Outlier detection in psychiatric epidemiology

MORVEN LEESE

INTRODUCTION

Outliers due to measurement or transcription errors are usually regarded as a nuisance in data analysis. Their detection is aimed at corrective action, either by eliminating them or by using methods that are robust to their presence. On the other hand, outliers can be useful and informative when they reflect real phenomena of interest. For example, an unusual cluster of observations on a particular patient might suggest a rare, hitherto unknown, syndrome. These two types of outlier will be familiar to all psychiatric epidemiologists, since problems of measurement and definition are common in psychiatry, compared to other branches of medicine. A third type, the outlier patient, practice or hospital, as used in medical audit, is less well known in psychiatry. Statistical methods for detecting outliers of this third type do not present fundamentally new problems. However, their identification may be linked to the construction of «league» tables, and this is an area in which new approaches are being developed, in particular Bayesian methods. This short review describes the range of techniques available for detecting outliers and how the league table approach is being viewed by statisticians and others.

A classic book on the subject of outliers, by Barnett & Lewis (1984), describes many methods, concentrating on the univariate case, and provides relevant tables. The simple definition stated by these authors is that an outlier is *an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*. In addition to the technical details of the tests, Barnett & Lewis discuss the general philosophy behind the tests. For example they distinguish between a contaminant which arises from a different population to that of interest and a genuine but extreme observation. A further distinction is between simple recording errors which result in out-of-range values (whose treatment is non-statistical) and extreme values caused by statistical factors. They also point out that outliers must always be defined in relation to a particular population, because the probability model underlying the main population affects whether a particular observation is regarded as an outlier. Thus a skewed distribution such as the log-normal may produce apparently discordant values which are in fact valid observations from the upper tails.

Deciding what to do with outliers once they have been detected is another important issue to be considered. Most authors recommend that each case should be considered on its merits rather than invoking a blanket rule, for example omitting any values falling outside ± standard deviations from the mean. In principle one might be able to extend one's statistical model to encompass such values but in practice this is rarely practical because the mechanism by which they were generated is usually unknown. The use of robust methods may be appropriate for data summaries, but for a more complex analysis omitting the outliers may be the only practical approach. In this case methods appropriate to missing-values might be used for analysing the remaining data.

UNIVARIATE TECHNIQUES

Genuine but unusual characteristics in an individual subject often occur together as clusters of variables and would be detected using multivariate methods. Some of these are discussed below. Univariate outliers on the other hand are often, although not

Indirizzo per la corrispondenza: Professor M. Leese, Community Psychiatry (PRiSM) and Special Hospital Psychiatry, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF (UK).
Fax +44 - 171- 277.1462.
E-mail: m.leese@iop.bpmf.ac.uk

https://doi.org/10.1017/S1121189X0000498X Published online by Cambridge University Press

always, the result of measurement error. If so they would usually be sought in initial data cleaning rather than as part of a substantive analysis. Basic techniques for univariate outlier detection are fairly straightforward and are built into most statistical software packages. SPSS (SPSS, 1995) for example, indicates potential outliers using very simple criteria (by printing out multiples of the interquartile range, and the highest and lowest few points). STATA (StataCorp, 1997) includes a «letter-value» display (Tukey, 1977) which indicates the proportions of the sample in tail areas of decreasing size starting at 1/2 (median), quartiles, 1/8 and so on up to 1/1024.

These are ad hoc guides but more formal approaches are also available. One such method, which is relatively easy to perform by hand, so long as the relevant tables are available, is based on a measure of discordancy defined as the «excess/range» statistic, i.e. the difference between the extreme point and the next most extreme as a proportion of the whole range (Dixon, 1953). A sequential or block-wise approach can also be adopted using similar «Dixon-type» statistics, where the most extreme outliers can be tested in presence of other potential outliers at either end of the distribution (the phenomenon known as «masking»). There are also versions tailored to specific distributions (such as exponential or gamma).

These by-hand calculations are thus somewhat more flexible than the routines available in standard software. A further drawback of many software routines is that they assume (without being very specific) that the data are normal. Before applying a specific test using a standard package, therefore, the distribution of the data needs to be considered and transformation (for example log-transformation) might be necessary. Of course in a particular case, especially with a small sample, distinguishing between a log-normal distribution and a normal distribution with one or two extreme points might be difficult, and prior knowledge about plausible distributions would be useful in this respect.

