OUTLIERS AND MULTILEVEL MODELS

John F. Bell & Eva Malacova[1]
University of Cambridge Local Examinations Syndicate
Bell.j@ucles.org.uk

Paper presented at RC33 Sixth International Conference on Social Science Methodology: Recent Developments and Applications in Social Research Methodology. Amsterdam, The Netherlands, August 16-20, 2004

This paper will consider some problems relating to outliers in multilevel models drawing on research into two data sets including progress in secondary and higher education and university admissions. In particular, it will be demonstrated that when a contaminating process generates outliers, they can lead to unnecessarily complex models. In these cases, it is necessary to consider carefully the nature of the data and the processes that are being modelled. For example, do variables measure the same thing for different sub-populations in the data set?

In one example, it will be demonstrated that such contaminating processes lead to an extremely complex model that completely masks the true relationship. The paper will also consider the importance of issues such as reproducibility of an analysis and the role of subjectivity and background knowledge in dealing with outliers.

Keywords: multilevel models, outliers, educational data

[1] Research and Evaluation Division, Assessment Directorate, University of Cambridge Local Examinations Syndicate, 1 Hills Road, Cambridge, CB1 2EU.
bell.j@ucles.org.uk

INTRODUCTION

In this paper, some issues relating to the effect of outliers on multilevel models will be considered. Outliers are a particular problem in multilevel modelling because their presence can greatly increase the complexity of the model (Langford and Lewis, 1998). Outliers can be broadly categorised into two types: data entry errors and contaminants. The first type of outliers is self-explanatory and the second type refers to data that have been generated by a different process from that of the majority of the data.

Although the detection of outliers uses mathematical methods (Barnett and Lewis, 1994), the way that they are dealt with may depend on reasoned but ultimately subjective judgement. Expert knowledge of the data and the processes involved may be needed. When outliers are found there are three methods of dealing with them: correction, omission and accommodation. Firstly, if the outlier has been generated by a mistake in data entry or in the construction of the data set (e.g. when merging files) then it may be possible to correct it. Sometimes this is the result of transcription errors and in others it might be possible to check the data with the respondents. If it cannot be corrected then it must be omitted. This is the second approach. This can also be applied to contaminants; there are two reasons for omitting them. Firstly, there may be too few to draw any conclusions about, for instance, fitting a dummy variable to remove one unit is not particularly useful. However, many multilevel analyses tend to be large. There is little difficulty in opting to omit 2 observations in a sample of 100 but it becomes a bit more problematic when omitting 200 observations from 10,000. The second reason is linked to the third method. Accommodation of outliers maintains the sample size but requires additional parameters to be fitted while omission reduces the sample size but might lead to a simple model. The decision which method is better depends on a number of factors. If the contaminating process is very different from the main process then the final model may be extremely complex and prove very difficult to report. Secondly, the contaminating process might not be relevant to the main purpose of the study and so modelling it is inappropriate. In both cases, omission may be the best solution. However, if the accommodation involves only adding a few additional parameters and/or is of some other relevance to the analysis then the method that accommodates outliers should be used.

All scientific research should be reproducible so it is particularly important that the researcher documents the corrections and omissions from the data set. For accommodation of outliers this is not a problem because the process of extraction and preparation of the data is part of the reported model. However, it is vital that whenever outliers are omitted from the analysis their omission is documented (the first example below is a clear example of this).

In an analysis there are two times when outliers can be detected: before the analysis with exploratory techniques and after the analysis using residuals. Exploratory analysis techniques and multilevel models have been discussed in Bell (2001). It can be very useful to consider outliers before fitting multilevel models. Outliers in multilevel models can greatly increase the complexity of multilevel models, which may then cause estimation problems. This is illustrated in the first example. In this example, the outliers are at level one and it illustrates how they can lead to spurious model complexity. In the second example, the outliers are at the second level and demonstrate how they can reveal interesting features of the data.

