Forward Search

Introduction

The forward search is a powerful robust statistical method for exploring the relationship between data and fitted models. It is a development of the methods described in the articles on **Residuals** and **Diagnostics** that aids the discovery of clusters of observations and previously unidentified important subsets of the data as well as revealing any groups of outliers.

In this article, we give examples of the use of the forward search for **regression** and **generalized linear models**. These applications, together with the material on **transformations** of data covered in the article **Fan Plot** and the extension to **nonlinear regression**, are described by Atkinson and Riani [1]. We also give an example involving **multivariate** data, a topic extensively covered in [3].

Regression and Residuals

The forward search orders the observations by closeness to the assumed model, starting from a small subset of the data and increasing the number of observations m used for fitting the model. **Outliers** and small unidentified subsets of observations enter at the end of the search.

We write the multiple regression model as

$$ym = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, $\boldsymbol{\beta}$ is the $p \times 1$ vector of parameters, and it is assumed that the additive errors of observation ε are independently distributed with constant variance σ^2. Also in (1), $\boldsymbol{X}$ is the $n \times p$ **matrix** of carriers, that is, of **explanatory variables** and perhaps functions of them, such as quadratics and interaction terms.

It is helpful to list the various stages of the forward search.

1. Notation. The vector of p parameters $\boldsymbol{\beta}$ is estimated by **least squares** applied to subsets of the observations. For an arbitrary subset of m observations, the estimate is denoted $\hat{\boldsymbol{\beta}}(m)$. For a subset $S^*(m)$ of size m chosen by the forward search, the estimate is written $\hat{\boldsymbol{\beta}}(m^*)$.

2. Starting the Search. The search starts from a small subset of size m_0; usually $m_0 = p$ or perhaps $p + 1$. To find the starting subset $S^*(m_0)$, we randomly select 1000 subsamples of size m_0. The initial subset $S^*(m_0)$ provides the least **median** of squares estimator $\hat{\boldsymbol{\beta}}(m_0^*)$, that is, it minimizes the median squared residual (Rousseeuw [5]) of the observations over the 1000 samples.

3. Moving Forward in the Search. When the m observations constituting $S^*(m)$ are used in fitting, the fitted values from the estimate $\hat{\boldsymbol{\beta}}(m^*)$ yield n least-squares residuals $\boldsymbol{e}(m^*)$. We order the squared residuals $\boldsymbol{e}^2(m^*)$ and take the observations corresponding to the $m + 1$ smallest as the new subset $S^*(m + 1)$. Usually, this process augments the subset by one observation, but sometimes two or more observations enter as one or more leave. This may also happen at the beginning of the search, where $S^*(m_0)$ is chosen to minimize the median squared residual, not to find the subset yielding the $m_0 + 1$ smallest squared residuals. Because of this very robust starting point and the form of the search, outliers, if any, tend to enter as m approaches n.

4. Monitoring the Search. If any quantity is of interest when it is calculated for the complete set of n observations, we can monitor its evolution during the forward search. In our example, we first look at a forward plot of the residuals $\boldsymbol{e}(m^*)$, scaled by the final estimate of σ. Examples of forward plots of other quantities of interest, such as estimates of the parameters $\boldsymbol{\beta}$ and σ^2 are given by Atkinson and Riani.

The analysis of the data on mandible length [6] in the article on **Goodness of Fit** using simple regression shows appreciable evidence of nonnormality of the residuals. The normal plot of the least-squares residuals in Figure 3 of the article "Goodness of Fit" shows three large negative residuals and two further residuals that are also rather large.

