

# Reflections on statistical modelling: A conversation with Murray Aitkin<sup>1</sup>

Murray Aitkin, John Hinde<sup>2</sup> and Brian Francis<sup>3</sup>

<sup>1</sup> School of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia.

<sup>2</sup> School of Mathematical and Statistical Sciences, National University of Ireland Galway, Galway, Ireland.

<sup>3</sup> Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom.

---

**Address for correspondence:** Murray Aitkin, School of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia..

**E-mail:** murray.aitkin@unimelb.edu.au.

**Phone:** (+61) 407 222 040.

**Fax:** (+61) 383 444 599.

---

**Abstract:** A virtual interview with Murray Aitkin by Brian Francis and John Hinde, two of the original members of the Centre for Applied Statistics that Murray created at Lancaster University. The talk ranges over Murray's reflections of a career in statistical modelling and the many different collaborations across the world that have been such a significant part of it.

---

**Key words:** Murray Aitkin; statistical modelling; Bayesian bootstrap

*Murray, it is a pleasure for us to have this opportunity to have this virtual conversation with you exploring your life in statistical modelling. You have had a long and very varied career working in many different institutions across the world and in this time you have had a number of significant collaborations that must have influenced your ideas and the developments that you are known for. At the same time you have had a great influence on many people, including the two of us who worked with you in Lancaster at a time of significant developments in statistical modelling. We owe you a lot for your generosity of ideas and time and for your infectious enthusiasm and passion that helped to shape our future careers and we are sure that the same is true for many others.*

*But let's start by going back to your early career and the experiences that shaped your passion for modelling and for an approach to hands-on statistics courses that are now common-place, but were not then.*

In 1969 I was appointed to a Senior Lectureship at the new Macquarie University in Sydney, jointly in the Schools of Behavioural Sciences and of Economic and Financial Studies. I was responsible for the content of the statistics courses in Behavioural Sciences, and

for the upper-level statistics courses in Economics. The University had a common non-mathematical course in Introductory Statistics for all Schools.

My principal concern was for a third year course in regression and its applications. For mathematically competent Statistics students this would not be a problem, though it would be mostly theoretical as the University computer centre had no statistical packages. For the Psychology students, this would be a real problem; many found the Introductory Statistics course off-putting, and had a real fear of mathematics. I postponed the idea of a course for the psychologists, until a new development: the University's computer centre bought and mounted the US National Bureau of Standards spreadsheet package OMNITAB.

*Well, that name sounds almost familiar, but I guess that statistical software looked very different in those days.*

Yes, OMNITAB was the precursor of the later commercial MINITAB package, and had two valuable features which made it usable by psychology students: a simple English control language, and a very well-thought-out (Gaussian) multiple regression (MR) routine. An unusual feature of the latter was a line-by-line decomposition of the residual sum of squares, in the order of the terms specified in the model: changing this order would change the decomposition, unless the terms were orthogonal. This very valuable feature was removed in the later MINITAB version, which gave only the residual sum of squares for the full fitted model. This was a regrettable act, as it contributed to the later vast confusion over ANOVA tables for non-orthogonal designs ([Aitkin, 1978](#)). This confusion alas continues.

*Ah, I can already see the germ of one of your life-long bugbears about the correct approach to variable selection in regression and later in glms. So, what did you see as the advantage of having early modelling software in designing a better, more useful course?*

It became clear to me that I could now write and give a regression course which would suit both statistics and psychology students, by concentrating on the model specifications and fitting in OMNITAB: the necessary mathematics would be only simple linear algebra and co-ordinate geometry, which I could arrange as an intensive four-week workshop before the course itself. There would be no need for algebraic sums of squares decompositions. This approach would be able to cover all the common model versions: multiple regression, ANOVA and ANCOVA. This is the point where I became a statistical modeller.

*This was a big departure from the style of courses at the time. How did you plan to bring this about and provide the necessary support for the students?*

The new approach would need full documentation: I would need to write a book on it connected to OMNITAB. For this to be possible I needed a break from my heavy teaching and consulting load. I applied for a Fulbright Senior Fellowship, to take up an offer to spend a year with Fred Lord's psychometric research group at the Educational Testing Service (ETS). I was fortunate to be successful with the application and was granted leave July 1, 1971 – June 30, 1972.

*I know that this was a very formative step in your career where you established many important and lasting contacts. Tell us a little about this time at ETS.*

The Psychometric group had only two permanent research staff: Fred Lord and Walter Kristof, and several programmers. Fred had a large visitors' programme which supported a remarkable number of visitors over its lifetime. The other visitors, at the time of my visit, were Ingram Olkin from Stanford, Leon Gleser (a former PhD student of Olkin's) from Johns Hopkins, and Michael Browne from South Africa. Karl Jöreskog came from Uppsala for part of the year. ETS was then establishing a Data Analysis group, which was headed by Don Rubin, who later attracted Paul Holland. My interactions with this large ETS group, and with the small Princeton statistics group, including John Tukey, Geoff Watson and Don McNeil, were very stimulating.

I had no responsibilities, except for giving an occasional seminar (all the visitors wanted to give them). I spent most of the year working on the textbook exposition of regression, ANOVA and ANCOVA through the general linear model, integrated with the OMNITAB system. The rest of the time was spent attending the seminars of others. At the end of the year I returned to Sydney, and began to develop the new Linear Models course, which started in 1973.

*So, your first real course in statistical modelling. How was this received?*

The course became a pre-requisite for the fourth honours year in psychology, and became known as the most difficult course in Psychology (there has to be one!). Despite its reputation and difficulty it was a real success, and the enrolment increased rapidly over the four years I gave it, from 25 to 65. These numbers were much larger than the numbers of third year Psychology students aiming for honours, plus the number of third year statistics students: students were coming from other areas as well. I wrote about this course in AMSTAT news ([Aitkin, 2013a](#)).