1. EXAMPLE 1: MODELLING PROGRESS IN HIGHER EDUCATION

This first example demonstrates how the complexity of a fitted model can be increased by failing to deal with a contaminating process. This research considered the progress of English university students from A-level examinations (usually taken at eighteen years of age) to the completion of their first degree. This work was described at the Amsterdam Multilevel Modelling Conference last year (Bell, 2003) and the conference paper gives a more complete account of the issues raised by this analysis. However, in the conference paper, only the final models were described and the issues discussed here were not considered. The analysis was based on archive data from the early 1990s extracted from the university statistical record, which contains information for all students in all of the old universities in the United Kingdom (i.e. they were universities prior to 1992). It was decided to use subsets of the data in this study. Obviously only students that had sat A-levels could be considered. It should be noted that the United Kingdom is made up of England, Scotland, Wales and Northern Ireland. The educational systems of these constituent parts vary. Although students from England, Wales and Northern Ireland attend Scottish institutions, both the Scottish school examination system and the university structure differ from those in the other parts. For this reason, it was decided that Scottish universities should be excluded from the analyses described in this paper. Students from Northern Ireland take A-levels but the organisation of the school system is very different from that of England and Wales and so the universities in Northern Ireland were also excluded. The objective of this analysis was to model a clear and well-specified process.

To measure prior attainment a score was derived from the A-level examination grades obtained by the candidates. Students take a wide variety of A-levels, usually of some relevance as a preparation for the course followed at university, although this varies from subject to subject. It would be surprising if the relationship between A-level score and university degree class was the same for all subjects. For this reason, it was decided to investigate the progress of students following the same subject. In this paper, only the results for English are presented.

At the end of their courses students are awarded degrees which are classified as follows: Fail, ordinary, and four classes of honours degrees, 3rd, 2.2, 2.1, and 1st. For the purposes of this paper, a binary dependent variable was formed taking the value 1 for a 1st or a 2.1 (sometimes referred to as a good degree) and 0 for all the other categories. For the purposes of this paper, the A-level grades obtained need to be converted into a score. When the data used in this paper were gathered students intending to go to university usually sat three to four A-level examinations (the fourth examination usually being A-level general studies). These examinations are taken after a two-year course over the ages of seventeen and eighteen. Traditionally pupils started secondary education at eleven and in what is known as the first form. The pupils taking A-levels are sometimes referred to as sixth formers (lower sixth-formers for first year and upper sixth-formers for the second, and not sixth and seventh-formers as might be expected). These examinations had five pass grades: A, B, C, D and E. For the purposes of this study, these A-level grades were converted into following scores: 0, 2, 4, 6, and 10. (This was the usual tariff at the time the data were collected and was designed to allow for another examination that was deemed to be the equivalent of half an A-level examination). For each student, the total of the best three A-level grades excluding general studies was calculated. The general studies A-level was excluded because not all schools entered students for this examination and it was not always used by universities. In this paper, the total of the best three A-level grades is will be referred to as the A-level score. For the purposes of modelling, this score was standardised with mean 0 and variance 1 and will be referred to as the standardised A-level score.

Although there is a number of different types of schools that students attended to take A-levels, for the purpose of this paper only the difference between state maintained and independent type of school will be considered by using a dummy variable.

In this case, the outliers were students with poor A-level results who nonetheless obtained good degree results (the reverse is a more common occurrence because academic ability at 18 is obviously not the sole determinant of success at university and because 'poor A-level result' students often don't get in). A closer inspection revealed that most of these cases were mature students. There are logical arguments as to why the relationship between A-level and degree progress of mature students is different from that of applicants straight from school. As the name implies, mature students are likely to be more dedicated and committed. However, more fundamentally there is the interpretation of the A-level score. The age of entry for the remaining students is usually eighteen or nineteen. Mature students were more likely to have relatively poor A-level examination results but this shortcoming has been compensated for by other qualifications and life experience in general. This means that they form a different population from the main body of the student population and the processes governing their selection and progress will be different. The A-level scores would not necessarily be expected to have the same predictive validity for mature and non-mature students. For this reason, a dummy variable was created to identify these students. The applicants straight from school have recently taken the examinations and the A-level scores measure their current attainment and potential. For a mature student, this is less likely to be the case. Their A-level grades are not necessarily a true reflection of their current achievement and potential.