This structure is apparent in the forward plot of the residuals in Figure 1. Units 165, 166, and 149 have large negative residuals throughout the search. Units 146 and 167 also have appreciable negative residuals for much of the search. Working backwards, the last units to join the search are these five, in order 165, 166, 149, 146, and 167. These are the five negative residuals visible in the $Q-Q$ plot in **Goodness of Fit**, which is of the unscaled version of the residuals at the end of the search in Figure 1. The forward search shows that, in this example, the residual plot,

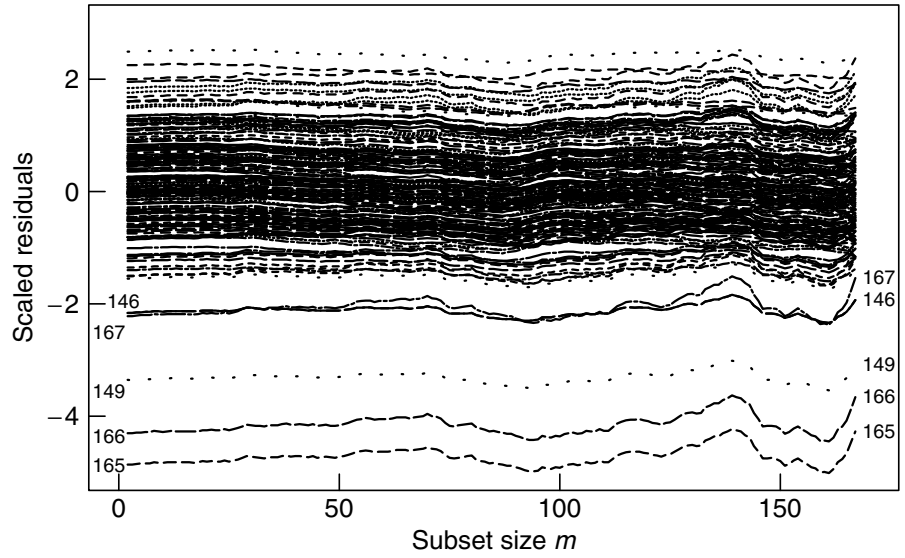

Figure 1 Mandible length data, first-order model: forward plot of scaled residuals. There are five large negative residuals for much of the search, but those for units 146 and 167 are masked at the end of the search

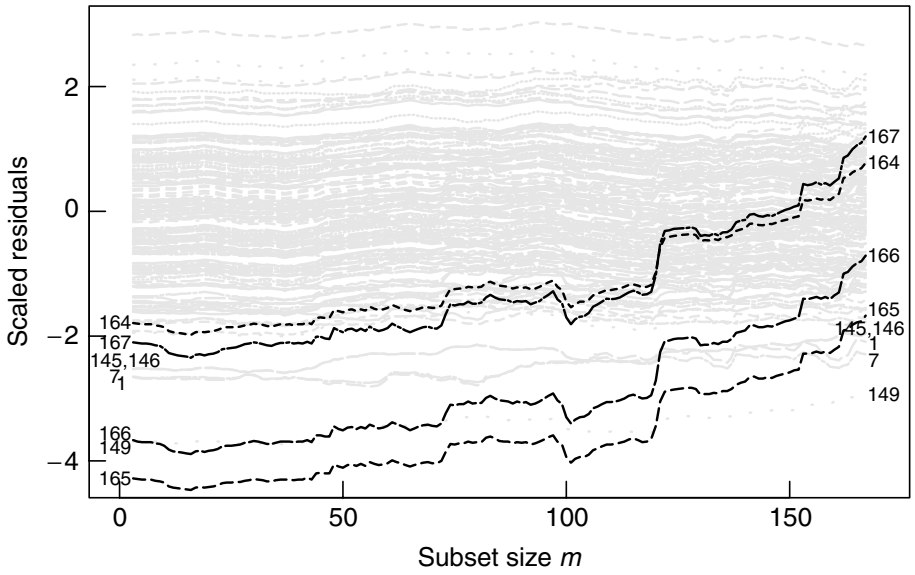

Figure 2 Mandible length data, second-order model, logged response: forward plot of scaled residuals. Four units, 164 to 167, behave differently from the rest, which have an approximately normal distribution

when all observations are fitted, identifies most of the structure of the residuals. The values of $\boldsymbol{x}$ and $\boldsymbol{y}$ for these units is clear from the scatter plot of Figure 5 of the article on **Diagnostics**. If the last three units, which are shown as open circles or crosses in the plot, are excluded, the straight line fitted to the data becomes such that unit 167 has an appreciable negative residual. As the forward plot shows, this residual is reduced when the last three units enter the subset. There is therefore some masking of the outlying nature of this unit.

