Computational Statistics and Data Analysis 104 (2016) 209–222

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust methods for heteroskedastic regression

Anthony C. Atkinson[a,*], Marco Riani[b], Francesca Torti[c]

[a] Department of Statistics, London School of Economics, London WC2A 2AE, UK

[b] Dipartimento di Economia, Università di Parma, Italy

[c] European Commission, Joint Research Centre, Competences Directorate, Text and Data Mining Unit, Ispra, Italy

HIGHLIGHTS

- Generalizes the standard model for heteroskedasticity in non-robust regression.
- Flexibility of the robust model shown on complex international trade data.
- Outperforms conventional "heteroskedastic robust" standard errors.
- Linked graphics provide insight into importance of individual observations.
- Provides publicly available Matlab code for very robust heteroskedastic regression.

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form 5 July 2016
Accepted 5 July 2016
Available online 15 July 2016

Keywords:
Forward search
Graphics
Heteroskedasticity robustness
Outliers
Scoring algorithm

ABSTRACT

Heteroskedastic regression data are modelled using a parameterized variance function. This procedure is robustified using a method with high breakdown point and high efficiency, which provides a direct link between observations and the weights used in model fitting. This feature is vital for the application, the analysis of international trade data from the European Union. Heteroskedasticity is strongly present in such data, as are outliers. A further example shows that the new method outperforms ordinary least squares with heteroskedasticity robust standard errors, even when the form of heteroskedasticity is mis-specified. A discussion of computational matters concludes the paper. An appendix presents the new scoring algorithm for estimation of the parameters of heteroskedasticity.

1. Introduction

We provide a new robust method for the analysis of heteroskedastic data with the linear regression model which is both efficient and has high breakdown point. Our development is driven by the need for diagnostic and robust exploration of international trade data in which consistently anomalous price–quantity relationships may indicate money laundering or tax fraud. An introduction to the vast scale of the problem is in The Economist (2014). Our example in Section 4 is of imports into the European Union. These data require careful analysis in order to detect the outliers and other properties which may be evidence of illegal behaviour. Because of the vast quantities of data involved, appropriate automatic methods of robust analysis are essential. We provide these by combining robustness with a form of weighted regression in which the weights modelling heteroskedasticity are also robustly estimated. Despite the specific motivation of our development, the resulting form of robust regression is both powerful and of generally applicability.

Methods for the robust analysis of homoskedastic regression data are well established (Maronna et al., 2006). Likewise methods for non-robust heteroskedastic regression analysis are widely described in econometrics (Greene, 2002, 2012). The

[*] Corresponding author.
E-mail addresses: a.c.atkinson@lse.ac.uk (A.C. Atkinson), mriani@unipr.it (M. Riani), francesca.torti@jrc.ec.europa.eu (F. Torti).

http://dx.doi.org/10.1016/j.csda.2016.07.002

methodological contribution of our paper is to provide a method for robust heteroskedastic regression which generalizes the form of heteroskedasticity described, in a non-robust context, by Harvey (1976). We give a link to publicly available Matlab code. Our subject matter contribution is to present the first robust heteroskedastic analysis of trade data. Our analysis reveals the presence of two appreciable outliers as well as the potential presence of data from more than one regression line.

In order to robustify heteroskedastic regression, it is helpful to divide the methods of robust regression into three categories. Following Riani et al. (2014b) these are:

1. Downweighting. M estimation and derived methods, as described in Maronna et al. (2006);
2. Trimming, as defined by Maronna et al. (2006, p. 132), includes Least Trimmed Squares (LTS) and Least Median of Squares (LMS) (Hampel, 1975; Rousseeuw, 1984) in which the amount of trimming is determined by a pre-specified factor and
3. Adaptive trimming. In the Forward Search (FS) (Atkinson et al., 2010) the observations are $(0, 1)$ trimmed, but the amount of trimming is found adaptively, providing high efficiency.

M estimation is the robust method most explored for heteroskedastic regression. An approach through weighting is in Chapter 4 of Carroll and Ruppert (1988). Further results are sketched by Welsh et al. (1994). The references to trimming are more recent. Cheng (2011) uses the model of Harvey (1976), combined with the FS and a trimmed likelihood for robust heteroskedastic regression. Neykov et al. (2012) use trimming in robust estimation of a general quasi-likelihood model, including the special case of heteroskedastic regression.

A major difficulty with downweighting for our intended applications is the absence of a clear relationship between individual observations and their effect on inferences drawn from the data. In addition, downweighting is also ruled out, as is non-adaptive trimming, by the comparative studies of Riani et al. (2014a,b). The results of Riani et al. (2014a) indicate, for finite sample sizes, that the adaptive downweighting associated with the FS provides robust parameter estimates with lower bias and variance for contaminated data than the other methods. In addition, Riani et al. (2014b) show that the FS leads to data analyses which are more informative about the pattern of departures from the linear model than are the other methods. Accordingly, we use the FS as the basic robust method for handling heteroskedasticity in regression. Johansen and Nielsen (2016a) discuss some theoretical aspects of the FS.

For the data which initiated our study, it is appropriate to assume that the conditional variance of the observations depends on a linear function of the explanatory variables in the regression, the parameters of which are to be estimated. We describe our model in Section 2 and comment on the relationship with Harvey's model.

Section 3 introduces the Forward Search (FS), which provides a robust, diagnostic fit to a single regression model. The search proceeds by fitting the model to subsets of the data of increasing size. Statistical properties of the method for outlier detection are also in Section 3, with the details of the procedure presented in Appendix B. In Section 4 the heteroskedastic FS is illustrated on an example of 1100 observations. Forward plots of residuals and parameter estimates as the subset size increases clearly demonstrate the properties of the method, including outlier detection. Weighted least squares can be considered as ordinary least squares in the space of weighted responses and explanatory variables. Comparison of results for the weighted and unweighted forms of the model leads to a clear understanding of leverage in weighted least squares.

In Section 5 we examine the stability of the weights calculated during the FS and the insensitivity of the search to starting values. We proceed further in Section 6 to two more statistical analyses of our example: the first illustrates the importance of using an analysis that accommodates heteroskedasticity. The second shows how using "brushing" to monitor the progress of the FS reveals important data structures.

We require a method of robust heteroskedastic regression also to be robust to the specification of the form of heteroskedasticity. A very general method (White, 1980) uses ordinary least squares (OLS) combined with "heteroskedastic robust" standard errors. In Section 7 we show how poorly this "heteroskedastic robust" procedure can perform when compared with a model with correctly specified heteroskedasticity. More importantly, we demonstrate that our method is never less efficient than OLS even when the skedastic relationship is incorrectly specified. We conclude with an indication of the computational efficiency of our method. Details of the scoring algorithm are in Appendix A. The emphasis in our paper is on the public provision of a method for robust heteroskedastic regression.

2. Heteroskedastic regression

2.1. Models for non-constant variance

The data that stimulated this research have a relatively simple structure, for which we can assume additive normal errors. Our model for heteroskedastic regression can be written

$$y_i = \beta^T x_i + \sigma_i \epsilon_i \quad (i = 1, \dots, n),$$

where the errors ϵ_i have a (homoskedastic) standard normal distribution, $\epsilon_i \sim \mathcal{N}(0, 1)$. We parameterize the variance function as

$$\sigma_i^2 = \text{Var } y_i = \sigma^2 \{1 + \exp(z_i^T \gamma)\}, \tag{1}$$

a form that is used in the analysis of pharmacokinetic data (Fedorov and Leonov, 2014).