*However, it seems that your year at ETS had given you the taste for travel and perhaps broadened your horizons on statistical research and the application of statistical modelling to different areas. Soon you were on the move again, how did this come about?*

In 1975 I applied for a one-year UK Social Science Research Council (SSRC) Professorship in Statistics Applied to the Social Sciences, to be held at the University of Lancaster in the Mathematics Department. It was renewable for a further two years. The aim of the post was to improve the applications of statistics in the social sciences. There was some possibility that the University could build on the work done to develop some form of continuing structure.

*So, what attracted you to this post and the challenges and opportunities that it might give you?*

I already had strong views about how the applications of statistics in the social sciences could, and should, be improved through a focus on modelling. I had spent considerable

effort at Macquarie on arguing for a different Introductory Statistics course, without any success. On the other hand the Linear Models course showed what could be done with a good statistical package and minimal mathematics at the regression modelling level.

At the interview I was asked what I would do if appointed. I had prepared a three-item list:

- i) I will give a seminar series in statistical modelling for the social sciences;
- ii) I will consult with quantitative people in each department to see how I can assist with their modelling problems;
- iii) I will continue to develop my own research [then in simultaneous inference].

This satisfied the committee. I was offered the position, and accepted it, from August 1976. Macquarie gave me a year's leave of absence.

*Tell us a little about the Department that you were joining in Lancaster. I'm sure that the statistics group at the time was not large.*

Of the 20 staff of the Lancaster Mathematics Department, there were only four in statistical areas. Emlyn Lloyd and David Warren were probabilists, Granville Tunnicliffe Wilson and Joe Whittaker were statisticians. Joe was very involved in generalised linear models (GLMs) and their implementation in GLIM (Generalised Linear Interactive Modelling), the recently developed and developing statistical package for fitting GLMs. It was difficult to use, and Joe wanted to train me on it, which he did.

*Ah, your introduction to GLIM and something that would really influence your teaching of statistics for many years to come. Now the father of GLIM was John Nelder and I know that you also had similar views on ANOVA tables and multiple partitionings of sums of squares.*

Yes, I soon saw that GLIM, crude as it then was, had far more promise than OMNITAB, which had no GLM possibilities and crude programming. Indeed, I met Nelder again; I had met him when he was visiting Sydney. I discussed with him the sum of squares partitioning problem with non-orthogonal data, and hit a sore point. He was angry at the insistence of some US statisticians that there should be only one ANOVA table, in which each effect is adjusted for all of the others, generally known as the "Type III sum of squares". My view of the multiple sequential tables was in accord with his, though his main argument was algebraic, over the "constraints" that had to be imposed on less than full-rank models. The dummy variable method I used in my course avoided all this difficulty. I decided to submit a paper for reading on this subject to JRSSA.

In April 1977 I attended the Spring ASA and Biometric Society regional meeting at Chapel Hill. I was a speaker in an afternoon panel session on the analysis of unbalanced data. Jim Goodnight from North Carolina State University was the Chair (before his creation

of SAS and its remarkable success). My talk was a version of the JRSSA paper which appeared in 1978 (Aitkin, 1978), arguing for multiple hierarchical analyses of unbalanced cross-classifications, as was made easy by OMNITAB. This was shouted down by the other panel members, who agreed that i) the multiple orthogonal decompositions were too complicated, and not proper hypothesis testing; ii) non-specialist users needed a single ANOVA table from which to draw conclusions. One US panel member said to me afterwards “You got it all wrong, fella” I said “Maybe”. Despite the appearance of my 1978 paper with discussion, the Type III system remains the recommended approach in the big statistical packages.

*Meanwhile, back in Lancaster how had you begun to address the project’s aim to improve the use of statistics in social sciences research?*

I set up the seminar series and visits to departments at Lancaster as I had proposed. These proved differentially profitable: Sociology and Geography were important application areas with interested people. I visited London and other centres regularly for RSS London and local group meetings.

*Now I know that one of these London trips for an RSS read paper was to have a significant impact on your future research and the course of your life.*

Absolutely, the outstanding statistical event of the 1976 year was the Dempster, Laird and Rubin RSS read paper on the EM algorithm (Dempster, Laird, and Rubin, 1977). The simplicity of the idea was amazing, and the applications vast, though the symbolic representation of it was at times confusing. It was clear to me that EM would change statistical analysis permanently, and that many applications could be programmed in GLIM — another reason for using it.

*Now one particular application of the EM algorithm is to the fitting of mixture models and I know that this has been a recurring aspect of many of your modelling developments in various different settings. But first, how did you come to establish a group at Lancaster and create the climate for the variety of work that the group undertook?*

My Fellowship was extended for the two additional years, and I resigned from Macquarie. I needed to think about the kind of permanent structure which might be considered by Lancaster. The conjunction of the extensive consulting I was doing in the University, the EM algorithm, and the continuing development of GLIM, suggested a research and consulting centre for the analysis of complex large-scale social data. Consulting alone would not support a research group: we would need to have large-scale research programme support as well, clearly from the SSRC. So I wrote two large proposals, one directed to the University, the other to the SSRC; each would support the case for the other.

*And how did this work out?*

Both the applications were successful. In these, as with my SSRC Fellowship, George Barnard played a major supportive role. The University established the Centre for Applied

Statistics (CAS) with me as Director, with one programming/consulting staff member, a part-secretary and a small establishment grant. The SSRC awarded the University a large programme grant for the Analysis of Complex Large-scale Social Data. This supported two research associates and substantial travel and maintenance expenses for distinguished visitors, for visits of four to six weeks. These would begin September 1 1979.