If all the data excluding dummy variable for mature students are analysed, then this results in the cubic regression model presented in Figure 1. The model is obviously very complicated. Every fixed term in the model is also random and this model is extremely difficult to interpret. There are better ways of presenting the results of MLwin (Bell, 2002) but the screen dump is a vivid representation of such complexity.

Figure 1: Final Model of Higher Education Progress Full data set (ignoring mature student effect)

It is not proposed to explain this model in detail. Instead, the exploratory analysis that should have been done before the main analysis begins is described. In logistic regressions, an outlier is either generated by a success when most of the cases near it are failures or vice versa. This is illustrated in Figure 2, which is a plot of the results from four different universities. Since the dependent variable can only take the values of 0 and 1, there would have been a large number of overlapping points if jittering (adding a small amount of random error) had not been used. This has resulted in two clouds of points at the top and bottom of the plot. A lowess smooth has been added to the plot. The mature students have been identified using crosses and non-mature students using circles. These examples show how the relationships vary from university to university. However, it is clear that the number of successful mature students with low A-level point scores determines the shape of the curves.

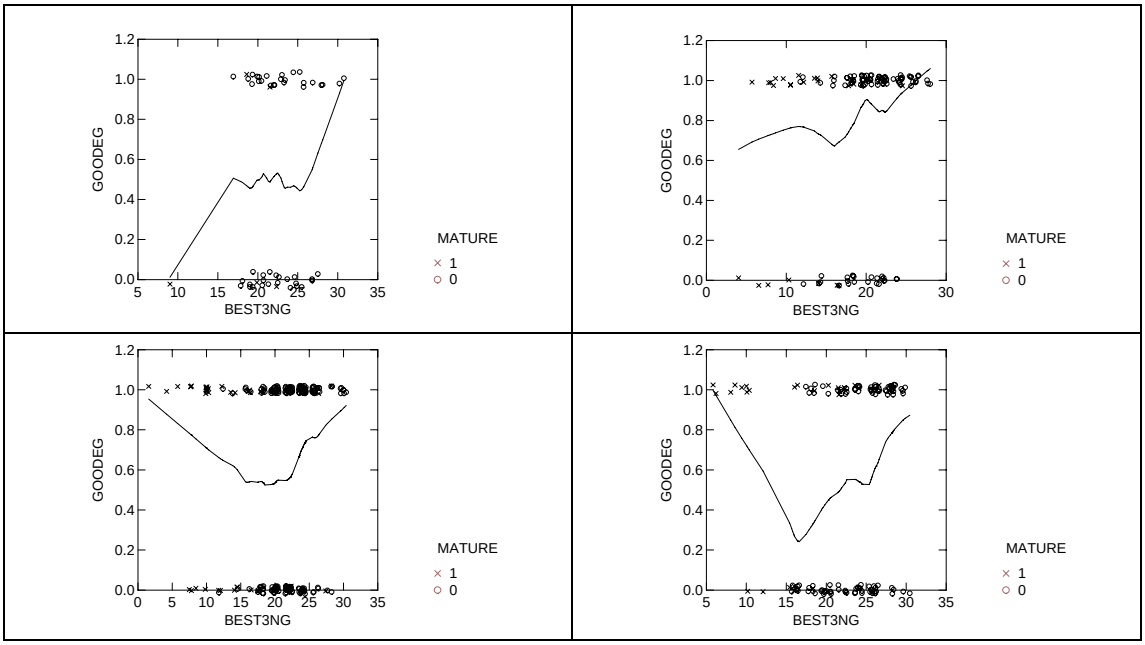


Figure 2: Relationship between probability of good degree and A-level scores

There are two approaches to dealing with the mature students: either they can be omitted or they can be accommodated. The final models from both of these processes are summarised in Table 1. Both are much simpler than figure 1. The model based on the omission of outliers is quadratic for standardised A-level score, while the model for accommodation of outliers is cubic for standardised A-level score.