In our example the variables x and z are identical. However, all elements of the two parameter vectors are distinct. Then the information matrix is block diagonal. For given γ estimation of β is by weighted least squares and σ^2 is estimated from the residual sum of squares. For given γ the weights are defined as

$$w_i = \sigma^2/\sigma_i^2 = \{g(z_i^T\gamma)\}^{-1}.$$

It is convenient to write

$$y_i^W = \sqrt{w_i}y_i, \qquad x_i^W = \sqrt{w_i}x_i \quad \text{and} \quad W = \text{diag}\{w_i\}.$$

Then

$$\hat{\beta}^W|\gamma = (X^TWX)^{-1}X^TWy,$$

although we shall usually notationally suppress the dependence of $\hat{\beta}^W$ on γ. Thus $\hat{\beta}^W$ is found by weighted regression of y on x or, equivalently, by unweighted regression of y^W on x^W. We find it informative to consider data analysis in both these spaces. Since $\hat{\beta}^W$ is found by least squares regression, it inherits the affine equivariance of least squares estimators.

It remains to estimate γ. Numerical maximization of the likelihood is one possibility. However, since we need to estimate the parameter for each subset of the data in the FS, of which there are almost n, we used a more efficient scoring algorithm, presented in Appendix A. The importance of efficient numerical methods of estimation increases if we need to simulate several thousand searches in order to establish distributional properties.

An important property of (1) is that, provided $\sigma^2 > 0$, the variance does not approach zero as $\exp(z_i^T\gamma) \to 0$. A combination of relatively constant variance, combined with regions of appreciable heteroskedasticity is common in some forms of data, including the data we analyse in Section 4.

A simpler model for heteroskedasticity with skedastic equation

$$\sigma_i^2 = \sigma^2 \exp(z_i^T\gamma), \tag{2}$$

for which the variance can go to zero, was introduced by Harvey (1976). The properties of heteroskedastic regression with (2), together with a scoring algorithm, are described and illustrated by Greene (2002, §11.7) and Greene (2012, p. 554–556). The algorithm is similar in structure to that in Appendix A. Although the algebraic expressions for (2) are easier to write down than those for model (1), the difference in computational complexity is negligible. Of course, the choice of a skedastic equation must depend on the data.

3. The forward search for weighted regression data

The forward search used in this paper starts from a very robust LMS fit with 50% breakdown point and then achieves greater efficiency by fitting the model to subsets of the data of increasing size.

For the moment we assume the weights w are known. In the weighted regression model for all n observations, $y^W = X^W\beta + \epsilon$, β is $p \times 1$ and the normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. The weighted least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y^W - \hat{y}^W = y^W - X^W\hat{\beta} = (I - H)y^W$, where $H = X^W\{(X^W)^T(X^W)\}^{-1}(X^W)^T$ is the 'hat' matrix, with diagonal elements h_i and off-diagonal elements h_{ij}. The residual mean square estimator of σ^2 is $s^2 = e^Te/(n-p) = \sum_{i=1}^n e_i^2/(n-p)$.

The FS algorithm starts from the LMS solution found by estimation on a selected subset of p observations. Observations in this subset are intended to be outlier free. If n is moderate and $p \ll n$, the choice of the initial subset can be achieved by exhaustive enumeration of all i_p distinct subsets of size p. The parameters of the model are estimated by least squares on each subset. Let the least squares residual for unit i $(i = 1, \ldots, n)$ from subset k be e_{ik}. We take as our initial subset the p-tuple $S^*(p)$ which satisfies

$$e^2_{[\text{med}],S^*(p)} = \min_k\{e^2_{[\text{med}],S_k(p)}\},$$

where $e^2_{[j],S_k(p)}$ is the jth ordered squared residual among e^2_{ik} $(i = 1, \ldots, n)$ and med is the integer part of $(n + p + 1)/2$. If the number of subsets is too large for exhaustive enumeration, we use some moderate number of samples like 1000.

Although the LMS estimate is highly robust it can have very low efficiency (Rousseeuw and Leroy, 1987). In Section 5 we present some numerical evidence of the consequences of this instability. We maintain the high breakdown point but increase efficiency by increasing the size of the subsets used in fitting, with the subset of size $m + 1$ chosen to be as close as possible to the model fitted to m observations. The introduction of outliers into the subset is diagnostically revealed by plots of residuals against subset size as well as formally by statistically tuned tests.

The size, m, of the subsets is such that $m_0 \le m < n$. Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X^W(m)$. 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. The $m \times m$ diagonal weight matrix has entries $w_i(m)$ calculated from the estimated parameter $\hat{\gamma}(m)$ of the skedastic equation. Residuals can be calculated for all observations including those not in $S^*(m)$. The n resulting least squares residuals are

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

The search moves forward with the augmented subset $S^*(m+1)$ consisting of the observations with the $m+1$ smallest absolute values of $e_i(m)$.

The consistency of the parameter estimates generated by the FS is proved by Cerioli et al. (2014) with further results on its robustness given by Johansen and Nielsen (2016b). These results arise from the use of maximum likelihood estimates in the FS with known weights. However, we also use maximum likelihood in estimation of the weights so that, in the absence of outliers, these estimates also converge to the population values. The numerical results of Section 5 show how surprisingly stable the estimated weights are over a large part of the search.

An alternative to the use of LMS for the starting subset of the search is Least Trimmed Squares, which has negligible effect on the subsequent properties of the FS. Further numerical results in Section 5 show that starting the search several times from random starting points leads to stable robust and efficient solutions. As we explain in Section 5, the robustness is generated by the search itself adding and deleting observations, not by the starting point.

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

$$r_i(m) = \frac{y_i^W - \hat{\beta}^T(m)x_i^W}{\sqrt{s^2(m)\{1+h_i(m)\}}} = \frac{e_i(m)}{\sqrt{s^2(m)\{1+h_i(m)\}}}, \tag{3}$$

where the leverage

$$h_i(m) = (x_i^W)^T[\{X^W(m)\}^T\{X^W(m)\}]^{-1}x_i^W.$$

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 $|r_{i\min}(m)|$, 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 $S^*(m)$.

In Section 4 we use diagnostic plots of the evolution of $|r_{i\min}(m)|$ with m in order to reveal the structure of the data. For formal testing we need a reference distribution for $r_i(m)$ in (3). 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 symmetrically truncated distribution, we take $s_T^2 = s^2(m)/c(m,n)$ as our approximately unbiased estimate of variance. In the robustness literature $c(m,n)$ is called a consistency factor (Maronna et al., 2006). See Riani et al. (2009) for a derivation from the general method of Tallis (1963).

The test statistic $|r_{i\min}(m)|$ is the $(m+1)$st ordered value of the absolute deletion residuals. To find its distribution we adapt the order-statistic argument of Riani et al. (2009) in which envelopes were required for the Mahalanobis distances arising in applying the FS to multivariate data. Here we are considering the absolute values of t distributed random variables and obtain the confidence level γ as

$$\gamma = 1 - F_{2(n-m),2(m+1)}\left\{\frac{m+1}{n-m}\left[\frac{1}{2T_{m-p}\{r_{\min}(m)\sigma_T(m)\}} - 1\right]\right\}, \tag{4}$$

for $m = m_0, m_0+1, \ldots, n-1$. In (4), F and T are the c.d.f.s of the F and T distributions. There is appreciable curvature in the plots of these envelopes for the minimum deletion residuals; as $m \to n$, the envelopes increase rapidly, since, even in the absence of outliers, large residuals occur at the end of the search.

