Econometric Applications of the Forward Search in Regression: Robustness, Diagnostics and Graphics

Anthony C. Atkinson
The London School of Economics, London WC2A 2AE, UK
a.c.atkinson@lse.ac.uk
and Marco Riani
Dipartimento di Economia, Università di Parma, Italy
mriani@unipr.it

October 12, 2004

Abstract

The paper illustrates the use of the forward search to provide a robust analysis of econometric data. The emphasis is on informative plots that reveal the inferential importance of each observation.

Keywords: Bad leverage point; fan plot; leverage; LTS; outliers; residuals; very robust methods

1 Introduction

Regression is one of the major statistical tools in applied economics and ordinary least squares (OLS) is the preferred method of estimation. Zaman, Rousseeuw, and Orhan (2001) report that few publications in economics apply robust methods in regression despite the well-known susceptibility of OLS to outliers and the difficulty of detecting outlying observations from the study of residuals. They accordingly advocate the use of high-breakdown robust regression, specifically LTS (least trimmed squares). In three examples they compare OLS and LTS parameter estimates and associated t-values. It is the purpose of the present paper to support the advocacy by Zaman et al. (2001) of the use of robust methods in econometrics. We supplement their analyses by use of the forward search (Atkinson and Riani...
The forward search starts from a very robust fit, either LTS or Least Median of Squares (LMS) (Rousseeuw 1984), which give very similar results as starting points. The initial fit is to the same number of observations, $p$, as there are parameters in the model. The search then steadily increments the number of observations used in the fit until all $n$ observations are fitted using OLS. In this way the parameter estimates move steadily from very robust to the highly efficient OLS. Plots of important quantities, such as residuals, parameter estimates, $t$-statistics and diagnostic tests during the search reveal not only which observations are outlying, but also the inferential effect of each observation; outliers may, or may not, effect the conclusions drawn from the data.

2 Example 1: Growth Study of De Long and Summers

The first example analysed by Zaman et al. (2001) is a regression example from De Long and Summers (1991) on data from 61 countries over the years 1960 to 1985. The aim was to investigate the relationship between equipment investment and economic growth over the period. The variables were:

$y$: GDP growth per worker  
$x_1$: labour force growth  
$x_2$: relative GDP gap  
$x_3$: equipment investment  
$x_4$: non-equipment investment.

De Long and Summers (1991) find a strong relationship between $y$ and $x_3$. The LTS analysis of Zaman et al. (2001) shows that one country (Zambia, observation 60) is “a very significant outlier”. Omission of this observation followed by OLS analysis of the remaining 60 observations shows that all variables including the constant increase in significance, only $x_1$ remaining below the 5% point.

With $p = 5$ (four variables and a constant) the number of subsets of size 5 is $61!/ (5!56!) = 5,949,147$. We use LMS and start our forward search with complete evaluation of all subsets. Although this is unnecessarily thorough, the procedure takes less than five minutes.

Figure 1(a) shows the forward plot of the residuals scaled by the estimates of $\sigma$ at the end of the search. Observation 60 stands out clearly as an outlier. Figure 1(b) is the forward plot of leverages. In general they decline with $m$ as the average value is $p/m$. An interesting feature is that unit 5 (Botswana) enters the subset when $m = 51$ with high leverage, which then declines.
We now turn to quantities of inferential interest. Figure 2(a) is a forward plot of the parameter estimates, except the non-significant $\tilde{\beta}_1$, scaled by their OLS values. We see that immediately before the end of the search all the values are greater than 1. The introduction of unit 60 at the end of the search causes the parameter estimates to decrease in magnitude as do the $t$ statistics calculated by deleting variables (Atkinson and Riani 2002) shown in the forward plot of Figure 2(b). This plot shows very clearly the way in which exclusion of unit 60 increases the significance of all three variables. We do not plot the intercept term.

