Introducing Prior Information into the Forward Search for Regression

Anthony C. Atkinson, Aldo Corbellini and Marco Riani

Abstract

The forward search provides a flexible and informative form of robust regression. We describe the introduction of prior information into the regression model used in the search through the device of fictitious observations. The extension to the forward search is not entirely straightforward, requiring weighted regression. Forward plots are used to exhibit the effect of correct and incorrect prior information on inferences.

1 Introduction

Methods of robust regression have been described in several books, for example [2,6,14]. The recent comparisons of [12] indicate the superior performance of the forward search (FS) in a wide range of conditions. 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

A.C. Atkinson (✉)
Department of Statistics, London School of Economics, London, UK
e-mail: a.c.atkinson@lse.ac.uk

A. Corbellini · M. Riani
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: aldo.corbellini@unipr.it

M. Riani
e-mail: mriani@unipr.it

T. Di Battista et al. (eds.), *Topics on Methodological and Applied Statistical Inference*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-319-44093-4_1

incorporated into FS for 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 in the absence of prior information, the FS uses least squares to fit the model to subsets of m observations, starting from an initial subset of m_0 observations. 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 specification of prior information and its incorporation into the FS is derived in Sect. 2. Section 3 presents the algebraic details of outlier detection with prior information. Forward plots in Sect. 4 show the dependence of the evolution of parameter estimates on prior values of the parameters. In the rest of the paper the emphasis is on forward plots of minimum deletion residuals which form the basis for outlier detection. These plots are presented in Sect. 4 for correctly specified priors and, in Sect. 4, for incorrect specifications. It is argued that use of analytically derivable frequentist envelopes is also suitable for Bayesian outlier detection when the priors are correctly specified. However, serious errors can occur with misspecified priors.

2 Prior Information in the Linear Model from Fictitious Observations

In the regression model without prior information $y = X\beta + \varepsilon$, 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. $N(0, \sigma^2)$.

In some of the applications in which we are interested, for example fraud detection [7], we have appreciable prior information about the values of the parameters. This can often conveniently 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 fictitious observations are always included in those used for fitting, their residuals being ignored in the selection of successive subsets.

There is one complication in combining this procedure with the forward search, which arises from the estimation of variance from subsets of observations. If we estimate σ^2 from all n observations, we obtain an unbiased estimate of σ^2 from the residual sum of squares. However, in the frequentist search we select the central m out of n observations to provide the mean square 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 [10] for a derivation from the general method of [15]. 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 [5,13].

In the Bayesian procedure, the n_0 fictitious observations are treated as a sample with variance σ^2. However, the m observations from the actual data come from a truncated distribution with 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, [9, 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 $c(m,n)^{-1/2}$. The residual mean square error from this weighted regression provides the estimate $\hat{\sigma}^2(m)$.

The prior information can also be specified in terms of prior distributions of the parameters β and σ^2. The details and relationship with fictitious observations are given by [4] as part of a study of Bayesian methods for outlier detection and by [3] in the context of the forward search.

3 Algebra for the Bayesian Forward Search

Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X(m)$. Weighted least squares on this subset of observations plus X_0 yields parameter estimates $\hat{\beta}(m)$ and $\hat{\sigma}^2(m)$, the latter on $n_0 + m - p$ degrees of freedom. Residuals can be calculated for all n observations including those not in $S^*(m)$. The n resulting least squares residuals are $e_i(m) = y_i - x_i^{\mathrm{T}}\hat{\beta}(m), \quad (i = 1, \dots, n)$.

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 $m_0 = 0$, since the prior information specifies the values of β and σ^2.

To test for outliers the deletion residuals are calculated for the $n - m$ observations not in $S^*(m)$. These residuals are

$$r_i(m) = e_i(m)/[\hat{\sigma}^2(m)\{1 + h_i(m)\}]^{0.5}, \tag{1}$$

where the leverage $h_i(m) = x_i^{\mathrm{T}}\{X_0^{\mathrm{T}}X_0 + X(m)^{\mathrm{T}}X(m)/c(m,n)\}^{-1}x_i$. Let the observation nearest to those forming $S^*(m)$ be $i_{\min} = \arg\min_{i \notin S^*(m)} |r_i(m)|$. To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{\mathrm{imin}}(m) = e_{\mathrm{imin}}(m)/[\hat{\sigma}^2(m)\{1 + h_{\mathrm{imin}}(m)\}]^{0.5}, \tag{2}$$

as a test statistic. If the absolute value of (2) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^*(m)$.