Finally, to avoid the problem of multiple testing (one outlier test for each value of m) we adapt the rule of Riani et al. (2009) to obtain a procedure for regression data with a samplewise size of around 1%. We run the FS monitoring the bounds from (4) until we obtain a "signal" indicating that observation $m^\dagger$, and therefore succeeding observations, may be outliers, because the value of the statistic lies beyond a threshold calculated from the bounds and depending on the value of m. The conventional envelopes shown, for example, in Fig. 2, consist roughly of two parts; a flat "central" part and a steeply curving "final" part. Our procedure for the detection of a "signal" takes account of these two parts. Details of this procedure are in Appendix B.

4. Example: International trade data

These international trade data record the transaction value and amount of imports of individual goods into the EU. For a specific quality of a good from an individual supplier there should be a straight line relationship between value and quantity, although the relationship may be different for different qualities and different suppliers. There may also be numerous outliers due to misrecording of the values of the two variables, or due to erroneous coding of goods. Interest is in detecting price–quantity relationships that are consistently anomalous; these may indicate money laundering or tax fraud.

The trade data with which we are concerned have non-negative values of the single explanatory variable x. Zero values in this case do not occur, but the data typically contain many values close to zero, for small transactions. An advantage of our

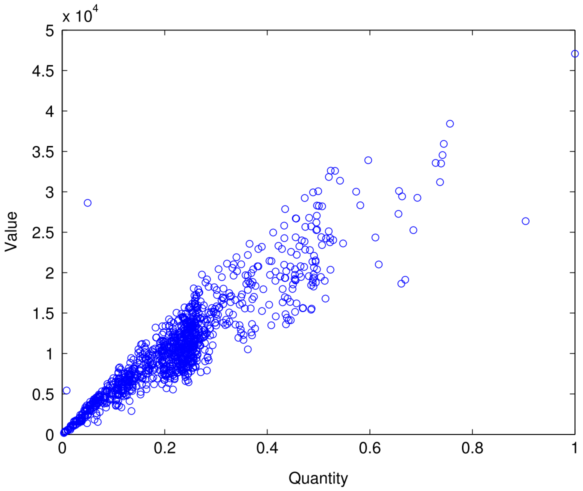

Fig. 1. Scatterplot of value against quantity (scaled to have a maximum of one) for 1100 transactions (imports into the EU). There are two appreciable outliers for low values of x.

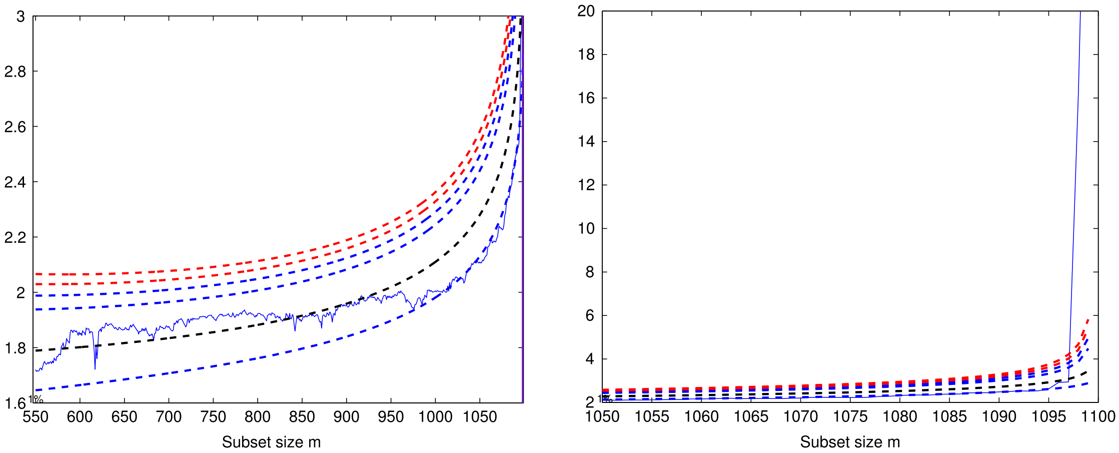

Fig. 2. Left-hand panel; forward plot of the values of the absolute minimum deletion residual $|r_{\text{imin}}(m)|$ from $n = 550$. Right-hand panel zoom for the last 50 observations, clearly showing the extreme nature of the outliers; 1%, 50%, 99%, 99.9%, 99.99% and 99.999% envelopes.

model for heteroskedasticity (1) compared to Harvey's (2) is that we avoid excessive weights for such small observations. We find it convent to reparameterize (1) as

$$\sigma_i^2 = \sigma^2(1 + \theta x_i^\alpha) \quad \text{so that } w_i = 1/\sigma_i^2,$$

which follows by putting $z^T = (1 \ \log x)$, when $\gamma^T = (\log(\theta) \ \alpha)$.

As an example we analyse the 1100 observations for an import into the EU plotted in Fig. 1. The data seem to have a simple structure, being apparently heteroskedastic with two clear outliers for low values of x. It is a characteristic of the human eye that it tends to find lines in the data for high values of x. Since economically important frauds tend to occur for such values, our statistical procedure should provide a firm distinction between outliers and fraudulent observations.

For numerical purposes and without loss of generality, we scale x to lie between 0 and 1, dividing by the maximum value of x. We fit a linear regression model with intercept since even small transactions often incur fixed costs, regardless of size. In this example we allow the scoring algorithm a maximum of 100 iterations and impose maximum values of 10 on the estimates of the parameters α and $\log(\theta)$, although these bounds are not, in fact, needed. The resulting forward search for heteroskedastic regression takes longer than the search for homoskedastic because we have to estimate θ and α for each subset size m. The forward plots of the minimum deletion residuals are in Fig. 2. The plot in the left-hand panel is from $n = 550$; the plot in the right-hand panel is a zoom, focusing on the last 50 observations to enter the search. The overall structure is clear; the right-hand panel shows the extreme nature of the two outliers. Otherwise, in both plots, there is no further evidence of appreciable outliers; for most of the search the observed values of the statistic lie close to the median of the distribution.

