only temporary. The distributed lag model is also used in operations research and engineering, such as in estimating hospital patient discharges (Y) from lagged values of previous admissions (X).

If the lags are unknown or difficult to guess, a more general lag structure is used so that

$$Y_t = \alpha + \sum \beta_i X_{t-i} + \varepsilon_t$$

There are too many parameters. The model is estimated by imposing restrictions on the parameters, such as

$$\beta_i = \beta_0 \varphi^j \qquad \text{for } 0 < \varphi < 1,\ j = 0, 1, 2,...$$

This is the Koyck (1954) lag structure, which results in declining values of β as the lag structure increases. The more distant values of X are given less weight in determining Y_t. To see this, we rewrite the model as

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \varphi X_{t-1} + \beta_0 \varphi^2 X_{t-2} + ... + \varepsilon_t.$$

The coefficients or weights decline for more distant values of X because $0 < \varphi < 1$. If we lag it by one period,

$$Y_{t-1} = \alpha + \beta_0 X_{t-1} + \beta_0 \varphi X_{t-2} + \beta_0 \varphi^2 X_{t-3} + ... + \varepsilon_{t-1}.$$

Multiplying both sides by φ and subtracting it from the first equation gives

$$Y_t - \varphi Y_{t-1} = \alpha(1 - \varphi) + \beta_0 X_t + v_t,$$

where $v_t = \varepsilon_t - \phi \varepsilon_{t-1}$. Thus,

$$Y_t = \alpha(1 - \phi) + \beta_0 X_t + \phi Y_{t-1} + v_t.$$

Observe that the distributed lags have been replaced by a lagged dependent variable Y_{t-1}. However, Y_{t-1} and v_t are correlated because Y_{t-1} depends on ε_{t-1}, and v_t also depends on ε_{t-1}. This violates the OLS assumption of independence between the independent variables and error term. We may estimate it using an instrument variable, but, as we have seen, such an instrument is hard to find.

For this reason, it is more practical to restrict the lags to a small number to avoid loss of degree of freedom, multicollinearity, and the need to find an instrumental variable. The loss of degree of freedom occurs because there will be many parameters (unknowns) to estimate. Multicollinearity may arise because lagged variables such as monthly interest rates (say, r_t) tend to move slowly over time so that r_t and r_{t-1} are likely to be highly correlated.

Pooled Regression

Suppose we have cross-sectional data on a sample of n firms, such as

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i, \qquad i = 1, ..., n.$$

Here, Y may be the firm's output, and X is the number of workers. As before, it is possible to add more independent variables, but we will use the above model to simplify the explanation. Other than organizations, we may also collect cross-sectional data from regions, individuals, and other entities. If we do not have sufficient data, it makes sense to collect the data over, say, a 3-year period ($T = 3$). The model becomes

$$Y_{it} = \beta_1 + \beta_2 X_{it} + \varepsilon_{it}, \; i = 1, ..., n; \; t = 1, ..., T. \tag{16.7}$$

There are two subscripts, one for the entity (i) and the other for time (t). The illustrative panel data is shown in Table 16.2. To simplify the notation, we often reset the year so that $t = 1$ for 2017, $t = 2$ for 2018, and $t = 3$ for 2019. Since there are 10 firms observed over a 3-year period, we have 30 observations. It is a short panel because $n > T$, which is common. In a long panel, $T > n$.

If the series are stationary or differenced stationary, we can use OLS to estimate Equation (16.7). This procedure is called *pooled regression* because it combines cross-sectional and time series data. However, the

Table 16.2 Illustrative panel data table.

Firm	Y	X	Year
1	10	5	2017
1	11	4	2018
1	15	6	2019
2	13	6	2017
2	17	8	2018
2	12	7	2019
...	...	...	...
10	15	5	2017
10	16	6	2018
10	18	8	2019

errors are likely to be autocorrelated or heteroscedastic. Hence, it is more common to use panel regression, which is discussed next.

Panel Regression

In pooled regression, we do not model the error term explicitly. If the firms are qualitatively different, then it is better to use a different intercept to reflect something unique to each firm (α_i), such as its management. The parameter α_i is called the firm-specific effect. If the sampling unit is an individual, it is called the individual-specific effect. Because of heteroscedasticity, we rewrite the model in Equation (16.7) as

$$Y_{it} = \beta_1 + \beta_2 X_{it} + u_{it},$$

where

$$u_{it} = \alpha_i + \varepsilon_{it}.$$

Here, ε_{it} is white noise, that is, normally independent series with a zero mean and variance σ^2. To capture the firm-specific effects, we use n dummy variables (D_i) so that

$$Y_{it} = \Sigma\alpha_i D_i + \beta_2 X_{it} + \varepsilon_{it}. \qquad i = 1, ..., n; t = 1, ..., T. \tag{16.8}$$

We have removed β_1 and used n dummy variables to avoid perfect collinearity. Equation (16.8) is called the *fixed effects model*. There are $n + 1$ parameters comprising α_1, ..., α_n and β_2. If $n = 10$ and $T = 3$, there are only 30 observations available to estimate 11 parameters, which is insufficient. Hence we need to increase n, T, or both to obtain more reliable estimates of the parameters.

The *random effects* or *error component model* replaces the error term ε_{it} in Equation (16.7) with

$$v_{it} = w_i + \varepsilon_{it}.$$

Here, v_{it} is assumed to be autocorrelated and heteroscedastic, and w_i is uncorrelated with X_{it}. As before, ε_{it} is white noise. This creates difficulty in estimating the model, which requires generalized least squares (GLS) estimation (see Chapter 15). It can be shown that the estimating equation is

$$Y_{it} - \gamma \bar{Y} = \beta_1(1-\gamma) + \beta_2(X_{it} - \gamma \bar{X}_i) + (v_{it} - \gamma \bar{v}_i)$$

with

$$\gamma = 1 - \left(\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_w^2} \right)^{\frac{1}{2}}$$

Variables with a bar refer to the mean. If $\gamma = 0$, it reduces to the pooled regression model. If $\gamma = 1$, it becomes the fixed effects model. In other words, the random effects model is a weighted version of both models. For more information on panel regression, see (Hsiao, 2003) and (Baltagi, 2013).

VAR Model

In Chapter 15, we showed how feedback may be modeled as an explicit system of simultaneous equations. A vector autoregression (VAR) model may also be used to examine feedback, but it does so without explicit modeling of how the variables are related. It lets the data determine the endogenous and exogenous variables as well as the lags.

Suppose Y depends on X as well as lagged values so that

$$Y_t = \beta_{10} + \beta_{12}X_t + \gamma_{11}Y_{t-1} + \gamma_{12}X_{t-1} + u_t$$

Similarly, to examine possible feedback, let

$$X_t = \beta_{20} + \beta_{21}Y_t + \gamma_{21}Y_{t-1} + \gamma_{22}X_{t-1} + v_t.$$

The time series are assumed to be stationary or differenced stationary, and the error terms are uncorrelated white noise. It is a first-order VAR model because, for simplicity, only a single lag is used. It is clearly possible to extend it to other lags. This system of structural equations may be written more compactly as

$$\begin{bmatrix} 1 & -\beta_{12} \\ -\beta_{21} & 1 \end{bmatrix} \begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \beta_{10} \\ \beta_{20} \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ v_t \end{bmatrix}.$$

That is,

$$\mathbf{BZ}_t = \boldsymbol{\beta}_0 + \mathbf{GZ}_{t-1} + \mathbf{u}_t.$$

Pre-multiplying by $\mathbf{B}^{-1}$ gives the reduced form

$$\mathbf{Z}_t = \mathbf{B}^{-1}\boldsymbol{\beta}_0 + \mathbf{B}^{-1}\mathbf{GZ}_{t-1} + \mathbf{B}^{-1}\mathbf{u}_t = \mathbf{a}_0 + \mathbf{A}_1\mathbf{Z}_{t-1} + \boldsymbol{\varepsilon}_t,$$

where $\mathbf{a}_0 = \mathbf{B}^{-1}\boldsymbol{\beta}_0$, $\mathbf{A}_1 = \mathbf{B}^{-1}\mathbf{G}$, and $\boldsymbol{\varepsilon}_t = \mathbf{B}^{-1}\mathbf{u}_t$. Written in full, the reduced form is

$$\begin{bmatrix} Y_t \\ X_t \end{bmatrix} = \begin{bmatrix} \alpha_{10} \\ \alpha_{20} \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

Each equation in the reduced form may be estimated using OLS, which is a major advantage of VAR models. The lack of theoretically-informed structural equations is sometimes viewed as a disadvantage. However, with the recent interest in big data and machine learning, it may be a strength.

Causality Test

Time series data may be used to test for predictive "causality" in a special sense: *X* "causes" *Y* if it fits the data better than if we regress *Y* on its past values alone. As before, it is assumed that both series are stationary or differenced stationary. In the first step, we regress

$$Y_t = \alpha + \sum_{i=1}^{p} \gamma_i Y_{t-i} + \varepsilon_t$$

and compute its residual sum of squares (RSS_R). This is the restricted model. The unrestricted model is

$$Y_t = \alpha + \sum_{i=1}^{p} \gamma_i Y_{t-i} + \sum_{j=1}^{m} \beta_j X_{t-j} + u_t$$

Let RSS_U be the unrestricted residual sum of squares. The null hypothesis is that *X* does not "cause" *Y*, that is, the βs are zero. Then, from Equation (16.3),

$$\frac{(RSS_R - RSS_U)/m}{RSS_U/(n-k)} \sim F(m, n-k).$$

Here, m is the number of restrictions, n is the number of observations in the unrestricted model, and k is the number of parameters in the unrestricted model. As a simple example, suppose we regress

$$Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$$

as the restricted model and obtain RSS_R. Assume that the unrestricted model is

$$Y_t = \alpha + \beta Y_{t-1} + \gamma X_{t-1} + u_t$$

and we compute RSS_U. Then, under H_0, there is only one restriction ($\gamma = 0$) and

$$\frac{(RSS_R - RSS_U)/1}{RSS_U/(n-3)} \sim F(1, n-3).$$

The above causality test is due to Granger (1969). It is based on predictability rather than causal mechanisms. Hence, strictly speaking, this "causality" test is correlative rather than causal.

Spectral Analysis

In spectral analysis, the data are analyzed in the frequency domain rather than in the time domain. Recall from high school math that, on the Cartesian graph,

$$y = A\cos(kx - \varphi)$$

traces out a cosine curve with amplitude A and phase shift $\varphi > 0$. The x-axis represents angle x, and k is a constant.

To sketch it out, if $kx - \varphi = 0$, $x = \varphi/k$ and $y = A\cos(0) = A$, which gives the first point in Fig. 16.2. If $y = 0$, $\cos(kx - \varphi) = 0$ and $kx - \varphi = \pi/2$, or $x = \varphi + \pi/2$, which gives the second point. Finally, if $x = 0$, $y = A\cos(-\varphi) = A\cos(\varphi)$ by property of the cosine function. This gives the third point, on the y-axis.

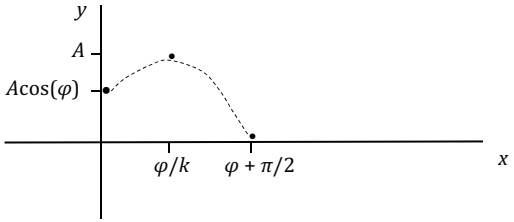

Fig. 16.2 Sketch of three points for $y = A\cos(kx - \varphi)$.

This function is nonlinear, and may be expanded as the sum of a sine function and a cosine function:

$$y = A[\cos(\varphi)\cos(kx) + \sin(\varphi)\sin(kx)] = a\cos(kx) + b\sin(kx),$$

where

$$A = \sqrt{a^2 + b^2};$$

$$\varphi = tan^{-1}\left(\frac{b}{a}\right).$$

To change the x-axis to time (t) instead of angle x, we replace x with ωt, where ω is the angular frequency given by

$$\omega = \frac{2\pi}{T} = 2\pi f,$$

where T is the period of oscillation and f is the frequency. We can replace x with $2\pi ft$ and write the above sum as

$$y = a\cos(2\pi fkt) + b\sin(2\pi fkt).$$

If there are n observations, we can approximate any time series by a sum of k sine and cosine functions so that

$$y_t = \sum_k \left\{ a_k \sin\left(2\pi \frac{k}{n} t\right) + b_k \cos\left(2\pi \frac{k}{n} t\right) \right\}. \qquad (16.9)$$

This is the Fourier representation of a series. The frequency f is equal to k/n, where k is a positive integer. For each value of k, we have a sine or

cosine curve of a particular frequency. Thus, a time series may be approximated by the sum of sine and cosine functions at different frequencies.

The unknowns are a_k and b_k, which may be computed from

$$a_k = \frac{2}{n}\sum_{t=1}^{n} y_t \sin\left(2\pi \frac{k}{n} t\right); \tag{16.10}$$

$$b_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos\left(2\pi \frac{k}{n} t\right). \tag{16.11}$$

The main interest is to determine which frequencies are important in contributing to the sum in Equation (16.9). This is done by examining the power *spectrum*, which is a plot of S_k against f where

$$S_k = \frac{1}{2}(a_k^2 + b_k^2). \tag{16.12}$$

It can be shown that (Priestley, 1982)

$$\sum_k S_k = Var(y_t).$$

In summary, a time series may be approximated by a sum of sine and cosine functions. Our interest is to pick out the main frequency or frequencies, and since $f = 1/T$, we may compute the corresponding periods. This is helpful in identifying the main periodicities of a time series. For example, if we have a 100-year series of GDP data, we can use spectral analysis to identify the underlying business cycles. There may be two types of cycles, namely the long waves or cycles of about 30 to 50 years and short cycles of about 5 to 10 years. To find the spectrum, we first compute a_k and b_k using Equations (16.10) and (16.11), respectively, for $k = 1, 2,..., n/2$, and then use the computed values to find S_k in Equation (16.12).