*What a great way to start out and a real vote of confidence in you and your vision for the Centre. Also, a very good basis to achieve that original aim of making an impact on statistics in the social sciences. So how did this all work out?*

Appointing CAS staff and SSRC research associates was the immediate task. We were very fortunate in being able to appoint Brian Francis to the CAS staff post and John Hinde and Dorothy Anderson to the SSRC research posts. Distinguished visitors had time commitments of their own, but given the five-year scope of the SSRC project, we were able to invite and support (alphabetically) George Barnard, Jim Berger, Darrell Bock, Steve Fienberg, Karl Jöreskog, Jack Kalbfleisch, Nan Laird, Richard Royall, Don Rubin (twice) and David Sprott (at his own expense while visiting Barnard). Other distinguished visitors came for short periods. The visit from Darrell Bock led to the ground-breaking paper ([Bock and Aitkin, 1981](#)) which began an intensive development of complex psychometric models and their maximum likelihood analysis.

*Yes, this is there we both came in and fell under your very positive influence. For young researchers the visitors programme gave us a chance to interact closely with some of the giants of the subject, especially when they had to share our office (physical space was at a premium). I remember that in addition to the modelling focus of the work there was an interest in statistical inference, particularly likelihood-based approaches and many of the visitors reflected this. However, the main drivers of the work were the modelling projects, tell us some more about these.*

Well, a ground-breaking CAS project was the reanalysis of the Teaching Styles and Pupil Progress ([Bennett, 1976](#)) study of Neville Bennett in the Lancaster School of Education. He had analysed the effect of the teaching style of 36 primary school teachers on the achievement of their pupils over one year, in reading, mathematics and English, by relating the child's achievement to the teacher's style of teaching through a least squares regression. The teaching style of the teachers had been assessed by a cluster analysis of 38 binary items on 468 teachers which had identified three clusters, labelled Formal, Informal or Mixed. The analysis was widely criticised because he had treated the student test scores within the same teacher as independent. The critics said he should, instead, have aggregated the pupil results for the teacher into pretest and test means, and then regressed "means on means" — the test mean on the pretest mean and the teacher style.

Our modelling reanalysis ([Aitkin, Bennett, and Hesketh, 1981](#); [Aitkin, Anderson, and Hinde, 1981](#)) used the latent class model to cluster the teachers, and the two-level variance component model to regress test score on pretest score and class model probability. The reanalysis attracted a great deal of attention and began an industry of variance component package development in several countries.

Additionally, a three-year research project was commissioned from the CAS by SOEC, the Statistical Office of the EEC, to investigate the value of statistical modelling of large-scale surveys of unemployment. Rob Healey and I showed that logistic regression modelling of very large cross-classified unemployment data gave a satisfactory model-based alternative to small-area estimation, even without allowing for the design effect. This was presented in a conference and a later paper: [Aitkin and Healey \(1984, 1987\)](#). The SOEC staff were piqued by our visualisation demonstration of a three-scale circular slide rule to show the combined main effects of age, geographical region and industry on the unemployment rate. There were separate slide rules for men and women.

*This European involvement and perhaps also your attendance at COMPSTAT 1982 in Toulouse led to another novel and interesting collaborative project with modelling as a key component. Can you tell us a little about that?*

The expanding French development of the analyse des données (correspondence analysis) was promoted as a (national) alternative general data analysis framework to the UK/US modelisation (statistical modelling). Henry Caussinus at Toulouse and I submitted a joint proposal to the CNRS (Centre National de la Recherche Scientifique) and the SSRC for a comparison of Anglo/American and French statistical analysis methods. This collaborative project ran for three years, with teams from each centre visiting the other alternately. The project report, in addition to giving comparative analyses of several two-way contingency table data sets, showed that the correspondence analysis singular value decomposition of the residuals from the main effect log-linear model was equivalent to fitting the log-linear model with main effects and multiplicative interactions. This brought the two approaches together in a common framework. In addition to the report, much of the work was described (in French) in the substantial joint paper [Aitkin, Francis, and Raynal \(1987\)](#).

*In some sense much of the work undertaken in the CAS at this time was heavily computational, but things were not always easy. I remember that when I began at Lancaster we were using GENSTAT remotely at the University of Manchester, an often slow and painful process. Of course, you and then, through your passion, the rest of us soon realised the possibilities of GLIM and so began a very exciting and creative time. What are your memories of the challenges, successes, and legacy of this time?*

An early frustration with statistical computing at Lancaster was that the University's small ICL 1906 was running under a business batch operating system which did not support multiple simultaneous users. This handicapped our use of GLIM, which had to run in batch, though it was designed (as in the title) for interactive use. With Brian Francis and Granville Tunnicliffe Wilson (Mathematics), I drafted an application to the Computer Board for workstations for statistical computing in teaching and research for use by the CAS and the Mathematics Department. This was successful with funding for 12 workstations that were eventually installed in the late 80s. The computing room in Mathematics was renamed the John Nelder Laboratory, and was opened officially by Nelder.

A major role of the CAS was to run short courses, especially on GLMs and their applications and analysis with GLIM. In 1982 we all attended the first GLIM conference, organised

by Bob Gilchrist in London, and then in 1985 Lancaster organised and hosted a second more international meeting, with invited and contributed papers by many European and US statisticians. These conferences evolved into the annual International Workshop on Statistical Modelling (IWSM) and, ultimately, the foundation of the Statistical Modelling Society.

