Making Sense of Statistical Methods in Social Research

Keming Yang

Making Sense of Statistical Methods in Social Research

Making Sense of Statistical Methods in Social Research

Keming Yang

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

First published 2010

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1S

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte Ltd
33 Pekin Street #02-01
Far East Square
Singapore 048763

Library of Congress Control Number: 2009931921

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-84787-286-9
ISBN 978-1-84787-287-6

Typeset by C&M Digitals (P) Ltd, Chennai, India
Printed in India at Replika Press Pvt Ltd
Printed on paper from sustainable resources

To Charles, Marie and Lixin

Contents

Lists of Figures, Tables and Abbreviations

Figures

Tables

Abbreviations

ANOVA	analysis of variance
BHPS	British Household Panel Study
CFA	confirmatory factor analysis
CFI	comparative fit index
CLA	cluster analysis
COA	correspondence analysis
CPI	consumer price index
DA	discriminant analysis
DE	design effect
df	degrees of freedom
EFA	exploratory factor analysis
ESS	European Social Survey
GSS	General Social Survey
HLM	hierarchical linear models (also the name of software)
ICC	intra-class correlation coefficient
IQR	inter-quartile range
IQV	index of qualitative variation
LATE	local average treatment effect
LGCM	latent growth curve models
MANCOVA	multiple analysis of covariance
MANOVA	multiple analysis of variance
MDS	multidimensional scaling
PCA	principal component analysis
PSID	Panel Study of Income Dynamics
PSU	primary sampling unit
QCA	qualitative comparative analysis
RMSEA	root mean square error of approximation
RMSR	root means square residual
s.d.	standard deviation
s.e.	standard error
SEM	structural equation models
SES	socio-economic status

SNA	social network analysis
SRS	simple random sampling
SUTVA	stable unit treatment value assumption
UKHLS	UK Household Longitudinal Study

Preface

This book is written for senior undergraduates, postgraduate students and junior researchers who have learnt some introductory statistics but are new to the practice of applying statistical methods in social research projects. I hope to help them not only to select the most appropriate method for their own research but also to be able to identify potential pitfalls. To achieve this, I offer a critical account and overview of the concepts that underlie the statistical methods popularly used in social science research, focusing on the logic for making sense of these methods in answering substantive research questions.

This book is a result of my experience of learning, teaching and using statistical methods in social research during the past 15 years. Although I earned a postgraduate degree in statistics and the title of 'Chartered Statistician' from The Royal Statistical Society, I am not — and I have no plan to devote myself to becoming — a methodologist who does research mainly on statistical methods. I am a sociologist who often finds statistical methods very useful. Numeric data usually cover a large sample of cases that represent an even larger population, the data are usually of high quality and available free of charge and finally, we can analyse them in many different ways and submit our analysis to some well established procedures.

The use of statistics in social research, however, has been highly controversial. At one extreme, statistics suffer many severe critiques: quantitative analysis is shallow since it does not touch the meaning of social actions; it cannot even scratch the surface of the richness of cultural values; it assumes people behave simply because of their attributes; it overlooks social interactions; it pretends to make causal arguments but actually it cannot and so on. It sounds as if we should stop using statistics and try something else, although no one has openly said that. At the other extreme, there are many 'quantitative researchers' who simply ignore these critiques and use statistical models as a routine. Such divided opinion is also reflected on the two sides of the Atlantic. While many social scientists in the USA take statistics as their default method, their British counterparts usually keep away from it, except for a handful of institutions. Courses of introductory statistics are compulsory in the American social science curriculum, while the UK's Economic and Social Research Council has long recognized the lack of statistics training among the British graduates.