MULTIVARIATE OUTLIERS

Outliers in a multivariate context are much more problematic. Not only do they share with univariate outliers all the problems of interpretation mentioned above but they usually require a more extensive analysis. Separate univariate analyses on each dimension, such as plotting histograms of the marginal distributions, would not be sufficient. An individual might have discordancies on many dimensions, each of which was too small to detect, yet whose sum was extreme. Although a total scale might reveal such outliers, a simple analysis such as this could be potentially misleading where the individual measures were correlated. In such a case any outliers might appear in dimensions comprising weighted combinations of variables, possibly with both positive and negative weights. This type of multivariate outlier is sometimes detectable by plotting the last few dimensions in a dimension-reducing technique such as principal components analysis (Joliffe & Morgan, 1992).

A further aspect of outliers which is of particular relevance in the multivariate case is that of «influence». Outliers in multivariate data may or may not have high impact on the statistical model in question depending on the dimension in which they appear. This point has been extensively studied in regression, where the concept of «leverage», as applied to an outlier residual from the regression model, can be used to measure the impact of a particular observation, for example in a combined measure such as Cook's distance (Cook, 1977). Influence statistics are also available for logistic regression (for example in Pregibon, 1981), and most software packages have a number of such statistics as part of regression diagnostics.

The visual approach based on principal components mentioned above has the advantage that it indicates the particular dimension(s) on which the outlier is manifest. An alternative approach, based on a single test statistic such as Hotelling's T_2, is less informative in this respect but is perhaps more convenient in that it provides a specific criterion for defining outliers. It does, however, require that the probability distribution of the statistic is known or can be estimated by simulation methods. Plotting the ordered values against the expected order statistics of the distribution is often useful here: in these plots the outlier(s) appear at the top or bottom ends of the plot, offset from the line through the main body of points. Gnanadesikan (1977) describes some of these plotting techniques and also considers the influence multivariate outliers can exert on a correlation matrix.

The problem of masking was mentioned in connection with univariate methods. As usual the cause of this (the presence of more than one outlier) is more difficult to detect and allow for in multivariate da-

https://doi.org/10.1017/S1121189X0000498X Published online by Cambridge University Press

ta. Although multiple outliers can be sought sequentially using recursive methods, this is not an ideal solution. However, a recent method (Hadi, 1994) has been proposed, based on a type of distance-based cluster analysis. It is suitable for reasonably large samples with at least 3p +1 cases, where p is the number of variables. A similar approach has been proposed by Hadi & Simonoff (1993) for detecting outliers in regression. These methods, which are highly computer intensive, have been incorporated in the software package STATA.

Three particular papers, which describe practical examples illustrating some of the points mentioned above, can be mentioned at this point. Liu & Weng (1991) discuss a number of the issues concerned with outlier detection in the context of a specific complex model in which there are both subjects, and also observations on subjects, to consider. Outlying subjects are detected using a sequential procedure based on Hotellings T^2, with simulation used to estimate p-values associated with the T^2 statistics; residuals from the model are used to identify outlying observations. Bacon (1995) discusses the detection of outliers in correlation matrices, a topic of importance in areas such as factor analysis and reliability assessment. The performance of a number of different distance measures distance are compared (Mahalanobis D^2), Comrey D (Comrey, 1985), a measure of correlational distance, and a new maximum likelihood measure proposed by the author. A third area concerns meta-analysis, where the outliers are parameter estimates from individual studies rather than subjects or observations. Outliers are clearly important here and are perhaps relatively likely to occur, given the lack of central control over data. Huffcutt & Arthur (1995) have developed a «sample-adjusted meta-analytic deviancy statistic» which reflects the influence of the particular study on the overall conclusions, taking account of sample size.

LEAGUE TABLES

The concept of an outlier hospital, practice or patient in medical audit and actuarial studies has recently become an issue, especially in the USA. For example, in that country hospitals may be especially compensated for any state-sponsored patients with particularly high costs; the latter clearly have to be identified according to some agreed rule. In many countries hospitals and consultants may be ranked by civil servants or consumer groups according to performance scores, and the extremes (best and worst) identified. While such financial or political motivations for identifying outliers are currently mainly confined to general hospitals, and use «hard» criteria such as mortality rates, the possibility that psychiatric service providers and consumers might eventually be ranked in this way should be considered. It is natural for the general public and officials to interpret the extremes as outliers, even if their scores are not particularly far removed from the main body of data.

