‘Ethical’ as well as ‘Radical’ Statistics[1]

Harvey Goldstein

All professional ethical codes stress the importance of honesty, personal integrity and the need to strive for objectivity. The American Statistical Association (ASA, 1999), for example, is clear that statisticians and those carrying out statistical analyses should “remain current in terms of statistical methodology: yesterday’s preferred methods may be barely acceptable today”. Thus, advances in knowledge can not only make previous technologies or methodology less efficient, but also that new knowledge can expose the hitherto hidden distortions and biases inherent in such previous technologies. New knowledge can make unethical what may previously have been considered acceptable procedure. This impact of knowledge is very clear in areas such as medicine, where, for example, the practice of patient bleeding may have been mainstream orthodoxy in the 18th century but would be considered highly unethical if used instead of treatments known to be scientifically effective in the 21st century.

I shall look at a particular set of evolving methodologies, those generally termed “multilevel models”, where I have some knowledge of techniques and applications. In particular, I would argue that this methodology has now reached a stage of maturity, as witnessed by its routine use and its incorporation into major statistical packages, which implies there is an ethical obligation to use it where appropriate. In other words, this methodology is indeed one that has made a large number of yesterday’s preferred methods “barely acceptable”.

I shall assume that readers are familiar with the basic methodology: if not there are introductory texts such as Hox (2002) and a more advanced text is Goldstein (2003). An excellent set of introductory materials is also available online at

http://www.cmm.bristol.ac.uk/learning-training/index.shtml .

If we have hierarchically structured data, and there are few real life situations where we do not, and we ignore the structure when modelling, then our inferences will be incorrect. Standard errors will tend to be too small, significance tests too optimistic and confidence

[1] This is based upon a chapter to appear in: A. T. Panter & S. K. Sterba (Eds.), Handbook of ethics in quantitative methodology. New York, NY: Taylor-Francis.

intervals too short. The size of such biases will depend upon the strength of the structure, but in general there is little (ethical) justification for ignoring it.

An important example of this in educational research was the early school effectiveness study in Inner London schools, entitled Fifteen Thousand Hours. This study obtained information from 2,000 children in 12 Secondary schools. The study made comparisons between school types, for example boys and girls schools, found non-significant differences, and concluded that such differences are of “negligible importance” (Goldstein, 1980). Yet, with a sample size of only 12 schools, it is hardly surprising that almost all comparisons will be non-significant.[2] The authors failed to appreciate this design problem and also made the common error of equating “non-significance” with “non-existence.” While this issue, often referred to as the “units of analysis problem”, was fairly well understood at that time and had been discussed in the methodological literature, it could be argued that this lapse should be regarded as merely *incompetent* rather than *unethical* behaviour. Yet, in their response to this point (Rutter et al., 1980), the authors refused to accept the strictures. Because that study turned out to be influential, appearing widely on student reading lists, a refusal to concede that there may have been a serious flaw could be considered by many to constitute a case where ethical norms were breached. This would not be in terms of deliberately providing a misleading description, but rather in terms of a failure to ensure that, as researchers, they were prepared properly to acknowledge current good professional practice. All of this was particularly unfortunate since the lessons for study design were obscured, and the importance of sampling adequate numbers of higher level units was not made clear to many researchers in this field.

The multilevel approach helps to shift the focus from the clustering simply being a convenient procedure to obtain a sample to a positive attempt to bring in ecological variables that are defined at the cluster level. Thus, in a recent study for the design of a large scale birth cohort study in the UK, the think tank (Longview, 2008) argued for a sample that consists of a nationally representative component together with a small number of tightly clustered samples in local areas or clustered around local institutions. The area samples would include all the births over a period of, say, a year, so that the characteristics of each child’s peer group could be measured, for example when they attend preschool facilities. The sample would obtain nationally representative data and the existence of a common set of variables

[2] A later analysis of a very similar population, but fitting a multilevel model to a large sample of schools, showed clear differences between boys, girls and mixed schools (Goldstein et al., 1993).

across the sample would allow the various subsamples to be linked. This linking can be done formally within the modelling framework, “borrowing strength” across the subsamples. In other contexts such designs are often known as matrix designs or rotation designs and they have many advantages in terms of efficiency as well as being able to combine local and national data (see for example, Goldstein, 2003, Chapter 6). In social research this is important because it begins to address potential criticisms of large scale empirical research on populations: that they ignore contextually relevant factors. The ability to combine large representative sample data with more intensive local data that are sensitive to local issues also begins to provide a way of drawing together large scale data sets and small scale studies such as those that collect detailed ethnographic data. Thus, the design possibilities for such studies become extended and this knowledge, as it becomes widely accepted, will exert an ethical pressure to consider these possibilities.

