Introducing Prior Information into the Forward Search for Regression

Un approaccio Bayesiano alla Forward Search

Anthony C. Atkinson and Aldo Corbellini

Abstract The forward search provides a flexible and informative form of robust regression. We describe two ways of introducing prior information into the regression model used in the search, either through fictitious observations or through prior distributions of the parameters. The relationship between the two methods is established. The extension to the forward search is not entirely straightforward, requiring weighted regression. Forward plots are used to exhibit the effect of prior information on inferences.
Abstract L'obiettivo del paper è mostrare come si può incorporare informazione a priori all'interno dell'algoritmo di forward search ed illustrare la performance di questo nuovo approccio.

Key words: consistency factor, fictitious observation, robust regression, weighted regression

1 Introduction

Methods of robust regression have been described in several books, for example Rousseeuw and Leroy (1987), Atkinson and Riani (2000) and Maronna *et al.* (2006). The recent comparisons of Riani *et al.* (2014b) indicate the superior performance of the forward search (FS). However, none of these methods includes prior information; they can all be thought of as developments of least squares. The purpose of the present paper is to show how prior information can be incorporated into FS for

Anthony C. Atkinson
Department of Statistics, London School of Economics, London WC2A 2AE, UK, e-mail: a.c.atkinson@lse.ac.uk

Aldo Corbellini
Dipartimento di Economia, Università di Parma, Italy e-mail: aldo.corbellini@unipr.it

regression and to give some results indicating the comparative performance of this Bayesian method.

In order to detect outliers and departures from the fitted regression model, FS uses least squares to fit the model to subsets of m observations. The initial subset of m_0 observations is chosen robustly, for example by least trimmed squares. The subset is increased from size m to $m+1$ by forming the new subset from the observations with the $m+1$ smallest squared residuals. For each m $(m_0 \leq m \leq n-1)$, we test for the presence of outliers, using the observation outside the subset with the smallest absolute deletion residual. The algebraic details of this procedure are described in §2. In §3 we introduce prior information into this forward procedure in the form of fictitious observations. The alternative form through the specification of prior distributions of parameters is introduced in §4 and related to statistics derived from the fictitious observations. The results of the FS are typically presented through a forward plot of quantities of interest as a function of m. In §5 we use such plots to elucidate the change in properties of the search with variation of the amount of prior information.

2 Algebra for the Forward Search

2.1 Tests in the Linear Model

In the regression model $y = X\beta + \varepsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T, and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ε_i are i.i.d. $N(0, \sigma^2)$.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, where $H = X(X^T X)^{-1} X^T$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2 = e^T e/(n-p) = \sum_{i=1}^n e_i^2/(n-p)$.

FS fits subsets of observations of size m to the data, with $m_0 \leq m < n$. Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X(m)$. Least squares on this subset of observations yields parameter estimates $\hat{\beta}(m)$ and $s^2(m)$, the mean square estimate of σ^2 on $m-p$ degrees of freedom. Residuals can be calculated for all observations including those not in $S^*(m)$. The n resulting least squares residuals are

$$e_i(m) = y_i - x_i^T \hat{\beta}(m). \tag{1}$$

The search moves forward with the augmented subset $S^*(m+1)$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$. To start we take $m_0 = p$ and search over subsets of p observations to find the subset that yields the LMS estimate of β. However, this initial estimator is not important, provided masking is broken.

To test for outliers the deletion residual is calculated for the $n-m$ observations not in $S^*(m)$. These residuals, which form the maximum likelihood tests for the

outlyingness of individual observations, are

$$r_i(m) = \frac{y_i - x_i^T \hat{\beta}(m)}{\sqrt{s^2(m)\{1+h_i(m)\}}} = \frac{e_i(m)}{\sqrt{s^2(m)\{1+h_i(m)\}}}, \tag{2}$$

where the leverage $h_i(m) = x_i^T \{X(m)^T X(m)\}^{-1} x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min}$ where

$$i_{\min} = \arg \min_{i \notin S^*(m)} |r_i(m)|.$$

To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual, namely $r_{imin}(m)$, as a test statistic. If the absolute value is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^*(m)$.