So, what shall we do with statistics? First, statistics may not be everyone's cup of tea, but nor should anyone who wants to apply statistical methods in social research be discouraged. Second, applying statistical methods in social research clearly demands far more than just being comfortable with numbers, maths or computer software. It is the sensible logic that counts. The key question is: How

can we make sense of statistical methods in social research? I do not think that we can reach a satisfactory answer simply by using variables related to social issues, a popular practice in statistics textbooks written for social science students. Statistical methods will remain statistical even after we rename variables or cases with popular sociological terms. If we want to better understand and apply statistical methods, then we need to closely examine the logic of each method and think hard why and when each of them will make sense. This book is about such logic.

A few words on what this book is not about. First, it is not about the philosophy or the epistemology of social research. Second, the reader should not expect a full coverage of the technical details of statistical analysis. The book can be used as a textbook for senior undergraduate students or postgraduate students in any social science discipline, but it is not the same as *Social Statistics: An Introduction Using SPSS for Windows*, *Statistics for the Behavioural and Social Sciences* or *Statistics: A Tool for Social Research*, etc., which offer a very good introduction to elementary statistics and popular statistics software such as SPSS. This book does not do those things. We will discuss specific techniques and we may even look carefully at some equations and models, but it is the logical and conceptual foundation of statistical methods, not merely these methods per se, that we shall focus on. Third, this book does not cover methods for analysing experimental data. It is for people who do observational studies, that is, there is little that they can do to intervene or control the phenomena under study.

Finally, let me take this opportunity to thank the people who have helped me deliver this monograph. I must firstly thank my editor, Patrick Brindle, for his encouragement, patience and critical comments. This book will not come out without his efforts. Jeremy Toynbee polished the whole draft by making many useful modifications and corrections, for which I am extremely grateful. I also very much appreciate the specific comments and suggestions, both complimentary and critical, of the nine anonymous reviewers of my book proposal. Roberto Franzosi, David Byrne and Malcolm Williams provided some valuable comments on the book proposal as well. Gareth Williams read the whole draft and provided many excellent suggestions, for which I am extremely grateful. My colleague Juan Morillas read the chapter on time-related methods, and I thank him for his useful comments.

As always, I see the value of my work in the eyes of my wife Lixin, my daughter Marie and my son Charles. They have made great contributions to the production of this book by allowing me the time to work on it. I dedicate the book to them as a token of my appreciation.

ONE

Introduction

Chapter Contents

The Status of Statistics in the Social Sciences

The history of social sciences after the Second World War can easily lead people to believe that statistical methods have enjoyed not only legitimacy but popularity (Raftery, 2001). First of all, some social scientists have made significant contributions by employing and developing statistical methods, for example, Paul Lazarfeld, Hubert Blalock, Otis Dudley Duncan, Leo Goodman, to name only a few of the most influential. For the past few decades, statistical methods have become so popular that, for some, it is the only tool in their research toolbox. In addition, some leading academic journals regularly publish papers based on sophisticated statistical methods. Institutionally, nearly all sociology, political science, and business departments in American universities now make learning statistics compulsory.

Nevertheless, there has always been a voice of caution, if not utter objection, to using statistical methods in the social sciences. Back in the mid-1950s, Hubert Blumer (1956) pointed out several problems with quantitative methods in general when used to understand group processes and cultural values. However, he did not offer an attractive alternative to statistical methods for constructing powerful models built on a large amount of data. More recent critiques have been highly specific and therefore more compelling, many coming from quantitative methodologists themselves, including Otis Duncan, Aage Sorenson, David Freedman, and Richard Berk.

At the core of the controversies is the connection between social theories and statistical models. A widely criticized bad practice is to turn every theory into a variety of linear regression models and to take the results as proof or disproof of the theory. We shall learn the details of such models in Chapter 7. For now, the reader may want to take a note that we need to exercise care when using statistical models and be cautious about what we can say based on the results.

Furthermore, social researchers can find many better uses for statistics other than just running models to support theories. Identifying what statistical methods are good at and not good at will be the task of this chapter.