If a multilevel analysis is envisaged, then there needs to be sufficient power to carry this out efficiently and data relevant to identifying and characterising higher level units has to be collected (i.e., unit and cluster identifiers such as school IDs and student IDs). I assume that in general the data analyst is also involved in design, although that will not always be the case, for example in secondary data analysis. Nevertheless, it will always be desirable that somebody with experience of data analysis is involved with the initial research design. So for practical purposes we can consider this to be the same person.

Real life data generally have a complex structure that is hierarchical and may also include cross classifications etc. It is ethically responsible for the data analyst to be aware of this, and also be concerned to make collaborators sensitive to this issue when a study is being designed so that there is sufficient power for required comparisons, especially those that involve higher level units. The data analyst will also have a role in formulating questions based upon what they know about the possibilities for data modelling. Thus, for example, the ability of multilevel models to model variation, as in the study of segregation, may not be immediately apparent to many researchers. Structuring a study to separate sources of variation may also be important for efficiency and understanding. Thus, O’Muircheartaigh and Campanelli (1998) cross classified survey interviewers by survey areas and were able to separate the between-interviewer variance from the between-area variance for various responses. Among other things this analysis allowed the “effects” of different interviewers to be estimated and is therefore able to inform more efficient survey design.

When it comes to modelling, the data analyst again has an ethical responsibility not only to seek the appropriate tools, but also to involve collaborators in understanding how they are being used and how results are to be interpreted. Likewise, the data analyst should be involved in the preparation of papers and reports that present results so that appropriate interpretations are communicated.

As in all statistical modelling, the analyst needs to be sensitive to the assumptions that are being made. Techniques for checking distributional assumptions using, for example, outlier analysis, are available (see for example Goldstein, 2003, Chapter 3). Sensitivity analyses can also be carried out where assumptions are systematically varied to view the effect on estimates. Where assumptions are not tenable, for example when a distribution cannot be assumed to be Gaussian, then as in traditional modelling, transformations or alternative model formulations may be possible.

A Case History: School League Tables

Starting in the 1980s, many educational systems, especially in the USA and the UK, began to experiment with the publication of examination results and test scores for schools and colleges. Visscher (2001) gives a history of international developments and a review of the debate. These league tables were designed for two principal purposes. The first was to monitor the performance of individual institutions so that “poorly performing” ones could be identified for further attention. At one extreme this involved their “formative” use as part of a “school improvement” programme where results were not published but used to inform individual schools of their possible strengths and weaknesses (Yang et al., 1999). At the other extreme they have been used directly in the determination of school funding and teacher remuneration (Dorn, 1998) and in decisions about school closure.

The second main purpose was to provide parents and students with information to guide school choice. In the UK this was explicitly stated in the so called “parents charter” issued by the John Major Government (DES, 1991), which encouraged parents to make use of the relative positions of (secondary) schools in tables of examination results. The implication was that those schools with higher average performance were educationally more effective.

These early uses of league tables were strongly criticised, especially by teacher unions and academics, on the grounds that average performance was strongly associated with achievement when students started school, and since schools were generally differentiated in terms of these initial achievements the final outcomes were in large part simply reflecting intake. It was argued that “value added” or “adjusted” performance was more appropriate, where account was taken of initial differences. To do this, models were constructed that were essentially multilevel, with students nested within schools (see Goldstein & Spiegelhalter, 1996 for a technical discussion). To some extent, policymakers took note of this criticism, so that adjusted league tables were introduced, and in England from 1995, it became official Government policy to move towards a ”value added” system.[3]

By 2003, value added tables for both primary and secondary stages of education were being published in England, alongside the unadjusted ones. Unfortunately, the media in general, while giving great prominence to the unadjusted or ‘raw’ tables, virtually ignore the “value added” ones, and the Government appears to be relatively unconcerned with this, leaving itself open to criticisms of complacency and even hypocrisy. The consequences for individual schools of being ranked low on such tables are fairly clear. Yet, in all this debate, the provisional nature of statistical modelling has largely been overlooked and the potential ‘unfairness’ to individual schools largely ignored. It is certainly the case that adjusted performance comparisons provide a “fairer” way to compare institutions, but they themselves are only as good as the data used to provide them and suffer from numerous drawbacks. Yet, many proponents of adjusted tables have either ignored or downplayed the limitations of the statistical models. Indeed, FitzGibbons and Tymms (2002) who carried out the pilot work for the English value added tables defend their use of “simple” methodology by stating that

> “The multi-level analysis, requiring special software and a postgraduate course in statistical analysis, was in contrast to the ordinary least squares analysis that could be taught in primary schools” (sic) and that “value added scores for departments or schools, correlated at worst 0.93, and more usually higher, up to 0.99 on the two (multilevel vs. ordinary least squares) analyses”.