*One of your great strengths, that we have both benefitted from, is your talent to inspire and mentor. But I recall also having other young recruits to the CAS in a sort of intern position and I think that the first of these was David Firth. How did this come about?*

An innovation of the SSRC was the establishment of two Junior Consulting Fellowships (JCFs), for new MSc graduates to obtain practical data analysis experience in one- or two-year attachments to a University statistical consulting group. I thought this was an excellent idea and applied from Lancaster for one of them. We were not successful, but our increasing consulting income could support one such position for one or two years, so I announced our own version of the JCF in a circular to the major UK statistics groups. Our JCF ran very successfully for three years, and we benefitted, in addition to the increased consulting work done, from stimulating interactions with David Firth, Mikis Stasinopolous and the other JCFs.

*Now certainly for the early years of the CAS there was a strong emphasis on likelihood methods and direct likelihood inference, as memorably championed by George Barnard, Jack Kalbfleish, and Richard Royall during their visits. But by the mid 80's there were the early hints of a more Bayesian perspective to some of your work. How did this begin?*

As interest increased among modellers in Bayesian extensions of statistical modelling, I became concerned with the misuse of priors in models with awkward likelihoods. An example was the “binomial( $N, p$ )” problem, in which we have a random sample of binomial successes  $r_i$ , but the number of trials  $N$  is unknown, as well as the common success probability  $p$ . [Aitkin and Stasinopoulos \(1989\)](#), reproduced in [Aitkin \(2010\)](#) §2.4.1, gave a major discussion of this problem with a full analysis of a small data set from Ingram Olkin’s work — he had identified the problem. The profile likelihood in  $N$  did not redescend for large  $N$  — it asymptoted to the Poisson maximised likelihood and was mostly flat. Any beta prior for the unknown  $p$  produced a strong mode in the posterior for  $N$ . Bayesians dismissed our analysis: the obvious solution was an informative prior for  $N$ , like  $1/N$ . This did not change the reality.

*At this point it must have seemed that your dedication and commitment to the CAS was paying off.*

Yes, the SSRC programme was extended for one further year, reviewed and praised by the SSRC (by then renamed to ESRC — the Economic and Social Research Council) as a major international success.

*And your passion for GLIM was also leading to promising things.*



Yes, following Nelder's retirement (1985) as Chair of the GLIM development committee, I was appointed to the Chair, and proposed a development of GLIM into a major statistical package with improved facilities.

*But the 1980s were not an easy time in the UK and things didn't develop as you and others might have hoped. Can you remind us how bad things got at that time?*

In 1980 the Thatcher government began a series of attacks on the UK University system which affected drastically many UK Universities, including Lancaster, whose funding was cut by 15% over three years. The subsequent splitting of the funding into teaching and research components, and the establishment of the Research Assessment Exercise to determine research funding, had further drastic effects. The CAS was particularly vulnerable as it depended heavily on soft research grant funds. By 1986 it had contracted from its maximum size of 8 in 1983 to 3, with no prospect of further funding, even for a six-month gap in Nick Longford's already approved, but delayed, SSRC Research Fellowship. Lancaster was not unique: the whole University system was seriously damaged, and would take many years to recover, if it did recover.

*And sadly, your vision for GLIM also ran into trouble.*

My attempt to develop GLIM into a more powerful modelling package was opposed by the RSS Council, and was abandoned, though Brian Francis later implemented in GLIM4 some of the facilities I had proposed. [The history of GLIM is given in detail in [Aitkin \(2018\)](#).] I left Lancaster for a sabbatical year in 1986 and did not return. However the GLIM book, which was part of the proposal to the SSRC, was completed and published: [Aitkin et al. \(1989\)](#).

*The next significant chapter in your life involved a return to Australia and also a shift in your inferential perspective. What prompted these changes?*

In 1990 I applied to the Australian Research Council (ARC) for a five-year Senior Research Fellowship, to develop a general likelihood theory of inference. To my surprise I was awarded the Fellowship, and took it up initially at the Australian National University (ANU), but after six months I moved it to the Department of Mathematics at the University of Western Australia (UWA). I had become concerned with the anomaly of the inconsistency of Bayesian and frequentist model comparisons (hypothesis testing) in the Gaussian distribution, where they should have been identical (with the usual diffuse priors), as they were for credible/confidence intervals.

*So, what was the novel approach that you developed?*

I first focussed on adapting Dempster's posterior distribution of the likelihood ([Dempster, 1974, 1997](#)) to the use of the posterior mean of the likelihood rather than the prior mean. The latter led to the integrated likelihood used in the Bayes factor, causing the inconsistency. This was unsuccessful (I was condemned for using the data twice), but towards the end of the Fellowship I took a different and successful (though disputed) approach, by gen-

eralising Dempster's comparison of a simple null hypothesis with a simple alternative, to the comparison of a composite null hypothesis with a composite alternative (Aitkin, 1997).

The central idea was that in model comparisons the likelihood (or deviance), instead of being summarised by a maximum or mean, had a posterior distribution, obtained by substituting the posterior draws of the model parameters into the likelihood (or deviance) function. Models were then compared by their deviance distributions, to assess their stochastic ordering corresponding to a preference ordering of the models. This idea was strongly opposed, indeed denounced, by many orthodox Bayesians.

*Well, those of us who know you well are familiar with the passion and strength that you hold your ideas and that you are not a man to be swayed by orthodoxy. But fellowships are only for a fixed period so what happened next?*

As the Fellowship neared its end, I applied for the Chair of Statistics at the University of Newcastle upon Tyne, in the newly re-established Department of Statistics.