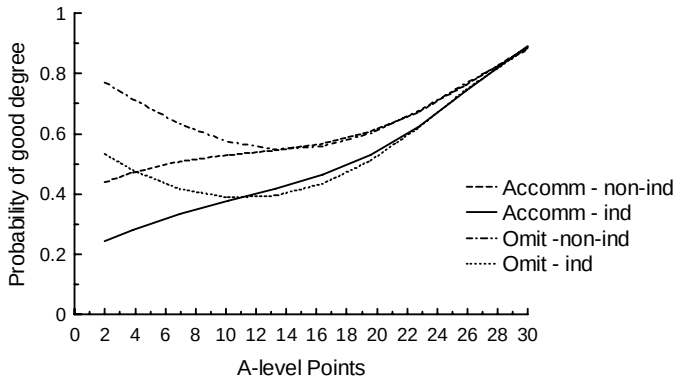
Table 1: Models of Higher Education Progress: Accommodating and omitting mature students

Model	Cons	A-level score	(A-level score)^2	(A-level score)^3	School: Ind=1
Accom.	0.44 (0.10)	0.43 (0.07)	0.24 (0.03)	0.06 (0.02)	-0.31 (0.08)
Omit	0.41 (0.10)	0.50 (0.10)	0.28 (0.06)	-	-0.37 (0.09)

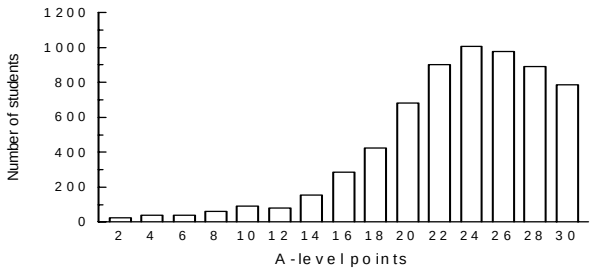
Model	(A-level)*ind	sex: male=1	year: 1993=1	Mature: yes=1	University var.
Accom	0.21 (0.08)	-0.14 (0.06)	0.15 (0.06)	0.72 (0.12)	0.16 (0.05)
Omit	0.25 (0.10)	-0.15 (0.06)	0.15 (0.06)	-	0.14 (0.05)

Note: A-level scores were standardised

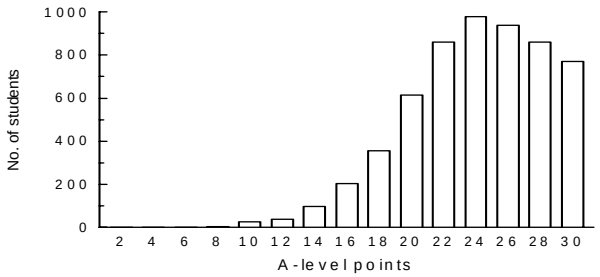
Although the models seem different in Table 1, they are actually very similar. This is illustrated by the probability curves in Figure 3(a). On the right hand-side of this plot there is little difference between the two models. On the left hand-side there is a difference. However, an inspection of the A-level points distributions for the whole set of data with accommodation and the subset created by omitting the mature students demonstrates that there was no difference of practical significance except were there is not much data to estimate the curve. There are hardly any data below an A-level score of 12 in either model, which is where the models differ (Figures 3(b) and (c)). This example has demonstrated an extreme case of how outliers at level 1 can lead to models involving random slopes. This example also illustrates the advantages of carrying out some exploratory analysis of multilevel data sets before fitting multilevel models. However, both approaches would lead to the same substantive conclusions about the relationship between progress and school type.



(a) Comparison of fitted curves for female students



(b) Distribution of A-level points including mature students



(c) Distribution of A-level points excluding mature students

Figure 3: Comparisons of the higher education models

EXAMPLE 2: CAMBRIDGE ADMISSIONS

As one of the world's leading universities, competition for places on the undergraduate courses at Cambridge university is intense and many applicants are rejected. In this example, the probability of success of Cambridge applicants for one particular subject is considered.