The units with large residuals identified in this analysis are not all of those plotted with open circles in Figure 5 of the article **Diagnostics**. One reason is that these were identified as being influential observations, rather than having large residuals. A second reason is that that analysis was for a logged response with a second-order model. Figure 2 shows the forward plot of the residuals from this model.

Four units are highlighted in Figure 2. If we ignore them, the forward plot of the residuals is virtually symmetrical throughout the search, with no other

appreciable outliers. The most negative residuals are those for units 149, 7, 1, 145, and 146. But these values do not change much during the search and, as the $Q-Q$ plot in Figure 1 of the article on Diagnostics shows, these are not particularly extreme values when compared with **order statistics** from a normal distribution. The four highlighted units in the figure are units 164 to 167. They are highlighted because their behavior is very different. Initially, they all have large negative residuals, but by the end of the search, the residuals are all appreciably smaller, two having become positive. These units are those for the four oldest fetuses. It seems as if the model toward the end of the search may be being altered by their presence and so produces small residuals. Certainly, this would not be surprising as such extreme points in $\boldsymbol{X}$ space will be leverage points, a property amplified by fitting a quadratic model. Figure 2 of the article on **Diagnostics** shows how extreme these leverage values are. A question we then have to consider is how the evidence for a quadratic model depends on these four units.

Forward Added Variable T Test

If the fitted model and data agree, the parameter estimates should be reasonably constant throughout the forward search. These estimates are **orthogonal** to the residuals used to order the entry of units into the subset $S^*(m)$. The same is not true of the estimate of σ^2, which, being the sum of squared residuals, increases during the search as increasingly outlying observations are included in the subset. As a result, the t tests (*see* **Student's t Distribution**) for the parameters in the linear model decrease dramatically during the forward search. We describe here an alternative form of search that provides information on the inferential effect of the units on the estimated linear model.

If the standard regression model (1) is rewritten as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{Q\theta} + \boldsymbol{w}\gamma + \boldsymbol{\varepsilon}, \qquad (2)$$

$\boldsymbol{Q}$ is the $n \times p-1$ matrix of carriers obtained by deleting the column $\boldsymbol{w}$ from $\boldsymbol{X}$. At the end of the search, the t test for the column of $\boldsymbol{X}$ corresponding to $\boldsymbol{w}$ from **multiple regression** on $\boldsymbol{X}$ is identically the added variable test described immediately after equation (11) in the article on **Residuals**. This is found by first regressing $\boldsymbol{y}$ and $\boldsymbol{w}$ on $\boldsymbol{Q}$ and then testing the regression through the origin of the resulting residuals of $\boldsymbol{y}$ on those of $\boldsymbol{w}$.

We adapt the added variable test to the forward search by dropping each column of $\boldsymbol{X}$ in turn to create $p-1$ vectors $\boldsymbol{w}$. We then use regression on each $\boldsymbol{Q}$ to provide a forward search from which $\boldsymbol{w}$ is excluded. We monitor the behavior of the added variable test for each $\boldsymbol{w}$, thus obtaining $p-1$ plots of t statistics from $p-1$ different forward searches: $p-1$ because we are not usually interested in testing hypotheses about the value of the constant in the regression model. Because we exclude $\boldsymbol{w}$ from the search, the t test for $\boldsymbol{w}$ has the correct distribution and increases during the search rather than decreasing. The details are in [2].