The outlying unit 60 also has an effect on the estimate of $\sigma^2$ shown in Figure 3(a), the increase contributing in part to the decrease in $t$ statistics at the end of the search apparent in Figure 2(b). The smooth shape of this plot suggests that we have ordered the data by closeness to the model. However, the plot of $R^2$ in Figure 3(b) is not smooth. The increase when $m = 51$ is caused by the introduction of unit 5, apparent in Figure 1(b). This is an example of a “good leverage point” that provides information appreciably reinforcing the parameter estimates from the units already in the subset.
We can plot any quantities of interest during the search. For example, Figure 4(a) is a forward plot of the Bowman-Shenton normality test, which is on two degrees of freedom. The upper panel shows how the value shoots up when the last unit is introduced, the lower panel plotting the significance ($p$ value). Figure 4(b) plots the skewness and kurtosis separately. Figure 1(a) shows how the residual for unit 60, which is, of course, only included in the calculations when $m = 61$, yields a skew distribution of residuals.

The outlying nature of unit 60 can be discerned by looking at the forward plot of the minimum deletion residual among those not in the subset, which is the standard outlier test (Atkinson and Riani 2000, p.23). Figure 5(a) shows how this jumps up at the end of the search when $m = 61$. However, to determine whether the jump at the end is significant requires the use of simulation envelopes. Figure 5(b) adds to Figure 5(a) envelopes calculated simulating 10,000 forward searches. Addition of these envelopes shows that this residual is larger than would
be expected in the last step of the search; even though a large residual is to be expected at this point, unit 60 lies above the 97.5% point of the distribution. It is an outlier.

Our analysis here corroborates that of Zaman et al. (2001). We show the importance of unit 60 on a variety of inferences. Furthermore, our plots show that there are no further observations with unexpectedly important properties. All the remaining 60 units contribute to similar inferences about the model. This conclusion is unchanged when we delete the non-significant $x_1$ from the model.

3 Example 2: Solow Model Applied to OECD Countries

The second example is from Nonneman and Vanhoudt (1996) who included a term in capital in an augmented Solow model. The variables were:

$y$: real GDP per capita of working age
$S$: average annual ratio of domestic investment to real GDP
$N$: annual population growth plus 5%

and the model

$$\log\left(\frac{y_t}{y_0}\right) = \beta_0 + \beta_1 \log y_0 + \beta_2 \log S + \beta_3 \log N + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$ (1)

There are 22 countries. For a very robust analysis to be stable, the number of observations should be appreciably greater than the number of parameters. It is often recommended that $n/p > 5$. Since here there are four parameters and 22 observations, we are perilously close to this border.

For a careful comparison of our results with those of Zaman et al. (2001) we used LTS with the residuals summed over $h = \lceil (n + p + 1)/2 \rceil = 13$ observations,
the initial subset of four observations being found by exhaustive enumeration. Figure 6(a) is a forward plot of the scaled residuals. The distribution is slightly skew, with observations 1 and 3, the last two to enter, being the most outlying for the whole search. There is no extreme outlier like observation 60 in Figure 1(a) and the pattern of residuals is relatively stable, the largest change arguably coming at \( m = 20 \). Figure 6(b) is the forward plot of leverages. This shows that observation 19 enters, with high leverage, at \( m = 20 \); hence the decrease in its residual in Figure 6(a). However, there is no evidence that this observation is outlying. The forward plot of minimum deletion residuals, Figure 7(a), shows a mild effect of observations 1 and 3, while the forward plot of modified Cook’s distance (Atkinson and Riani 2000, p.25, 34), Figure 7(b), indicates only a slight overall change in the parameter estimates. The plot of \( t \) statistics calculated by variable deletion, Figure 8(a), shows that the coefficient of \( x_2 \) increases steadily during the search to reach a final value of 3.22; \( x_1 \) is highly significant throughout. Only \( x_3 \) is not quite significant, with a final \( t \) value of \(-1.97\). None of these conclusions seems sensitive to individual observations; although a slight effect of observation 3 can
In (1) the response has been log transformed. Figure 8(a) is a “fan” plot, that is a forward plot of the $t$ statistics for the score test for transformation. This shows that the data should certainly be transformed, both the square root and no transformation being rejected. The log transformation is acceptable, even though observation 3 does cause a decrease in the value of the statistic to $-2.24$ With logged explanatory variables it makes sense to use a logged response and this analysis shows that this transformation is supported by the data, except for some effect of observation 3.