*So, a return to the UK, but I guess to a somewhat different climate in universities from when you left.*

Newcastle University, with a new Vice-Chancellor and a statistically sympathetic Deputy Vice-Chancellor, had reversed the common disastrous pattern of Universities closing statistics departments by combining them with mathematics, and had also created a new Statistical Consultancy Service with a new Director, to be appointed, and two consulting staff positions for three years. I was appointed to the Chair, and soon found that the Directorship of the Consultancy Service attracted a very small number of applicants: statisticians at this level could make a lot more in commercial consulting. I decided to take this on myself for three years, and was released from undergraduate teaching. I modelled the service on the CAS service: if the applicant had funding, or could apply for it, we charged the appropriate fee; if not then we gave advice without charging or analysing data.

*In many ways then a return to the heady days of Lancaster and also lots of exciting statistical modelling work, including with one of the contributors to this special issue.*

At the Orvieto meeting of the IWSM, I was asked by Cecilia Vitiello, a statistician at La Sapienza University in Rome, to be the host for one year of two La Sapienza colleagues, Roberto Rocci who had finished his PhD, and Marco Alfò who was a cotutelle PhD student. I agreed to supervise them: I was then in the middle of EM projects in GLMs and GLMMs. Roberto did a post-doc project on GLM measurement error models: Aitkin and Rocci (2002) and Marco began a series of papers on GLMMs: Alfò and Aitkin (2000).

*And also a chance to further develop the Bayesian model comparisons work.*

Indeed, I joined the Bayesian research group, set up by Richard Boys, and involved Richard in my heterodox Bayesian treatment of model comparisons. The orthodox Bayesian insistence on integrating the likelihoods of the competing models over their priors did not work, despite many attempts to make it work. It was clear to me that Dempster's original idea

for comparing a simple null hypothesis with a composite alternative hypothesis could be extended generally to the comparison of multiple alternative incompletely specified models. Eventually, Richard and I and my PhD student Tom Chadwick published a paper on the general case (Aitkin, Boys, and Chadwick, 2005). This was the point where I was identified, and criticised, as a heterodox Bayesian. The good performance of the “posterior deviance” in determining the number of components in a finite mixture was demonstrated much later in a social network study (Aitkin, Vu, and Francis, 2015).

*But it seems that all good things come to an end and once again you suffered from the attacks on statistics as an independent subject and an erosion of support for academic statistical consultancy services.*

My term as the Consultancy Service Director concluded, and under the new management it now had to be self-funding. I knew this could not work, as the Engineering School already had a successful commercial statistical consultancy service set up by Barrie Wetherill. A new Director was appointed, and the Service was amalgamated soon afterwards with the commercial service. Its academic research support function disappeared.

Also, under the new Vice-Chancellor all departments had been abolished again, including the Department of Statistics, which was integrated with Mathematics in a School of Mathematics and Statistics. Back to square one. It was time to consider other possibilities.

*It seemed now that serendipity played a role in what was to shape the next stage of your life and the direction of your research. Can you tell how this happened?*

I had decided to look for visiting or permanent positions in the US. The American Institutes for Research (AIR), a company I had not heard of, advertised a position as Chief Statistician in Washington DC to advise the US Office of Education’s National Center for Education Statistics (NCES) on its statistical issues. The list of required skills was long, but I satisfied most of them. I applied, was interviewed and was offered the job, a 5-year fixed term appointment. I accepted. I applied for a leave of absence from the University, which was granted.

In addition to consulting with the senior staff of NCES, I gave a “brown-bag” series of lunchtime seminars at AIR, which shared office space in the same building with NCES, but on different floors. The series was on the design-based and model-based theories of analysis, and specific model-based methods like finite mixtures. It was attended by both senior NCES staff and junior AIR staff. This caused some friction with the NCES’s Chief Statistician, who followed the no models philosophy.

*But once again events intervened.*

Yes, the September 2001 attacks on the World Trade Center in New York and the Pentagon in Washington changed permanently the US political system, quite apart from the huge destruction and deaths they caused. After two years away, Newcastle refused to extend my leave of absence, and we decided to return.

*However, your time in the US was far from wasted and did lead to a new phase of research that you have been pursuing to this day.*

The (Acting) Commissioner of NCES Gary Phillips, a former Professor of Education Statistics, was very disappointed at my departure, and commissioned me to write a groundbreaking project, which he would fund for a year. I did so, on the Bayesian bootstrap ([Rubin, 1981](#)).

*Tell us a little about the basic idea here.*

The idea of the general multinomial model was not new: [Hartley and Rao \(1968\)](#) and [Ericson \(1969\)](#) had used the idea for both maximum likelihood and Bayesian analyses without a parametric model assumption. The general criticism of the Bayesian bootstrap by orthodox Bayesians at the time was that the non-informative improper Dirichlet prior was old hat and inappropriate: the Dirichlet Process prior (DPP) should be used instead. To me at the time, as now, the DPP was the solution to a different problem: how to make an inference about the entire population underlying the sample data. The DPP required the specification of a prior kernel density and a spawning parameter, and gave as output random draws from an infinite Dirichlet-weighted mixture of the specified kernel densities.

The aim of the Bayesian bootstrap was more modest: to make inferences about weighted functions of the multinomial parameters, like the mean, variance, median, percentiles in general, and regression coefficients. The problem with the Dirichlet prior was that any informative prior was informative, not only about the parameters, but also about the whole structure of the population. I expected that some form of weakly informative Dirichlet would do better than the non-informative Dirichlet, but found the opposite: any informative Dirichlet biased the posterior away from the likelihood information. The problem of where to place the informative prior parameters was insoluble, since the population support was unknown.

The irony of the arguments over priors was that it was the likelihood, not the prior, that excluded the parameters of the unsampled values — there was no data information about them. An analyst using an informative prior which gave weight to unobserved values had no basis for assigning the prior weights — they were entirely a creation of the analysts's mind.