4 Example 1: Correct Prior Information

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. These simulations are intended to complement the analysis of [3] based on the Windsor housing data introduced by [1]. In these data there are 546 observations on regression data with four explanatory variables and an intercept, so that $p = 5$. Because of the invariance of least squares results to the values of the parameters in the regression model, we simulated the responses as independent standard normal variables with all regression coefficients equal to zero. The explanatory variables were likewise independent standard normal, simulated once for each set of simulations, as were the fictitious observations providing the prior. We took $n = 500$ in all simulations reported here and repeated the simulations 10,000 times.

Figure 1 shows forward plots of the parameter estimates when there is relatively weak prior information ($n_0 = 30$). Because of the symmetry of our simulations in the coefficients β_j, the left-hand panel arbitrarily shows the evolution of $\hat{\beta}_3$. From the simulations all other linear parameters give indistinguishable plots. The plot is centred around the simulation value of zero with quantiles that decrease steadily and smoothly with m. The right-hand panel is more surprising: the estimate of σ^2 decreases rapidly from the prior value of one, reaching a minimum value of 0.73 before gradually returning to one. The effect is due to the value of the asymptotic correction factor $c(m, n)$ which is too large. Further correction is needed in finite samples. Reference [8] use simulation to make such corrections in robust regression, but not for the FS.

The differing widths of bands in the two panels serve as a reminder of the comparative variability of estimates of variance. Reference [3] give the plot for stronger prior information when $n_0 = 500$. With equal amounts of prior and sample information at the end of the search, the bands for $\hat{\beta}_3$ are appreciably more horizontal than those of Fig. 1. However, the larger effect of increased prior information is in estimation

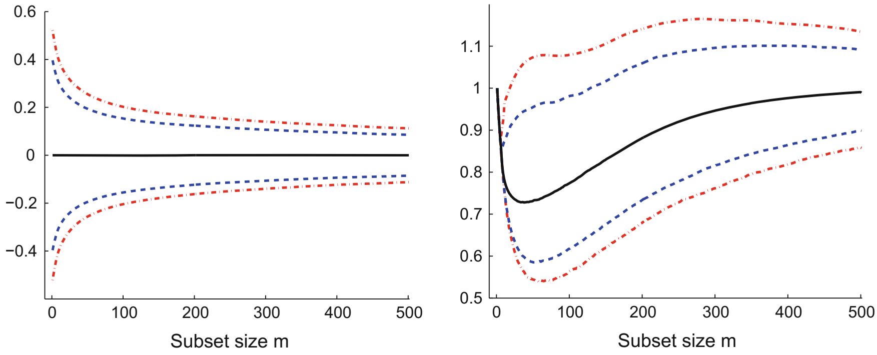

Fig. 1 Distribution of parameter estimates when $\beta_3 = 0$ and $\sigma^2 = 1$. *Left-hand panel* $\hat{\beta}_3$, *right-hand panel* $\hat{\sigma}^2$; weak prior information ($n_0 = 30$; $n = 500$). 1, 5, 50, 95 and 99 % empirical quantiles

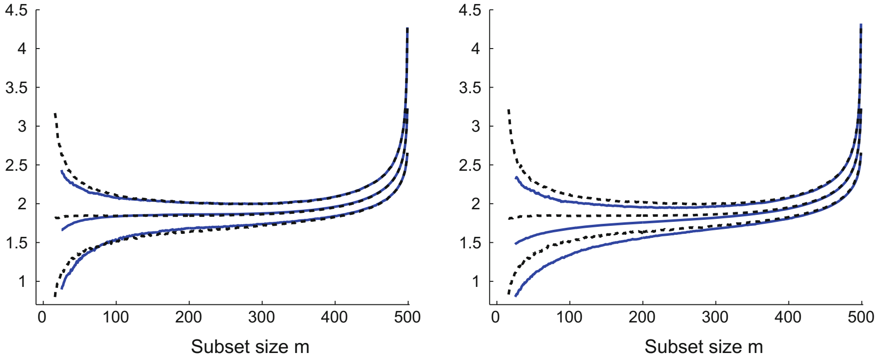

Fig. 2 The effect of correct prior information on forward plots of minimum deletion residuals. *Left-hand panel*, weak prior information ($n_0 = 30$; $n = 500$). *Right-hand panel*, strong prior information ($n_0 = 500$; $n = 500$), 10,000 simulations; 1, 50 and 99 % empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

of σ^2, which now has a minimum value of 0.97 and appreciably narrower bands for the quantiles.

