Robust and Efficient Multivariate Calibration

M. Riani[1] and S. Salini[2]

[1] University of Parma, via Kennedy 6, 43100 Parma, Italy
[2] University of Milan, via Conservatorio 7, 20122 Milano, Italy

Keywords: Forward search, LMS, masked outliers, calibration.

1 Introduction

Multivariate calibration uses an estimated relationship between a multivariate response Y (of dimension q) and an explanatory vector X (of dimension p) to predict unknown X in future from further observed responses. If the prediction sample is inconsistent with the calibration data, it is a prediction outlier (Martens and Naes, 1989). One of the main problems in multivariate calibration is the detection of multiple outliers causing the so-called masking effect (several outliers can interact in a complicated way to strengthen or to cancel each other's influence).

A single outlier can easily be detected by the methods of deletion diagnostics in which one observation at a time is deleted, followed by the calculation of new parameter estimates and residuals. With two outliers, pairs of observations can be deleted and the process can be extended to the deletion of several observations at a time. This is the basic idea of multiple deletion diagnostics. A difficulty both for computation and interpretation is the explosion of the number of combinations to be considered. A similar approach is based on the repeated application of single deletion methods (backward methods). However, such backwards procedures can fail due to masking. An alternative is to employ robust procedures. Up to now very little has been written about robust calibration (Kistos and Müller 1995; Riu and Rius 1995; Cheng and Van Ness, 1997) though considerable has been written about robust regression. In addition, although robust estimators can sometimes reveal the structure of the data, they do so at the cost of downweighting or discarding some observations. Finally, if the calibration experiment is made up of different subsets, the use of robust estimators will tend to produce a centroid which lies in between different groups. In this last case prediction will be strongly determined by the size of the subsets which make up the calibration experiment. The purpose of this paper is to present multivariate calibration methods which are able to detect and investigate those observations which differ from the bulk of the data or, more generally, to identify subgroups of observations. We are concerned not only with the identification of atypical observations, but also with the effect that they have on parameter estimates, on inferences about models, and on their suitability. In this paper particular attention will be paid to the "forward search" approach (Atkinson and Riani, 2000; Atkinson, Riani and Cerioli, 2004). In this method we start with a fit to very few outlier free observations and then successively fit larger subsets. We thus order the observations by closeness to the fitted model. As a result, not only are outliers and distinct subsets of the data discovered, but the influential effect of these observations is made clear.

2 Some examples

As an illustration of the suggested approach we apply the forward search to a data set coming from a certification laboratory (in this example $q = 2$ and $n = 152$). The experiment is performed in order to determine the strength of concrete blocks. Calibration problem arises since the standard method to determine exactly the grade strength is destructive. The alternative method gives $q > 1$ measures obtained thought a sclerometer. The initial subset of dimension r say $S^{(r)}$ to initialize the forward search is found using the intersection of units having the smallest LMS squared residuals

considering each of the two responses independently. In symbols for each response j, $S^{(p)}_{\mathbf{c}^*,j}$ satisfies

$$e^2_{[\text{med}],S^{(p)}_{\mathbf{c}^*,j}} = \min_{\mathbf{c}}[e^2_{[\text{med}],S^{(p)}_{\mathbf{c},j}}], \tag{1}$$

where $e^2_{[k],S^{(p)}_{\mathbf{c},j}}$ is the k th ordered squared residual among $e^2_{i,S^{(p)}_{\mathbf{c},j}}$, in the regression which considers the j-th variable as response, $i = 1, \ldots, n$, $\mathbf{c}$ is a collection of p units (the number of $\mathbf{c}$ collections is $\binom{n}{p}$) and med is the integer part of $(n+p+1)/2$. The initial subset is associated with the k units whose residuals at maximum have the r-th position ($r \leq n/2$) among $e^2_{[1],S^{(p)}_{\mathbf{c}^*,j}}, \ldots, e^2_{[n],S^{(p)}_{\mathbf{c}^*,j}}$, $j = 1, 2, \ldots, q$. The search progresses from subset size m to $m+1$ by selecting the smallest $(m+1)$ scaled Mahalanobis distances (MD) (Atkinson, Riani and Cerioli, 2004) from multivariate regression. This algorithm proceeds up to when all units are included in the subset ($m = k, k+1, \ldots, n$). Figure 1 shows the monitoring of MD during all steps of the forward search. This plot clearly shows that the sample is made of different subgroups of data. Note that this feature is completely masked both at the beginning (when we use highly robust estimators) and at the end of the search when we use MD based on maximum likelihood estimators. Only the monitoring of MD in different steps enables to see the existence of subgroups of observations. In the initial steps of the search and from step $m = 55$ the centroid which is used to estimate the covariance matrix lies in between different groups preventing us to discover the real structure of the data set. The groups of units highlighted by the forward search correspond to blocks of concrete that have different composition and are produced from different fixtures.

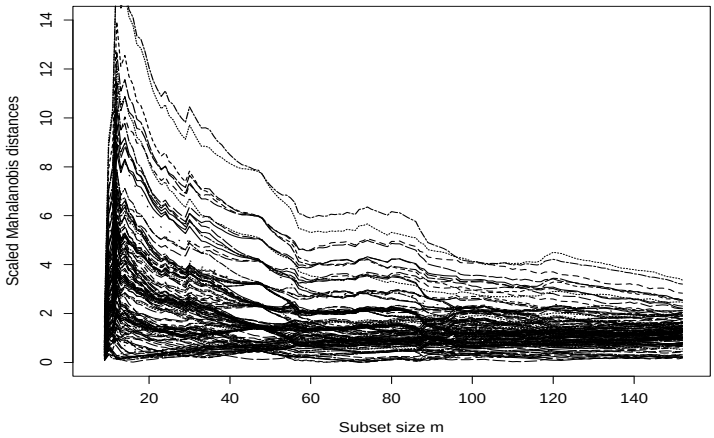

FIGURE 1. Concrete Data: forward plot of scaled Mahalanobis distances based on residuals of multivariate regression

References

Atkinson A.C. and M. Riani (2000). *Robust Regression Diagnostics*, Springer Verlag, New York.

Atkinson A.C, M. Riani and A. Cerioli (2004). *Exploring Multivariate Data With the Forward Search*, Springer Verlag, New York.

Cheng C.L. and J.W. Van Ness (1997). Robust Calibration, *Technometrics* 39, 401–411.

Kistos M.L. and C.H. Müller (1995). Robust Linear Calibration, *Statistics*, 27, 93–106.

Martens H. and T. Naes (1989). *Multivariate Calibration.* Wiley & Sons, New York.

Riu J. and F.X. Rius (1995). Univariate Regression Models With Errors in Both Axes, *Journal of Chemometrics* 9, 343–362.