We start, in Figure 3, with a forward plot of the added variable t test for regression of untransformed mandible length on **gestational age**. The plot shows a steady upward trend to a very significant value of 35.90. There is no sign of the importance of individual observations such as the units giving large residuals in Figure 1; evidence for the regression is spread throughout the data.

Figure 4, for regression of log $\boldsymbol{y}$ on a quadratic in age is similarly well behaved. The value of t_1, the t test for regression on age, rises steadily to 18.08, while that for t_2 for the quadratic term decreases to -12.55. The leverage points 164 to 167, which are such a notable feature of Figure 2, do not enter at the end of either of the added variable searches on which the plots in Figure 4 are based. The plot shows no evidence that these four units are responsible for the quadratic term in the model. Despite the appearance of Figure 2, the evidence of curvature

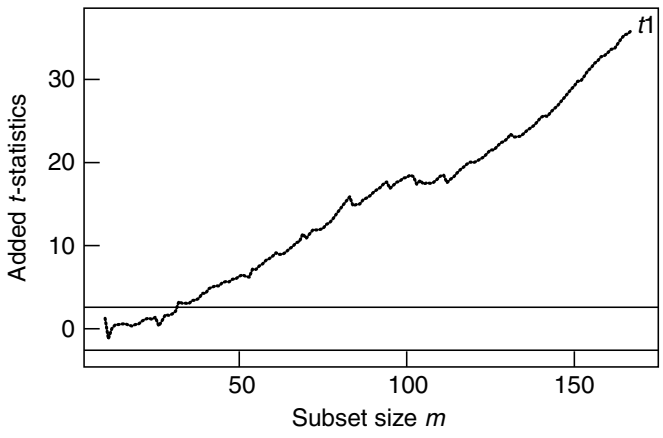

Figure 3 Mandible length data, first-order model: forward plot of added variable t test t_1 for regression on age. Evidence for the regression is spread throughout the data

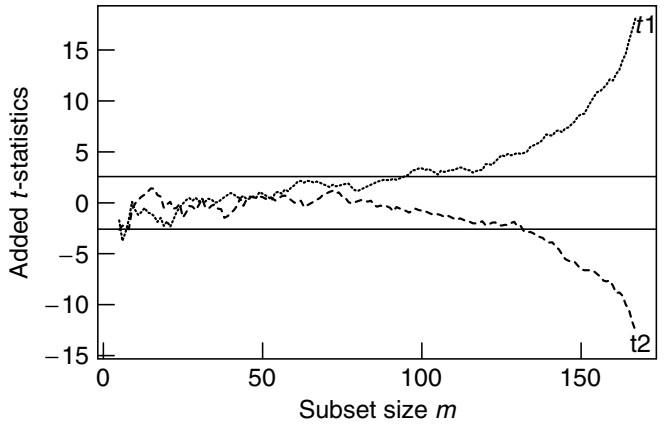

Figure 4 Mandible length data, second-order model, logged response: forward plot of added variable t tests t_1 and t_2 for regression on age and its square. Evidence for the regression is again spread throughout the data

in the relationship with a logged response is spread throughout the data.

Our analysis thus shows that taking a logged response combined with a quadratic model produces residuals, which have an approximately normal distribution, with four leverage points, the residuals for which change appreciably during the search. These four units are not influential for the choice of terms in the linear model. However, they might be influential for the choice of the transformation. But the forward plot of the test for transformation in Figure 5 of the article on the **Fan Plot** shows that this is not the case. Thus, these procedures provide no evidence for the suggestion mentioned by Royston and Altman that the fetuses with an age greater than 28 weeks were different from the younger ones.

Generalized Linear Models

The structure provided by the theory of generalized linear models allows us to apply the forward search to, particularly, **gamma**, **Poisson**, and **binomial** data in a manner analogous to that used for multiple linear regression. Chapter 6 of Atkinson and Riani [1] contains theory and examples.