The analyses described in this paper are to illustrate the effect of outliers on multilevel models and it is not intended to be a definitive analysis of the process. As such, the analyses were restricted to a subset of the applicants. It was decided to consider only those applicants who were in the final year of schooling or had obtained their A-levels in the previous year. For the purposes of the models described in this paper, only applicants with A-level and GCSE results could be used. However, it should be noted that many of the applicants would not have sat their A-levels when the decision about their application was made; teachers'

forecasts of the grades the applicants will get are used in place of results. A-levels are the examinations usually used for university admissions (Bell, Malacova and Shannon, 2003), while GCSEs (General Certificate of Secondary Education) are taken two years earlier. Students usually take eight or more of GCSEs that cover a wide range of different subject areas (Bell, 2001). The objective of the study was to investigate whether GCSE score could be used as an additional piece of evidence in the admissions process. A-level results are not of much use because there are more applicants than places who have at least three grades A's at A-level. Although there are older applicants with such examination results, it was decided to exclude them for the reasons stated in the previous example.

In this paper, school attainment at GCSE and A-level is used to model the probability of success in the application of candidates. Success is defined as being made a conditional or unconditional offer for a place on the undergraduate course in Cambridge. The variables used in this analysis are given in Table 2. The proportion of successful applicants from independent schools is higher than for those coming from state schools. This is an obvious issue to be investigated. Using information from the examination boards' database, applicants were classified as coming from the independent or state sector depending on the institution where they sat their A-levels. The applications to Cambridge are based on the college system and this generates a complex multilevel model structure. Applicants can either apply to a specific college or they can make an open application. Open applicants are allocated to individual colleges. They go to colleges which happen in that year to have fewer applications per place in the particular subject than the other colleges. Once allocated to a college their applications will be treated exactly the same as any other to that college. However, because open applicants tend to be less successful doubts are sometimes expressed about the treatment of their applications (out of forty "open" applicants, only four were successful). There is also a 'pool' process, which is designed to give a second chance for applicants to re-apply if their chosen college has more suitable applicants than places. Some schools have more experience of the Cambridge applications system than others do and this could, in theory, have an effect on the success of the applicants. A measure based on the number of applicants coming from a particular centre but applying to any course was used to investigate this. This distribution of experience variable is not related to school type – there are state-sector institutions with considerable experience of the Cambridge admissions procedure.

Table 2: Explanatory variables used in the modelling

Measure	Type
A-level Performance	Dummy variables: 30 points=1, < 30 points =0
Mean GCSE	Mean GCSE, (mean GCSE)^2, (mean GCSE)^3
Sex	Dummy variable: Male=1, female =0
School type	Dummy variable: Independent=1, state =0
Open	Dummy variable: Open = 1, college = 0
School's Experience of Cambridge applications	Dummy variables: High, medium and low levels of experience

The multilevel models were fitted using MLwin. Initially models were fitted using Quasilikelihood estimation. This uses the mean and variance properties associated with binomial distributions to define the covariance structure which is then fitted using IGLS/RIGLS (Goldstein, 1995). Although this form of parameter estimation is rapid, the estimates are biased (Snidjers and Bosker, 1999). The final models reported in this paper were obtained with Monte Carlo Markov Chain (MCMC) methods (Browne, 2002) which give better estimates but are more computationally intensive. The parameter estimates for the

fitted models are presented in Table 3. Not all the models fitted have been reported (indeed in the actual analysis more variables were considered such as social class and ethnicity). The first model is the null multilevel model. From this model, it can be observed that the college level variation is not significant when no other factors are considered. The second model includes all the significant explanatory variables except those relating to A-level and GCSE attainment. For model II, open applicants were much less likely to obtain a place but applicants from independent schools and/or experienced schools (those with eleven or more applicants to Cambridge in the year in question) were more likely to obtain a place. In model III, measures of prior attainment have been included. After controlling for attainment the only other significant explanatory variable was high levels of school experience of Cambridge applications.