The parameter estimates form an important component of the forward plots of minimum deletion residuals. The plots of these residuals, which are the focus of the rest of this paper, are the central tool for the detection of outliers 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 desired 1 % size, is given by [10] for multivariate data and by [11] for regression. The procedure has two stages, in the second of which envelopes are required for a series if values of n. The left-hand panel of Fig. 2 shows the envelopes for weak prior information ($n_0 = 30$), together with those from the FS in the absence of prior information. Unlike the Bayesian envelopes, those for the frequentist search are found by arguments based on the properties of order statistics. In this panel the frequentist and Bayesian envelopes agree for all except sample sizes around 100 or less. In the right-hand panel the prior information is stronger, with $n_0 = 500$. The upper envelopes for procedures with and without prior information agree for the second half of the search. 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 in the frequentist search. In general, the agreement in distribution of the statistics is not of central importance, since the envelopes apply to different situations. One important, although expected, outcome is the increase in power of the outlier tests that comes from including prior information, which is quantified by [3]. Also important is the agreement of frequentist and Bayesian envelopes towards the end of the search, which is where outlier detection usually occurs. This agreement allows us to use the frequentist envelopes when testing for outliers in the presence of prior information. Such envelopes can be calculated analytically, avoiding the time consuming simulations that are needed when envelopes for different values of n are required.

5 Example 2: Incorrect Prior Information

In the housing data analysed by [3], there is evidence of incorrect specification of the prior values of some parameters. The effect of misspecification of σ^2 is easily described; estimates of β remain unbiased, although with a changed variance compared with those when the specification is correct. The estimate of σ^2 also behaves in a smooth fashion; initially close to the prior value it moves steadily towards the sample value.

The effect of misspecification of β is more complicated since both $\hat{\beta}$ and $\hat{\sigma}^2$ are affected. There are two effects. The effect on $\hat{\beta}$ is to yield an estimate that moves from the prior value to the sample value in a sigmoid manner. Because of the biased nature of $\hat{\beta}$, the residual sum of squares is too large and $\hat{\sigma}^2$ rapidly moves away from its correct prior value. As sample evidence increases the estimate gradually stabilises and then moves towards the sample value. There are then two conflicting effects on the deletion residuals; an increase due to incorrect values of β and a reduction in the residuals due to overestimation of σ^2. Plots illustrating these effects on the parameter estimates are given by [3]. Here we show the effect of misspecification of β on envelopes like those of Fig. 2.

Our interpretation of Fig. 2 was that the frequentist envelopes could be used for outlier identification with little change of size or loss of power in the outlier test compared with use of the envelopes for the correctly specified prior. We focus on this aspect in interpreting the envelopes from an incorrectly specified prior.

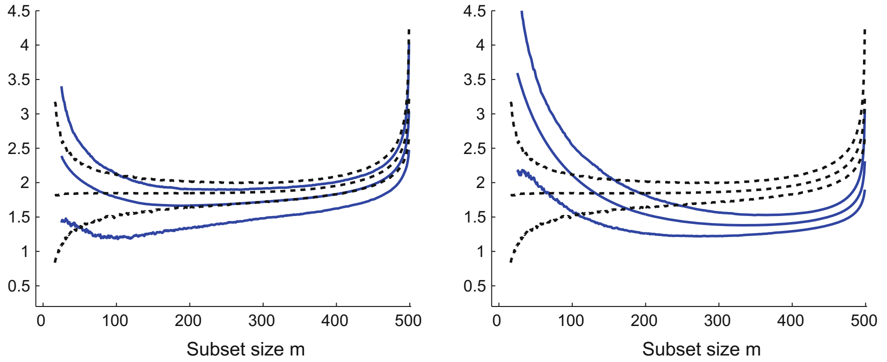

Fig. 3 The effect of incorrect prior information on forward plots of minimum deletion residuals; $\beta_0 = 1.5$. *Left-hand panel*, $n_0 = 6$, *right-hand panel*, $n_0 = 100$, 10,000 simulations; 1, 50 and 99 % empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