In generalized linear models, we have a response $\boldsymbol{y}$, a vector of linear predictors with elements $\boldsymbol{\eta} = \boldsymbol{x}^T\boldsymbol{\beta}$, and a link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ connecting the two. In the article on **Residuals**, the deviance D, the analogue of the residual sum of squares in regression, was written as

$$D = \sum_{i=1}^{n} d_i^2, \tag{3}$$

where d_i^2 is the contribution of the ith unit to the total deviance. The deviance residual was then defined as

$$r_{Di} = d_i \text{ sign } (y_i - \hat{\mu}_i). \tag{4}$$

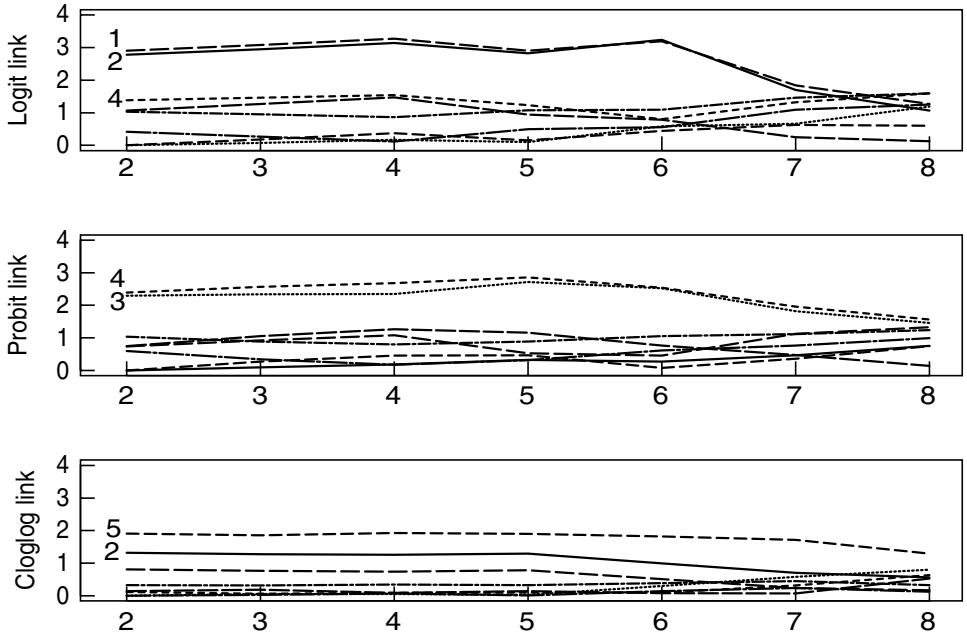

Figure 5 Bliss's beetle data: absolute values of deviance residuals as the subset size increases: (a) logit, (b) probit and (c) complementary log-log links

To extend the forward search to generalized linear models, we replace the least-squares residuals e_i with the deviance residuals r_{Di}. Then, as before, when m observations are used in fitting, the optimum subset $S^*(m)$ yields n deviance residuals $\boldsymbol{r}_D(m^*)$. We order the squared residuals $\boldsymbol{r}_D^2(m^*)$ and take the observations corresponding to the $m+1$ smallest as the new subset $S^*(m+1)$.

For the regression models in the previous sections, we looked at forward plots of residuals and of t tests for components of the linear predictor. As well as problems about individual outliers and the correct form of the linear predictor, there is also a need in generalized linear models to specify the correct form of link function. In the articles on **Goodness of Fit** and **Residuals**, analyses are given of Bliss's beetle data. These are binomial data in which the probability of success θ_i at dose level x_i is modeled by the link function $g(\theta_i) = \eta_i$. The analysis used the **logistic** link

$$g(\theta) = \log \frac{\theta}{1-\theta}. \tag{5}$$

There was evidence that this link was not satisfactory for these data. Alternative links are the probit

$$g(\theta) = \Phi^{-1}(\theta), \tag{6}$$

where Φ is the cdf of the standard normal distribution, and the complementary log–log link

$$g(\theta) = \log\{-\log(1-\theta)\}. \tag{7}$$

(*see* **Quantal Response Models**).