Our analysis differs from that of Zaman et al. (2001), who make use of a plot of LTS residuals against robust leverages (Rousseeuw and van Zomeren 1990). The leverages are calculated using the Minimum Covariance Determinant (MCD) (Rousseeuw and Van Driessen 1999) that again depends on $h$. This plot can be obtained using the S-Plus function `tsreg`, when $h$ can be specified. When $h = 13$, the minimum value allowed by the function, we obtain Figure 9(a). The ordering of residuals is the same as at the beginning of the search in Figure 6(a), only the scaling being different. The plot is somewhat different from Figure 9(a), obtained when $h = 16$. Now units 1, 19 and 22 both have large absolute LTS residuals and high leverages. They can therefore be classified as “bad” leverage points and Zaman et al. (2001) recommend deletion of these three units.

An interesting side-effect of the use of LTS is the increase in the centre of the plot of minimum deletion residuals in Figure 7(a) compared with the plot for a search starting with LMS in Figure 5(b). This slight hump, which has few inferential implications, is centred on the value of $h$, here 13.

This analysis shows how very robust methods may be unstable when $h$ is not large relative to $p$. It also shows the problems that may arise when robust estimation of leverages is determined solely by the matrix $X$ without reference to the
Figure 9: Solow model: plots of LTS residuals and MCD leverages; (a) $h = 13$ and (b) $h = 16$

response $y$ in selecting the subset of observations for estimation. The effect is particularly sharp in the case of unit 19. In both Figures 9(a) and (b) this is classified as a “bad” leverage point. Yet Figures 6(a) and 8(a) show that, although this is a leverage point, its inclusion strengthens the existing inferences; it is “good”. Although leverage points in regression may both be outlying and have small least squares residuals when $m = n$, there is, as we have seen, nowhere for such difficult units to hide during the forward search.

4 Example 3: Loyalty Cards

The third example of Zaman et al. (2001) analyses data from Benderly and Zwick (1985) on returns from US stocks. The data cover 28 years and our conclusions agree with those of Zaman et al. (2001), in particular that 1979 and 1980 are influential years for estimation of the parameters. Since $p = 3$, the data have much the same dimension as Example 2. Instead of discussing these data further we turn to an appreciably larger set, to illustrate the use of the forward search when $n > 500$.

There are 509 observations on the behaviour of customers with loyalty cards from a supermarket chain in Northern Italy. The data are themselves a random sample from a larger database. The sample of 509 observations is available at www.riani.it/loyalty. The variables are:

$y$: the amount, in euros, spent at the shop over six months

$x_1$: the number of visits to the supermarket in the six month period

$x_2$: the age of the customer

$x_3$: the number of members of the customer’s family.

The data need transformation to achieve constant variance. Figures 10(a) is the fan plot for the Box-Cox power transformation of the response. At the end of
the search \( \lambda = 1/3 \) is acceptable, but it is rejected around 15 observations from the end. At this point a value of 0.4 is acceptable, as it has been throughout the search, although the last 23 observations are clearly different, causing this value to be rejected as the search progresses. For smaller data sets, such as those analysed by Box and Cox (1964) \((n = 27 \& 48)\) a coarser grid of \( \lambda \) values suffices, but here, 0.5 and 0, for example are both clearly rejected. We work with \( \lambda = 0.4 \) for the rest of our analysis.