This led me later ([Aitkin 2008](#); [Aitkin 2010](#), Chapter 4) to generalise the Bayesian bootstrap to stratified and clustered designs, and regression modelling. With these generalisations, the Bayesian bootstrap analysis could handle 90% of surveys with ignorable sample designs in official statistics, finally solving the previously intractable problem of model-based analysis depending on the assumed probability distribution for the response.

*In the coming years you were able to show the feasibility and power of these ideas. How did this happen?*

Following this study, the AIR asked my wife Irit and me to examine the current design

and analysis of the giant National Assessment of Educational Progress (NAEP), the largest US national survey of educational attainment. Initially this work was done through a research contract with the University of Newcastle, but after one year I retired early from Newcastle and we moved to Melbourne. Here Irit and I became personal contractors on the investigation, which developed steadily over more than 10 years.

*Tell us a little more about the context of this study and developments to the analysis that you were able to make.*

The survey structure was formidable: a multi-stage sample design with over-sampling of rare groups, a multiple imputation process generating plausible values for the latent student abilities underlying the test item responses, and a detailed set of child, teacher and school covariates mandated by Congress. Though there were two nested levels in the survey design, only the design effect for the nesting of schools in the PSU (primary sampling unit) level was accounted for. The nesting of students in schools was handled by fitting the schools (more than 400 of them) as fixed effects in a very large linear predictor. This assumed zero correlation between responses from children in the same school. With many other interaction variables, the model size was more than 1200, too large for computation, and the number of model terms was reduced to a smaller number of principal components.

Our reanalysis concluded with a hierarchical – four-level – fully model-based maximum likelihood analysis for the binary test item responses, in which a psychometric model for the student's latent ability (replacing the imputed abilities) became an additional model level in the multi-level structure. The complexity of the GLMM four-level variance-component model with covariates at all levels was unprecedented. We had to search for possible implementations which could handle this level of complexity, and with the help of Sophia Rabe-Hesketh with GLAMM and Jeroen Vermunt with Latent Gold we were able to fit the model with standard errors.

*And what was the impact of this work?*

The consequence of not allowing for the school design effect in the NCEES analysis was that their standard errors were under-stated for covariates at the student level and over-stated for those above the student level. Our investigation culminated in a short book ([Aitkin and Aitkin, 2011](#)) on the theories of analysis, and the comparative analyses of three large NAEP mathematics surveys. Soon afterwards the NCEES analysis policy, of not allowing model-based analyses of NCEES surveys, changed. Researchers applying for permission to do model-based analysis on the surveys were not only permitted, but also supported with model-based software designed for the analysis. The software was designed originally by Jon Cohen, the Chief Scientist of AIR, but under the previous NCEES philosophy its use had not been allowed.

Another advance for statistical modelling!

*But you are never a man to be just working on a single project and this (final?) move back to Australia led to another significant collaboration. How did this come about?*

Soon after my arrival in Melbourne I gave a seminar on the use of the posterior deviance for Bayesian model comparisons. After the few questions had ended, a student came up to me and said

“Your method of Bayesian model comparisons is completely wrong. You have to use the Bayes factor”

I thought that this was pretty cheeky, but such certainty is uncommon in a student, especially one in a different field, so I took him seriously, and we had a very long discussion. Charles Liu was a PhD student in psychology, but had become a Bayesian through the influence of a notable Adelaide Bayesian psychologist. I convinced him of the validity of the deviance approach, and he became an enthusiast. He found a recent paper in a major mathematical psychology journal which used the Bayes factor to choose the best-supported model from a set of two-parameter GLMs for successful performance in a recall study. The Bayes factor approach supported best the model with the lowest maximised likelihood, but the highest integrated likelihood, because its likelihood contours were very wide, while those for the highest maximised likelihood were very tight. We published [Liu and Aitkin \(2008\)](#) a major paper on this example, showing that all methods of model comparison other than the Bayes factor gave the best-supported model as the one with the highest maximised likelihood. Charles finished his PhD and moved to Boston. We later extended our assessment of current Bayesian methods to prediction: [Aitkin and Liu \(2018\)](#). My book on the general theory with applications ([Aitkin, 2010](#)) was reviewed unfavourably in an extraordinary 15 journal pages by three distinguished Bayesians, on the grounds that it wasn't Bayesian, and wasn't useful. My response was published ([Aitkin, 2013b](#)) in 12 pages!

*There was also new work with one of us.*

The social network group in psychology was very active, and with Brian Francis I was successful in an Australian Research Council-funded investigation of community structure in “two-mode” networks with binary ties. Working with Duy Vu, the research fellow on the grant, we showed that the latent class model for community structure was readily fitted and the number of communities could be identified by the comparison of posterior deviances for the several models being compared. One remarkable aspect of the two-mode analysis was that we could identify component structure in the Noordin Top terrorist network, which an exhaustive analysis of the derived one-mode network by ([Everton, 2012](#)) had failed to identify ([Aitkin, Vu, and Francis, 2017](#)).

*I know that you also continue to be very active and are working on a number of topics. One loose end from the NAEP study was the issue of missing data.*

Indeed, our maximum likelihood analysis of the NAEP was restricted to complete cases. This was mainly because the original NCES analysis with which we compared the model-based analysis had been done under this restriction, so the comparison was based on the same data structure. However we were concerned by the NCES philosophy of no imputation for missing covariate data, since this required models for the incomplete covariates. (The

imputation of plausible values for the student abilities was excluded from this restriction.) We had planned further analyses using full-information multiple imputation methods. We are currently working on this extension, using MCMC with the latent class mixture of conditionally independent distributions for the incomplete covariates.