2.2 Estimation of the Variance

In order to test for outliers we need a reference distribution for $r_i(m)$ in (2). If we estimated σ^2 from all n observations, the statistics would have a t distribution on $n-p$ degrees of freedom. However, in the search we select the central m out of n observations to provide the estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, let the variance of the symmetrically truncated normal distribution containing the central m/n portion of the full distribution be $\sigma_T^2(m)$. See Riani *et al.* (2009) for a derivation from the general method of Tallis (1963). We take as our approximately unbiased estimate of variance $s_T^2 = s^2(m)/\sigma_T^2 = s^2(m)/c(m,n)$. In the robustness literature $c(m,n)$ is called a consistency factor (Riani *et al.*, 2014a).

3 Prior Information from Fictitious Observations

In some of the applications in which we are interested, for example fraud detection (Perrotta and Torti, 2010) we have appreciable prior information about the values of the parameters. This can sometimes be thought of as coming from n_0 fictitious observations y_0 with matrix of explanatory variables X_0. Then the data consist of the n_0 fictitious observations plus n actual observations. The search in this case now proceeds from $m = 0$, when the fictitious observations provide the parameter values for all n residuals from the data. The search then continues as outlined above but with the fictitious observations always included in those used for fitting, their residuals being ignored in the selection of successive subsets.

There is one complication in this procedure. The n_0 fictitious observations are treated as a sample with variance σ^2. However, the m observations from the ac-

tual data are again from a truncated distribution of m out of n observations and so asymptotically have a variance $c(m,n)\sigma^2$, which must be adjusted before the two samples are combined. This becomes a standard problem in weighted least squares (for example, Rao 1973, p. 230). Let y^+ be the $(n_0+m)\times 1$ vector of responses from the fictitious observations and the subset and let the covariance matrix of these observations be $\sigma^2 G$, with G a diagonal matrix. Then the first n_0 elements of the diagonal of G equal one and the last m elements have the value $c(m,n)$. In the least squares calculations we need only to multiply the elements of the sample values of y and X by $c(m,n)^{-1/2}$.

4 Prior Distribution of Parameters

Chaloner and Brant (1988), as part of a study of Bayesian methods for outlier detection, specify prior information for a linear regression model in terms of parameter values. For comparability between the two approaches, we describe these prior parameter values in terms of the fictitious observations of the previous section.

Let $\tau = 1/\sigma^2$. The prior distribution of τ is gamma, that is $p(\tau) \propto \tau^{(a-1)}e^{-b\tau}$. The mean is a/b and the variance a/b^2. To find a and b, we consider the distribution of s_0^2, the customary mean square estimator of σ^2 from the n_0 observations. The degrees of freedom are $\nu = n_0 - p$. Let $S = \nu s_0^2$, that is the residual SS, distributed as $\sigma^2\chi^2_\nu$. Then it follows that $\text{var}\,(s_0^2) = 2\sigma^4/\nu$. We require $\text{var}\,\hat{\tau}$, where $\hat{\tau} = 1/s_0^2$. We use the standard Taylor series expansion, $\text{var}\,g(x) = \{g'(x)\}^2\text{var}(x)$, with $g(x) = 1/x$, $x = s_0^2 = 1/\hat{\tau}$, so that $\text{var}(\hat{\tau}) = \text{var}(x)/x^4 = 2\tau^2/\nu$. Then the relationships for the mean and variance of the gamma distribution of τ yield

$$a = \nu/2 = (n_0 - p)/2 \quad \text{and} \quad b = \nu/(2\hat{\tau}) = \nu s_0^2/2 = S/2.$$

We now use the prior information for the linear model to obtain an expression for S.

Prior information for the linear model is given as

$$R = X_0^T X_0 \quad \text{and} \quad \hat{\beta}_0 = R^{-1}X_0^T y_0,$$

so that

$$S = y_0^T y_0 - \hat{\beta}_0^T R\hat{\beta}_0.$$

5 Examples

To explore the properties of FS including prior information, we use simulation to provide forward plots of the distribution of quantities of interest during the search. In these simulations we kept the prior values fixed, rather than repeatedly simulating them using fictitious observations. Figure 1 shows plots of the estimate of σ^2. In the left-hand panel the prior information is overwhelming, with $n_0 = 1,000$, whereas