Spatial Regression

Spatial regression may be used if there are spatial spillovers. Suppose a country is divided into six regions (Fig. 16.3), numbered 1 to 6. Let Y_i be the variable of interest in region i, such as its unemployment rate. Clearly, Y_1 depends on its own exogenous variables X_1 and X_2, such as

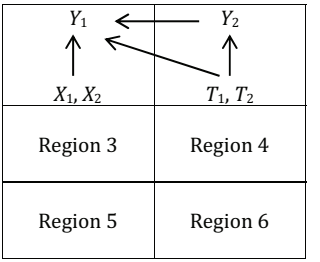

Fig. 16.3 Spatial interaction in a country with six regions.

its structure of industry and resource endowment. For simplicity, we will restrict the discussion to two independent variables to avoid carrying too many symbols. Y_1 may also depend on Y_2, the unemployment rate in a neighboring region, as well as the exogenous variables T_1 and T_2 that affect Y_2.

Putting these effects together and adding the error term, we have the *general spatial regression model*:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{T}\boldsymbol{\theta} + \mathbf{u}; \qquad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}.$$

Here, $\mathbf{y}$ is the $n \times 1$ vector of observations, ρ is a spatial dependence parameter, $\mathbf{W}$ is an $n \times n$ weight matrix, $\mathbf{X}$ is the usual $n \times k$ design matrix comprising the Xs, $\boldsymbol{\beta}$ is the $k \times 1$ vector of parameters, $\mathbf{T}$ is the $n \times m$ design matrix consisting of the Ts, $\boldsymbol{\theta}$ is an $m \times 1$ vector of parameters, and $\mathbf{u}$ is an $n \times 1$ error term. There is also spatial dependence in the errors with λ as the parameter, and ε is the usual $n \times 1$ white noise vector.

The weight matrix $\mathbf{W} = [w_{ij}]$ is specified prior to the regression. In most cases, we assume only neighboring regions have spillover effects. Hence, the weight matrix is

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

The matrix is formed row by row. It takes the value 1 if the regions are neighbors and 0 otherwise. Hence, $w_{ij} = 1$ if regions i and j are neighbors, and zero otherwise. Along the main diagonal, $w_{ii} = 0$. The assumption that only neighbors have spillover effects may not be correct. For example, Region 6 is not a neighbor of Region 1, but it can affect the unemployment rate in Region 1 if it is a poor region. The unemployed from Region 6 will migrate to other regions to seek employment, particularly to the primate (biggest) city in the country (Lynch, 2004).

There are too many parameters in the general spatial regression model. The practical spatial models simplify by setting some combinations of ρ, λ, and $\boldsymbol{\theta}$ to zero. For example, setting λ and $\boldsymbol{\theta}$ to 0 and $\mathbf{0}$, respectively, gives the *spatial lag model*

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{16.13}$$

The unknowns are ρ and $\boldsymbol{\beta}$. Equation (16.13) may be written as

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{I}$ is the identity matrix. Hence, the reduced form is

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}.$$

This representation is interesting because it shows that the new error term is no longer white noise. Further, since

$$(\mathbf{I} - \rho\mathbf{W})^{-1} = \mathbf{I} + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + ...,$$

the original spatial dependence based on $\rho\mathbf{W}$ does not describe the entire dynamic process.

Since ρ is unknown, it has to be estimated by rewriting the model as

$$\mathbf{y} = [\mathbf{x} \quad \mathbf{w}\mathbf{y}]\begin{bmatrix}\boldsymbol{\beta} \\ \rho\end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{Z}\boldsymbol{\phi} + \boldsymbol{\varepsilon}. \tag{16.14}$$

OLS cannot be used to estimate $\boldsymbol{\phi}$ because $\mathbf{W}\mathbf{y}$ is correlated with $\boldsymbol{\varepsilon}$. Hence we need to use the instrumental variable or the 2-stage least squares methods, as discussed in Chapter 15. Again, the problem is to find an appropriate instrumental variable. For more details on spatial regression models, see (Ward and Gleditsch, 2018) and (Chi and Zhu, 2019).

Rank-deficient Models

Thus far, our discussion on the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

assumes that $\mathbf{X}$ is of full rank so that the normal equations

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

has a unique least squares solution $\mathbf{b}$. This is assured if $\mathbf{X}$ is of full rank and $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ is invertible. There are classes of problems where $\mathbf{X}$ is not of full rank or said to be rank-deficient (Hansen, 1998). A simple example is when there are insufficient observations to estimate the k parameters in $\boldsymbol{\beta}$.

Before discussing rank-deficient systems, it is helpful to understand some basic concepts. An equation such as

$$X + Y = 1$$

has infinite solutions (Fig. 16.4). Every point on the line is a possible solution. To obtain a unique solution, we impose certain restrictions. There are two simple solutions, namely, (0, 1) and (1, 0), where the line cuts the Y and X axes, respectively.

A third solution is to impose the restriction that the solution has minimum distance or Euclidean norm to the origin; that is, we minimize

$$\sqrt{X^2 + Y^2}$$

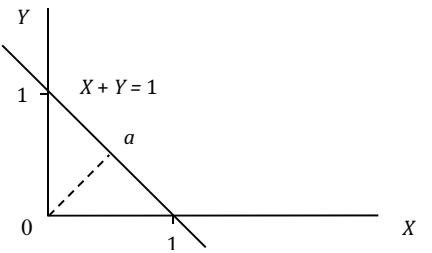

Fig. 16.4 Possible solutions to $X + Y = 1$.

This is the same as minimizing

$$L = X^2 + Y^2 = X^2 + (1 - X)^2.$$

Thus

$$\frac{dL}{dX} = 2x + 2(1 - X)(-1) = 0.$$

The solution is $X = ½$, and hence $Y = ½$. The minimum norm solution is shown as point *a* in Fig. 16.4.

The *rank* of a matrix is the number of independent rows (or columns). To find the rank of a matrix, we perform the following elementary row operations to transform it to echelon matrix form:

- interchanging any two rows;
- multiplying each element of a row by a non-zero scalar; and
- adding a non-zero multiple of one row to another.

For example, to find the rank of

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 0 \\ 1 & 1 \end{bmatrix},$$

we first interchange rows 1 and 3 to make the first row start with 1:

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 2 & 3 \end{bmatrix}.$$

Next, we multiply row 1 by (−1) and add it to row 2 to obtain

$$\begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 2 & 3 \end{bmatrix}.$$

The third step is to multiply row 1 by (−2) and add it to row 3:

$$\begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}.$$

Next, multiply row 2 by (−1) to obtain

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Finally, multiply row 2 by (−1) and add it to row 3 to obtain

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Thus rank (**A**) = 2 because it has two independent rows. A matrix has full rank if its rank equals the number of rows. In this case, **A** is rank-deficient.

We are now ready to discuss the linear regression model if **X** is rank-deficient. One possible solution is to *re-parameterize* the model by removing one or more parameters. For example, let

$$Y = \alpha + \beta X_1 + \lambda X_2 + \varepsilon.$$

If the matrix **X** is rank-deficient, we may impose a restriction on β to remove the deficiency. This requires external information on the value of β. Suppose we know, from previous studies, that $\beta = 2$. Then the transformed model is

$$Y - 2X_1 = \alpha + \lambda X_2 + \varepsilon.$$

Letting $y = Y - 2X_1$, we are left with a new regression model with only two parameters, α and λ. The number of restrictions will depend on the

rank deficiency; for example, if $\mathbf{X}$ has a rank deficiency of 2, then we require 2 restrictions.

An alternative approach is to rewrite the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\mathbf{d} = \mathbf{R}\boldsymbol{\beta}$$

where $\mathbf{R}$ is an $r \times k$ matrix, and $\mathbf{d}$ is an $r \times 1$ vector where r is the number of restrictions. For our example,

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \lambda \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$[2] = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \lambda \end{bmatrix}.$$

Observe that $\mathbf{R}\boldsymbol{\beta} = \mathbf{d}$ is exactly $\beta = 2$. The enlarged system may be compactly written as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{d} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{R} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \mathbf{0} \end{bmatrix}. \tag{16.15}$$

This may be written as $\mathbf{z} = \mathbf{H}\boldsymbol{\beta} + \mathbf{u}$, and OLS may then be used. Hence, we just need to add the restriction as an additional row to the original model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The solution to Equation (16.15) is

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{\mathrm{T}}\mathbf{d}). \tag{16.16}$$

To conduct significance tests, we require

$$\begin{aligned} Var(\mathbf{b}) &= Var[(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{R}^{\mathrm{T}}\mathbf{d})] \\ &= Var[(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] \\ &= Var[(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}] \\ &= \sigma^2\mathbf{M}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{M}^{\mathrm{T}}. \end{aligned}$$

The derivation assumes that $\mathbf{R}^{\mathrm{T}}\mathbf{d}$ **and** $\mathbf{X}\boldsymbol{\beta}$ are fixed and do not affect $Var(\mathbf{b})$. Further, by OLS assumption, $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ where $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{M}$ equals $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{X}^{\mathrm{T}}$ and is introduced for

notational convenience. For other approaches to solving rank-deficient systems using the Lagrange multiplier method, singular value decomposition, and the generalized inverse, see (Myers and Milton, 1991; Hansen, 1998; Webber and Skillings, 2000).

References

Baltagi, B. (2013) *Econometric analysis of panel data*. New York: Wiley.

Box, G. and Jenkins, G. (1976) *Time series analysis: Forecasting and control*. Oakland, California: Holden-Day.

Chi, G. and Zhu, J. (2019) *Spatial regression models for the social sciences*. Oakland, California: Sage.

Durbin, J. (1960) Estimation of parameters in time series regression models. *Journal of the Royal Statistical Society*, **22**, 139–153.

Engle, R. and Granger, C. (1987) Cointegration and error correction representation, estimation, and testing. *Econometrica*, **50**(2), 251–276.

Fuller, W. (1976) *Introduction to statistical time series*. New York: Wiley.

Granger, C. (1969) Investigating causal relations by econometric models ad cross-spectral methods. *Econometrica*, **37**(3), 424–438.

Granger, C. and Newbold, P. (1974) Spurious regressions in econometrics. *Journal of Econometrics*, **2**, 111–120.

Hansen, P. (1998) *Rank-deficient and discrete ill-posed problems*. Philadelphia: SIAM.

Hsiao, C. (2003) *Analysis of panel data*. Cambridge: Cambridge University Press.

Johansen, S. (1995) *Likelihood-based inference in cointegrated vector autoregressive models*. London: Oxford University Press.

Kmenta, J. (1986) *Elements of econometrics*. New York: Macmillan.

Koyck, L. (1954) *Distributed lags and investment analysis*. Amsterdam: North-Holland.

Lynch, K. (2004) *Rural-urban interaction in the developing world*. London: Routledge.

Myers, R. and Milton, J. (1991) *A first course in the theory of linear statistical models*. Boston: PWS-Kent.

Priestley, M. (1982) *Spectral analysis and time series*. New York: Elsevier.

Ward, M. and Gleditsch, K. (2018) *Spatial regression models*. Oakland, California: Sage.

Webber, D. and Skillings, J. (2000) *A first course in the design of experiments*. New York: CRC Press.

This page intentionally left blank

Chapter 17

Machine Learning

The Machine Learning Approach

Machine learning (ML) dates back to the 1980s or even earlier (Lovell, 1983) when it was called data mining, artificial intelligence (AI), or statistical learning. It is also called predictive analytics because of the focus on developing algorithms to learn from experience (i.e. training data set) to make useful predictions. It is distinct from another subset of AI, called deep learning (Goodfellow et al., 2016), which deals with the more complicated problem of representation. In machine learning, the inputs or features (*X*s) and output (*Y*) are known. In deep learning, a challenge is how to extract entity features, such as automatically detecting houses in an aerial photograph to replace the tedious process of manually digitizing a map.

Nowadays, ML is considered a subset of AI, the science of teaching machines to think. Prior to the 1990s, there were few applications of ML because of limited computing power. Nowadays, it is used in many areas such as banking, smart cities, building performance, education, security, medicine, sales, transport, and social media (Kuhn and Johnson, 2018; Wlodarczak, 2020). The popularity stems in part from the increase in computing power, better data quality, development of data science, and the availability of a large number of free data analytics software on the Internet. More recently, the rise of digital twin technology (Nath and van Schalkwyk, 2021) has increased the need for data-driven knowledge to complement physics-based knowledge to adjust the parameters or processes to optimize the physical system.

The ML approach is inductive. It starts from the data (i.e. "big data") rather than theory or reason. Recall that, in the deductive approach, we

start by developing a theory and then test it using empirical evidence. As discussed in Chapter 1, science is an endless cycle of induction, deduction, and abduction. Hence, the application of ML still requires theory to guide us on what to measure. It may be a preliminary framework rather than a well-developed theory.

There are downsides to ML. It is not easy for beginners to learn, and requires big data that may be difficult or costly to acquire. Measurability bias (Muller, 2018) refers to the tendency to prefer options that are easily measured. For example, measuring teaching ability using student feedback scores can lead to perverse incentives, such as encouraging the lowering of academic standards to avoid negative feedback. A further problem is that what is not measured is often ignored.

Since machines use past data to learn and then predict, the assumption is that the causal mechanisms in the past continue to operate in the present and future. This may be true of some processes but not of others. This is why it is not possible for machines to beat the stock market, where share prices follow random walks (Malkiel, 2020) because of unpredictable external shocks. Similarly, it is not possible to predict random events like a lottery.

A related problem with ML is over-fitting, where the model fits the existing data well but predicts the future poorly (Hawkins, 2004). This is called the generalizability problem. The opposite case is under-fitting where, for example, we fit a straight line such as $Y = \alpha + \beta X$ to a curve. Overfitting occurs if there are too many parameters, such as $Y = \alpha + \beta X + \lambda X^2 + \theta X^3 + \phi X^4$. This higher-order polynomial equation will fit the curve nicely but is unlikely to predict well.

There are also ethical and security issues with big data research. The issues with data collection include lack of consent and loss of privacy. The inappropriate use of data, particularly by a small group of people, is also a major concern. For example, if the model predicts poorly, its use is highly questionable. Finally, data may be leaked or fall into the wrong hands.

The purpose of this chapter is to provide an overview of ML algorithms and some intuitive explanations. It is beyond the scope of this chapter to deal with all the mathematical intricacies and plethora of

software on ML. There are many good texts on machine learning, such as (Hastie et al., 2001) and (Haykin, 2009).