The top left-hand plot of Fig. 3 shows the scatterplot of the data with the two outliers plotted as crosses. These are, of course, the outliers that were evident to the naked eye in Fig. 1. The other three panels provide forward plots of the estimates

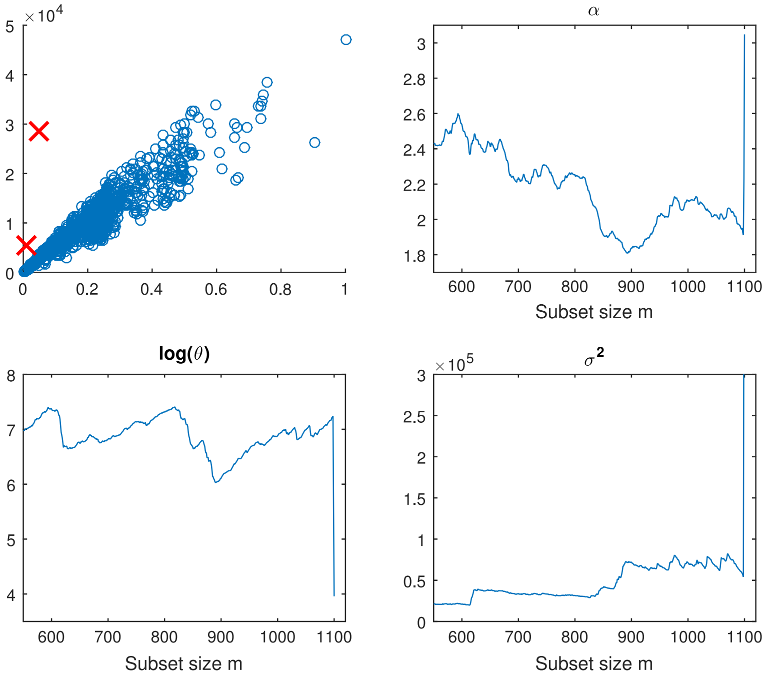

Fig. 3. Top left-hand panel: scatterplot of the data with the last two observations to enter the search indicated by ×. Top right-hand panel: forward plot of $\hat{\alpha}$. Bottom panels, forward plots of $\log(\hat{\theta})$ and $\hat{\sigma}^2$. There is a dramatic effect of the two outliers on both these estimates at the end of the search.

of the parameters. The values of $\hat{\alpha}$ in the upper right-hand panel are comparatively stable, trending down from around 2.5 to 1.8 and then up again towards 2, before there is a sudden change at the end of the search; the inclusion of the outliers causes a sudden change in the relationship between x and the variance σ_i^2. These values are typical of those for $\hat{\alpha}$ we have found in the analysis of numerous datasets; almost invariably, the estimates lie between 1.5 and 3 until outliers or other non-null data structures are included in the subset $S^*(m)$.

For much of the search the larger values of θx^α are appreciably greater than one. However, with the settings we use for the scoring algorithm, the estimates of both θ and σ^2 are well defined. At the end of the search, the two outliers enter the subset and reduce the relationship between the variance σ_i^2 and the value of x. At step $m = n - 2$ the value of $\log \hat{\theta}$ is 7.24, reducing to 3.96 when $m = n$; the estimate of σ^2 increases in compensation from 0.054×10^6 to 2.837×10^6. Although, in general, comparing variance estimates from differently weighted regressions may not be meaningful, the stability of the weights revealed in Section 5 allow us to conclude that these large changes, particularly in the estimate of σ^2, clearly indicate the incorporation of outliers into the subset of data used in parameter estimation.

We now look at the prediction intervals for a new observation $Y(x)$ in the original (unweighted) space for which $\hat{\beta}^W$ provides the best estimate of β, with variance $\sigma^2(X^TWX)^{-1}$, if the weights are known. In the original space the variance of a new observations is σ_i^2 and the prediction is $x^T\hat{\beta}^W$; accordingly, the variance of prediction is $\sigma_i^2 + \sigma^2 x^T(X^TWX)^{-1}x$. Fig. 4 shows these prediction limits. The parameter estimates used to calculate the limits exclude the last two observations and the limit gives an interval with a nominal content of 99%. In this scale the plots of the limits are virtually straight lines. The two outliers lie well outside the limits. A few other observations also lie just outside.

This variance calculation ignores any effect of estimation of the parameters θ and γ in the weights. Fig. 5 shows the results of 5000 simulations of envelopes for the absolute minimum deletion residuals when θ and γ are estimated compared with simulated envelopes in the absence of heteroskedasticity. For a sample of this size, there is no discernible difference between the two sets of 50% and 99% envelopes, except at the very beginning of the search. Agreement with the lower limit is not quite so good, but it is large values of the residuals that are of interest. For smaller sample sizes and more extreme quantiles the agreement is not quite so good. However, for our example, the agreement is such that the effect of estimation can be ignored.

We now consider estimation in the weighted scale of x^W and y^W, which enables consideration of leverage in the weighted least squares fit. Fig. 6 is the very different plot of the prediction limits in this weighted scale. Now the prediction is $(x^W)^T\hat{\beta}^W$. For large $\hat{\theta}$ as we have here, $\sigma_i^2 \propto x_i^\alpha$, so that $w \propto x^{-\alpha}$ and $x^W \propto x^{1-\alpha/2}$. With $\hat{\alpha}$ close to 2, x^W is virtually constant, as the figure shows. The prediction variance is now $\sigma^2\{1 + (x^W)^T(X^TWX)^{-1}(x^W)\}$. As the figure shows, the prediction limits are virtually horizontal and only very slightly curved.

This new scale is that of homoscedastic estimation. The two outliers, which are for small values of x, where the variance σ_i^2 is relatively small, are seen to be more extreme in Fig. 6 than in Fig. 4. The other interesting feature of Fig. 6 is that,

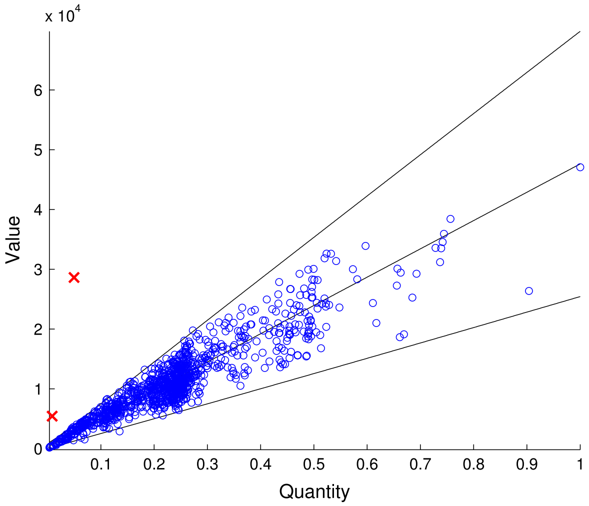

Fig. 4. 99% prediction intervals for a new observation in the original space of the observations using parameter estimates from $m = n - 2$; × the two outliers.

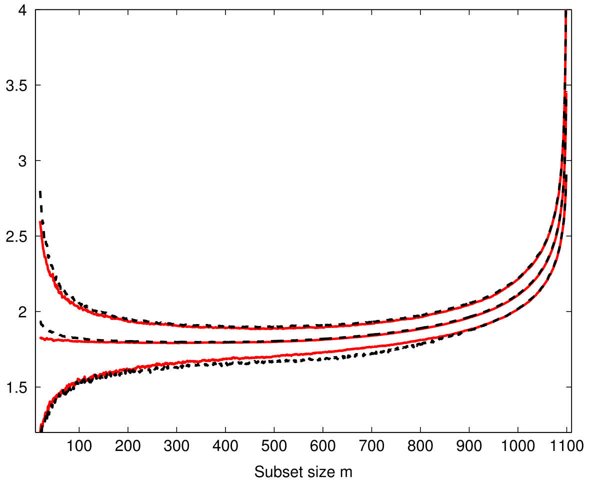

Fig. 5. Comparison of empirical envelopes. 1%, 50% and 99% envelopes based on 5000 simulations. Continuous lines, homoskedastic search; dashed lines, estimated heteroskedasticity.

compared to Fig. 4, the observations cluster at the right-hand end of the plot, that is for high values of x^W. In fact, in the weighted space used for homoscedastic estimation, there is a series of leverage points for low values of x^W, which include one outlier.