Table 3: Parameter estimates for a range of models
(estimate – normal font, s.e. in italics)

Candidate level Parameters

Model	Constant	Open=1	Ind=1	Exp4=1	GCSE	$(GCSE)^2$	3A's
I	-0.60 *0.07*						
II	-0.86 *0.11*	-1.84 *0.59*	0.29 *0.14*	0.51 *0.13*			
III	70.28 *1.80*			0.30 *0.15*	-22.06 *0.46*	1.64 *0.04*	2.72 *0.31*
IV	75.81 *1.61*				-23.68 *0.32*	1.75 *0.03*	2.87 *0.32*

College Parameters

Model	College – B =1	College – C=1	College variance
I			0.03 *0.03*
II			0.07 *0.06*
III			0.16 *0.10*
IV	1.11 044	1.90 0.58	0.05 *0.05*

Although in model III the college level variance component is not significant the value is fairly large and has increased compared with the non-attainment models. In such circumstances it can make sense to check the college level residuals for outliers. The college level residuals were plotted against rank (Figure 4). In this plot, there is evidence to suggest that three of the colleges could be considered as outliers - one with a lower level of success and two with a higher level of success. Dummy variables were created for each of these colleges and a further model was fitted. After some experimenting only the parameters for colleges with higher levels of success were significant. However, in model IV that includes these parameters the high school experience effect was not present.

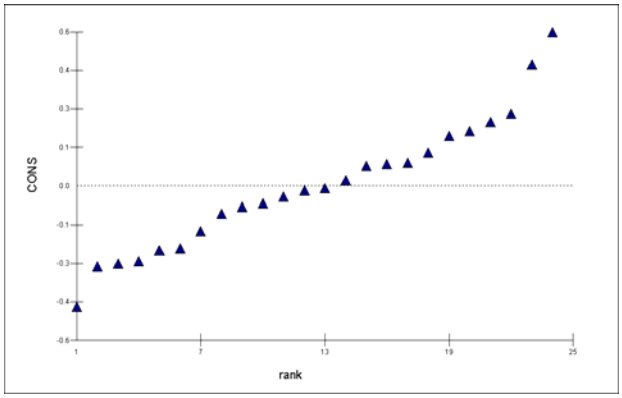
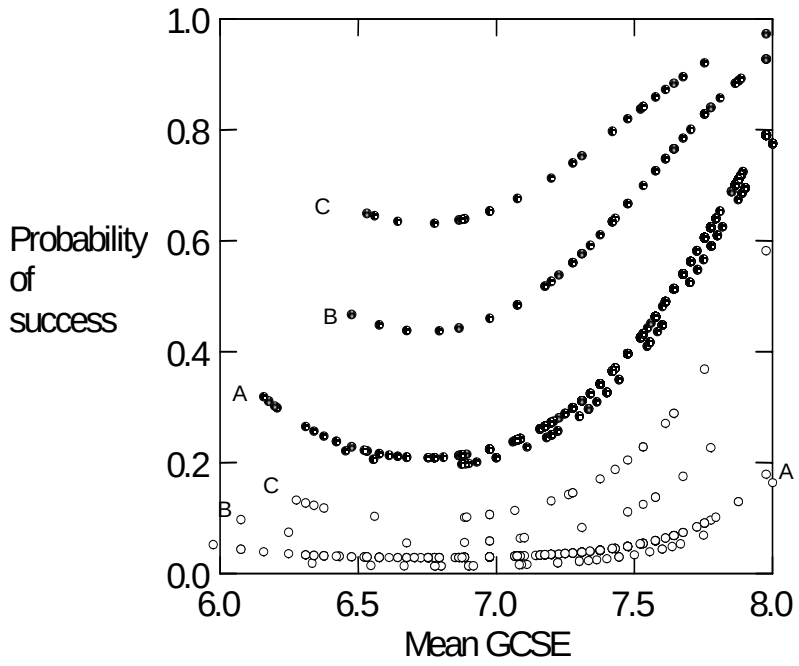