Classification of ML Algorithms

There are three types of ML, namely,

- supervised learning;
- unsupervised learning; and
- reinforcement learning.

In supervised learning, the functional relation is $Y = f(X_1,..., X_k)$, where Y is the output variable and the Xs are input variables. If there are two or more output variables, the models include the simultaneous equations model, vector autoregression model (see Chapter 15), canonical correlation model, multivariate analysis of variance (MAN0VA), or multivariate analysis of covariance (MANCOVA). If there is one output variable, the main algorithms in supervised learning are

- regression models if Y is continuous; and
- classification models if Y is discrete.

The latter include logistic regression, random forest classifier, naive Bayes classifier, and support vector machine. Some algorithms, such as random forest, neural networks, and regression, can handle both continuous and discrete cases. The learning is "supervised" because past data are used to train the model to discover data patterns before using it for prediction or classification.

In unsupervised learning, there is no output variable. Hence, the emphasis is on correlations, such as in cluster analysis and correlative associations (e.g. recommender systems). Dimensionality reduction techniques such as principal components analysis or PCA, while technically not "learning" models, are often classified here because there is no output variable.

Finally, reinforcement learning models contain user or agent feedback, such as in computer gaming, robotics, and self-driving cars. These dynamic feedback ML models will not be discussed here.

Machine learning algorithms can be mathematically complex. Some of the common and simpler ones are discussed below.

Regression

Recall that the linear regression model may be written as

$$Y = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon.$$

Here, Y is a continuous variable, and the model may be used for understanding the relations between the Xs and Y and for predicting Y. Since Chapters 14, 15, and 16 have covered many issues concerning this model, they will not be repeated here.

Logistic Regression

Recall from Chapter 15 that the logistic regression uses the logistic curve

$$Y = \pi = \frac{1}{1 + e^{-(\alpha + \beta x)}}.$$

If Y is a binary variable indicating whether a person has lung cancer and X is the number of cigarettes smoked by the person, the predicted value of Y may be interpreted as the probability of having lung cancer. If it is greater than 0.5 or some other criterion value, the model classifies this person as having lung cancer. The model may be extended to include more explanatory variables.

Random Forest Classifier

You are thinking of visiting Beijing or Venice and asking m friends for their opinions. If there are k variables to consider, such as the cost of the air ticket, cost of hotels, food, places to visit, and so on, each friend will only use a few of these variables before making a suggestion. You then use majority voting to make a decision. In essence, this is how the random forest classifier works (Table 17.1).

A decision tree consists of a root (first) node, leaf nodes, and branches (lines) (Fig. 17.1). A decision is made at each node, and the branches

Table 17.1 Analogy between travel and the random forest classifier.

Travel analogy	Random forest classifier
Destination (Beijing, Venice)	Output Y (e.g. 0 or 1)
Friend	Decision tree
Group of m friends	Forest with m decision trees
k variables	k variables
Gives suggestions based on a few variables	Makes decisions using a few variables
Majority wins	Majority wins

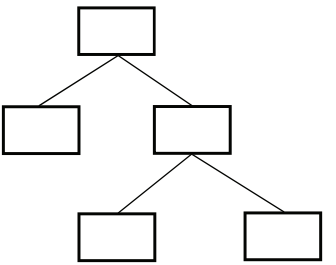

Fig. 17.1 A decision tree.

indicate possible decisions. If there are no more decisions to be made, the tree ends with a leaf node.

Suppose a lender is interested in predicting mortgage loan defaults (Table 17.2). There are n borrowers (e.g. 10,000), and the explanatory variables (Xs) are the determinants of default, such as

- characteristics of the borrowers, for example, age (X_1), gross monthly income (X_2, in \$'000), marital status (X_3; 1 if married; 0 if single), number of dependents, wealth, and amount of other outstanding loans;
- loan characteristics, for example, loan amount, loan to value ratio, loan period, and loan interest rate; and
- macroeconomic factors, for example, changes in house prices, interest rate, inflation, and recessions.

There are 14 explanatory variables, so $k = 14$. There may be missing data, so k may be smaller than 14. The outcome Y is a dummy variable; $Y = 0$ if there is no default, and $Y = 1$ if a default occurs.

Table 17.2 Data table on mortgage defaults.

Borrower ID	X_1	X_2	X_3	...	X_k	Outcome (Y)
1	25	2	0			1
2	30	3	0			1
3	45	4	1			0
4	50	6	1			0
5	55	8	1			0
...						
n	30	3	1			0

Table 17.3 Data set to train the first tree.

Borrower ID	X_1	X_2	X_3	Outcome (Y)
1	25	2	0	1
2	30	3	0	1
3	45	4	1	0
4	50	6	1	0
5	55	8	1	0

To generate the first tree, we use random sampling to generate the data to train the tree. If there are insufficient data, sampling with replacement (*bootstrapping*) may be used. Sampling with replacement is similar to drawing from a hat and putting it back before making the next draw. A drawback of bootstrapping is that it may artificially inflate the importance of rare observations.

For ease of exposition, I will use the first five data points (Table 17.3); in practice, each training data set contains many more points. Each record contains only three independent variables, similar to Pareto's principle that a few key factors cause most events. A rule of thumb is to use $\sqrt{k}$ variables. If there are data for $k = 11$ variables, then $\sqrt{k}$ is about 3 when rounded to the nearest integer. If a new set of data is drawn for the second decision tree, it will have a different combination of explanatory variables, such as X_1, X_4, and X_5.

The next step is to select the root node of the tree for the data set. It will be X_1, X_2, or X_3. The criterion is based on the idea of *information gain* (*IG*) between two data sets (*A* and *B*):

$$IG = E_A - E_B.$$

Here *E* stands for *entropy*, which is a measure of randomness (disorder) in the data. If a basket has 10 balls of different colors, the entropy (E_A) is high because there is a lot of randomness in the data. If a second basket has the same number of balls but only two colors (Blue and Red), then E_B is lower than E_A. The issue, then, is how to compute *IG* and select the root node with the highest *IG*.

Entropy is computed using

$$E = -\sum p_i \log_2(p_i).$$

Here, p_i is the probability, and the log to base 2, which is customary in the digital communications theory, may be computed using

$$\log_a(b) = \frac{\log_c(b)}{\log_c(a)}.$$

Normally, $c = e$, the base of natural log. It is easier to just use the log base 2 calculator on the Internet. To understand the formula for *E*, we rewrite it as

$$E = \sum p_i \log_2\left(\frac{1}{p_i}\right).$$

The formulas are equivalent because

$$-\log(x) = \log(1/x).$$

The log function is necessary to sum up the product of probabilities to avoid a measure that multiplies probabilities, which will result in a very small number. For example, if a bag contains 3 Red, 4 Blue, and 5 White balls, the probability of obtaining a Red, then Blue, and finally a White

ball with replacement is (3/12)(4/12)(5/12), which is a small number. Observe that entropy is inversely related to probability. If the bag has 9 Red balls and a White ball, the probability of picking a Red ball is 0.9, which is high. The data is less random, which means E should be low. Finally, the first term p_i may be conceptualized as a weighting scheme to find an average value of entropy for the data set.

For the data set S = (ID1, ID2,..., ID5} in Table 17.3, there are 2 defaults (Y = 1) and 3 non-defaults. Hence

$$E(S) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971.$$

The next step is to select an attribute, e.g. X_3.

X_3	Y
0	1
0	1
1	0
1	0
1	0

For singles (X_3 = 0), there are two defaults and

$$E(S_{Singles}) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0.$$

Similarly,

$$E(S_{Married}) = -\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right) = 0.$$

Thus, for X_3, IG = 0.97 – 0 – 0 = 0.97. This value is then compared with IG from X_1 and X_2, and the variable with the highest IG is selected as the root node. This tedious process is then repeated for each branch until there are only leaf nodes. For this example, X_3 is the root node, and the tree is shown in Fig. 17.2.

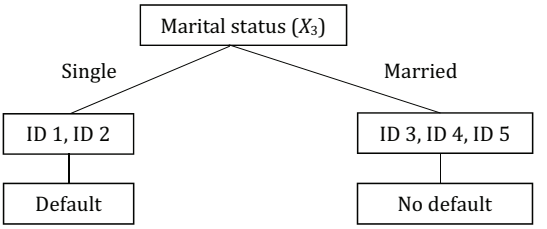

Fig. 17.2 Decision tree from the first data set.

Intuitively, X_3 is the root node because, looking at Table 17.3, it contains only 0s and 1s. It is the least random of the three variables. However, with many more data points, it becomes harder to use this visual approach.

Finally, for a new data point, we pass it through all the trees and select the decision based on majority voting. The tree above predicts a default if the potential borrower is single, which is unrealistic, given that it is constructed from only five data points. In practice, we need to use computer software to run a random forest classifier. The algorithm is popular because it generally works well. However, building a forest with many trees is computationally demanding.

Naive Bayes Classifier

The naive Bayes classifier is simpler to implement because it uses aggregate data to compute probabilities using Bayes' Theorem. For any two independent events, A and B,

$$P(A \text{ and } B) = P(A)P(B).$$

For example, the probability of obtaining two sixes in two consecutive tosses of a fair die is

$$P(6 \text{ and } 6) = (1/6)(1/6).$$

Independence means that the outcome of one event does not affect another. This is the "naive" assumption of this classifier. For the Bayes' part, the conditional probability of A given B has occurred is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}.$$

For example, the probability of obtaining a king of spade, given that it is a king, is

$$P(King\ of\ Spade|King) = \frac{P(King\ of\ Spade\ and\ King)}{P(King)}$$
$$= \frac{1/52}{4/52} = \frac{1}{4}.$$

Essentially, conditional probability shrinks the sample space from 52 to 4 cards, and the probability of obtaining a king of spade is 1/4. Shifting $P(B)$ to the left gives

$$P(B)P(A|B) = P(A \text{ and } B).$$

Further, by switching A and B,

$$P(A)P(B|A) = P(A \text{ and } B).$$

Hence,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This is Bayes' theorem. It is more intuitive to rewrite it as

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}.$$

Here H is the hypothesis, and e is the evidence. Thus, $P(H)$ is the prior probability of the hypothesis, $P(e|H)$ is the likelihood of the evidence given the hypothesis, and $P(H|e)$ is the posterior probability of the hypothesis given the evidence. Thus, Bayes' theorem updates our prior belief with the evidence (Lee, 2004).

Let us apply it to classify spam and non-spam emails (Table 17.4). The data set consists of 100 training emails, of which 25 are classified as

Table 17.4 Data from 100 emails.

	Spam (S)	Non-spam (N)
Total	25	75
"Discount"	20	5
"Delivery"	15	0

spam, and 75 are non-spam (sometimes called "ham"). Of the 25 spam emails, 20 contain the word "Discount," and 15 contain the word "Delivery." Similarly, of the 75 emails classified as non-spam, 5 contain the word "Discount," and none contain the word "Delivery." What is the probability that an email containing the word "Discount" (D) is a spam?

By Bayes' theorem,

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)}.$$

Here, $P(S) = 25/100 = 0.25$, and $P(D|S) = 20/25 = 0.8$. An email with the word "Discount" may be a spam *or* non-spam. Hence,

$$P(D) = \frac{25}{100}\left(\frac{20}{25}\right) + \frac{75}{100}\left(\frac{5}{75}\right) = 0.25.$$

The first term is the product of the probability of getting a spam (25/100) and the probability of an email containing the word "Discount," given that it is a spam (20/25). Similarly, the second term is the probability of obtaining a non-spam (75/100) multiplied by the probability of an email containing the word "Discount," given that it is not a spam (5/75). Thus,

$$P(S|D) = \frac{P(D|S)P(S)}{P(D)} = \frac{0.8(0.25)}{0.25} = 0.8.$$

Of course, we need to set a cut-off probability criterion (e.g. 0.90) to classify the email as spam or non-spam.

Although it is easy to implement, the naive Bayes classifier is not as popular because of its naivety, that is, the assumption of independent events may not hold true, resulting in sub-optimal predictions.

Support Vector Machine

A support vector machine (SVM) classifier uses a hyperplane to classify n data points into two groups. There are k features (Xs) such as age, income, and so on, and one output variable (Y). The output variable is designated as 1 or −1; $Y = 1$ if a borrower defaults on a mortgage loan and $Y = -1$ if there is no default (Table 17.5).

In the simple case of just two input variables (X_1 and X_2), the hyperplane is the solid line that separates the two classes, shown as circles and squares, in Fig. 17.3. The dashed lines are drawn using support vectors, that is, points on the boundary. For simplicity, we assume the classes are separable, that is, there is no data point on the "street" within the boundaries. A classification error (e_i) occurs if such a point exists.

The expression $\mathbf{w}^T\mathbf{x} - b = 0$ is the equation of a line because

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Thus,

$$\mathbf{w}^T\mathbf{x} - b = w_1X_1 + w_2X_2 - b = 0,$$

where b is a scalar, called the bias. Clearly, there are many lines that can separate the two classes. SVM selects the support vectors that maximize the distance d, the total margin or width of the street. The margin is $d/2$.

Since $\mathbf{w}$ is perpendicular to the lines, it is possible to compute d by simple geometry. For a point $\mathbf{x}_0$ on the bottom line,

$$\mathbf{w}^T\left(\mathbf{x}_0 + d\,\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) - b = 1.$$

Table 17.5 Data table for SVM.

ID	X_1	X_2	...	X_k	Y
1	25	2			1
2	35	3			1
...					−1
n	50	5			−1

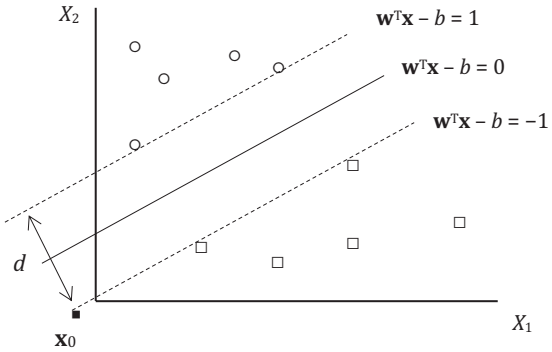

Fig. 17.3 Illustration of a SVM.