Figure 10(b) is the forward plot of scaled residuals. At the beginning of the search there is a central majority and some 30 observations with large negative residuals with a wide range of values. 23 of these observations enter sequentially at the end of the search, causing a rapid change in many residuals, including the positive values at the top of the plot. The last 23 units to enter the search are highlighted; their values move together, showing they form a group. These are the same observations as were causing the change in the test for \( \lambda = 0.4 \) in Figure 10(a) from 0.305 to −3.83. The group of observations is shown as outlying by the forward plot of minimum deletion residuals and as influential by the plot of modified Cook’s distances; neither plot is given here. The inferential effect of these observations is clear in Figure 11(a), the forward plot of the \( t \) statistics found by deletion of variables. Although the value for \( x_1 \) continues to increase steadily in significance to 50.0, that for \( x_3 \) decreases at the end of the search, whilst remaining a significant 6.92. The effect of the identified subset on \( x_2 \) is more drastic. The coefficient is 3.09, significant at the 1% level, at \( m = 480 \), but is a nonsignificant 1.12 at the end of the search. This subgroup that we have identified is having an appreciable effect both on the preferred transformation and on the terms in the linear model.

The observations we have found are outlying in an interesting way. Figure 11(b) is the scatterplot of the transformed response \( y^{0.4} \) against \( x_1 \) (frequency)
with the last 23 units to enter the forward search highlighted. We have identified a group that is spending less than would be expected; these people will be important in any further modelling of the data.

5 Discussion

Our graphics-rich procedure extends and amplifies earlier robust analyses. Zaman et al. (2001) present many possible explanations for the low take-up rate of robust regression in economics, including the suggestion that computer time is greater than for OLS. However, even if a robust analysis of the kind we have described were to take 5 minutes, compared with 1 second for OLS, this is a negligible difference when compared with the time required both to collect the data and to study and assimilate the implications of the results of the analysis for the models being studied.

S-Plus software for the regression methods described in Atkinson and Riani (2000) is available at the website for that book www.riani.it/ar.

We are grateful to Professor Mehmet Orhan for making available electronic versions of the data used in Zaman et al. (2001).

References


THIS APPENDIX IS NOT INTENDED FOR PUBLICATION, UNLESS THE REFEREES FEEL THAT ITS VALUE JUSTIFIES THE EXTRA LENGTH. OUR INTENTION IS TO MAKE IT AVAILABLE ON THE WEB AS A SUPPLEMENT TO OUR PAPER

6 Appendix: Theory

6.1 Ordinary Least Squares (OLS)

In the regression model \( y = X\beta + \epsilon \), \( X \) is an \( n \times p \) full-rank matrix of known constants, with \( i \)th row \( x_i^T \), and \( \beta \) is a vector of \( p \) unknown parameters. The errors \( \epsilon_i \) are assumed i.i.d. \( N(0, \sigma^2) \).

With \( \hat{\beta} \) the least squares estimator of \( \beta \) the vector of least squares residuals is \( e = y - \hat{y} = y - X\hat{\beta} = (I - H)y \), where \( H = X(X^TX)^{-1}X^T \) is the ‘hat’ matrix, with diagonal elements \( h_{ii} \). The mean square estimator of \( \sigma^2 \) is \( s^2 = e^T e / (n-p) = \sum_{i=1}^{n} e_i^2 / (n-p) \).

6.2 The Forward Search

Let \( \mathcal{M} \) be the set of all subsets of size \( m \) of the \( n \) observations. The forward search fits subsets of observations of size \( m \) to the data, with \( m_0 \leq m \leq n \). In regression we take \( m_0 = p \) and use LMS or LTS to find the initial subset.

Let \( S_{(m)} \in \mathcal{M} \) be the optimum subset of size \( m \), for which the matrix of regressors is \( X_{(m)} \). Least squares applied to this subset yields parameter estimates \( \hat{\beta}(m^*) \) and \( s^2(m^*) \), the mean square estimate of \( \sigma^2 \) 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

\[ e_i(m^*) = y_i - x_i^T \hat{\beta}(m^*). \]  

The search moves forward with the subset \( S_{(m+1)} \) consisting of the observations with the \( m + 1 \) smallest absolute values of \( e_i(m^*) \).