To check the effect of these leverage points, we trimmed all observations with $x^W < 0.011$, leaving 1081 units including only one of the two original outliers. Such trimming is suggested, in a non-robust context, by Davidian and Carroll (1987). Fig. 7 shows the new prediction interval with parameter estimates from 1080 units, with the outlier excluded. The nonlinear structure of the fitted model in the weighted space is now evident. However, in the original space the fit is still linear and we receive a plot very much like Fig. 4. An important feature of the plot is the concentration of units with values of x^W close to 0.041. These units provide virtually no information on the skedastic relationship and account for the variable parameter estimates evident in Fig. 3.

5. Stability of the search and insensitivity to the starting point

We first illustrate the extreme stability of the weights $w_i(m)$ for a search starting from an LMS subset. The plot in Fig. 8 shows the weights for a randomly selected subset of 100 observations, plus, represented by heavier lines, the two outliers. There is virtually no fluctuation in the weights for these last 50 observations. The trajectories of the weights in the plot do not cross until the very end of search when there is a modification in all weights as the two outliers are included.

The FS used in this paper starts from an LMS subset of 2 subsets. This assures high breakdown, but is inefficient and unstable. As m increases the efficiency, and so the stability, of the estimates increase. Fig. 8 shows a forward plot of absolute

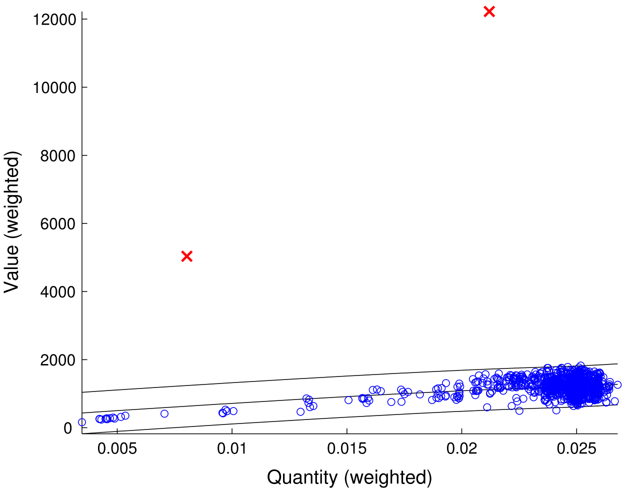

Fig. 6. 99% prediction intervals for a new observation in the weighted space of x^W and y^W using parameter estimates from $m = n - 2$. The two outliers × are revealed as very extreme.

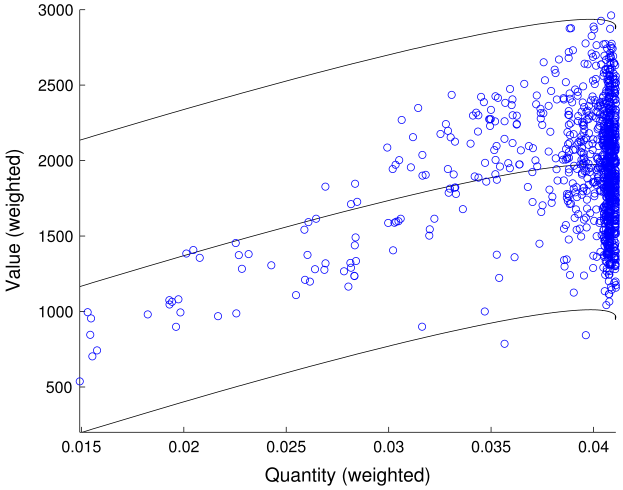

Fig. 7. 99% prediction intervals for a new observation in the weighted space of x^W and y^W; leverage points excluded, leaving 1081 units. Parameter estimates from 1080 observations. There is a concentration of units around $x^W = 0.041$.

minimum deletion residuals for 300 searches through the international trade data starting from randomly chosen subsets of size 2. The plot initially shows much fluctuation but, by halfway through the data, all 300 plotted trajectories have converged to a single trajectory, identical to that obtained using the LMS starting subset shown in the left-hand panel of Fig. 2. The convergence occurs because at each step in the FS we reorder all n residuals; as a consequence observations can both enter and leave the subset $S^*(m)$. Once two trajectories have converged they will have the same subset and so cannot diverge. This stability is much like that we have observed for random start searches in unweighed regression and multivariate analysis (Atkinson and Riani, 2007); convergence happens between half and two thirds of the way through the search (see Fig. 9).

6. Further data analysis

The analysis in Section 4 clearly shows the presence of heteroskedasticity in the data. To emphasize the importance of the heteroskedastic analysis of such data we reanalyze the data with a homoskedastic model. The resulting plot of fitted values, prediction intervals and outliers is shown in Fig. 10. This analysis finds 196 outliers. As the plot shows, these come by trimming away the larger observations which lie away from the fitted line.

The contrast with the FS is informative about the structure of the data. Fig. 11 shows the effect of brushing the FS when m is around 1000. The dotted units are those not included in the subset at this point. As well as the two outliers, they include two extreme lines of observations, one high, one low. Although these observations have not been declared as outliers, it would be informative to return to the data and to identify these two sets of units according to country of origin and importer.

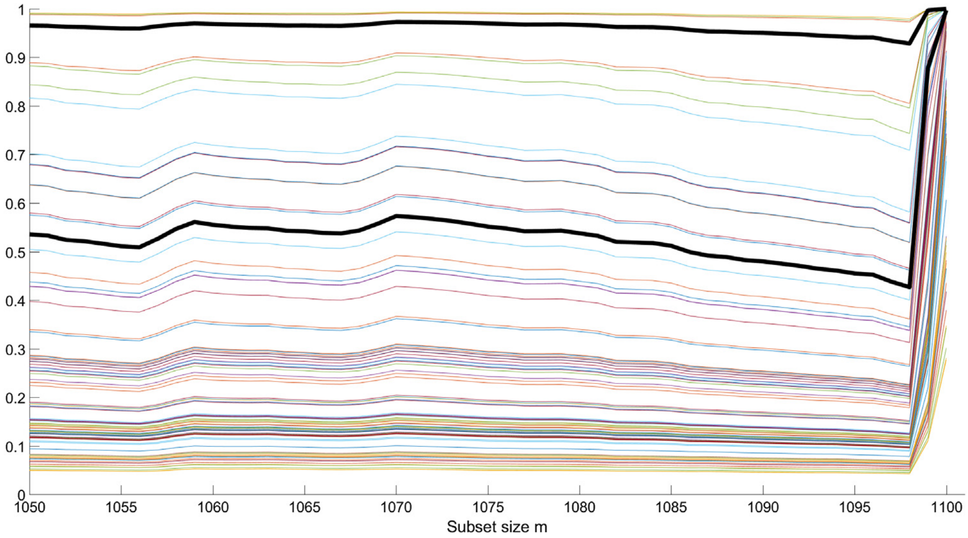

Fig. 8. Forward plot of weights $w_i(m)$ in the last 50 steps of the search for 100 randomly selected observations and, in heavy lines, the last two observations to enter the subset.

