Statistics for Social and Behavioral Sciences

Statistical Modeling of the National Assessment of Educational Progress

Springer

Statistical Modeling of the National Assessment of Educational Progress

Statistics for Social and Behavioral Sciences

Series Editors:
S. E. Fienberg
W. J. van der Linden

For other titles published in this series, go to
http://www.springer.com/series/3463

Murray Aitkin • Irit Aitkin

Statistical Modeling of the National Assessment of Educational Progress

Springer

Murray Aitkin
Department of Mathematics and Statistics
University of Melbourne
Melbourne Victoria 3010
Australia
murray.aitkin@unimelb.edu.au

Irit Aitkin
Department of Mathematics and Statistics
University of Melbourne
Melbourne Victoria 3010
Australia
irit.aitkin@unimelb.edu.au

ISBN 978-1-4419-9936-8 e-ISBN 978-1-4419-9937-5
DOI 10.1007/978-1-4419-9937-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011927940

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This book is a description of research and data analysis carried out by the authors with substantial funding from the National Center for Education Statistics (NCES) and the Institute of Education Sciences (IES), divisions of the U.S. Department of Education, in partnership with the American Institutes for Research (AIR). The purpose of this work was to evaluate a new approach to the analysis and reporting of the large-scale surveys for the National Assessment of Educational Progress (NAEP) carried out for the NCES.

The new approach was based on a full statistical and psychometric model for students' responses to the test items, taking into account the design of the survey, the backgrounds of the students, and the classes, schools and communities in which the students were located.

The need for a new approach was driven by two unrelated issues: the demands for *secondary analysis* of the survey data by educational and other researchers who needed analyses more detailed than those published by NCES, and the need to accelerate the processing and publication of results from the surveys.

The modeling approach is complex and computationally intensive, but less so than the existing methods used for these surveys, and it has the twin advantages of *efficiency* in the statistical sense – making full use of the information in the data – and *optimality*: given the validity of the statistical model, this form of analysis is superior to any other non-Bayesian analysis in terms of precision of the estimates of group differences and regression coefficients of important variables.

The use of a full statistical model avoids the ad hoc methods that are otherwise necessary for the analysis of the data. It is dependent, for successful adoption, on efficient computational implementations in generally available software. Developments in this area have been rapid in the last ten years: we began our analyses in 2003 using Gllamm in Stata (Rabe-Hesketh and Skrondal 2005); see the Website.

```
http://www.gllamm.org
```

By 2008 we were able to use the very fast Latent Gold program which makes large-scale model fitting straightforward for NAEP data sets; see the Website.

```
http://www.statisticalinnovations.com/products
/latentgold_v4.html
```

The following chapters, apart from the first, are set out in a sequence representing the main aspects of the NAEP surveys and the development of methods for fitting the increasingly complex models resulting from the incorporation of these aspects. The content of the book is drawn mostly from our NCES research reports, which are described briefly in Chapter 4. We generally do not give references to specific reports in the text, as the chapters draw from many of the reports. The full reports themselves are available on our Website, as described in Chapter 4.

The models and analysis approach are illustrated with detailed results from two NAEP surveys. The first is from the 1986 national NAEP mathematics test and includes results on the set of 30 items from the Numbers and Operations: Knowledge and Skills subscale, for age 9/grade 3 children. The "explanatory" regression model fitted is quite small and was chosen to nearly replicate the tables of "reporting group" variables published by NCES for this survey. We extended this analysis to all 79 test items on three scales.

The second survey is from the 2005 national NAEP mathematics test and includes results on the set of 70 items from the Numbers and Operations scale for age 10/grade 4 children. We fitted a much larger regression model with variables from the student, teacher and school questionnaires. We analysed the California and Texas state subsamples with more complex item response models.

Chapter 1 is an introduction to the current theories of data analysis used for large-scale surveys. It may surprise non-statistician readers to find that there are major disputes within the statistics profession about the role of statistical models in official (national government) survey analysis. We describe the critical theoretical issues that divide the several theories, and give an indication of the extent to which each theory is used in current official practice.

Chapter 2 describes the current method of analysis of NAEP surveys. This has changed several times; we give the analysis that was used for the 1986 survey, which we use as an illustration in later chapters, and note the changes that have occurred since then. The design and analysis of the 1986 survey were very complicated, and we have omitted aspects of the design that are not critical to the analysis. Some complex sections (for example, jackknifing) have been described at length because these are critical for the comparison with our approach.

Chapter 3 sets out the psychometric models used in the NAEP analysis, gives some extensions of them using mixture distributions for student ability, discusses the survey designs used in the surveys, and gives the multilevel model representation of the designs.