Recall that a vector divided by its own length $\|\mathbf{w}\|$ is a unit vector. By expanding this expression and using $\mathbf{w}^T\mathbf{x}_0 - b = -1$, we obtain

$$d = \frac{2}{\| \mathbf{w} \|}.$$

Hence, maximizing d is equivalent to minimizing the length of the weight vector $\mathbf{w}$. Hence, the SVM algorithm minimizes $\|\mathbf{w}\|$ subject to

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i - b \geq 1 \qquad & \text{if } Y_i = 1 \\ \mathbf{w}^T\mathbf{x}_i - b \leq 1 \qquad & \text{if } Y_i = -1 \end{aligned}$$

These constraints specify that each data point ($\mathbf{x}_i$) must lie on the correct side of the boundaries.

The algorithm may be extended to include cases where the classes are not separable, resulting in some classification errors. It is also possible to consider nonlinear functions so that the hyperplane becomes

$$\mathbf{w}^T\phi(\mathbf{x}) - b = 0.$$

These extensions will not be discussed here.

Cluster Analysis

We begin our discussion on unsupervised learning with cluster analysis. The data table is similar to Table 17.5, except that there is no output

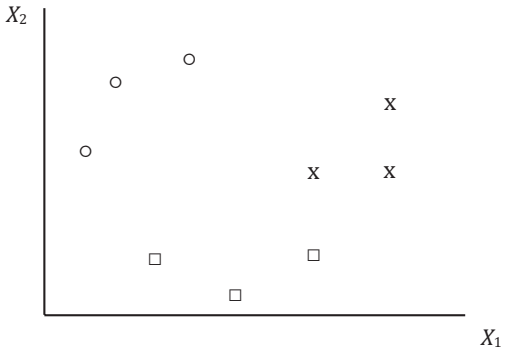

Fig. 17.4 Data for cluster analysis.

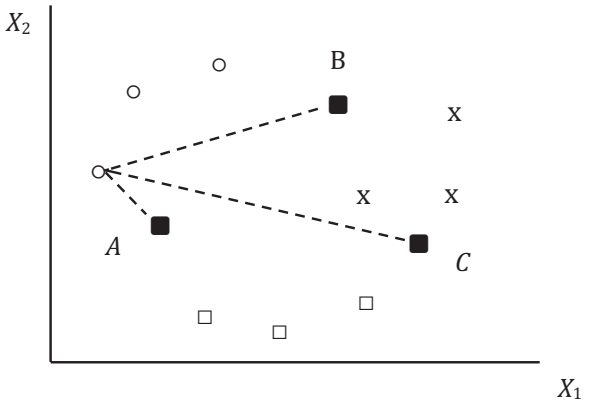

Fig. 17.5 Random assignment of centroids.

variable (Y). For example, given a data set of countries on k variables, the objective is to cluster them into K groups that share similar features. In Fig. 17.4, we show the logic of K-means cluster analysis using just two variables, X_1 and X_2. There are 9 data points (e.g. countries or individuals), and, intuitively, they can be grouped into 2 or 3 clusters. If there are 2 clusters, it could be circles in one cluster and the rest in another cluster. Or there may be 3 clusters, one for each symbol.

The first step is to determine K, the number of clusters. Suppose $K = 3$, and this number will be adjusted later. Next, we set cluster centroids randomly, as shown by the black boxes (Fig. 17.5).

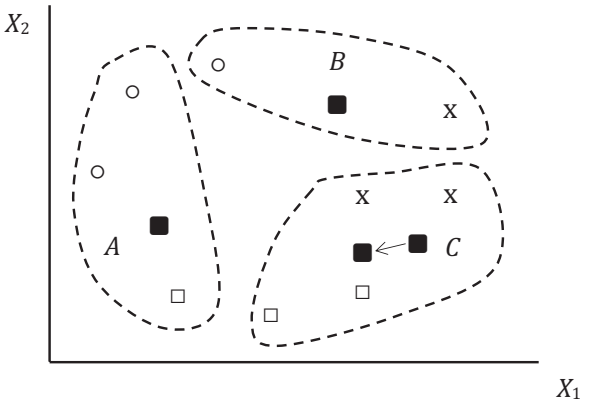

Fig. 17.6 Formation of initial clusters.

The third step is to compute the distance of each data point to the centroids and assign the point to the cluster with the shortest distance. As shown in Fig. 17.5, the point belongs to cluster *A*.

After running through all points, we have the initial clusters based on the shortest distance (Fig. 17.6).

In step 4, we compute the centroid of the initial clusters. For example, for cluster *C*, the shift from the initial to the new centroid is indicated by an arrow. We then repeat steps 3 and 4 until the cluster distribution does not change.

A weakness of this algorithm is that the initial random assignment of centroids matters. Hence, we need to run the algorithm a few times using different random assignments and select the final clusters with the minimum total distance, the sum of distances from each point to its centroid. It can be seen that the clusters in Fig. 17.6 are not optimal because the distances within each cluster are relatively large. A better set of clusters is to group the data using the symbols (Fig. 17.7).

Finally, we need to apply the algorithm for other values of *K* to select the optimal number of clusters. This is found from the elbow diagram, which shows how the sum of squared distances for all points,

$$S = \sum d_i^2,$$

varies with *K* (Fig. 17.8). Here, $K = 3$ because the decline in *S* is marginal between $K = 3$ and $K = 4$.

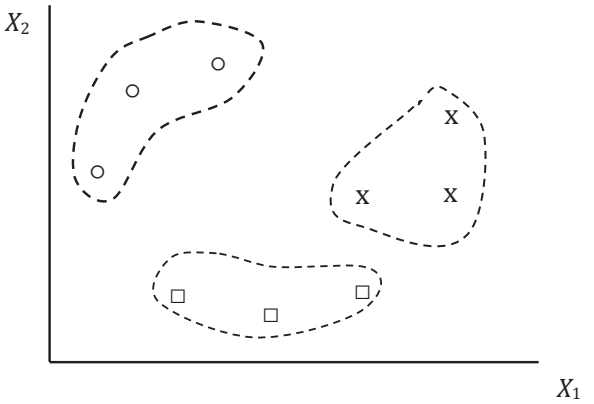

Fig. 17.7 A better set of clusters.

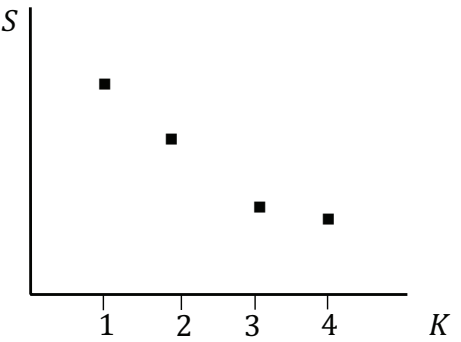

Fig. 17.8 Use of the elbow diagram to determine the value of *K*.

Principal Components Analysis

Principal components analysis (PCA) is a dimensionality reduction technique. Another similar technique is factor analysis, which is exploratory if there are no restrictions on the parameters, and "confirmatory" if there are restrictions. Factor analysis is discussed in the next section. The data structure for PCA is similar to Table 17.5, except that there is no output variable. The basic idea is to reduce the k variables to a few principal components. For example, a lender may have a lot of loan applications data on income, gender, education, marital status, current residential status (renting or owning a place), age, current employment, savings, debt, and the number of credit cards. Similarly, a human resource manager may

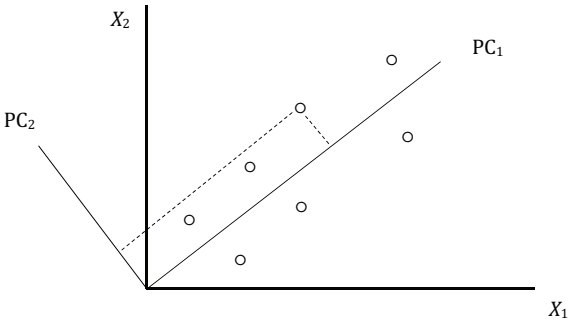

Fig. 17.9 First and second principal components.

have data for a particular job (e.g. project manager) on an applicant's age, gender, race, education level, Grade Point Average, working experience, current salary, expected salary, communication skills (rated from job interview), writing skills, and quality of one's resume.

The PCA technique uses correlations to reduce the k variables to a smaller set. In Fig. 17.9, we plot the data for seven individuals on two variables (X_1 and X_2). If they are highly correlated as drawn, it is possible to define a new axis (first principal component) that captures most of the variation in the data. The second principal component is orthogonal to the first one, which, in this case, does not capture much variation in the data. Each data point is projected onto the PCs, as illustrated for a data point. Hence, the data in two dimensions (X_1 and X_2) can be summarized into one dimension based on the first principal component (PC) if they are highly correlated.

Mathematically, the projection is given by

$$\mathbf{Z} = \mathbf{XA},$$

where $\mathbf{Z}$ is an n x k matrix of principal components, that is, $\mathbf{Z} = [\mathbf{z}_1 \ldots \mathbf{z}_k]$ and each $\mathbf{z}_i$ is a principal component, $\mathbf{X}$ is the n x k data matrix, and $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_k]$ is a k x k matrix of loadings or weights. Each $\mathbf{a}_i$ is a column vector of $\mathbf{A}$. The data matrix consists of data on n individuals on the k variables (i.e., the Xs). At this point, only $\mathbf{X}$ is known. Thus, $\mathbf{z}_1 = \mathbf{Xa}_1$, $\mathbf{z}_2 = \mathbf{Xa}_2$, and so on.

To find the first principal component, we maximize the variation in $\mathbf{z}_1$, that is, we maximize

$$\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1$$

subject to $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$. This restriction is necessary for the projection to work; if the elements of $\mathbf{a}_1$ are unbounded, the projection does not work. Since $\mathbf{z}_1 = \mathbf{X}\mathbf{a}_1$, its transpose is $\mathbf{a}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}$. The Lagrangean is

$$\varphi = \mathbf{a}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1 - \lambda_1(\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 - 1).$$

Thus,

$$\frac{\partial\varphi}{\partial\mathbf{a}_1} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0.$$

That is,

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1 = \lambda_1\mathbf{a}_1. \tag{17.1}$$

Hence, λ_1 is the first eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{a}_1$ is the corresponding eigenvector. Further,

$$\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1 = \mathbf{a}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1 = \mathbf{a}_1^{\mathrm{T}}\lambda_1\mathbf{a}_1 = \lambda_1.$$

Hence, maximizing $\mathbf{z}_1^{\mathrm{T}}\mathbf{z}_1$ is equivalent to maximizing λ_1.

To find the second PC, we maximize

$$\mathbf{z}_2^{\mathrm{T}}\mathbf{z}_2 = \mathbf{a}_2^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_2$$

subject to

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 = 1 \; and \; \mathbf{z}_2^{\mathrm{T}}\mathbf{z}_1 = 0.$$

The second condition requires the PCs to be orthogonal; recall that two vectors **x** and **y** are orthogonal if $\mathbf{x}^{\mathrm{T}}\mathbf{y} = 0$. Using Equation (17.1), it may also be written as

$$\mathbf{z}_2^{\mathrm{T}}\mathbf{z}_1 = \mathbf{a}_2^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_1 = \mathbf{a}_2^{\mathrm{T}}\lambda_1\mathbf{a}_1 = 0.$$

Since λ_1 is not zero, then

$$\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1 = 0.$$

The Lagrangean is

$$\varphi = \mathbf{a}_2^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_2 - \lambda_2(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_2 - 1) - \mu(\mathbf{a}_2^{\mathrm{T}}\mathbf{a}_1)$$

Hence,

$$\frac{\partial\varphi}{\partial\mathbf{a}_2} = \mathbf{2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 - \mu\mathbf{a}_1 = 0.$$

It can be shown that $\mu = 0$, giving

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a}_2 = \lambda_2\mathbf{a}_2. \tag{17.2}$$

Thus, λ_2 is the second largest eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. A plot of eigenvalues against the number of PCs is called a scree plot. It may be used to determine the number of PCs. In Fig. 17.10, the number of PCs is 3; thereafter, the decline is marginal.

To summarize, the algorithm for PCA is as follows:

- fill in the n x k data matrix $\mathbf{X}$;
- form $\mathbf{X}^{\mathrm{T}}\mathbf{X}$;
- find the eigenvalues λ_1, ..., λ_k and the corresponding eigenvectors $\mathbf{a}_1$, ..., $\mathbf{a}_k$ of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$;
- use the scree plot to determine the number of PCs; and
- interpret the data.

In Table 17.6, there are nine variables (X_1,..., X_9). The eigenvalues are $\lambda_1 = 2.3$, $\lambda_2 = 1.5$, and so on. If the scree plot shows that there are only two

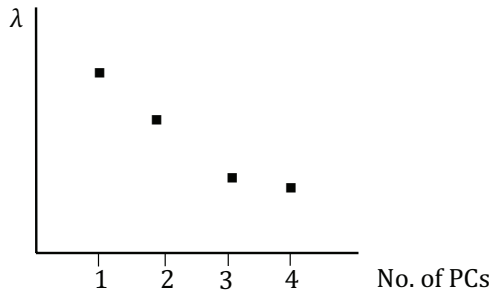

Fig. 17.10 A scree plot.

Table 17.6 Template for PCA output.

Eigenvalue	2.3	1.5	...	0.1
Variable	PC_1	PC_2	...	PC_9
X_1	0.10	0.10		
X_2	**0.40**	0.03		
X_3	**0.51**	0.04		
X_4	0.09	0.12		
X_5	0.11	0.15		
X_6	0.04	**–0.40**		
X_7	0.07	**–0.60**		
X_8	0.05	–0.10		
X_9	0.13	0.15		

PCs, it can be seen that the loadings are high on X_2 and X_3 for PC_1. For PC_2, the loadings are high in absolute terms for X_6 and X_7. The final step is to interpret what are PC_1 and PC_2. For example, in the job application case, if X_2 is the Grade Point Average and X_3 is the performance in the interview test, then PC_1 represents academic ability. Hence, some subjectivity is involved in naming the PCs.