*And you are also turning your attention to machine learning that I know you first worked on in your Newcastle days. What are your concerns about the recent developments here?*

Google states forthrightly (Web ref 23/05/2021):

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

In fact, the human intervention occurs at the beginning. Machine learning is a category of procedures which involve programming to achieve the aims of the designers/programmers. Machines can learn only what they are programmed to learn: the aim of a Machine learning procedure is fixed by its programming. This may involve fitting a model by maximum likelihood, or a Bayesian analysis, or an analysis by a computational algorithm without a model or a likelihood. It may be performing many analyses with different models or algorithms, and choosing the best, or averaging them, or listing them with the preference order of the models. A characteristic feature of many Machine learning procedures is the number of options that have to be specified to tune the procedure to the data. Default values are generally available to release the analyst from having to specify them.

Hype seems to be a feature of machine learning developments and advertising. Machine learning procedures have been overhyped in the past, as in the early neural network literature. Convergence failures of the optimisation methods in even simple neural networks led to the temporary abandonment of these methods. They were examined by (Foxall, 2001) in his PhD thesis at Newcastle. A reformulation of the neural network model as a latent variable model resolved the convergence difficulty, which was due to a mis-specification of the objective function (Aitkin, and Foxall, 2003), and provided an EM algorithm for maximisation of the latent variable model likelihood.

My current concern is with the proliferation of model-free machine learning procedures into large-scale data analysis, especially the current excitement with “deep learning”, a multi-level extension of the simple neural network model. Some of these procedures, like the frequentist bootstrap for generating the precision of a sample estimator of a population parameter or function, have model-based alternatives like the Bayesian bootstrap, which is based on the always-true multinomial distribution of Hartley and Rao (1968).

*Related to this, do you have anything to say on Breiman’s “two cultures” that essentially contrast algorithmic machine learning with what we would commonly understand as statistical modelling?*

I see a different two cultures, reflecting the opposed views of Neyman and Fisher on models and likelihood. Neyman did not accept the need for probability models or likelihood: the Gauss-Markov and Central Limit Theorems were all that was needed. Fisher thought models and likelihood fundamental, and the repeated sampling principle redundant.

Many machine learning procedures do not have existing model-based alternatives. For some of these, model-based procedures can be derived; for others a different form of model specification may be needed. Machine learning procedures which have no model basis cannot be evaluated for appropriateness: without them procedure comparisons would require extensive simulation studies which cannot be universal.

The development of model-based alternatives to current and future machine learning procedures would be an important contribution to statistical analysis. In such developments the Bayesian analysis would maximise the value of their contribution. The Bayesian bootstrap shows how important the always-true multinomial distribution can be.

*We've seen that throughout your career you have been concerned with teaching statistics, with a special focus on courses aimed at the non-mathematician reflecting your own inferential approaches. Where did this passion come from and what's driven your interest over the years?*

In my Science degree at Sydney, I took *Theory of Statistics*. It was interesting and well taught. I particularly liked the practical data analysis. Probability and probability distributions were easy and neat, but inference was mystifying, in the definition:

The standard error of a statistic is its standard deviation in its sampling distribution.

What could this mean? There was only the one sample — where did the others come from? They were hypothetical. Much later I discovered that Fisher dismissed the repeated sampling principle as irrelevant to inference. He showed that the likelihood gave the probability distribution of the sufficient statistics and the MLEs, and any ancillary statistics. You did not need any concept of repeated sampling to get the result. Fisher's problem was that he dismissed Bayes analysis (as did Neyman), so did not have a usable theory. Fiducial distributions worked only with location/scale models. But Neyman's repeated sampling worked (since it was hypothetical), and became the standard theory. This puzzle troubled me for many years.

Introductory courses for non-mathematical students were always difficult, for the students and the teachers. The standard course was mystifying to many students. The negative aspects of the subject and its teaching became notorious. This had a lasting effect on my response to people who asked me what I did. When I said statistician, there was always one of two standard replies:

- i) This was the course I hated most at Uni.



ii) I was never any good at maths.

— as though I was attacking them. I could only say, as I did, that it was the worst-taught course in the University. So to me, the need to understand why the courses were so bad, and how to improve them, was always pressing. Statisticians had been guilty of bad teaching of bad courses. I OWED future students a good and well-taught course.

*So what should such a course look like?*

I finally found how to do this for non-math students in my last course at Newcastle. I had returned from Washington in mid-year and was given two courses to teach in the first semester. The interesting one was *Statistics in the Modern World*, a name invented by a former staff member who had left. It was a very small (10 students) once-only course for Humanities and Arts students which would disappear in the restructuring of the Engineering School programme. The course had no notes, only a sheet of paper with the idea that students should read the newspaper at home and bring a story about, or involving statistics, to be discussed in the class. No staff member wanted to teach it or know about it, and it would never be given again.

This was the perfect opportunity to do exactly what I wanted. I used the StatLab book ([Hodges, Krech and Crutchfield, 1975](#)) with the data base of families involved in the Child Health and Development Study conceived and directed by Professor Jacob Yerushalmy, in the School of Public Health at the University of California, Berkeley.

I introduced this study to the students, and gave them three research questions they were to answer:

- What proportion of mothers and fathers in the population were smoking at the diagnosis of pregnancy?
- If mothers and/or fathers were smoking at the diagnosis of pregnancy, did they have smaller babies than the parents who weren't smoking?
- Is the intelligence of the child at age 11 related to parents' smoking at birth or at age 11, or to the birthweight of the baby?

The students generated random samples of different sizes from the database by throwing actual dice and, using the binomial distribution, drew conclusions about smoking, birthweight and intelligence.

Students found this absorbing and the course was a great success for everyone. I'm now writing a new book for CRC Press updating this approach, with a much wider range of examples, for mathematically capable students, in an integrated Bayesian/frequentist framework including a detailed discussion of the Fisher/Neyman arguments. I hope in a second volume to extend the range of models and data structures needed for Data Science and Machine Learning students.