Chapter 4 summarises the main conclusions from our extensive simulation studies, which showed the improvement in precision and the reduction in bias resulting from the fully model-based analysis of small-scale models compared with the current approach. References to the full reports on this work are given there.

Chapter 5 sets out the series of analyses we used with the range of models from Chapter 3 for the 30-item scale from the 1986 math test for age 9/grade 3 children. Chapter 6 extends these analyses to the full set of 79 items on the test. Chapter 7

applies more complex analyses to the 2005 national NAEP subsample for Texas for age 10/grade 4 children. Chapter 8 applies the same analyses to the 2005 subsample for California for age 10/grade 4 children.

Chapter 9 discusses the results of the analyses and draws conclusions about the benefits and limitations of fully model-based large-scale survey analysis.

Acknowledgements

Murray's interest in psychometric modeling began with his post-doctoral position with Lyle Jones at the Psychometric Laboratory, University of North Carolina at Chapel Hill in 1966–67, and developed substantially from his visiting year as a Fulbright Senior Fellow in Frederic Lord's psychometric research group at the Educational Testing Service (ETS), Princeton in 1971–72.

In his large-scale research programme on EM algorithm applications to incomplete data problems at the University of Lancaster 1979–85, Murray developed with Darrell Bock (Bock and Aitkin 1981) an EM algorithm for the 2PP model. This algorithm has been very widely extended to other psychometric models.

Murray returned to ETS as a Visiting Scholar in 1987–88. Here he reviewed (Aitkin 1988) the extent to which hierarchical variance component modeling, incorporating the survey design in additional levels of the model, could be used for the analysis of NAEP data, and the possible information that it could provide. He noted that fitting the 3PL model in a full hierarchical model was beyond the capabilities of available programs, and suggested variations to the E step of the EM algorithm that could give an approximate analysis.

Our joint interest in NAEP developed with Murray's appointment as Chief Statistician at the Education Statistics Services Institute (ESSI), American Institutes for Research, in Washington, D.C. in 2000–2002, as a senior consultant to NCES. It continued through a series of research contracts with NCES through AIR in subsequent years.

We have benefited greatly from discussions and interactions with many staff members at NCES, particularly Andrew Kolstad, Steve Gorman, and Alex Sedlacek, and are grateful for the ongoing support of Peggy Carr, NCES Associate Commissioner for Assessment. The outline of statistical theories in Chapter 1 has greatly benefited from two "brown-bag lunch" seminar series that Murray gave to NCES and AIR staff in Washington; we appreciate the comments and feedback from participants in these seminars. We much appreciate the many discussions with current and former senior staff at AIR, particularly Gary Phillips (a former Deputy and Acting Commissioner of NCES), Jon Cohen, Eugene Johnson, Ramsay Selden, and Laura Salganik. We are particularly grateful for administrative support from Laura Salganik and from Natalia Pane, Janet Baldwin-Anderson, and Linda Schafer at ESSI and its successor NAEP ESSI. We much appreciate the many other welcoming and helpful people at these institutes supporting the work of NCES.

We thank John Mazzeo and Andreas Oranje at the Educational Testing Service for help with data access and interpretation, Kentaro Yamamoto at ETS and Charles Lewis at Fordham University for discussions on psychometrics, and Chan Dayton, Bob Lissitz, and Bob Mislevy at the University of Maryland for helpful discussions. We thank Sophia Rabe-Hesketh and Jeroen Vermunt for many technical discussions that have helped greatly to clarify aspects of the analyses in Gllamm and Latent Gold, and Karl Keesman for much help with Stata.

The very large-scale simulations needed for the evaluation of competing methods were run in Melbourne on the cluster computers of the Victorian Partnership for Advanced Computing (VPAC). We thank Chris Samuel and Brett Pemberton of the VPAC staff for much help. We particularly thank Michael Beaty of the School of Mathematics and Statistics of the University of Newcastle, UK, for detailed and continuing help with many computing aspects of our work, and our son Yuval Marom for a great deal of help with programming.

At the University of Melbourne, we are grateful for the support and help of Pip Pattison, Henry Jackson, and Bruce Ferabend in the Department of Psychology, Peter Hall and Richard Huggins in the Department of Mathematics and Statistics, and Dirk van der Knijff in the High Performance Computing unit.

We extend our special thanks to Sue Wilson for constant encouragement and support.

In preparing the final version of this book, we were greatly helped by suggestions and advice from Steve Fienberg on the first draft, which was changed substantially. Any remaining errors or obscurities are entirely our responsibility.

Murray Aitkin — Melbourne
Irit Aitkin — December 2010

This work has been funded by federal funds from the U.S. Department of Education, National Center for Education Statistics, and Institute for Education Sciences under various contracts and grants. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education, National Center for Education Statistics, or AIR, nor does mention of trade names, commercial products, or organisations imply endorsement by the U.S. Government or AIR.