We explore these three possible link functions by looking at forward plots of absolute deviance residuals, which will indicate whether the unsatisfactory nature of the logistic link was caused by a few outliers or whether there is a systematic lack of fit. Figure 5 shows plots of absolute deviance residuals from forward searches for three models in which the explanatory variable is log(dose) and the three links are the logit, probit and complementary log–log. The observations are numbered from the lowest dose level to the highest. For the logit link observations, 1 and 2 are the last two to be included in the forward search. The crossing of the lines at the end of the plot in the top panel of Figure 5 shows that the inclusion of observations 1 and 2 seems noticeably to affect the ordering of the residuals. With the probit link units 3 and 4 (the last two to be included) seem to be different from the rest of the data: they are badly predicted by models in which they are not included. However, the residuals from the forward search with the complementary log–log link in the bottom panel of the figure show no such behavior; all residuals are smaller than two throughout, and relatively constant. Since the scale parameter is not estimated, it is possible to make such absolute comparisons of the residuals across different models, even if they come from different link families.

The conclusion from Figure 5 is that the complementary log–log link is satisfactory and that the other two are not. This conclusion is not dependent on a few observations, but is spread throughout the data. To sharpen and quantify this general impression based on forward plots of residuals, we now consider the goodness of link test, introduced in the article on **Goodness of Fit**. This provides a test for the adequacy of each link from the t test for the inclusion of the constructed variable $\hat{\boldsymbol{\eta}}^2$ in the linear predictor. The constructed variable plot in Figure 8 of the article on **Goodness of Fit** indicates rejection of the logistic link when all observations are used in fitting. We use forward plots of the test statistics to test three links and to see whether the conclusions are based on all observations.

Figure 6 shows a forward plot of the goodness of link test, the order of introduction of the observations, as in Figure 5, being different for the three links. For the logit and probit links, these plots show evidence of lack of fit at the 5% level, which is indicated by the statistic going outside the bounds in the plot. Although, it is inclusion of the last two observations

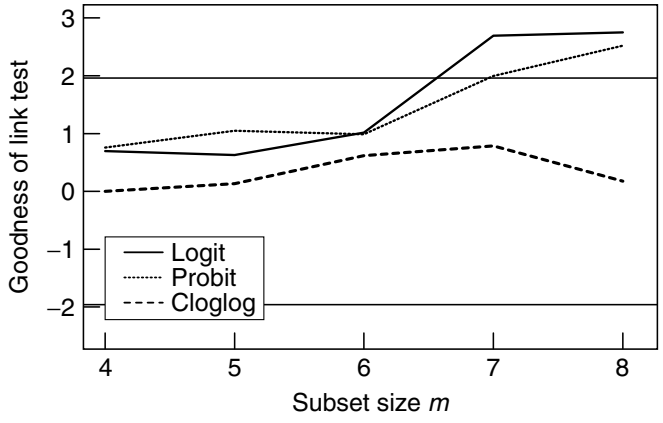

Figure 6 Bliss's beetle data: forward plot of the goodness of link test. Only the complementary log-log link is satisfactory

that causes the values of the statistic to become significant, it is clear from the steady upward trend of the plots that lack of fit is due to all observations. The plot for the complementary log–log link shows no evidence of any departure from this model. This plot also shows that unit 5, which is the one with the biggest residual for the complementary log–log link and the last to be included in this forward search, has no effect on the t value for the goodness of link test.

This analysis shows that, of the three links considered, only the complementary log–log link is satisfactory. The plot of fitted values for the logistic link in Figure 6 of the article on **Residuals** relates this finding to individual observations. The fitted dose response curve for this symmetrical link fits badly in the center of the experimental region, whereas, as Figure 6.36 of Atkinson and Riani [1] shows, the asymmetric complementary log–log link provides an appreciably better fit over the whole range of $\boldsymbol{x}$ values.