*Finally, in one sentence how would you summarise your current take on statistical modelling?*

I have reached the level of ultra-heterodox Bayesian, through this promotion of the Bayesian bootstrap. I hope to go further ...

*Murray, thanks very much for this fascinating insight into your impressive career and the lasting influence that you have had on statistical modelling. We wish you well for a continuing active and productive life and hope to meet up with you again in the not too distant future, maybe at an IWSM.*

## References

- Aitkin, M. (1978) The analysis of unbalanced cross-classifications (with Discussion). *Journal of the Royal Statistical Society A*, **141**, 195—223.
- Aitkin, M., Anderson, D.A. and Hinde, J.P. (1981) Statistical modelling of data on teaching styles (with Discussion). *Journal of the Royal Statistical Society A*, **144**, 419—461.
- Aitkin, M., Bennett, N. and Hesketh, J. (1981) Teaching styles and pupil progress: a reanalysis. *British Journal of Educational Psychology*, **51**, 170—186.
- Aitkin, M. and Healey, A.R. (1984) Mathematical modelling of the EEC Labour Force Survey. in *Recent Developments in the Analysis of Large-Scale Data Sets*. Office for Official Publications of the European Communities, Luxembourg, 23—50.
- Aitkin, M. and Healey, A.R. (1987) Statistical modelling of the EEC Labour Force survey: a project history. in *The Statistical Consultant in Action*, eds. Hand, D.J. and Everitt, B.S. University Press, Cambridge, 171—179.
- Aitkin, M., Francis, B.J. and Raynal, N. (1987) Une étude comparative d'analyses des correspondances ou de classifications et des modes de variables latentes ou de classes latentes. *Revue de Statistique Appliquée*, **35**, 53—82.
- Aitkin, M. and Stasinopoulos, M. (1989) Likelihood analysis of a binomial sample size problem. in *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, eds. Gleser, L.J., Perlman, M.D., Press, S.J. and Sampson, A.R. Springer-Verlag, New York, 399—411.
- Aitkin, M., Anderson, D.A., Francis, B.J. and Hinde, J.P. (1989) *Statistical Modelling in GLIM*. Clarendon Press, Oxford.
- Aitkin, M. (1997) The calibration of P-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood (with Discussion). *Statistics and Computing*, **7**, 253—272.
- Aitkin, M. and Rocci, R. (2002) A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, **12**, 163—174.

- Aitkin, M. and Foxall, R. (2003) Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing*, **13**, 227—239.
- Aitkin, M., Boys, R.J. and Chadwick, T. (2005) Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing*, **15**, 217—230.
- Aitkin, M. (2008) Applications of the Bayesian bootstrap in finite population inference. *Journal of Official Statistics*, **24**, 21—51.
- Aitkin, M. (2010) *Statistical Inference: An Integrated Bayesian/Likelihood Approach*. Boca Raton: CRC Press.
- Aitkin, M. and Aitkin I. (2011) *Statistical Modeling of the National Assessment of Educational Progress*. Springer, New York.
- Aitkin, M. (2013a) Algebra and Statistics. *Amstat News*, October 2013, **436**, 24—25.
- Aitkin, M. (2013b) Comments on the review of Statistical Inference. *Statistics and Risk Modeling*, **30**, 121—132.
- Aitkin, M., Vu, D. and Francis, B. (2015) A new Bayesian approach for determining the number of components in a finite mixture. *Metron*, **73**, 155—176.
- Aitkin, M., Vu, D. and Francis, B. (2017) Statistical modelling of a terrorist network. *Journal of the Royal Statistical Society A*, **180**, 751—768.
- Aitkin, M. and Liu, C.C. (2018) Confidence, credibility and prediction (with discussion by Little and Welsh and response). *Metron*, **76**, 305—320.
- Aitkin, M. (2018) *A history of the GLIM statistical package*. *International Statistical Review*, **86**, 275—299.
- Alfö, M. and Aitkin, M. (2000) Random coefficient models for binary longitudinal responses with attrition. *Statistics and Computing*, **10**, 275—283.
- Bennett, S.N. (1976) *Teaching Styles and Pupil Progress*. London: Open Books.
- Bock, R.D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika*, **46**, 443—459.
- Dempster, A.P. (1974) The direct use of likelihood in significance testing. In *Proceedings of the Conference on Foundational Questions in Statistical Inference*, eds. O. Barndorff-Nielsen, P. Blaesild and G. Sison, 335—352.
- Dempster, A., Laird, N., and Rubin, D. (1977) *Maximum likelihood from incomplete data via the EM algorithm*. *Journal of the Royal Statistical Society B*, **39** 1—38.
- Dempster, A.P. (1997) The direct use of likelihood in significance testing. *Statistics and Computing*, **7**, 247—252.
- Ericson, W.A. (1969) Subjective Bayesian models in sampling finite populations (with Discussion). *Journal of the Royal Statistical Society B*, **31**, 195—233.

- Everton, S.F. (2012) *Disrupting Dark Networks*. University Press, Cambridge.
- Foxall, R. (2001) *Statistical modelling of artificial neural networks*. PhD thesis, University of Newcastle-upon-Tyne.
- Hartley, H.O. and Rao, J.N.S. (1968) A new estimation theory for sample surveys. *Biometrika*, **55**, 547—557.
- Hodges, J.L., Krech, D. and Crutchfield, R.S. (1975) *Statlab: An Empirical Introduction To Statistics*. McGraw Hill Higher Education
- Liu, C.C. and Aitkin, M. (2008) Bayes factors: prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, **52**, 362—375.
- Rubin, D. (1981) The Bayesian bootstrap. *Annals of Statistics*, **9**, 130—134.