Contents

Chapter 1
Theories of Data Analysis and Statistical Inference

1.1 Introduction

Every survey, in any field, begins conceptually with a *population list*, a *sampling plan* or *sample design* by which an appropriate sample is to be drawn from the population, a *measurement instrument* specifying the information – *response* variables and *covariates* or *explanatory variables* – to be obtained from the sampled population members, and an *analysis plan* by which the response variables, and their relation to the covariates or explanatory variables, are to be analysed.

In the NAEP surveys that we describe and analyse, the population list is of school students of several ages and grades, the sampling plan is a complex clustered and stratified design, and the measurement instrument is a set of test items measuring achievement in mathematics or another subject (the response variables in many analyses, including ours) and a set of questionnaire items describing students, their home background, and teacher and school characteristics that we use as covariates for achievement, though many may be response variables in other analyses.

We use from now on the term *covariates*, rather than explanatory variables, as the NAEP surveys we discuss are *observational studies* in which the issue of *causality* – of whether variation in the covariates *causes* or *explains* variations in the outcome variables – cannot be assessed from the surveys, as these are not experimental studies involving *randomisation* of students to classes or to educational and family contexts.

The analysis plan is the subject of this book, which discusses in this chapter the different philosophies in the statistics profession about how data from such studies should be analysed. Our view of analysis is *model-based*: defined by a full statistical model – in contrast to the current analysis of these surveys, which is a mixture of model-based and *design-based*: defined by hypothetical replications of the survey design.

In describing and discussing the important differences in these approaches, we adapt the discussion in Chapter 1 of Aitkin (2010) and use several very simple examples that, however, make clear the importance of the philosophical differences.

M. Aitkin, I. Aitkin, *Statistical Modeling of the National Assessment of Educational Progress,* 1
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-1-4419-9937-5_1, © Murray Aitkin and Irit Aitkin 2011

1.2 Example

We have a simple random sample of size 40 from a finite population of 648 families and for each family record the family income for the previous tax year. From this sample, we wish to draw an inference about the *population mean* family income for that tax year. How is this to be done? The sample of incomes, reported to the nearest thousand dollars, is given below.

```
        Family income, in units of 1000 dollars
-------------------------------------------------
26 35 38 39 42 46 47 47 47 52 53 55 55 56 58 60 60
60 60 60 65 65 67 67 69 70 71 72 75 77 80 81 85 93
96 104 104 107 119 120
-------------------------------------------------
```

Theories of data analysis and inference can be divided into two classes: those that use the *likelihood function* (defined below) as an important, or the sole, basis for the theory and those that do not give the likelihood any special status.

Within the first class, there is a division between theories that regard the likelihood as the *sole* function of the data that provides evidence about the model parameters and those that interpret the likelihood using other factors.

Within the second class, there is a division between theories that take some account of a *statistical model* for the data and those based exclusively on the properties of estimates of the parameters of interest in repeated sampling of the population. Comprehensive discussions of the main theories can be found in Welsh (1996) and Lindsey (1996), to which we refer frequently. We illustrate these theories with reference to the income problem above.

1.3 Statistical models

Theories that use the likelihood require a *statistical model* for the population from which the sample is taken, or more generally for the *process* that generates the data. Inspection of the sample income values shows that (in terms of the measurement unit of \$1000) they are *integers*, as are the other unsampled values in the population. So the population of size N can be expressed in terms of the *population counts* N_J at the possible distinct integer values of income Y_J or, equivalently by the *population proportions* $p_J = N_J/N$ at these values.

A (simplifying) statistical model is an *approximate representation* of the proportions p_J by a *smooth probability distribution* depending on a small number of *model parameters*. The form of the probability function is chosen (in this case of a large number of distinct values of Y) by matching the cumulative distribution function (cdf) of the probability distribution to the empirical cdf of the observed values. Figure 1.1 shows the empirical cdf of the sample values. A detailed discussion of this process is given in Aitkin et al. (2005) and Aitkin et al. (2009). We do not give

details here, but the matching process leads to the choice of an approximating continuous cdf model $F(y\,|\,\lambda)$ and corresponding density function $f(y\,|\,\lambda) = F'(y\,|\,\lambda)$; the probability p_J of Y_J is approximated by $F(Y_J + \delta/2\,|\,\lambda) - F(Y_J - \delta/2\,|\,\lambda)$, where δ is the measurement precision (which equals 1 in the units of measurement). When the variable Y is inherently discrete on a small number of values, as with count data, the values p_J are approximated directly by a discrete probability distribution model.