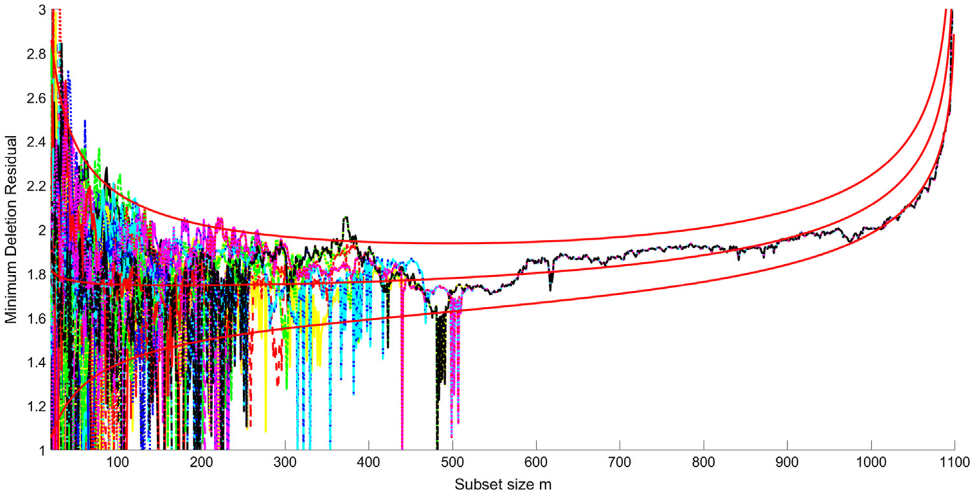

Fig. 9. Forward plot of minimum Mahalanobis distances from 300 forward searches starting from randomly chosen subsets of size 2. All trajectories have converged to the single robust search by half way through the search.

7. Robustness to heteroskedastic specification

Our procedure assumes that the skedastic model is correctly given by (1). A standard practice in econometrics is to use the ordinary least squares estimator $\hat{\beta}_{OLS}$ from unweighted regression. If heteroskedasticity is present, the covariance matrix of $\hat{\beta}_{OLS}$ is estimated from a heteroskedasticity consistent sandwich estimator suggested by White (1980). We first compare the efficiency of our procedure with that of the heteroskedasticity-consistent estimate of the efficiency of OLS which avoids precise specification of the relationship. Because the relationship does not have to be specified, this procedure is often described as robust. We then evaluate efficiencies when the skedastic relationship is mis-specified and White's procedure should indicate that OLS has relatively improved properties.

In the presence of heteroskedasticity $\hat{\beta}_{OLS}$ is unbiased, but has an inflated variance. With Σ the variance–covariance matrix of the observations,

$$\mathrm{var}(\hat{\beta}_{OLS}) = (X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}.$$

To estimate this matrix for independent observations, White takes $\sigma_i^2 = e_i^2$, where the e_i are the residuals from OLS. Let $\widehat{\Sigma}_W = \mathrm{diag}\ e_i^2$, then

$$\widehat{\mathrm{var}}(\hat{\beta}_{OLS}) = (X^TX)^{-1}X^T\widehat{\Sigma}_W X(X^TX)^{-1}. \quad (5)$$

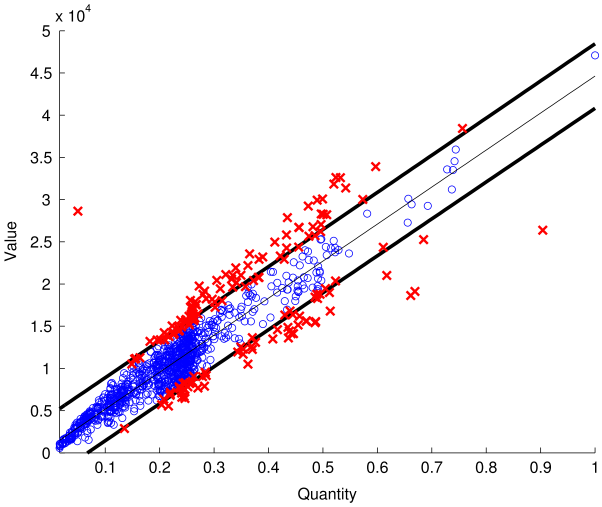

Fig. 10. 99% prediction intervals for a new observation and 196 outliers when a homoskedastic analysis is erroneously employed.

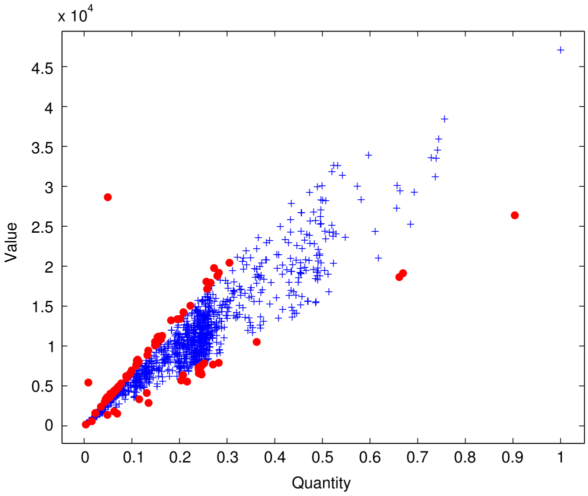

Fig. 11. Brushing the FS; +, units in the subset when m is about 1000.

The efficiency of the estimators is compared through the estimated generalized variance of the estimate of β. We confine our comparisons to Harvey's model (2) from fitting which we obtain weights $w_i^H = 1/\hat{\sigma}_i^2$, giving a weight matrix $\widehat{W}_H$. The generalized variance is proportional to the determinant of the covariance matrix. In comparing OLS and Harvey's model as estimators of β, the efficiency is thus

$$1/\{|(X^TX)^{-1}X^T\widehat{\Sigma}_WX(X^TX)^{-1}| \times |X^T\widehat{W}_HX|\}^{1/p}, \qquad (6)$$

since the covariance matrix from Harvey's model is $(X^T\widehat{W}_HX)^{-1}$. In (6) the dimension of β is p. Raising the product of determinants to the power $1/p$ gives an efficiency measure that is proportional to the number of observations. An efficiency of 50% indicates that twice as many observations are needed with the inefficient estimator as with the efficient estimator to get the same amount of information about the values of the parameters β.

If the skedastic equation is correctly specified, OLS will lose efficiency. Our purpose is to present some quantitative values for this decline. But also, and more importantly, we investigate the efficiency of OLS using White's procedure when the data do not come from the parametric model we fitted.

Throughout we use Harvey's model (the results for our more general model are very similar and are not given here for lack of space) with a constant and three explanatory variables so that $p = 4$. The linear model generating the data is simulated with three non-negative independent variables distributed as $|\mathcal{N}(0, 1)|$, kept constant for the 2000 simulations for each sample size. These are also the variables in the skedastic equation, so that $\sigma_i^2 = \exp \gamma^T x_i$. We compare several values for the γ_j, with the all three values the same in each case. In the simulations the values of the coefficient β for the linear model are irrelevant, although in our simulations they were set equal to three. The additive errors ϵ_i come from the standard normal distribution.