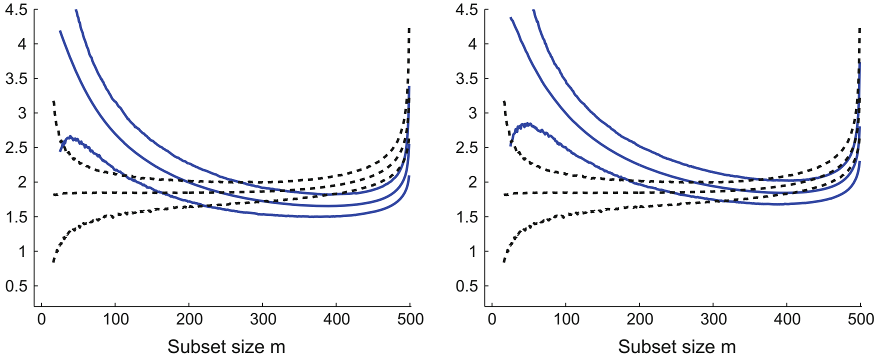

Fig. 4 The effect of increased incorrect prior information on forward plots of minimum deletion residuals; $\beta_0 = 1.5$. *Left-hand panel*, $n_0 = 250$, *right-hand panel*, $n_0 = 350$, 10,000 simulations; 1, 50 and 99 % empirical quantiles. *Dashed lines*, without prior information; *heavy lines*, with prior information

In the simulations all values of β were incremented by 1.5. In the left-hand panel of Fig. 3 we take $n_0 = 6$. Initially the envelopes lie above the frequentist bands, with a longer lower tail. Interest in outlier detection is in the latter half of the envelopes, for which the true envelopes lie below the frequentist ones; the residuals tend to be smaller and outliers would be less likely to be detected even at the very end of the search. In the right-hand panel, n_0 has been increased to 100. The result is to increase the size of the residuals at the beginning of the search. However, in the second half, the correct envelopes for this prior lie well below the frequentist envelopes; although outliers would be even less likely to be detected than before, the series of residuals lying well below the envelope would suggest a mismatch between prior and data.

Figure 4 shows two further forward plots of envelopes of minimum deletion residuals but now with greater prior information. In the left-hand panel $n_0 = 250$ and in the right-hand panel the value is 350. The trend follows that first seen in the right-hand panel of Fig. 3. In the first half of the search the envelopes continue to rise above the frequentist bands—very large residuals are likely at this early stage, which will provide a signal of prior misspecification. However, now, the envelopes for the right-hand halves of the searches are coming closer together. Particularly for $n_0 = 350$, there are unlikely to be a large number of residuals lying below the frequentist bands, although outliers will still have residuals that are less evident than they would be using the correct envelope.

This discussion suggests that forward plots of deletion residuals can provide one way of detecting a misspecification of the prior distribution. Similar runs of too small residuals can also be a sign of other model misspecification; they can occur, for example, in the frequentist analysis of data with beta distributed errors under the assumption of normal errors. The analysis of the housing data presented by [3] provides examples of the effect of prior misspecification on forward plots of minimum deletion residuals.

References

1. Anglin, P., Gençay, R.: Semiparametric estimation of a hedonic price function. J. Appl. Econ. **11**, 633–648 (1996)
2. Atkinson, A.C., Riani, M.: Robust Diagnostic Regression Analysis. Springer, New York (2000)
3. Atkinson, A.C., Corbellini, A., Riani, M.: Robust Bayesian regression. Submitted (2016)
4. Chaloner, K., Brant, R.: A Bayesian approach to outlier detection and residual analysis. Biometrika **75**, 651–659 (1998)
5. Johansen, S., Nielsen, B.: Analysis of the Forward Search using some new results for martingales and empirical processes. Bernoulli **22** (2016, in press)
6. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, Chichester (2006)
7. Perrotta, D., Torti, F.: Detecting price outliers in European trade data with the forward search. In: Palumbo, F., Lauro, C.N., Greenacre, M.J. (eds.) Data Analysis and Classification. Springer, Heidelberg (2010)
8. Pison, G., Van Aelst, S., Willems, G.: Small sample corrections for LTS and MCD. Metrika **55**, 111–123 (2002)
9. Rao, C.R.: Linear Statistical Inference and its Applications, 2nd edn. Wiley, New York (1973)
10. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc., Ser. B **71**, 447–466 (2009)
11. Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D.: Monitoring robust regression. Electron. J. Stat. **8**, 646–677 (2014)
12. Riani, M., Atkinson, A.C., Perrotta, D.: A parametric framework for the comparison of methods of very robust regression. Stat. Sci. **29**, 128–143 (2014)
13. Riani, M., Cerioli, A., Torti, F.: On consistency factors and efficiency of robust S-estimators. TEST **23**, 356–387 (2014)
14. Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection. Wiley, New York (1987)
15. Tallis, G.M.: Elliptical and radial truncation in normal samples. Ann. Math. Stat. **34**, 940–944 (1963)