6.3 Robust Estimation and the Start of the Search: LMS, LTS and MCD

The search starts from a subset of \( p \) observations \( S_{(p)} \) that is chosen by repeated random sampling or complete enumeration to provide a very robust estimator of the regression parameters. For Least Median of Squares (LMS, Rousseeuw 1984)
the subset of $p$ observations is found minimizing the scale estimate
\[ \sigma^2(h) = e^2_{[h]}(p), \]
where $e^2_{[k]}(p)$ is the $k$th ordered squared residual for some subset $S^{(p)}$ and $h$ is the integer part of $(n + p + 1)/2$, corresponding to ‘half’ the observations when allowance is made for fitting. For Least Trimmed Squares the sum of squares
\[ S(h) = \sum_{i=1}^{h} e^2_{i}(p) \]
is minimized, where now $h \geq [(n + p + 1)/2]$. Likewise, the Minimum Covariance Determinant (MCD) estimator of the variance-covariance matrix $\Sigma$ of $X$ finds the minimum of $|\hat{\Sigma}(h)|$ by searching over subsets of $X$ (Rousseeuw and Van Driessen 1999).

Provided any masking is broken, the search is insensitive to the starting method.

6.4 Minimum Deletion Residual

To test for outliers the deletion residual is calculated for the $n - m$ observations not in $S_{i}^{(m)}$. These residuals are
\[ r^*_i(m^*) = \frac{y_i - x^T_i \hat{\beta}(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}} = \frac{e_i(m^*)}{\sqrt{s^2(m^*)\{1 + h_i(m^*)\}}}, \]  \hspace{1cm} (3)

where $h_i(m^*) = x^T_i \{X(m^*)^T X(m^*)\}^{-1} x_i$; the leverage of each observation depends on $S_{i}^{(m)}$. Let the observation nearest to those constituting $S_{i}^{(m)}$ be $i_{\min}$ where
\[ i_{\min} = \arg \min_{i \notin S_{i}^{(m)}} |r^*_i(m^*)| \]
the observation with the minimum absolute deletion residual among those not in $S_{i}^{(m)}$. To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual $|r^*_{i_{\min}}(m^*)|$.

6.5 Tests for Coefficients Deleting Variables

To test the significance of the coefficient $\beta_j$ we need the estimate $\hat{\beta}_j(m^*)$ to be independent of the search. Accordingly partition the model as
\[ y = X \beta + \epsilon = X_{(j)} \beta_{(j)} + x_j \beta_j + \epsilon, \]
where $X_{(j)}$ is $n \times (p - 1)$. The search is performed using residuals from regression on the $p - 1$ variables in $X_{(j)}$, but we test $\beta_j$ at each step of the search from the regression on all $p$ variables, for $j = 2, \ldots, p$ (Atkinson and Riani 2002).
6.6 The Modified Cook Distance

To test the effect of the inclusion of observation \( i_{\text{min}} \) on the vector parameter estimate \( \hat{\beta}(m^*) \) we use the modified Cook distance

\[
C_i = \left\{ \frac{m + 1 - p}{p} \frac{h_{i_{\text{min}}}(m^*)}{1 - h_{i_{\text{min}}}(m^*)} \right\}^{1/2} \left| r_{i_{\text{min}}}(m^*) \right|
\]


6.7 Fan Plot of Score Test for Box-Cox Transformation

To test the transformation parameter \( \lambda \) in the Box and Cox power transformation we expand the normalized transformation (Box and Cox 1964) in a Taylor series to obtain a constructed variable \( w \). The approximate score test is the \( t \)-test for regression on \( w \) in the multiple regression of the normalized transformation on \( X \) and \( w \). The fan plot shows the values of this \( t \)-test during a series of forward searches, one for each value of \( \lambda \) being tested (Atkinson and Riani 2000, p.85, 89).