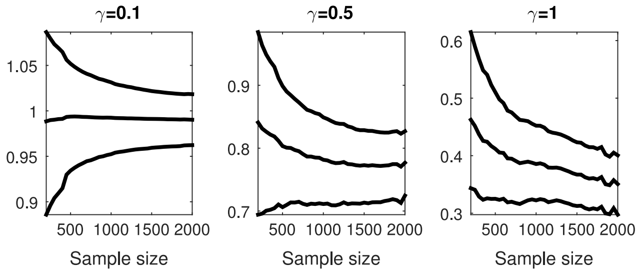

Fig. 12. Relative efficiency of OLS and Harvey's model (6) for estimation of parameters β as heteroskedasticity increases. The data are generated from Harvey's model. 2000 observations, smoothed output.

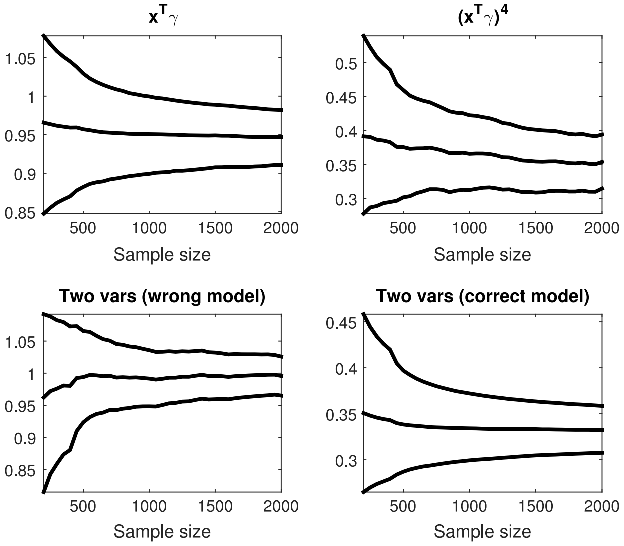

Fig. 13. Relative efficiency of OLS and Harvey's model (6) for estimation of parameters β when data are generated by some other model. Upper panels, $\sigma_i^2 = x^T\gamma$ and $(x^T\gamma)^4$. Lower panels, data generated in two groups with standard deviations in the ratio 1:10. This form of Harvey's model is fitted in the lower-right panel. 2000 observations, smoothed output.

The left-hand panel of Fig. 12 shows the relative efficiency (6) for slight heteroskedasticity generated with $\gamma = 0.1$ for sample sizes $200, 250, \ldots, 2000$. Also included on the plot are 10% and 90% simulation bands. For this slight heteroskedasticity the ratio of σ_i at all $x_i = 2$ relative to that when all $x_i = 0$ is 1.35. As the panel shows, the efficiency of OLS is virtually steady, with a value of 0.99 at $n = 2000$. For $\gamma = 0.5$ and 1.0, the ratios of standard deviations at all $x_i = 0$ and 2 are 4.448 and 20.09. However, with the folded normal distribution of the x_i, very few observations will be this extreme in all three variables. The centre and right panels of the plot show how the efficiency decreases for these larger values of γ. At $n = 2000$ the values are 0.78 and 0.35. A further interesting feature is that, as γ increases, the plots take longer to reach a steady value. Clearly the efficiency for $\gamma = 1$ is still decreasing at $n = 2000$.

We now consider what happens to the efficiency when Harvey's model is fitted, but the data are generated by a different heteroskedastic relationship. In the top left-hand panel of Fig. 13 the data are generated with the skedastic relationship $\sigma_i^2 = x^T\gamma$ for which OLS has an efficiency of at least 0.95. In the top right-hand panel, instead of $x^T\gamma$ we take $(x^T\gamma)^4$. Now OLS has the much lower efficiency of 0.35 when $n = 2000$.

Before discussing these results, we turn to the lower panels of Fig. 13. In the left-hand panel the heteroscedasticity is produced solely by a random half of the observations having additive errors with a standard deviation ten times that of the remainder. For this situation OLS has an efficiency very close to one (0.997) relative to fitting Harvey's model with the three explanatory variables which are unrelated to the variance. Finally, the lower right-hand panel shows what happens when one of the terms in Harvey's model is at two-levels for the two variances, thus perfectly modelling the heteroscedasticity. OLS then has an efficiency of 0.33.

The very surprising feature of these examples is that OLS is never more efficient than fitting Harvey's model. Fig. 12 shows the loss of efficiency when Harvey's model is correct. As the heteroskedasticity increases, the relative efficiency of OLS decreases. However, in the top panels of Fig. 13 Harvey's model, which is incorrect, is still more efficient than OLS. This arises as the heteroskedasticity in the generated data increase with x, although not in the way specified by the fitted model. Nevertheless, this fitted model provides a useful, and in the case of the right-hand panel, a very useful, approximation to the heteroskedasticity. In the bottom left-hand panel, with just two variances, OLS is no more efficient than Harvey's model.

But, when the two groups are correctly modelled, the efficiency of OLS falls to 0.33. Further simulations show that, when the ratio of standard deviations is 1:50, fitting the wrong model leaves the efficiency at one, but, when the correct form of Harvey's model is fitted, the efficiency of OLS drops to 0.078. The overall conclusion is that, over a wide range of scenarios, we have found that fitting a model for heteroskedasticity, even if it is incorrect, is never less efficient than OLS and usually much more efficient.

The use of functions of residuals in weighted regression, including the weights used by White in (5), is discussed by many authors, including Carroll and Ruppert (1982), who explore slight misspecifications in the skedastic model. We however use a fully parameterized model to provide an efficient method for very robust heteroskedastic regression.

8. Computation

All the new routines for the analysis of heteroskedastic data described in this paper have been written in MATLAB and have been integrated inside the FSDA toolbox for MATLAB, owned jointly by the University of Parma and the Joint Research Centre of the European Commission (Riani et al., 2012). This new software library, which extends MATLAB and its Statistics Toolbox to support the robust and efficient analysis of complex datasets, affected by different sources of heterogeneity, is freely downloadable from the websites http://www.riani.it/MATLAB and http://fsda.jrc.ec.europa.eu. All routines are publicly available and do not call `dll` or external code.

Each file contains a set of readily executable examples which can be immediately executed. Particular attention has been devoted to profile each segment of code in order to choose the fastest option. To run a full forward search and produce, for example, the plot in Fig. 2 (which requires the estimation of parameters about 1100 times) takes less than 15 s using an Intel(R) Core(TM) i7-4900MQ CPU 2.80 GHz (8 CPUs). With the same number (1100) of observations, but ten explanatory variables, the time increases to 20 s. Increasing the number of observations to 2000, but still with 10 explanatory variables causes an increase to 85 s. In these timings the number of variables in the regression model and in the skedastic equation are identical.

The convention adopted inside the toolbox is to add the letter "H" at the end of the traditional routine for homoskedastic data. For example, routine `FSRmdr`, which computes the minimum deletion residuals for homoskedastic errors, becomes `FSRHmdr` for heteroskedastic errors. Similarly, routine `FSR` (forward search in regression) which computes the automatic procedure for outlier detection using the rules described in Appendix A, has become `FSRH`. For each heteroskedastic specification the default is to use an estimation procedure based on the scoring algorithm. However, with just a single explanatory variable it is also possible to use a grid search algorithm. We have provided in the header of each `.m` file a full description of the input and output arguments of each routine and a corresponding HTML documentation which is fully integrated in the help system of the latest release of MATLAB, namely 2016a. All forward plots produced are brushable. That is, use of our optional argument `databrush`, makes it possible to select a set of steps during the forward search and to see the units which enter the subset in those steps highlighted in the scatter plot matrix of y against each column of X.