The controversial status of statistics in social research is evident in the UK. Initially, the quantitative wave seemed not to have spilled over to British social sciences – academic publications are highly discursive and qualitative, and only a handful of sociology departments make the learning of quantitative methods compulsory. Although it is not true that all British social scientists shy away from statistical methods, I believe it is safe to say three things about 'quantitative social researchers' in the UK:

(1) Most researchers are clustered in a handful of institutions, including Essex, Lancaster, Manchester, Oxford, and Surrey.
(2) Instead of being sociologists or political scientists, many are 'policy researchers'. They work on issues that are connected to government policies, such as education, poverty, employment, ethnicity, election turnout and so forth, and are concerned more with the implications of their research results for policies than for the growth of knowledge.
(3) Most are specialists on the collection and management of a large data set, such as the British Household Panel Study, British Social Attitudes Survey, British Crime Survey.

What all this means is that although there are some strongholds of quantitative methods in the UK, in most institutions such methods are not integral parts of sociology. Consequently, when voices lamenting the lack of quantitative skills in British social sciences are raised, such as those of the Economic and Social Research Council (ESRC) and Royal Statistical Society (RSS),[1] most often they are those of statisticians. It would be much easier to improve the quality of statistical analysis if sociologists themselves joined the debate.

Institutional initiatives assume that this is purely an issue of training. It is unclear, however, how social scientists in the UK view statistics in the first place. It will be very hard to improve the situation if it is an attitude problem. Why do most British social researchers shy away from statistics? Is it because they know that they are not mathematically competent and are put off by the difficulties of learning statistics? If this is the case, then it is simply a training problem. There is another possibility, however, that they believe that the limitations of statistics are too serious for it to be useful. The most perilous situation, in my view, would be one in which established social scientists in the UK discourage their students from learning and using statistics *for reasons other than the accepted limitations of statistics*, such as rejecting statistical methods as an example of positivism, thereby depriving social science students of the opportunity to learn how to use statistics carefully and thoughtfully.

All in all, the status of statistics in social sciences is not as secure and widely accepted as it initially appears. It is important to point this out at the beginning,

[1]In its First Report of Session 2004–05 to House of Commons Science and Technology Committee, ESRC was 'deeply concerned by the skills shortages afflicting, in particular, the quantitative branches of social science' (p. 33). The current Chief Executive of ESRC, Ian Diamond (2006), expressed his personal concern in a cover report of *RSS NEWS*.

especially for those who are about to learn and use statistical methods seriously. It may sound disheartening, but it is more helpful to tell a sad truth than a happy lie. Most importantly, we should address the question of what statistical methods can (or cannot) do for social research.

Before doing that, it is important to point out that the limitations of statistics should never be confused with problems that are caused by bad practices. Improper use of a tool should not lead to the judgment that the tool is useless. It is not fair to ask statistics to do something that it is not designed to do, and it is even more unfair to claim that it is the fault of statistics while the researcher is a fault. It is counterproductive to focus only on the limitations of statistical methods, ignoring situations in which these methods are of great utility. Such a negative attitude can easily lead the novice to believe that statistics is ill suited to social research and should not be used at all. To completely dismiss statistics from social research is not the solution. Let us think about the limitations and the utilities of statistical methods in specific terms, and then we shall know how to use them properly and responsibly.

My Approach

I take a pragmatic view to the application of statistical methods in the social sciences. To my mind, social researchers should not spend much time on philosophical or epistemological issues. Some may object, feeling that I am distracting students from 'the deeper issues'. My reply would be to let us do some research before talking about philosophical problems. If it turns out that we cannot proceed without sorting out those abstract issues, then it will not be too late to consider them; otherwise, it takes an unnecessarily long period of time to reach any useful results. Social researchers should spend more time developing new skills and trying them out in real research than they spend considering the philosophical background to those skills. We come to philosophical issues only when we have to.