There has been some criticism of this «league table» approach on general grounds, such as the difficulty of fairly adjusting for risk factors, and the concentration on individuals rather than on general improvements in health (see for example Hofer & Hayward, 1996). There are also statistical criticisms of the simplistic ranking methods which are often used; such criticisms have also been voiced in other fields, for example in UK educational statistics, where league tables are also common. The attribution to chance of a particular position is rarely admitted by officials but can be demonstrated by statistical methods. Bayesian analysis can be used to throw light on this, and a discussion of the Bayesian analysis of league tables is given by Goldstein & Spiegelhalter (1996). A recent letter by Langford (1997) shows an example in which such methods were used to find mean ranks with 95% confidence intervals. Here, the hospital ranked first out of 17 (in mortality rates) has mean empirical Bayes rank 2.8, with 95% confidence interval (1-9). The comparison between the estimated mean rank (2.8) with the rank (1) actually observed in one particular year shows that its relative position could well be improved on a future occasion, and the width of the confidence interval is an indication of the level of uncertainty. (In fact instability is often observed in practice, as the ordering of league tables typically changes from year to year.)

SUMMARY

Outliers may be of interest in their own right or they may merely be distractions from the point in question, and a hindrance to generalisation. In the medical field, including psychiatry, concise summary statistics and parsimonious models have tended

https://doi.org/10.1017/S1121189X0000498X Published online by Cambridge University Press

to be the main aim of many studies. However, partly motivated by the requirements of medical audit and other political and financial considerations, the detection of outliers as an end in itself is becoming a subject of interest in the public domain, and this may extend from medicine in general to psychiatry. The recognition of uncertainty in ranks using statistical methods can place the labels «best» and «worst», when applied to hospitals and consultants, into perspective, and here new developments in Bayesian methods will be important. Other areas of current development include computer intensive methods for multiple, multivariate outliers and for outlier detection tailored to specific situations such as correlational models in factor analysis and reliability studies, and in meta-analysis. These areas are likely to be of particular interest to psychiatric epidemiologists because of the complex nature of their data.

REFERENCES

Bacon D.R. (1995). A maximum likelihood approach to correlational outlier identification. *Multivariate Behavioral Research* 20, 125-148.

Barnett V. & Lewis T. (1984). *Outliers in Statistical Data.* Wiley: New York.

Comrey A.L. (1985). A method for removing outliers to improve factor analytic results. *Multivariate Behavioral Research* 20, 273-281.

Cook D.R. (1977). Detection of influential observations in linear regression. *Technometrics* 19, 15-18.

Dixon W.J. (1953), Processing data for outliers. *Biometrics* 9, 74-89.

Gnanadesikan R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* Wiley: New York.

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

Hadi A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society*, Series B 56, 393-396.

Hadi A.S. & Simonoff J.S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association* 88, 1264 -1272.

Hofer T.P. & Hayward R.A. (1996). Identifying poor quality hospitals. Can mortality rates detect quality problems for medical diagnosis? *Medical Care* 34, 737-753.

Huffcutt A.I. & Arthur W. (1995). Development of a new outlier statistic for meta-analytic data. *Journal of Applied Psychology* 327-334.

Joliffe I.T. & Morgan B.J.T. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research* 1, 69-95.

Langford I.H. (1997). Bayesian analysis should be used instead of league tables of performance. *Letter BMJ* 314, 73-74.

Liu J-P. & Weng C-S. (1991). Detection of outlying data in bioavailability/bioequivalence studies. *Statistics in Medicine* 10, 1375-1389.

Pregibon D. (1981). Logistic regression diagnostics. *Annals of Statistics* 9, 705-724.

StataCorp (1997). *Stata Statistical Software: Release 5.0.* Stata Corporation: College Station, Texas.

SPSS (1995). *SPSS for Windows Release 7.0.* SPSS Inc.: Chicago.

Tukey J.W. (1977). *Exploratory Data Analysis.* Addison Wesley Publishing Company: Reading, MA.

https://doi.org/10.1017/S1121189X0000498X Published online by Cambridge University Press