Multivariate Data

With multivariate observations, we replace the squared residuals used in the forward search for regression and generalized linear models with the squared **Mahalanobis distances**

$$d_i^2(m^*) = \{\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}(m^*)\}^T \hat{\boldsymbol{\Sigma}}^{-1}(m^*)\{\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}(m^*)\}, \tag{8}$$

where $\hat{\boldsymbol{\mu}}(m^*)$ and $\hat{\boldsymbol{\Sigma}}(m^*)$ are estimates of the mean and **covariance matrix** of the observations based on the subset $S^*(m)$. These distances are used for ordering the observations and for determining how we move forward in the search. We use the robust bivariate boxplots of Zani et al. [7] to determine an initial subset, which is not outlying in any two-dimensional plot of the data. The content of the contours is adjusted to give an initial subset of the required size. Once we have some idea of the structure of the data, we start the search with subsets that seem potentially interesting.

As an example with some expected and some unexpected structure, we look at readings on six dimensions of 200 Swiss bank notes, 100 of which may be genuine, and 100 forged. All notes have been withdrawn from circulation and classified by an expert, so some of the notes in either group may have been misclassified. Also, the forged notes may not form a homogeneous group. For example, there may be more than one forger at work. The data, and a reproduction of the bank note, are given by Flury and Riedwyl [4, pp. 4–8].

Figure 7 is a forward plot of Mahalanobis distances scaled by the estimate of $\boldsymbol{\Sigma}$ at the end of the search. The search starts with 20 observations on notes believed genuine. In the first part of the search, up to $m = 93$, the observations seem to fall into two groups. One has small distances and is composed of observations within or shortly to join the subset. Above these there are some outliers and then, higher still, a concentrated band of outliers, all of which are behaving similarly. The plot clearly shows the difference between the genuine notes and the forgeries. Toward the end of the search, there is evidence that the group of forgeries is not homogeneous.

The structure of the group of forgeries is also readily revealed by the forward search. Figure 8 is a forward plot of the scaled Mahalanobis distances just for the forgeries. In the center of the plot, around $m = 70$, this shows a clear structure of a central group, one outlier from that group and a second group of 15 outliers. As successive units from this cluster enter after $m = 85$, they become less remote and the distances decrease. By the end of the search there is appreciable masking, so that the group of 15 observations is no longer clear from the plot of the Mahalanobis distances. Under such

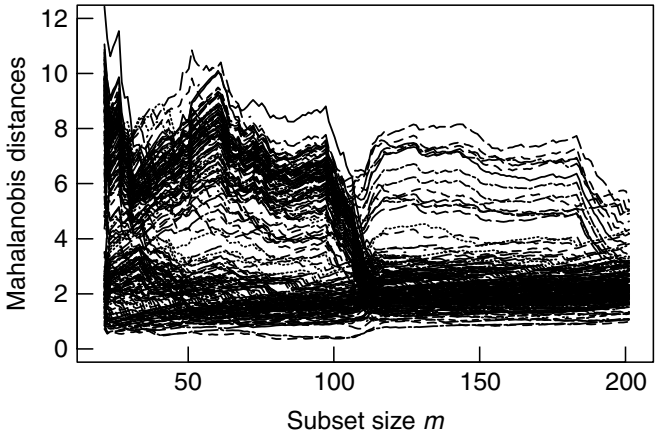

Figure 7 Swiss Banknote Data, all 200 observations: forward plot of scaled Mahalanobis distances starting with 20 notes believed to be genuine. The two groups are clear, but a third group seems to appear toward the end of the search