In empirical work, I believe that social researchers should adopt a more balanced attitude towards statistical methods. Statistics should not be used automatically but carefully and appropriately. This means that we must consider the context in which the data were produced and the implications of our statistical analysis for the substantive conclusions that we can make. For this reason, it is very hard to be a social scientist, because it is a considerable challenge for one person to produce creative research designs, to be well read, to be competent in employing statistical methods, and to be able to make sharp observations based on the data gathered. Similarly to any other method, statistical methods have their limitations, but I seriously doubt that one can understand – let alone criticize – their limitations effectively without actually having used them in real research. It is only through careful learning and working with statistics on specific problems that we can identify the limitations and benefits of using statistical methods.

My students usually make two general complaints about statistics: first, statistics is not relevant, and, second, statistics is too hard. Both are understandable, but they can be easily countered. For relevance, just browse the large number of publications

on social issues. Is statistics hard? Yes, and it will remain hard forever if you keep telling yourself and everyone else that 'I am not a math person' or 'I am not here to learn statistics'. What I have found absolutely unacceptable is to connect the above two points together: 'statistics is irrelevant because it is hard'. If you are not prepared to learn statistical methods, please apply qualitative methods – many prominent social scientists have made great contributions without using statistics at all. However, it is unfair to claim that statistics is useless or too hard to learn for the sake of justifying your choice of qualitative methods.

Overview

While planning this book, I have tried my best to employ a logical structure, gradually moving from simple topics to the more complicated ones, more specifically:

- from general issues to more specific topics;
- from data collection to data analysis;
- from univariate (one variable) to bivariate (two variables) to multivariate (three or more variables) statistics;
- from descriptive statistics to inferential statistics;
- from one-level models to multilevel models;
- from cross-sectional (one time point) to longitudinal (multiple time points) models;
- from variable-oriented methods to case-oriented methods;
- from manifest (observed) variables to latent variables.

The reader is strongly recommended to read the whole book in its present order unless you feel absolutely confident of selecting or skipping any particular chapter. Most people should have no difficulty of understanding the first five chapters, but for those without any background in statistics it is a good idea to read an introductory statistics text before moving on to Chapters 6–11.

After this general introduction, we shall discuss a few more specific issues pertinent to the status of statistical methods in social research in Chapter 2. What can they do? What can they not do? What general principles must we follow in order to use them properly?

From Chapter 3, our journey of learning specific concepts and techniques starts with the target of statistical analysis, that is, the case-by-variable data matrix. It is crucial to have a proper understanding of cases and variables before learning any special method for analysing them. The most important issue here is a variable's level of measurement. We should not be obsessed with it, but it is nevertheless true that many statistical tools are created by considering the level of measurement. Therefore, our choice of a particular tool will often heavily depend on it. Later in the Chapter 4, I offer an overview of statistical methods based on our discussion of variables. The final section of Chapter 3 will contain some basic but important rules for using statistics in social research.

Where do the data come from? We cannot analyse data until we examine the data collection process. As most data for social research are collected from sample

surveys, we shall take a closer look at the idea of sampling in Chapter 4. The difference between population and sample might seem obvious, but many researchers are not really aware of the effects of sampling designs and sampling errors. We will spend some time on sampling issues, but the key objective of Chapter 4 is to help the reader understand how sampling procedures affect subsequent statistical analyses.

Knowing the effects of sampling is also a first step toward learning the logic of statistical inference – saying something about the population based on the information collected from only a part of it (the sample). Using the example of measuring and estimating one important phenomenon, we will learn in Chapter 5 why we can say something about the population parameters with statistics produced from only one sample.

Today, social researchers are rarely satisfied with estimating the magnitude of a single variable, no matter how important it is. They study several variables at the same time in order to say something about their relationships, such as looking for the direction of the relationship and measuring its strength, and testing the robustness of the relationship across different situations. Things can appear quite complicated due to the demand of using a specific method for each combination of two types of variables. The relatively large number of ways of describing and representing relationships often perplexes students. Which method should be used? In Chapter 6, I identify the situations in which a particular method should be used and discuss the logic of why that particular method is the right choice.