In fact, the high correlations quoted result from the fact that only variance component models were fitted by these authors so that schools varied solely in terms of their intercept terms. In fact, schools are known to be differentially effective (see for example Yang et al., 1999), their “value added” scores varying according to the intake achievement, gender and other student level factors. To understand the role of such factors, it is essential to fit more complex multilevel models that include both an intercept and slope terms to reflect differential school effects. If this is done, the misleading claims made by the above authors do not stand up to careful examination (Yang et

[3] In the UK the four constituent countries have separate jurisdiction over education. Thus, by 2009 only England still published school league tables and, for example, Scotland had never instituted their publication.

al., 1999). This case is an illustration of an ethical failure to understand the true complexity of the system being studied, so that over-simple models are used that do not reflect important aspects of the data. The above quotations also reflect a rather worrying antagonism that some researchers exhibit towards the use of complex models on the grounds that "simple models will do the same job." In fact, simple models often do not 'do the same job'. This kind of intellectual philistinism towards sophisticated quantitative modelling is as ethically reprehensible as it is scientifically blinkered. I am not, of course, advocating model complexity for the sake of it, but I am arguing in favour of modelling at a level of complexity that seeks to match the complexity of the real life data being analysed.

Some guidelines

Let me try to formulate some general guidelines for analysis, design, interpretation and reporting drawn from the above discussion.

Using mathematical or statistical models to describe complex systems has always been a kind of catch-up process. As our methodological tools and data collection facilities become more sophisticated, they can uncover more of the complexity that lies within natural or social systems. Unfortunately, all too often, researchers are confused by a perceived (and often justifiable) need to present findings in an accessible, as simple as possible, form to non-experts, with the need to carry out research using complex techniques that are *only* accessible to experts. The challenge for the experts is not to simplify their techniques but to simplify their explanations of those techniques. I would suggest that multilevel modelling has reached a stage of development and accessibility that should mandate its routine use for modelling complex hierarchical structures, and examples have been presented to show how an understanding of multilevel modelling can improve our understandings and generally advance research.

One implication is that not only researchers but also those who train researchers, largely in universities, should incorporate such modelling techniques as routine. It is quite interesting that there is little emphasis in existing ethical codes of, for example, the APA, ASA, RSS, and ISI organizations on the role of methodological educators. Yet this is unfortunate because it is such individuals and the materials they produce who will have a large influence on the conduct of research and scholarship.

Another implication is that those carrying out research have a responsibility to remain abreast of developments in both methodology and its applications. I would argue that, given access to the Internet, there are now adequate opportunities for this to happen, using the web resources that have been mentioned. Professional societies also play an important role here in providing continuing professional development activities, in the form of materials, workshops and meetings. Radstats too needs to understand these issues, especially by providing an interface between professionals and non-professional users of data.

References

American Statistical Association, (1999). *Ethical Guidelines for Statistical Practice.* http://www.amstat.org/about/ethicalguidelines.cfm

DES (Department for Education and Science) (1991). *The Parent's Charter: You and your child's education*, DES, London.

Dorn, S. (1998). "The Political Legacy of School Accountability Systems." education policy analysis archives **6**(1): 1-33.

FitzGibbon, C. T., & P. Tymms (2002). Technical and ethical issues in indicator systems: doing things right and doing things wrong. *Education Policy Analysis Archives, 10*, 1-26.

Goldstein H, Rasbash J, Yang M, Woodhouse G, Pan H, Nuttall D & Thomas S. (1993). A multilevel analysis of school examination results. Oxford Review of Education **19** (4) 425-433

Goldstein, H. and D. J. Spiegelhalter (1996). "League tables and their limitations: statistical issues in comparisons of institutional performance." Journal of the Royal Statistical Society, A. **159**: 385-443.

Goldstein, H. (2003). *Multilevel statistical models (3rd Ed.).* London, Edward Arnold.

Goldstein, H. In "Critical notice of Fifteen Thousand Hours". (1980). J. Child Psychology & Psychiatry, 21, 363-369.

Hox, J. (2002). *Multilevel analysis, techniques and applications.* Mahwah, New Jersey, Erlbaum.

Longview, (2008). Scientific case for a new cohort study. http://www.longviewuk.com/pages/reportsnew.shtml

O'Muircheartaigh, C. and Campanelli, P. (1999). "A multilevel exploration of the role of interviewers in survey non-response." Journal of the Royal Statistical Society, A. **162**: 437-446.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A (1980). In Critical notice of Fifteen Thousand Hours. (1980). *Journal of Child Psychology & Psychiatry, 21*, 363-369.

Visscher, A. (2001). "Public school performance indicators: problems and recommendations." Studies in Educational Evaluation **27**: 199-214.

Yang, M., H. Goldstein, Rath T, Hill N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education, 25*, 469-483.

Harvey Goldstein
Centre for Multilevel Modelling
University of Bristol
h.goldstein@bristol.ac.uk