Figure 4: College Level Residuals

Interpreting the parameter estimates of model IV is difficult and it is easier to understand the results if plots of the predicted probability are used. In Figure 5, the predicted probabilities for applicants to be offered a place in Cambridge in the categories of the best three A-levels being grades As or less (and low experience and non-outlying college) against mean GCSE are presented. The upper line with filled symbols represents the predicted probability of success for those applicants who obtained at least three grades As at A-level, while the lower line with unfilled symbols represents the predictions for applicants with less than three grades As (the majority having AAB). For applicants with less than three grade As, only those applicants with exceptional GCSEs (i.e. the majority at the top grade of A*) have a chance of successfully being offered a place. Model IV also includes the college effects for Colleges B and C. It is clear that for any given level of prior attainment the probability of a successful application is much greater for these colleges.



(B and C are the two outlier colleges and A is the result for all colleges).

Figure 5: Predicted probabilities of successful application including college effects

The differences between models III and IV suggest that the issue of college choice and school experience needs a more detailed analysis. Note that the 'pooling' process has not been

modelled in this analysis. This is the process used to ensure the best applicants get a place by exporting from oversubscribed colleges to undersubscribed colleges. (http://www.cam.ac.uk/cambuniv/undergrad/statistics/). Further analyses are necessary including the effect of the pools is necessary before a definitive explanation of the process that generate models III and IV is available. It is worth noting that apart from this, none of the other variables (sex, school type, social class, ethnic origin or region) used in this analysis proved to be significant except for those relating to the attainment of the applicants.

DISCUSSION

In the first example, it has been demonstrated that the existence of outliers can have serious consequences when fitting multilevel models and that a lot of time and trouble can be saved by carrying out some exploratory analyses (Bell, 2001). The second example illustrated that outliers are not necessarily a nuisance but they can be used in gaining an understanding of the processes under consideration.

In multilevel models, the existence of outliers in some groups in any given level of the hierarchy may change relationships within a group and so lead to spurious between-group differences. In addition, outliers at one level may change the relationships at other levels too. This means that an iterative process of fitting a range of models, removing outliers and then re-fitting these models may be necessary.

There are two important issues to be considered when dealing with outliers. Firstly, the selection of the final data set to be analysed must be documented. This ensures that the analyses can be reproduced if needed. Secondly, an understanding of the processes and issues under consideration is necessary. It is very useful to have a clear understanding of exactly what the processes being modelled are.

References

Barnett V. and Lewis T. (1994) *Outliers in statistical data* (3rd edition). New York: John Wiley.

Bell, J.F. (2001) Visualising multilevel models: the initial analysis of data. Third International Conference on Multilevel Analysis, Amsterdam, April. [On-line] UK; Available: http://www.ucles-red.cam.ac.uk Accessed: 26 July 2004

Bell, J.F. (2001) Patterns of subject uptake and examination entry 1984-1997. *Educational Studies*, 27(2), 201-219.

Bell, J.F. (2002) On the presentation of the results of multilevel analysis. Society for Multivariate Analysis in the Behavioural Sciences, Tilburg, The Netherlands, July. [On-line] UK; Available: http://www.ucles-red.cam.ac.uk Accessed: 26 July 2004

Bell, J.F., Malacova, E. and Shannon, M. (2003) *The changing pattern of A-level uptake.* Paper presented at the British Educational Research Association Conference, University of Edinburgh, Edinburgh. [On-line] UK; Available: http://www.ucles-red.cam.ac.uk Accessed: 26 July 2004

Browne, W. J. (2002) *MCMC estimation in MLwiN*. London, Institute of Education. [On-line] UK; Available: http://multilevel.ioe.ac.uk Accessed: 26 July 2004

Goldstein, H. (1995) *Multilevel Statistical Models*. London, Edward Arnold: New York, Wiley.

Langford, I. H. and Lewis, T. (1998) Outliers in multilevel models (with discussion). *Journal of the Royal Statistical Society, A, 161*, 121-160.

Snijders, T.A.B. and Bosker, R.J. (1999*)* *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.