Figure 1.2 shows the cdf of a normal distribution with the same mean (67.1) and standard deviation (22.4) as the sample income data, superimposed on the empirical cdf. Figure 1.3 shows the same cdfs, but on the vertical probit scale of $\Phi^{-1}(p)$. On this scale it is clearer that the income sample has some degree of *skew*, with a longer right-hand tail of large values, so an approximating model with right skew might be appropriate. The gamma, lognormal, and Weibull distributions are possible choices. However, to establish which of several possible models is most appropriate for sample data requires advanced model comparison methods, which we discuss in later chapters.

Here we will assume that the normal distribution with parameters μ (the mean) and σ (the standard deviation) is a reasonable model, where μ is the *parameter of interest* and σ is a *nuisance parameter* – we want to draw conclusions about the parameter of interest, μ, but the model depends as well on the nuisance parameter σ:

$$f(y\,|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

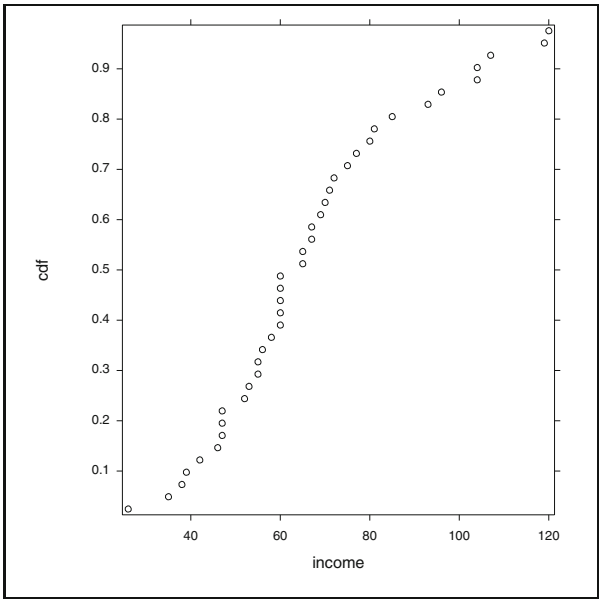

Fig. 1.1 cdf of sample income data

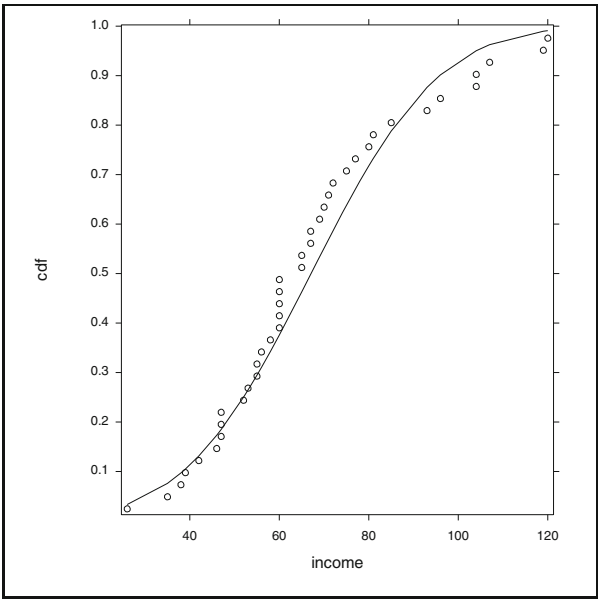

Fig. 1.2 cdfs of sample income data and normal distribution

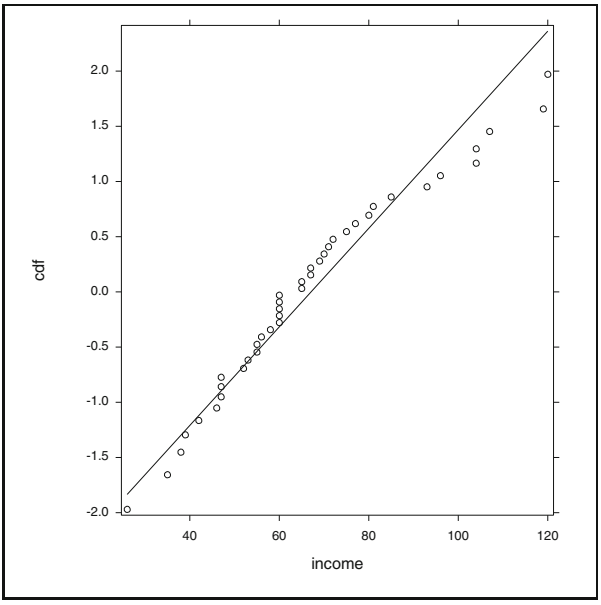

Fig. 1.3 cdfs, probit scale