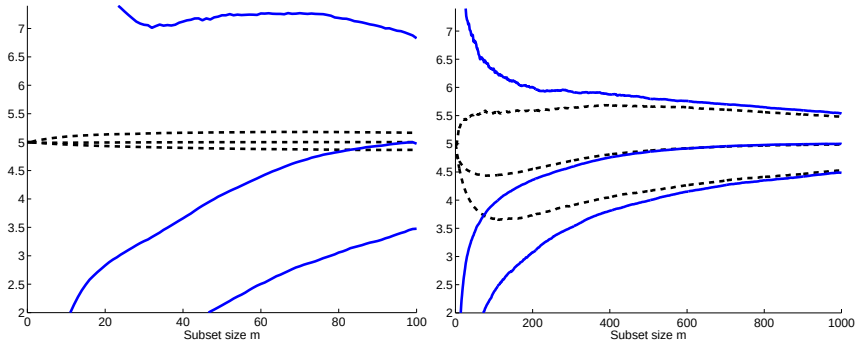

Fig. 1 Forward plots of estimates of σ^2. Left-hand panel, strong prior information ($n_o = 1,000; n = 100$). Right-hand panel, weak prior information ($n_o = 100; n = 1,000$). 10,000 simulations; 1%, 50% and 99% empirical quantiles. Dotted lines, with prior information; heavy line, without prior information

n is only 100. The distribution of the estimate is shown by dotted lines; it is little affected by the sample information. The continuous lines in the figure are for the estimate of σ^2 that uses only sample information. Although, by the end of the search, the median estimate is close to the population value of five, there is great variability, particularly at the start of the search, where the skewed nature of the scaled chi-squared distribution of the estimate is evident.

In the right-hand panel the prior information is appreciably reduced, with $n_0 = 100$ and $n = 1,000$. The estimate using prior information is always less variable than that based solely on sample information, although they have virtually converged by the end of the search. For small values of m the estimate including prior information is much more precise, although the distribution is slightly skewed, reflecting the asymptotic nature of the action of the correction factor.

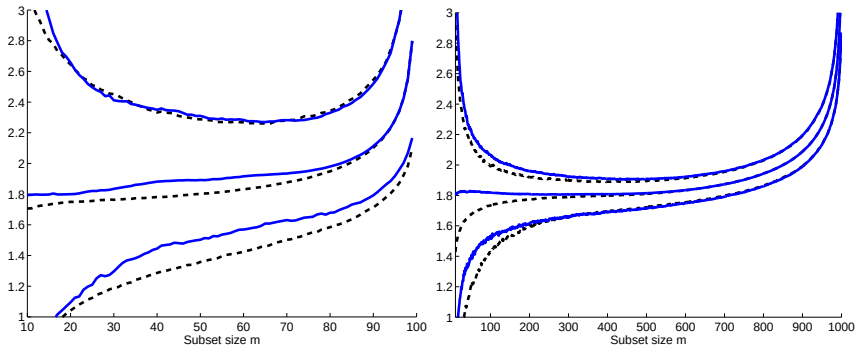

Fig. 2 Forward plots of minimum deletion residuals. Left-hand panel, strong prior information ($n_o = 1,000; n = 100$). Right-hand panel, weak prior information ($n_o = 100; n = 1,000$). 10,000 simulations; 1%, 50% and 99% empirical quantiles. Dotted lines, with prior information; heavy line, without prior information

The forward plots of minimum deletion residuals form the central tool for outlier testing in the FS. Outliers are detected when the curve for the sample values falls outside a specified envelope. The actual rule for detection of an outlier has to take account of the multiple testing inherent in the FS (once for each value of m). One rule, yielding powerful tests of the correct size is given by Riani *et al.* (2009). The left-hand panel of Figure 2 shows the envelopes for strong prior information. The upper envelopes for procedures with and without prior information agree. For the 1% and 50% quantiles the values of the statistics in the absence of prior information are higher than those in its presence, reflecting the increased prevalence of smaller estimates of σ^2, division by which give larger values of the statistics. In the right-hand panel, with weaker prior information, the two sets of curves agree for all except sample sizes around 100 or less. However, it is not the agreement in distribution of the statistics that is important, since the envelopes apply to different statistics, but rather the importance is the increase in power of the tests that comes from including prior information. We look forward to demonstrating this feature in our future research.

References

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

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–659.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester.

Perrotta, D. and Torti, F. (2010). Detecting price outliers in European trade data with the the forward search. In F. Palumbo, C. Lauro, and M. Greenacre, editors, *Data Analysis and Classification*. Springer-Verlag, Heidelberg.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications, 2nd edition*. Wiley, New York.

Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

Riani, M., Cerioli, A., and Torti, F. (2014a). On consistency factors and efficiency of robust S-estimators. *TEST*. (In press).

Riani, M., Atkinson, A. C., and Perrotta, D. (2014b). A parametric framework for the comparison of methods of very robust regression. *Statistical Science*, **29**. (In press).

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, **34**, 940–944.