Factor Analysis

Factor analysis is similar to PCA. It is a dimensionality reduction technique. The model is given by

$$\mathbf{x} = \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon}. \tag{17.3}$$

Here, $\mathbf{x}$ is a k x 1 vector of standardized variables, $\mathbf{L}$ is a k x g matrix of factor loadings, $\mathbf{f}$ is a g x 1 vector of uncorrelated common factors, and $\boldsymbol{\varepsilon}$ is the k x 1 error vector. Note $\mathbf{x}$ is standardized, that is, we subtract each element (x_i) from its mean and divide by its standard deviation. The distributional assumptions are

$\mathbf{x} \sim (\mathbf{0}, \mathbf{R})$ where $\mathbf{R}$ is a k x k correlation matrix;
$\mathbf{f} \sim (\mathbf{0}, \mathbf{I})$ where $\mathbf{I}$ is a g x g identity matrix; and
$\boldsymbol{\varepsilon} \sim (\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is a k x k diagonal matrix.

Written in full, the equation is

$$\begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & l_{1g} \\ \vdots & \ddots & \vdots \\ l_{k1} & \cdots & l_{kg} \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_g \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{bmatrix}.$$

Now,

$$Var(\mathbf{x}) = E(\mathbf{Lf} + \boldsymbol{\varepsilon})(\mathbf{Lf} + \boldsymbol{\varepsilon})^{\mathrm{T}} = E(\mathbf{Lff}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}) + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} = \mathbf{LL}^{\mathrm{T}} + \boldsymbol{\Psi}.$$

Since $Var(\mathbf{x}) = \mathbf{R}$, we have

$$\mathbf{R} = \mathbf{LL}^{\mathrm{T}} + \boldsymbol{\Psi}. \tag{17.4}$$

Thus,

$$Var(x_i) = \left(\sum_{j=1}^{g} l_{ij}^2 \right) + \psi_i. \tag{17.5}$$

The first term containing the sum is called the *commonality*. Hence, the factor analysis algorithm is as follows:

- standardize **x** and compute **R** using pairwise correlations between x_i and x_j;
- find **L** from the spectral decomposition of **R**; and
- compute ψ_i using Equation (17.5).

The spectral decomposition of **R** is given by

$$\mathbf{R} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}} + \cdots + \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$$

where the λs are the eigenvalues and the **u**s are corresponding eigenvectors of **R**. Retaining only g terms and using Equation (17.4),

$$\mathbf{L} \approx \left[\sqrt{\lambda_1}\mathbf{u}_1 \cdots \sqrt{\lambda_g}\mathbf{u}_{\mathrm{g}} \right].$$

The algorithm may be found in many standard statistical software. Suppose 500 respondents rated the relative importance of k characteristics of houses using a 5-point scale. The characteristics (Xs) include price, plot size, number of rooms, location, distance to bus stops, distance to transit, security, tenure, age, design, neighbors, pool, distance to parks and

Table 17.7 Template for factor analysis output.

	Loadings (l_{ij})				
Variable	$\mathbf{f}_1$	$\mathbf{f}_2$	$\mathbf{f}_3$	Communality	ψ_i
X_1	**0.85**	0.30	0.20	0.85	0.15
X_2	**0.70**	0.13	–0.22	0.55	0.45
X_3	**0.75**	0.10	0.02	0.57	0.43
X_4	0.21	**0.75**	0.15	0.63	0.37
X_5	0.10	**0.70**	0.14	0.52	0.48
X_6	0.11	**0.85**	0.22	0.78	0.22
X_7	0.32	0.33	0.30	0.30	0.70
...					
X_k	0.22	0.19	–0.23	0.14	0.86
λ_i	6.0	3.0	1.0		
λ_i/k	0.50	0.25	0.08		

recreation areas, noise, air quality, and so on. The template for factor analysis output is given in Table 17.7.

Here, we consider three factors ($\mathbf{f}_1$, $\mathbf{f}_2$, and $\mathbf{f}_3$). The first factor has high positive loadings on the first three variables and low loadings elsewhere (not shown). The second factor has high positive loadings on X_4, X_5, and X_6. The third factor has low loadings and will not be considered. Note that "high" loading is subjective, and it is the absolute value that matters and not whether they are positive or negative. The communality is the row sum of squares of loadings, and this is computed for the first row, that is, $0.85^2 + 0.30^2 + 0.20^2 = 0.85$. The error ψ_i is then computed using 1 – communality.

The next step is to interpret the two factors. Here $\mathbf{f}_1$ represents house characteristics, and $\mathbf{f}_2$ represents location attributes. Interpretation is subjective and may not be easy. The second last row shows the eigenvalues associated with each factor. They are divided by k to determine the contribution of each factor to the total variance (see Equation (17.4)).

Associative Methods

Associative methods predict using correlations. For example, recommender systems are widely used in retailing and social networks, such as

- "Other movies you may enjoy" (Netflix);
- "Books frequently bought together" (Amazon);
- "People you may know" (Facebook);
- "Recommended videos" (YouTube);
- "Recommended songs" (Spotify); and
- "Jobs you may be interested in" (LinkedIn).

For illustrative purposes, a small set of $n = 6$ transactions for four books, A to D, is shown in Table 17.8. For example, buyer #5 bought books A and B.

Books C and D have fewer sales. Books A and B are frequently bought together and may be used as a basis to recommend book B to buyer #6. It is possible to compute frequency statistics and apply the minimum *support* criterion, such as

$$f(C)/n = 2/6 > 30\%$$

for book C, where $f(.)$ stands for frequency. Similarly,

$$f(A \text{ and } B)/n = 3/6 > 30\%.$$

A weakness of this simple recommender system, called *user-user collaborative filtering*, is that buyer #5 may not like the books after reading them. A solution is to use a simple rating system, such as the number of stars (Table 17.9).

The next step is to compute the row mean and center the data (Table 17.10). A positive value is interpreted as a "like," and a negative value is a "dislike." Books that are not rated have zero values, and these are shown in bold.

Table 17.8 Transactions for books.

Buyer	*A*	*B*	*C*	*D*
1	x	x		x
2			x	
3	x	x	x	
4				x
5	x	x		
6 = *n*	x			

Table 17.9 User rating for books.

Buyer	*A*	*B*	*C*	*D*	Mean
1	4	5		3	4
2			3		3
3	1	4	4		3
4				5	5
5	3	3			3
6 = *n*	5				5

Table 17.10 Centered user rating for books.

Buyer	*A*	*B*	*C*	*D*
1	0	1	**0**	−1
2	**0**	**0**	0	**0**
3	−2	1	1	**0**
4	**0**	**0**	**0**	0
5	0	0	**0**	**0**
6 = *n*	0	**0**	**0**	**0**

To recommend a book to a buyer, we must find similar buyers. A measure of similarity is the cosine of the angle between two vectors, treating each row as a vector representing a buyer. Recall that, for any two vectors $\mathbf{x}$ and $\mathbf{y}$, the cosine of the angle between them is given by

$$\cos(\theta) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\| \mathbf{x} \| \| \mathbf{y} \|}.$$

For example, to compute the similarity between buyers #1 and #3, we first find

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \\ 1 \\ 0 \end{bmatrix} = 1.$$

Further,

$$\| \mathbf{x} \| = \sqrt{(1^2 + 0 + 0 + (-1)^2)} = \sqrt{2}; \text{and}$$
$$\| \mathbf{y} \| = \sqrt{-2^2 + 1^2 + 1^2 + 0} = \sqrt{6}.$$

Hence,

$$\cos(\theta) = \frac{1}{\sqrt{2}\sqrt{6}} = 0.29.$$

The similarity between the two buyers is low because $\cos(\theta)$ as a measure of the correlation lies between 0 to 1.

In *item-item collaborative filtering*, we compute the similarity between two items (books) by treating each column in Table 17.10 as a vector. The cosine of the angle between them may then be computed. If a buyer buys a book, the system can recommend a similar book based on this measure. However, it may be simpler, in this case, to use descriptors rather than collaborative filtering. For example, if a buyer buys a book on "Machine Learning," it is easier to recommend another highly rated book with the same or similar title.

References

Goodfellow, I., Bengio, Y., and Courville, A. (2016) *Deep learning*. Cambridge: MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The elements of statistical learning*. New York: Springer.

Hawkins, D. (2004) The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, **44**(1), 1–12.

Haykin, S. (2009) *Neural networks and learning machines*. New York: Pearson.

Kuhn, M. and Johnson, K. (2018) *Applied predictive modeling*. New York: Springer.

Lee, P. (2004) *Bayesian statistics*. New York: Wiley.

Lovell, M. (1983) Data mining. *Review of Economics and Statistics*, **65**(1), 1–12.

Malkiel, B. (2020) *A random walk down Wall Street*. New York: W. W. Norton.

Muller, J. (2018) *The tyranny of metrics*. London: Oxford University Press.

Nath, S. and van Schalkwyk, P. (2021) *Building industrial digital twins*. Birmingham: Packt.

Wlodarczak, P. (2020) *Machine learning and its applications*. London: Routledge.

This page intentionally left blank

Chapter 18

Meta-analysis

What is Meta-analysis?

Scientific meta-analysis is a form of data analysis where the data are not the usual "observations" on different entities such as individuals, households, organizations, and governments. Instead, the analyst analyzes the research findings of published works on a particular issue. It is a quantitative synthesis of research findings to increase the precision of parameter estimates and, where possible, resolve discrepancies (Lipsey and Wilson, 2000; Cooper, 2016).

Similar approaches include bibliometric analysis and a systematic literature review (Ball, 2017). The scope of analysis may include the identification of progress in knowledge, technological development, publication trends, new and fruitful areas of research, the leading researchers in the field, the changing government policies, the roles of different stakeholders in the production of knowledge, and the impact of such research. The stakeholders include private funding agencies, government agencies, international agencies, researchers, affiliated organizations, fellow scientists, and publishers.

Bibliometric analyses are clearly useful in carrying out the above functions. However, if used wrongly, they can lead to perverse results, such as the inappropriate reliance of such analyses as an indicator of research quality (Gingras, 2016).

In this chapter, we will focus on the use of scientific meta-analysis to resolve parametric estimates and research issues.

Steps in Meta-analysis

The steps in meta-analysis are:

- formulation of research question or issue that needs resolution;
- development of screening criteria;
- conduct of literature review;
- screening for suitable studies;
- abstraction of data;
- synthesis; and
- publication of results.

I will use the example of the effect of class size on student achievement to illustrate these steps except the last one, which is straightforward.

Identifying the Issue

The issue of class size and student achievement is hotly debated (Mishel and Rothstein, 2002; Chingos, 2013; Harfitt, 2015; Jepsen, 2015). It is a popular policy with teachers to reduce teaching load and with parents for a better quality of education. It is also largely supported by educators and administrators, except that it is an expensive policy strategy. Assuming there is just one teacher per class, reducing the class size from 30 to 20 students for 8 classes in a cohort requires 12 teachers, up from 8 teachers or an increase of 50%. This is then multiplied by the number of cohorts in a school.

The issue that needs resolution is whether the reduction in class sizes leads to better student performance. The latter is often measured using performance on standardized tests and, in the long term, wages.

Screening Criteria

The screening criteria to shortlist the empirical studies depend on the issue. In this case, the meta-analysis could cover all countries, a group of countries (e.g. OECD or developing countries), or a particular country. Another criterion is the education level (grade), such as primary school, secondary school, and so on.

A possible third criterion is methodology, such as covering only studies that adopted a particular research design, such as experiments. Finally, there may be other considerations, for example,

- whether to restrict the period of study, such as after the use of online learning;
- whether to include unpublished reports, such as those produced by public agencies;
- whether to include conference papers and dissertations;
- minimum sample size; and
- other sample characteristics.

The other sample characteristics may include gender, race, urban or rural schools, private or public schools, and so on.

Literature Review

The literature review is similar to that of other research. It is done using a search engine such as Google Scholar. The relevant articles are extracted from the databases, saved as files, and then shortlisted using the above screening criteria. If desired, the VOSviewer may be used to provide graphic images of bibliometric results such as publication trends, citation analysis, relations among keywords, and researcher networks.

Data Extraction

The next step is to summarize the data in the form of a table, such as that shown in Table 18.1. If the meta-analysis involves parameter estimates, it

Table 18.1 Template for summarizing data.

Author(s)	Year	Source	Grade	Sample size	...	Findings
A	2010	*Journal X*	Primary	200		No effect
...						
Z	2020	*Journal Y*	Secondary	500		Significant effect

Table 18.2 Template for summarizing estimated parameter(s).

Author(s)	Year	Source	Sample size	Estimated parameter	Standard deviation	...	Comments
A	2010	*Journal X*	200	0.8	0.2		Small sample
...							
Z	2020	*Journal Y*	500	1.1	0.1		Too high

is usual to provide the mean and standard deviation as well as the possible reasons for the variations in the "Comments" column (Table 18.2).

Synthesis

The synthesis involves trying to reconcile the different findings or estimated parameter values. For example, in a meta-analysis of the income elasticity of housing demand (e_y), de Leeuw (1971) examined cross-sectional evidence and concluded that e_y is about 0.8 to 1.0 for renters and "moderately higher" for homeowners. Harmon (1988) also tried to reconcile the differences.

There are many reasons why there are different estimates, resulting in highly technical parameter adjustments that are beyond the scope of this chapter. Hence, only brief comments are provided below.

One reason for the differing estimates is the context; for example, the education environment in developing countries is different from that of developed countries. The facilities, quality of teachers, incentives, resources, student motivation, and so on are likely to differ substantially. Similarly, housing markets are primarily local and differ across cities within a country and across countries. The demand side is mostly driven by fundamental factors such as the local population and income growth while the supply side depends on public housebuilding, private profitability, the availability of land, zoning regulations, building standards, the housing development approval process, and the productivity of the construction industry. The presence of highly connected cities, high population mobility, and well-developed national housing finance systems can result in national housing cycles (Leamer, 2007). There are also global housing cycles, such as the sustained rise in house prices in many

countries during the 2000s before the subprime crisis of 2007/8 (Igan and Loungani, 2012).