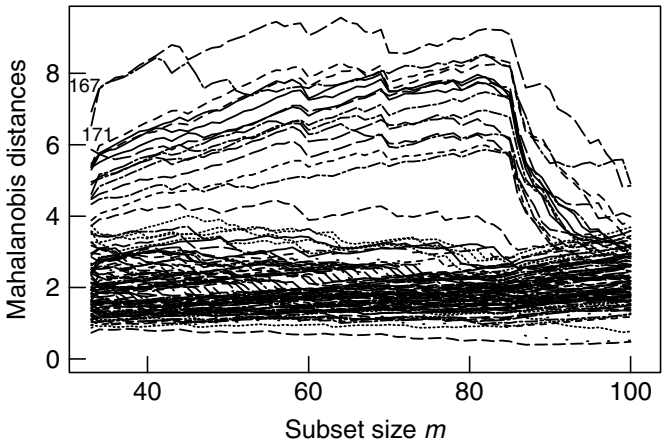

Figure 8 Swiss Banknote Data, 100 notes classified as forgeries: forward plot of scaled Mahalanobis distances. Toward the end of the search, there seems to be a group of 15 observations and a further single outlier

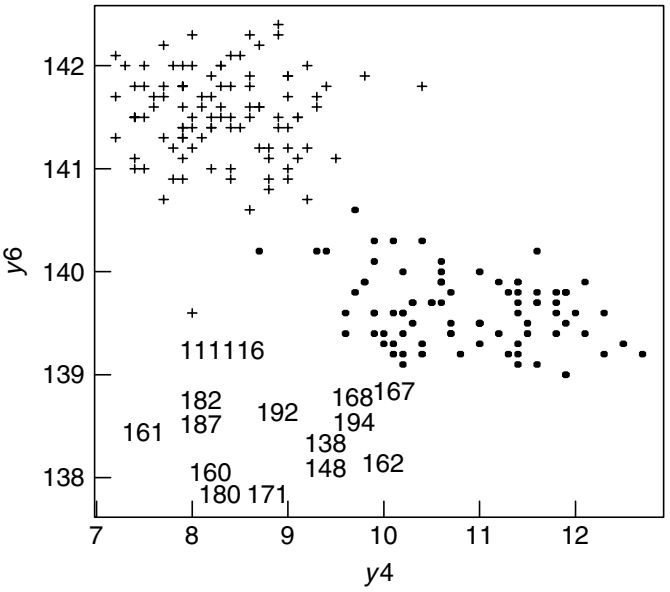

Figure 9 Swiss Banknote Data: scatterplot of y_6 against y_4. The "genuine" notes are marked with crosses; the labeled units are the last 15 to enter the search

conditions, the deletion methods described in the article on **Diagnostics** are likely to fail to reveal the structure.

In this example, the forward search clearly indicates not only the presence of two groups of notes, but also an unexpected subset of 15 observations, showing that the group of forgeries is not homogeneous but consists of two subgroups. Once attention has been drawn to the existence of this structure, it is possible to find it in the data. Figure 9 is one of the 15 different panels of the scatterplot matrix for these six dimensional data and by far the most revealing. The last 15 observations to enter the subset are numbered: the other forgeries are shown by filled circles and the "genuine" notes by crosses. It seems that one genuine note has been misclassified.

The entries in this article show various ways in which the forward search can elucidate the structure of data and, in the case of the third example, reveal unexpected subsets. A fuller analysis of the data on Swiss banknotes, together with numerous other applications of the forward search to multivariate data, are described in [3].

References

[1] Atkinson, A.C. & Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer-Verlag, New York.

[2] Atkinson, A.C. & Riani, M. (2002). Forward search added variable t tests and the effect of masked outliers on model selection, *Biometrika* **89**, 939–946.

[3] Atkinson, A.C., Riani, M. & Cerioli, A. (2003). *The Forward Search in Multivariate Data Analysis*. Springer–Verlag, New York. (In preparation).

[4] Flury, B. & Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach*. Chapman and Hall, London.

[5] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association* **79**, 871–880.

[6] Royston, P.J. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion), *Applied Statistics* **43**, 429–467.

[7] Zani, S., Riani, M. & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis* **28**, 257–270.

(*See also* **Model Checking**; **Model, Choice of**)

A.C. ATKINSON & MARCO RIANI