In Chapter 7, by looking at the relationships among three or more variables, we enter the world of multivariate statistical methods. Perhaps the most popular method is multiple linear regression and its generalizations. Although statisticians have tried to invent flexible models so that we can always have a model suitable for a particular situation, there have been growing criticisms of using such models in social research. Again, a key issue centres on the function of these models: what are they supposed to do? Most users would say that the models should 'explain' the relationship between the variables that we are interested in. But is that the right thing to expect from the models? Even if it is, what do we mean by 'explain'?

All the above methods are used to analyse data collected at one particular time point. Time, of course, is significant in social research. The challenge, however, is to incorporate the temporal dimension explicitly and meaningfully in our analysis. In Chapter 8 we shall learn a few methods that in one way or another take time or temporary order seriously. Without going into technical details, this chapter presents the similarities and differences between these methods by clearly laying out the situation in which the social researcher may find it useful to apply one of the selected methods.

There has been a call to move away from variable-oriented to case-oriented methods in social sciences. In Chapter 9, I show that in addition to qualitative methods there are 'case-oriented' statistical methods. I use the word 'oriented' purposefully because I believe that cases and variables are interdependent on each other and that we should not create another artificial division between research methods. The major difference is that case-oriented methods look more

carefully at the relations among the cases, while variable-oriented methods pay special attention to the relations among the attributes of the cases. It would be simplistic to say that one is better than the other.

Most of the methods discussed in Chapters 4 to 9 are designed, or will only work properly, for manifest (observed) variables. Many variables, however, cannot be directly measured, or even when they can be measured, there is a large amount of error. The source of such errors can be either conceptual or practical, or both. In Chapter 10, I introduce some methods that are exclusively designed to analyse latent (unobserved or unobservable) variables. The first thing to keep in mind is the appropriateness of the variables that are deemed as 'latent' before any of the methods is to be used. It would be pointless to apply these methods if the variables were incorrectly identified as latent in the first place.

In the final chapter we come to the most difficult topic of this book – causal analysis in observational studies. It is relatively straightforward to demonstrate causal relations in controlled experiments as we can manipulate the initial conditions so that the effect of the interested cause will stand out. Most social science studies are observational because such manipulation is almost always infeasible. Therefore, in social sciences, causal relations are proposed first of all based on theories, knowledge, logic and even common sense. Then, evidence is collected to support these causal arguments. Many criticisms have focused on the practice of using linear regression and structural equation models to make causal inference, so we shall try to identify what these problems are. To go beyond those models, some statisticians have developed counterfactual statistical methods for inferring causal effects in observational studies. We cannot sufficiently cover the details of these methods in a single section of one chapter, but the basic ideas will be introduced.

TWO

The Use of Statistical Methods in Social Research

Chapter Contents

What Statistics Is Not Good At

Describing Unique Phenomena in Great Detail

Statistics is a tool for discovering meaningful information from a large amount of numeric data. It is most useful for obtaining concise and precise information about a large number of cases. Cases may come in many different forms: groups of human beings, buffalos, crops, microchips, accidents, web pages and so forth. When it is more important to know the characteristics of these cases as a whole than to learn about each particular unit, statistics starts to shine. It is simply too hard for human brains to detect any meaningful patterns in a large matrix of numbers. Statistics comes to the rescue with a few numbers and equations that summarize the patterns.

Conversely, statistics is a very clumsy tool when the interest is in the details of one or very few unique cases the idiosyncrasies of which can be represented in many different aspects. For example, anthropologists try to understand the uniqueness of a very small number of cases in a particular context. They routinely carry out this kind of work by staying in a unique community for years, taking extremely detailed field notes, and finally writing up what Clifford Geertz calls 'thick description' (1973: 5–10). In sociology, beginning with Max Weber, there has been a long tradition of understanding meanings, interpretations, values and contexts. Comparative studies on a small number of cases – the so-called 'small-N' studies – have been an attractive research method to many students of social, historical and political sciences.

For example, why have some former communist nations been successful in transforming their economies while others failed? There are not many former