Second, analysts may use different methodologies, which makes it difficult to compare the results. Even if the same research design, such as an experiment, is used, the sample sizes may differ substantially, there may not be adequate randomization, performance tests are not properly administered, and so on. Similarly, studies of housing demand use different functional specifications such as a linear regression model $y = \alpha + \beta x + \varepsilon$, a double-log model $\log y = \alpha + \beta \log x + u$, or a system of demand and supply equations. Even with the same data set, the estimates of β will differ. Analysts also use different variables for theoretical and practical reasons. The theoretical reasons may be institutional, such as the varying taxes and subsidies to the housing sector as well as the role of government in the provision of public housing. On the practical side, variables may be omitted for lack of reliable data. Measurement errors also loom large in housing studies. The same concept, such as house price, may be measured in many ways as an index depending on data availability and how to adjust for quality differences (Silver, 2012).

Finally, de Leeuw (1971) observed that estimates from micro-data obtained from individuals or households differed from that of aggregate macro-data. This is called the aggregation problem in economics (Green, 2016). Similarly, time series and cross-section evidence are likely to differ because of differing data sets and conceptual differences in estimating elasticity over time and at a point in time. Another puzzle is different elasticities for renters and homeowners (Mayo, 1981).

In summary, it is unwise to just take the mean value and standard deviation of different parameter estimates. It is necessary to determine the sources of bias and, if possible, make the necessary adjustments.

References

Ball, R. (2017) *An introduction to bibliometrics*. Berlin: Elsevier.

Chingos, M. (2013) Class size and student outcomes: Research and policy implications. *Journal of Policy Analysis and Management*, **32**(2), 497–532.

Cooper, H. (2016) *Research synthesis and meta-analysis*. Thousand Oaks: Sage.

De Leeuw, F. (1971) The demand for housing — A review of cross-section evidence. *Review of Economics and Statistics*, **53**(1), 1–10.

Gingras, Y. (2016) *Bibliometrics and research evaluation: Uses and abuses*. Massachusetts: MIT Press.

Green, J. (2016) *Aggregation in economic analysis*. New Jersey: Princeton University Press.

Harfitt, G. (2015) *Class size reduction*. New York: Springer.

Harmon, O. (1988) The income elasticity of demand for single-family owner-occupied housing: An empirical reconciliation. *Journal of Urban Economics*, **24**, 173–185.

Igan, D. and Loungani, P. (2012) Global housing cycles. *IMF Working Paper*, WP/12/217, 1–56.

Jepsen, C. (2015) Class size: Does it matter for student achievement? *IZA World of Labor*, **90**, 1–10.

Leamer, E. (2007) *Housing IS the business cycle*. Kansas City: Federal Reserve Bank of Kansas City.

Lipsey, M. and Wilson, D. (2000) *Practical meta-analysis*. Thousand Oaks: Sage.

Mayo, S. (1981) Theory and estimation in the economics of housing demand. *Journal of Urban Economics*, **10**, 95–116.

Mishel, L. and Rothstein, R. (Eds.)(2002) *The class size debate*. Washington, DC: Economic Policy Institute.

Silver, M. (2012) Why house price indexes differ: Measurement and analysis. *IMF Working Paper*, WP/12/125, 1–38.

CHAPTER 19

Concluding Your Study

Format

After the completion of data analysis, the next step of the research process is to develop the conclusion of your study. It consists of the following sections:

- Summary;
- Contributions and implications;
- Limitations;
- Recommendations; and
- Suggestions for future research.

We will discuss each section below.

Summary

The Summary section, which comprises about two pages, recapitulates the rationale for the research, the research question, scope, and objectives. Thereafter, provide a statement of the research framework, model, or hypothesis as appropriate, the research design, and the methods of data collection.

The next paragraph of the Summary section presents the main findings from the data analysis. Be concise; do not go into details of data analysis such as *t* tests, and so on. It is also important to go back to the objectives and state whether they have been met. Sometimes, they are not met, and you should provide good reasons.

The Summary section is not the place to introduce anything new, such as a new variable you have not considered, new interpretations, or new data.

Contributions and Implications

This is the "so what?" section to highlight the theoretical and practical *contributions* of the study. These may include

- identifying and solving a new problem;
- introducing a fresh perspective to an old problem;
- modifications of an existing theory;
- applying an existing theory to a different environment; and
- discovery of new facts.

It is clear from these points that it is important to state your contribution in relation to existing literature. Avoid overstating your case when challenging an existing theory or bridging a research gap in knowledge. The argument has to be persuasive, but you will lose credibility with unreasonable claims.

After discussing and making a strong impression on your contributions, the next step is to consider the *implications* of your research. The implications may be theoretical or practical, such as suggestions to include new variables or modify a public policy.

Limitations

The limitations section provides an opportunity to be reflective. The first limitation of your study may be *philosophical*. For example, you may be interested in causal laws, but it may not apply to human behavior.

The second limitation may be theoretical. There may be some theoretical or conceptual problems. For example, cultural theories of economic development cannot fully explain why certain communities are more innovative. There may be omitted variables and dynamic feedback effects.

The third limitation is *methodological*, such as in research design, sampling, and methods of data collection. For example, a comparative design does not prove causality, and a case study cannot be used to generalize its findings to other settings. To save time or cost, you may have used a biased or small sample. Finally, there may be missing, imprecise, or incomplete data.

The final limitation concerns the *analysis of data*. For example, as discussed in Chapters 15 and 16, there are many issues in regression analysis, and failure to address these issues adequately may be a limitation of the study. You should anticipate possible objections and indicate how you have tried to overcome these limitations. The reader may find your analysis unpersuasive or may have an alternative theory.

Recommendations

Where appropriate, you may recommend certain actions based on your findings. For example, you may recommend a change in public policy. In this case, you should be clear on

- the benefits and costs;
- feasibility;
- implementation issues; and
- evaluation of its impact.

For more details on policy evaluation, see Nagel (2002).

Suggestions for Further Research

The final section of the concluding chapter may contain suggestions on how to extend the work in future.

A good source of such suggestions is the limitations of your study. Here, you can suggest one or two ways to overcome these limitations. When reviewing the literature, you may also come across a related research gap that is not considered in your study. Similarly, in analyzing the data, it may strike you that there is a different way to analyze the data. Being creative is part of research (Ulibarri et al., 2019).

References

Nagel, S. (Ed.) (2002) *Handbook of public policy evaluation*. Thousand Oaks: Sage.

Ulibarri, N. et al. (2019) *Creativity in research*. London: Cambridge University Press.

This page intentionally left blank

CHAPTER 20

The Research Report

Format for Research Report

The format for the research report depends on whether it is a dissertation, thesis, business research report, conference paper, or journal article. For example, a dissertation, thesis, or business research report should include the following:

- Title fly page;
- Title page;
- Letter of transmittal (for business research report);
- Letter of authorization (for business research report);
- Table of contents;
- List of figures and tables (optional), abbreviations (if necessary), and table of court cases (where applicable);
- Acknowledgments;
- Summary (or abstract);
- Body;
- Conclusion;
- Appendix; and
- References.

The *title fly* contains the title of the research report. The title should be concise. Terms such as "A case study," "The relationship between" and "its effects on" are redundant. For example, "Income and housing demand" is a better title than "A regression analysis of the effect of income on housing demand."

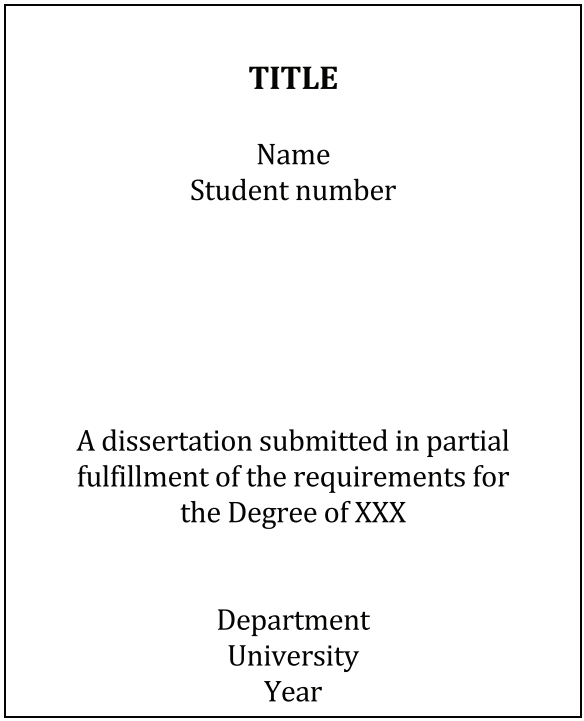

Fig. 20.1 Title page for a dissertation.

The *title page* includes the title of the report, the organization for which the report was prepared, the author(s), and the date (see Fig. 20.1).

In a business research report, the *letter of transmittal* releases the report to the recipient (Fig. 20.2). The *letter of authorization* includes the person(s) responsible for the project, background, scope, research methodology, fees, deadlines, and other information.

The *table of contents* (Fig. 20.3) includes headings for acknowledgments and the abstract as well as appendixes and references. The lists of figures, tables, court cases, abbreviations, illustrations, schedules, or maps are optional. Align the subheadings and ensure that they do not dangle. For example, subsection 2.2.2 is required if there is a subsection 2.2.1. There should not be too many levels. Three or four digits should suffice, for example, 2.1.1 or 2.2.1.1.

NAME OF ORGANISATION
Recipient's name
Position
Organization
Date

Dear XXX,
Subject
Here is my report on ... which was prepared according to your letter of authorization dated...

We found that...

The report recommends that...

Thank you for the opportunity to...

Sincerely,
XXX
Name and position

Fig. 20.2 Letter of transmittal.

The *acknowledgments* page is not included in a dissertation submitted for examination so that examiners do not know the supervisor(s). The page appears only in the final report.

The *summary* or *abstract* summarizes the research problem, objectives, hypothesis (or framework, as appropriate), methodology, main findings, and implications. This should take about two pages.

As shown in Fig. 20.3, Chapter 1 outlines the research problem, objectives, scope, and organization of study. State the research problem clearly, and explain why it is interesting and important. For instance, the problem may arise from conflicting views about an issue. Then state why its resolution is important.

Chapter 2 provides the literature review and develops the hypothesis, model, or framework for the study. Focus on the key concepts rather than on the authors. Further, do not review every detail about the problem. As shown in Fig. 20.3, the chapter ends with the hypothesis, model, or research framework.

CONTENTS

Fig. 20.3 Table of contents.

Chapter 3 provides the methodology, and an example is given below:

3 Research methodology

3.1 Research design
3.2 Sampling
    3.2.1 Population
    3.2.2 Sampling frame
    3.2.3 Sampling method
    3.2.4 Sample size
3.3 Methods of data collection
    3.3.1 Questionnaire
    3.3.2 Pretest
    3.3.3 Interview
    3.3.4 Response rate
3.4 Data collection and processing
    3.4.1 Selection of interviewers
    3.4.2 Equipment
    3.4.3 Training
    3.4.4 Supervision
    3.4.5 Data processing

The exact headings for this chapter depend on the research design adopted. For instance, if it is a case study, then the subsections on sampling are not required. However, you need to explain why you select a case study design, such as to probe into an issue and why you select particular cases. Recall that considerations for case selection may include typical cases, unique cases, falsifying cases, and so on.

The methods of data collection section are where you define and discuss measures for the variables. For example, in a survey design, it may be a questionnaire that uses a Likert rating scale. In regression analysis, this is where you define the variables and provide details on how they are measured. For instance, you may measure the "monthly mortgage interest rate" using the quoted mortgage rates of the four major banks in the

country in the middle of each month. In this case, the subheadings for Section 3.3 are as follows:

3.3.1 Dependent variable
3.3.2 Independent variable 1
Etc.

For example, if you are regressing house price on income, interest rates, and so on, then the subheadings should be

3.3.1 House price
3.3.2 Income
3.3.3 Interest rates
Etc.

In the subsections, explain how you measure each concept, such as house price. Often, a house price index is used, and details of its constructions are required because of possible biases.

Chapter 4 contains the data analysis. The headings will depend on the type of data. If it is a qualitative (interpretive) approach, the "data analysis" is largely textual and may spread over several chapters. For instance, in content analysis, the data analyses include the selection, mining, and integration of textual content into a coherent picture.

In quantitative studies, it is advisable to conduct exploratory data analysis to get a sense of the data patterns before applying the more sophisticated statistical methods. Extract, not reproduce, the main items from the printout from statistical software packages (see Table 20.1).

The final chapter of the report contains the *conclusion* and *recommendations*. Use the guidelines given in Chapter 19.

The *appendix* includes charts, detailed calculations, graphs, general tables, data collection forms, maps, and other support material. There is no excuse for sloppy bibliographies, so pay attention to detail.

Format for Journal Articles

The format for journal articles vary across journals. A standard structure is as follows:

Table 20.1 Regression results.

	Dependent variables		
Independent variables	Y	ΔY	$\log(Y)$
Constant	3.63(8.83)***	0.16(2.38)**	0.02(0.82)
T	0.11(3.05)***	—	—
V	1.36(3.54)***	—	—
ΔV	—	1.93(3.19)***	2.31(1.17)
$\log(M)$	—	0.61(6.42)***	0.58(4.91)***
R-square	0.86	0.67	0.54
F-statistic	23.45***	23.24***	7.75*

Figures in brackets refer to t values.
*, **, *** indicate significance at 0.1, 0.05 and 0.01 levels, respectively.

Title
Author(s)
Affiliation
Abstract
Keywords

Introduction
Literature review
Hypothesis, model, or framework
Methodology
Results and discussion
Conclusion
References

Unlike the *Summary* in a research report or thesis, the *Abstract* is brief. It summarizes the paper in about 100 words.

The *Introduction* provides the context, the research question, and justification. The latter often specifies the research gap that the paper seeks to bridge within the current body of knowledge. This section may end with a description of how the paper is organized.

The *Literature review* provides greater detail on what is known, the research gap, and the development of the hypothesis, model, or

framework. For a journal paper, cite only the main works by focusing on the development of the key ideas.

The *Methodology* may consist of subsections on the research design, sampling, methods of data collection, the field data collection process, and data processing. It is important to justify your choices, such as why a particular research design or sampling method is selected. This section is called by different names (e.g. methods), which may be confusing.

The *Results and discussion* section contains the data analysis, followed by a discussion of how the results compare with previous studies. This section ends with a discussion of the theoretical, empirical, or practical implications of your findings.

Finally, in the *Conclusion*, provide a brief summary of the issue, hypothesis, methodology, results, and implications.

The Writing Process

Resist the temptation to start writing at the beginning of the research project. Begin with a *brief outline* of the table of contents to provide a structure for the study.

As work progresses, fill the table of contents with *brief notes* on the research problem, scope, and so on. These notes can be in point form. Do not worry if you are not able to specify these items at this stage; they will become clearer as the research progresses. For example, the literature on urban regimes may look like this:

Literature review

Pluralist theory
- Local urban power and decision-making is dispersed (Dahl, 1961)

Elite theory
- Power is concentrated in the business community (Hunter, 1953)
- Tends to overstate the power of business

Urban growth machine
- Exerts a high degree of influence through a coalition of businesses, real estate, lenders, and possibly labor unions (Molotch, 2007)

- Focuses on local economic development through tax breaks and infrastructure investment to counter urban decline
- Limits to distributional policies because of mobility of capital (Peterson, 1981)

Regime theory
- Cooperation of city's top officials and businesses in a *stable* alliance to get things done (Stone, 1989)
- Different types of regimes: Developmental, maintenance, and progressive, with different ideologies, discourses, and strategies

Critique of regime theory
- US-centric and may not exist elsewhere (Keating, 1991)
- Neglect of external influences on cities (Imbroscio, 2010)
- Excessive focus on economic development projects (Pagano and Bowman, 1995)

Write the *first draft* of your research report only after you have completed the data analysis. Thereafter, there will be much rewriting. Do not expect your supervisor to be your proofreader and, even if you write reasonably well, it is helpful to ask someone to proofread your draft for style, logical errors, typographical errors, readability, and so on. Be realistic; proofreading is tedious, and you may wish to ask competent people to proofread only one or two chapters rather than the entire work. The downside is that proofreaders do not get a real feel of the entire work, even if they have the table of contents.

Writing Style

There are many reference books on writing styles, such as Strunk and White (1979), Day and Gastel (2011), and Szuchman (2013). What follows is a summary from different sources.

(a) *Voice*

Use the *active voice*. In the past, many journals encouraged the use of the passive voice because it is impersonal. Thus, write

"I *interviewed* the project manager."

rather than

"The project manager *was interviewed* by me."

Nowadays, the active voice is increasingly used but do not do it excessively. For example, use "The sample size is 50" rather than "I took a sample of 50 students."

(b) *Tenses*

Use the *past tense* to
- describe your methodology, for example, "We *interviewed*…"; or
- refer to past studies, for example, "Marx (1886) *argued* that…"

Use the *present tense* to
- express general truths, for example, "birds *fly*"; or
- discuss your findings, for example, "From the results, I *suggest* that…"

If the time is indefinite, use the *present perfect tense*:

He *has gone* to school.

If there are two actions, use the *past perfect tense*:

If he *had done* what management wanted, they would have promoted him.

The words *did*, *shall*, *will*, and *would* are associated with the present tense:

I *did* not *go* home yesterday.

I *will do* what is required.

(c) *Wordiness*

Cut out unnecessary words by editing it ruthlessly, and never use a long word when a short one will do. For example, the following paragraph is wordy:

> Each and every country has its own unique system of registration of contractors. However, the selection criteria that form the basis of the

> various forms of registration may differ from country to country. The main objective of a system of registration of contractors is to ensure that there are suitable contractors to tender for projects. To tender for projects, contractors should have all the necessary qualifications and be competent in what they are doing...

Better:

> The registration of contractors ensures that only qualified contractors tender for projects. Countries differ in their criteria for registration.

Omit needless words. For example, refer to the list below for some common redundancies (left side) and how they may be shortened (right side):

absolutely essential	essential
add together	add
advance warning	warning
as to whether	whether
at this point in time	now
basic fundamentals	basic
brand new	new
cancel out	cancel
close proximity	proximity
combine into one	combine
combined total	total
complex maze	maze
consensus of opinion	consensus
contributing factor	factor
end result	result
for the purpose of	for
in the event that	if
in the first instance	first
in the majority of the cases	most
investment purpose	investment
lag behind	lag
may or may not	may
mixed together	mixed
new initiative	initiative

new innovation	innovation
occasional irregularity	irregularity
on a weekly basis	weekly
owing to the fact that	since
past experience	experience
personal opinion	opinion
plan ahead	plan
reason why	reason
revert back	revert
serious crisis	crisis
small minority	minority
summarize briefly	summarize
take action	act
the manner in which	how
the reason why is that	because
with the exception of	except

Break *long sentences* into shorter ones, but vary your sentences to avoid monotony. In general, a sentence should not exceed 30 words or about two lines of spacing. A long sentence is hard to read, and may contain too much information or too many ideas for the reader to digest. Rewriting your manuscript is hard work, but remember that wordiness will not attract the intelligent reader.

(d) *Qualifications*

Take a stand when you need to, and qualify your statements if necessary. Do not pepper the report with too many qualifications such as "seem," "apparent," "maybe," "perhaps," "occasional," or "generally." Call a spade a spade; write, "We interviewed 20 shoppers" rather than "20 people" who may not be shoppers.

(e) *Spelling*

Use a consistent style of English, for example, American or British spelling. Non-native speakers tend to mix them up unconsciously, such as "film" and "movie." Here are some examples:

American	British
neighbor	neighbour
center	centre
realize	realise
program	programme
skeptic	sceptic
aluminum	aluminium
gray	grey
kerb	curb
maneuver	manoeuvre
mold	mould
polyethylene	polythene
toward	towards
aging	ageing

Some words have similar meanings:

mutual fund	unit trust
sales tax	value-added tax
amortization	depreciation
receivables	debtors
by-laws	articles of association
real estate	land and buildings

(f) *Metaphors*

A metaphor is a figure of speech. It is a word or phrase that compares two things that are not alike but have something in common.

Metaphors such as "the *heart* of the city" provide useful mental images, but they may blind us as well. The city has no "heart" to keep "pumping," whether "healthy" or "ailing." The use of metaphors in writing is inevitable. If properly used, a metaphor serves as an important rhetorical device to persuade the reader. Examples of business metaphors include housing *bubbles*, supply *chain*, demand *curve*, competitive *race*, *lemon* markets, *spiking* prices, *creeping* inflation, and starting on a wrong *footing*.

(g) *Other rules of grammar*

Apostrophes can be confusing for the non-native speaker. In American usage, James'*s* book is acceptable.

For quotation marks, Americans use double quotes for "nice" instead of 'nice' and place the full stop within quotes, such as

This dog is "cute."

For a series of terms, there is an optional comma before "and" such as

red, blue, and green.

Use "which" to indicate alternatives, for example,

Which unit is your flat?

Otherwise, use "that" as in

Beware of dogs *that* bite.

Subject-verb agreements can be tricky. If "one of" is used, use the plural form:

He is *one of* those *boys* who *are*...

For *each*, *either*, *everyone*, *everybody*, *nobody*, and *someone*, use the singular verb, as in

Everybody *loves* Raymond.

For *none*, use a singular verb if it means "no one," as in

None of us *is* going.

If we want to mean more than one thing or person, use the plural verb, for example,

None *are* so capable that *they* can do it by themselves.
None of us *are* going.

The word "*any*" may take a singular or plural verb, for example,

Any of the answers is acceptable.

Additional nouns that connect to a singular subject do not change verb agreement, for example,

James, *as well as* John, *likes* to jog.
James, *together with* John, *likes* to leave first.

In *collective nouns*, use the singular, as in

The team *is* strong.

Do not verb nouns. For example, banks lend money rather than loan money. The variable *X* has an impact on *Y*, and not *X* impacts *Y*. You conduct experimental trials, but do not trial experiments.

Place adverbs as close as possible to the verb it modifies, for example, the first sentence is better:

I will *happily* assist you.
I will assist you *happily*.

There is a difference in meaning between these two sentences:

He *works only* on Sundays.
He *only works* on Sundays.

In the first sentence, he does not work on other days. In the second sentence, he does nothing else on Sundays.

Avoid *sexist or racist language*. Use "supervisor" rather than "foreman." Often, the masculine includes the feminine, that is, use "him" and not "his or her," as in the school regulation, "A student should not dye his hair." The plural form is better, for example, "Students should not dye their hair." Avoid "he or she" or the ugly "(s)he."

If abbreviations are used, spell it out the first time, for example, "The Housing and Development Board (HDB)…" If necessary, provide a list of abbreviations.

Writing Form

(a) *Tables, charts and diagrams*

Number tables, charts and diagrams consecutively in each chapter (for example, **Figure 4.2** Trends in house prices.). Place table headings at the top or bottom of the table.

Place tables, charts, and diagrams just after the paragraph where you refer to them in the text. You may reference them by the expression "As shown in Chart 1.2, …" or by using brackets, for example, "ABC Corporation relies on internal rather than external financing (see Table 1.2)." Avoid reifications such as "The graph *points to* the fact that…" Graphs do not point, suggest, or show anything.

Use *single line spacing* with a smaller font size (for example, 9 or 10 points) in table cells. As far as possible, do not break tables across pages. If this is not possible, then repeat the Table captions on the next page for truncated tables, for example, "Table 1.2 Output by various industries (continued)."

(b) *Pagination*

Number the preliminary pages (before the first chapter) in lower case Roman numerals (for example, i, ii, iii, and so on) centered at the bottom of the page. The title page is not numbered. Number the body of the report, starting from the first page of the first chapter, consecutively.

(c) *Footnotes and endnotes*

Avoid footnotes and endnotes. These side comments, located at the bottom of the page or end of the chapter, respectively, tend to disrupt the flow of the narrative.

(d) *Quotations*

Block quotations should be typed single-line spacing and properly referenced by author and page, for example, according to Hije (1980),

> The "organization problem" concerns the nature of labor as a quasi-commodity, one that is *reflexive* and active. Unlike machines, the worker can, to some extent, select the level of effort required. (p.2) (emphasis added)

Alternatively, place the source after the passage:

> The "organization problem" concerns the nature of labor as a quasi-commodity, one that is *reflexive* and active. Unlike machines, the worker can, to some extent, select the level of effort required. (emphasis added)
>
> (Hije, 1980, p.2)

Do not block shorter quotations, for example,

> As Hije (1980) observed, "those who have trouble with..." (p.9).

Alternatively, place the page number in front, as in

> As Hije (1980, p.9) observed, "those who have trouble with..."

Use quotations sparingly. It is better to write the report in your own words. Avoid popular quotations because they look stale. Avoid useless words when citing the literature, for example,

> David Hije (1980) in his study noted that…

Here, both "David" and "in his study" are unnecessary.

(e) *Abbreviations*

ibid. Short form for *ibidem* (in the same place). Often used in conjunction with a footnote where the same source is cited more than once, for example, *Ibid.*, p.24.

op. cit. Short form for *operato citato* (in the work cited). Often used in footnotes where we cite the source previously but not immediately preceding, for example, we cite source A, followed by source B, and then A again.

cf. Compare, as in cf. Hocking (1978).

Chap.	Chapter.
ed.	Editor or edition.
Eds.	Editors.
2nd ed.	Second edition.
et al.	And others, from Latin *et alii*. Mostly used in bibliographic references, for example (Smith *et al*., 1982).
ff.	And in the following pages, as in p.86ff.
n.	Footnote, as in n.4.
pp.	Pages, as in pp.34–5.
passim.	In various places in the text.
sic.	So or thus. Inserted in square brackets to show the original source contains an obvious error. For example, "The dog bark [*sic*]…"
v.	Versus. Used in court cases but conventionally translated as "and."
vols.	Volume, as in 4 vols.
Vol.	Volume, as in Volume 4.

(f) *Citations and references*

Various systems of referencing are available. The recommended system for the physical sciences is the number system (for example, [1], [2], and so on). For the social sciences, use the Harvard Referencing System or the American Psychological Association (APA) style.

The author, year, and, where appropriate, page numbers may appear in the text in the following manner:

> Garde (1968) found that...However, other studies (Vix, 1974; King, 1978, pp. 34) reported…

If two or more publications are cited, the one that is published earlier is cited first (i.e. "Vix" comes before "King" because his paper appeared in 1974, while King's paper appeared in 1978). List the references alphabetically by author's surname at the end of the research report:

> Garde, K. (1968) *Project failures*. London: Wiley.
> King, H. (1978) *Project risks*. New York: McGraw-Hill.
> Viz, N. (1974) *Project risk management*. New York: Macmillan.

If you cite two or more works of the same author, list them in chronological sequence. For works published in the same year, use Jones (1968a), followed by Jones (1968b).

If there are more than two authors, all the names will appear in the bibliography but "et al." (and others) is used in the text, for example, (Smith et al., 1983).

Examples of bibliography

A book by a single author:

> Fox, D. (1969) *Managerial economics*. London: Longman.

A book or technical report by more than one author:

> Strunk, W. and White, E. (1979) *The elements of style* (3rd ed.). New
> York: Macmillan.

For handbooks, (Vols. 1–3) replaces (3rd ed.); for translated book, (D. Smith, Trans.) replaces (3rd ed.); for technical report, (Report No. 12-1234) replaces (3rd ed.). Indent the second line beginning with "York."

Edited book:

> Hall, P. (Ed.) (1966) *Von Thunen's isolated State*. Oxford: Pergamon.

Chapter in an edited book:

> Stone, P. (1965) The prices of building sites in Britain. In P. Hall (Ed.),
> *Land values* (pp.12–27). London: Heineman.

Journal article:

> Pite, D., and Tesa, C. (1981) The crisis of our time. *Journal of Environmental Housing*, *23*(3), 123–141.

Newspaper article, no author:

CIDB perceives strong growth for construction sector. (1993, December 17) *The Straits Times*, p. 47.

Use "pp.1, 25." for discontinuous article.

Newspaper article, with author:

Tan, T. S. (1993, December 12) URA to auction 12 sites in Jurong. *The Straits Times*, p. 36.

Conference paper:

Unpublished:

Brent, B. (1983, May) *Valuation of hotels*. Paper presented at the meeting of the Society of Valuers, Melbourne, Victoria.

Published:

Brent, B. (1988) Valuation of hotels. In E. Dave (Ed.), *Proceedings of the Third International Symposium on Valuation* (pp. 3–9). Vancouver: Zeti Press.

Unpublished manuscript:

Jameson, K. (1993) *Testing concrete strength*. Unpublished manuscript.

Dissertation or thesis

Lim, K. (2000) *Lean building construction*. Unpublished undergraduate dissertation, Department of Building, National University of Singapore.

Replace "undergraduate" with "master" or "doctoral" and "dissertation" with "thesis" where appropriate.

Web site

Global Concrete Network. (2010 Jan). *New understanding of concrete*. New Jersey: GCN. Retrieved 20 March, 2010 from the World Wide Web: http://gcn.org/concrete.html.

Do not segment the bibliography into journal articles, dissertations, and so on. Place them together.

(g) *Citing legal authorities*

For court cases, the parties to a decision are underlined or indicated in italics but not the connecting "v.," for example, "In *Donoghue* v. *Stevenson*

[1932], the House of Lords decided that...” The reference to this case appears in the table of court cases (after the list of figures and tables, not with the bibliography) as:

TABLE OF CASES

Donaldson v. Hemmett (1901) 11 Q.L.J. 35 23
Donoghue v. Stevenson [1932] A.C. 562 28
Etc.

The names of the parties involved are not in italics when they appear in the table. The year of the case is given in square or round brackets according to the status of the law report or journal, followed by the source document, which may be abbreviated (A.C. here, which stands for Appeal Cases, Privy Council and House of Lords, England), but a separate list of abbreviations must appear after the table of court cases. The volume (11) and starting page (35) are also given.

For statutes, brackets may be used in text, for example, “The Chief Surveyor or any Government surveyor authorized by him may undertake field and office checks on the title survey work of a registered surveyor or a licensed corporation or partnership.” (Land Surveyors Act, Cap 156, Revised edition, 1991, s. 36(1)). Alternatively, write “The Land Surveyors Act, Cap 156, Revised Edition, 1991, s. 36(1) stipulates that the Chief Surveyor…”

Lawyers often use footnotes in legal citation. For example, “The employee undertakes that he is reasonably skilled,[1] that he will…” appears in the main text and the footnotes are given below.

[1]*Harmer* v. *Cornelius* (1858) 5 CB (NS) 236.
[2]Land Surveyors Act, Cap 156, Revised edition, 1991, s. 36(1).

Sometimes, the footnotes refer only to the section of the Act, for example, in the main text, “The Land Surveyors Act, Cap 156, Revised Edition, 1991[1] stipulates that the Chief Surveyor…”

[1]Section 36(1).

(h) *Units of measurement and numbers*

Use the metric system unless tradition dictates otherwise (for example, per square foot (psf) in quoting property prices). Use the singular form, as

in 3cm rather than 3cms, and do not use the period unless cm appears at the end of a sentence. The choice of writing 3cm (no space in the abbreviation) or 3 cm (with space in the abbreviation) depends on the prevailing house style. If the original measurement is in imperial units, the converted metric equivalent appears in brackets, as in 3ft (0.91m). For areas and volumes, use m^2 or m^3 rather than sq m or cu m.

Write out numbers below ten, for example, three houses. Exceptions include ratios (12:1), page numbers, and money ($3). When the number is large, there are several options such as "2.3 million" or "2.3×10^{-5}." Do not begin sentences with numbers; if you must, spell it out, such as "Twenty shoppers jumped the queue." For multiple units such as meters per second, use "m/s" or "ms^{-1}."

Decimals should reflect how precise the item is measured. For example, land areas are expressed to one decimal place (for example, $90.5m^2$), and the coefficient of determination R^2 is expressed to two decimal places. Decimals should be consistent, such as 10.2, 23.1 and 0.0 when they appear in tables, and aligned properly.

(i) *Mathematical symbols and equations*

Mathematical symbols are in italics, such as $y = f(x)$. Note f is also in italics but not the brackets. Avoid unnecessarily complex mathematics. If necessary, place long and complicated derivations in an appendix. Letters representing vectors (lower case) and matrices (upper case) are in boldface, for example, $\mathbf{v}$ and $\mathbf{V}$, and not in italics.

Use subscripts to enhance clarity rather than add complexity. Subscripts in expressions such as x_t should be in italics unless they are numerals, for example, x_1. Similarly, superscripts are in italics unless they are numerals. This rule applies to vectors and matrices, for example, $\mathbf{A}_1$ and $\mathbf{A}_t$. The subscripts are not in boldface letters.

All equations are indented or centered and, if required, place equation numbers on the left or right margin, for example,

$$R = 2 \log (1 + g) + 3. \tag{8.3}$$

Refer to it as "Equation (8.3)" or "Eq. (8.3)." Note the full stop after the equation, that is, an equation is part of a sentence. If the

equation is in the middle of your sentence, use a comma. For example, the equation is

$$R = 2 \log (1 + g) + 3, \tag{8.3}$$

where R is rent and g is the rate of rental growth. Use "log" instead of "ln" to represent the natural log because "log n" is clearer than "ln n" or "lnn." If desired, leave a blank space between the operator and operand to enhance clarity. Similarly, note the spacing in "$1 + g$."

Use Greek letters to represent population parameters and the English alphabet to represent sample estimators. For example, b is an estimator of β.

(j) *Bullet list*

The modern version of a bullet list has fewer punctuation marks; for example,

The reasons are as follows:

- benefits
- costs
- risks

The traditional version treats it like a sentence; for example,

The reasons are as follows:

- benefits;
- costs; and
- risks.

Each bullet starts with a lower case letter unless it contains a sentence. The colon is not necessary if it is not an independent clause; for example,

The reasons include

- benefits;
- costs; and
- risks.

Sometimes, the comma is used instead of a semicolon. An "independent clause" is one that can stand alone as a sentence.

(k) *Layout*

For theses and dissertations, it is common to use 1.5 line spacing for text and single line spacing for the table of contents, table entries, and quotations. Use a margin of 25 mm except for the left margin, which is slightly bigger to facilitate binding. Center page numbers near the bottom of the page. Print on both sides of the paper in the interest of sustainability. You should check your university's requirements on the layout.

Illustrations should be neatly drawn and clearly labeled. Where a drawing is complex, use color to indicate different lines. Otherwise, use black printing ink. Enhance the clarity of your work by using diagrams and tables.

References

Day, R. and Gastel, B. (2011) *How to write and publish a scientific paper*. Westport, Connecticut: Greenwood Press.

Strunk, W. and White, E. (1979) *The elements of style*. London: MacMillan.

Szuchman, L. (2013) *Writing with style: APA style made easy*. Belmont, California: Wadsworth.

Appendix

Table A.1 Critical points for chi-square distribution.

	α(%)			
Df	10	5	2.5	1
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.09
9	14.68	16.92	19.02	21.67
10	15.99	18.31	20.48	23.21
11	17.28	19.68	21.92	24.72
12	18.55	21.03	23.34	26.22
13	19.81	22.36	24.74	27.69
14	21.06	23.68	26.12	29.14
15	22.31	25.00	27.49	30.58
16	23.54	26.30	28.85	32.00
17	24.77	27.59	30.19	33.41
18	25.99	28.87	31.53	34.81

(*Continued*)

Table A.1 (*Continued*)

	α(%)			
Df	10	5	2.5	1
19	27.20	30.14	32.85	36.19
20	28.41	31.41	34.17	37.57
21	29.62	32.67	35.48	38.93
22	30.81	33.92	36.78	40.29
23	32.01	35.17	38.08	41.64
24	33.20	36.42	39.36	42.98
25	34.38	37.65	40.65	44.31
26	35.56	38.89	41.92	45.64
27	36.74	40.11	43.19	46.96
28	37.92	41.33	44.46	48.28
29	39.09	42.56	45.72	49.59
30	40.26	43.77	46.98	50.89
40	51.81	55.76	59.34	63.69
50	63.17	67.50	71.42	76.15
60	74.40	79.08	83.30	88.38
70	85.53	90.53	95.02	100.43

Table A.2 Critical points for *t* distribution.

Df	α(%)			
	10	5	2.5	1
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681
13	1.350	1.771	2.160	2.650
14	1.345	1.761	2.145	2.624
15	1.341	1.753	2.131	2.602
16	1.337	1.746	2.120	2.583
17	1.333	1.740	2.110	2.567
18	1.330	1.734	2.101	2.552
19	1.328	1.729	2.093	2.539
20	1.325	1.725	2.086	2.528
21	1.323	1.721	2.080	2.518
22	1.321	1.717	2.074	2.508
23	1.319	1.714	2.069	2.500
24	1.318	1.711	2.064	2.492
25	1.316	1.708	2.060	2.485
26	1.315	1.706	2.056	2.479
27	1.314	1.703	2.052	2.473
28	1.313	1.701	2.048	2.467
29	1.311	1.699	2.045	2.462
30	1.310	1.697	2.042	2.457
∞	1.282	1.645	1.960	2.326

Table A.3 Critical points for F distribution ($\alpha = 5\%$).

	n_1						
n_2	1	2	3	4	5	6	7
1	161.5	199.5	215.7	224.6	230.2	234.0	236.8
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33
40	4.04	3.23	2.84	2.61	2.45	2.34	2.25
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01

Table A.3 (*Continued*)

	n_1						
n_2	8	9	10	20	30	40	∞
1	239	241	242	248	250	251	254
2	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	8.85	9.81	8.79	8.66	8.62	8.59	8.53
4	6.04	6.00	5.96	5.80	5.75	5.72	5.63
5	4.82	4.77	4.74	4.56	4.50	4.46	4.37
6	4.15	4.10	4.06	3.87	3.81	3.77	3.67
7	3.73	3.68	3.64	3.44	3.38	3.34	3.23
8	3.44	3.39	3.35	3.15	3.08	3.04	2.93
9	3.23	3.18	3.14	2.94	2.86	2.83	2.71
10	3.07	3.02	2.98	2.77	2.70	2.66	2.54
11	2.95	2.90	2.85	2.65	2.57	2.53	2.40
12	2.85	2.80	2.75	2.54	2.47	2.43	2.30
13	2.77	2.71	2.67	2.46	2.38	2.34	2.21
14	2.70	2.65	2.60	2.39	2.31	2.27	2.13
15	2.64	2.59	2.54	2.33	2.25	2.20	2.07
16	2.59	2.54	2.49	2.28	2.19	2.15	2.01
17	2.55	2.49	2.45	2.23	2.15	2.10	1.96
18	2.51	2.46	2.41	2.19	2.11	2.06	1.92
19	2.48	2.42	2.38	2.16	2.07	2.03	1.88
20	2.45	2.39	2.35	2.12	2.04	1.99	1.84
25	2.34	2.28	2.24	2.01	1.92	1.87	1.71
30	2.27	2.21	2.16	1.93	1.84	1.79	1.62
40	2.18	2.12	2.08	1.84	1.74	1.69	1.51
60	2.10	2.04	1.99	1.75	1.65	1.59	1.39
120	2.02	1.96	1.91	1.66	1.55	1.50	1.25
∞	1.94	1.88	1.83	1.57	1.46	1.39	1.00

n_1: Numerator degrees of freedom.
n_2: Denominator degrees of freedom.

Table A.4 Areas under the standard normal distribution.

Z	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.000	.004	.008	.012	.016	.020	.024	.028	.032	.036
0.1	.040	.044	.048	.052	.056	.060	.064	.068	.071	.075
0.2	.079	.083	.087	.091	.095	.099	.103	.106	.110	.114
0.3	.118	.122	.126	.129	.133	.137	.141	.144	.148	.152
0.4	.155	.159	.163	.166	.170	.174	.177	.181	.184	.188
0.5	.192	.195	.199	.202	.205	.209	.212	.216	.219	.222
0.6	.226	.229	.232	.236	.239	.242	.245	.249	.252	.255
0.7	.258	.261	.264	.267	.270	.273	.276	.279	.282	.285
0.8	.288	.291	.294	.297	.300	.302	.305	.308	.311	.313
0.9	.316	.319	.321	.324	.326	.329	.332	.334	.337	.339
1.0	.341	.344	.346	.349	.351	.353	.335	.358	.360	.362
1.1	.364	.367	.369	.371	.373	.375	.377	.379	.381	.383
1.2	.385	.387	.389	.391	.393	.394	.396	.398	.400	.402
1.3	.403	.405	.407	.408	.410	.412	.413	.415	.416	.418
1.4	.419	.421	.422	.424	.425	.427	.428	.429	.431	.432
1.5	.433	.435	.436	.437	.438	.439	.441	.442	.443	.444
1.6	.445	.446	.447	.448	.450	.451	.452	.453	.454	.455
1.7	.455	.456	.457	.458	.459	.460	.461	.462	.463	.463
1.8	.464	.465	.466	.466	.467	.468	.469	.469	.470	.471
1.9	.471	.472	.473	.473	.474	.474	**.475**	.476	.476	.477
2.0	.447	.478	.478	.479	.479	.480	.480	.481	.481	.482
2.1	.482	.483	.483	.483	.484	.484	.485	.485	.485	.486
2.2	.486	.486	.487	.487	.488	.488	.488	.488	.489	.489
2.3	.489	.490	.490	.490	.490	.491	.491	.491	.491	.492
2.4	.492	.492	.492	.493	.493	.493	.493	.493	.493	.494
2.5	.494	.494	.494	.494	.495	.495	.495	.495	.495	.495
2.6	.495	.496	.496	.496	.496	.496	.496	.496	.496	.496
2.7	.497	.497	.497	.497	.497	.497	.497	.497	.497	.497
2.8	.497	.498	.498	.498	.498	.498	.498	.498	.498	.498
2.9	.498	.498	.498	.498	.498	.498	.499	.499	.499	.499
3.0	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499

The table shows the area under the curve from the center ($Z = 0$) to Z. For example, the area from $Z = 0$ to $Z = 1.96$ is 0.475.

Index