Finally, all datasets used in this paper and all those dealing with heteroskedastic data contained in the various editions of the book of Greene (we cite the 5th and 7th) have been added to the repository of regression datasets contained in the FSDA toolbox.

Appendix A. The scoring algorithm

With the skedastic equation given by (1) the loglikelihood of the n observations is $L(\beta, \sigma^2, \gamma)$

$$= -\frac{1}{2}\sum_{i=1}^{n}\left\{\log(2\pi) + \log\sigma_i^2 + (y_i - x_i^T\beta)^2/\sigma_i^2\right\}$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\left(\log\{2\pi\} + \log\sigma^2 + \log\{1+\exp(z_i^T\gamma)\} + (y_i - x_i^T\beta)^2/\left[\sigma^2\{1+\exp(z_i^T\gamma)\}\right]\right). \qquad (A.1)$$

Because of the block diagonal nature of the information matrix, we only require derivatives w.r.t. γ. The score vector is

$$S(\gamma) = \frac{\partial L}{\partial\gamma} = -\frac{1}{2}\sum_{i=1}^{n} z_i\left[\frac{\exp(z_i^T\gamma)}{1+\exp(z_i^T\gamma)} - \frac{(y_i - x_i^T\beta)^2}{\sigma^2}\frac{\exp(z_i^T\gamma)}{\{1+\exp(z_i^T\gamma)\}^2}\right]$$

$$= \frac{1}{2}\sum_{i=1}^{n}\frac{z_i\exp(z_i^T\gamma)}{1+\exp(z_i^T\gamma)}\left[\frac{(y_i - x_i^T\beta)^2}{\sigma^2\{1+\exp(z_i^T\gamma)\}} - 1\right].$$

A second differentiation yields the observed information as

$$\mathcal{I}(\gamma) = -\frac{\partial^2 L}{\partial\gamma\partial\gamma^T} = \frac{1}{2}\sum_{i=1}^{n} z_i z_i^T\left[\frac{\exp(z_i^T\gamma)}{\{1+\exp(z_i^T\gamma)\}^2} + \frac{(y_i - x_i^T\beta)^2}{\sigma^2\{1+\exp(z_i^T\gamma)\}^3}\exp(z_i^T\gamma)\{\exp(z_i^T\gamma) - 1\}\right].$$

To obtain the expected information we find $E(y_i - x_i^T\beta)^2$ which yields

$$\begin{aligned} I(\gamma) &= E\{\mathcal{I}(\gamma)\} = \frac{1}{2}\sum_{i=1}^{n} z_i z_i^T \left[\frac{\exp(z_i^T\gamma)}{\{1+\exp(z_i^T\gamma)\}^2} + \frac{\exp(z_i^T\gamma)\{\exp(z_i^T\gamma)-1\}}{\{1+\exp(z_i^T\gamma)\}^2}\right] \\ &= \frac{1}{2}\sum_{i=1}^{n} z_i z_i^T \left[\frac{\exp(z_i^T\gamma)}{1+\exp(z_i^T\gamma)}\right]^2 \\ &= \frac{1}{2}\sum_{i=1}^{n} z_i z_i^T \left[\frac{1}{1+\exp(-z_i^T\gamma)}\right]^2. \end{aligned}$$

For the scoring algorithm, it is convenient to write

$$q_i = \left[\frac{\exp(z_i^T\gamma)}{\{1+\exp(z_i^T\gamma)\}}\right]^2 \quad \text{and} \quad Q = \text{diag}\{q_i\}.$$

The parameter estimate at iteration $k+1$ follows from that at iteration k by using the Fisher scoring algorithm

$$\begin{aligned} \gamma_{k+1} &= \gamma_k + \delta I(\gamma_k)^{-1} S(\gamma_k) \\ &= \gamma_k + \delta(Z^T Q_k Z)^{-1}\sum_{i=1}^{n}\frac{z_i\exp(z_i^T\gamma_k)}{1+\exp(z_i^T\gamma_k)}\left[\frac{(y_i - x_i^T\beta_k)^2}{\sigma_k^2\{1+\exp(z_i^T\gamma_k)\}} - 1\right], \end{aligned} \tag{A.2}$$

since the 2 and 1/2 cancel. Here $\delta(<1)$ is a step-length parameter, intended to avoid divergence of the algorithm.

It is clear from (A.2) that the updated value of $\hat{\gamma}$ is computed by adding the vector of estimated coefficients in the least squares regression of $(y_i - x_i^T\beta_k)^2/[\sigma_k^2\{1+\exp(z_i^T\gamma_k)\}] - 1$ on $z_i^W = z_i\exp(z_i^T\gamma_k)/\{1+\exp(z_i^T\gamma_k)\}$. Convergence occurs when the derivative is zero. Given γ_{k+1}, the parameters β and σ^2 are re-estimated by weighted least squares and the algorithm continues until sufficient accuracy is obtained. In our computations we stop when

$$\|d_{k+1} - d_k\|^2/\|d_k\|^2 < 10^{-8},$$

where $d_k = (\beta_k^T\ \gamma_k^T)^T$. The remaining detail is to determine the starting values for the iteration. In the absence of prior information the simplest procedure is to use OLS for β and to take γ as the coefficients from the regression of $ne_i^2/\sum_{i=1}^n e_i^2 - 1$ on z_i, where e_i is the raw residual from OLS regression.

Appendix B. Detection of a signal in the FS for regression

The envelopes in plots such as Fig. 2 give the pointwise distribution of the absolute minimum deletion residuals $|r_{\text{imin}}(m)|$ defined in Section 3. The following rules provide for the detection of a signal, indicating the presence of one or more outliers, for a test with an approximate samplewise size of 1%.

There are four conditions, the fulfilment of any one of which leads to the detection of a signal.

- In the central part of the search we require 3 consecutive values of $|r_{\text{imin}}(m)|$ above the 99.99% envelope or 1 above 99.999%;
- In the final part of the search we need two consecutive values of $|r_{\text{imin}}(m)|$ above 99.9% and 1 above 99%;
- $|r_{\text{imin}}(n-2)| > 99.9\%$ envelope;
- $|r_{\text{imin}}(n-1)| > 99\%$ envelope. In this case a single outlier is detected and the procedure terminates.

The final part of the search is defined as $m \geq n - \left[13\,(n/200)^{0.5}\right]$, where here [ ] stands for a rounded integer. For $n = 1100$ the value is slightly less than 3% of the observations.

References

Atkinson, A.C., Riani, M., 2007. Exploratory tools for clustering multivariate data. Comput. Statist. Data Anal. 52, 272–285. http://dx.doi.org/10.1016/j.csda.2006.12.034.

Atkinson, A.C., Riani, M., Cerioli, A., 2010. The forward search: theory and data analysis (with discussion). J. Korean Stat. Soc. 39, 117–134. http://dx.doi.org/10.1016/j.jkss.2010.02.007.

Carroll, R.J., Ruppert, D., 1982. A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. J. Amer. Statist. Assoc. 77, 878–882.

Carroll, R.J., Ruppert, D., 1988. Transformation and Weighting in Regression. Chapman and Hall, London.

Cerioli, A., Farcomeni, A., Riani, M., 2014. Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter. J. Multivariate Anal. 126, 167–183.

Cheng, T.-C., 2011. Robust diagnostics for the heteroscedastic regression model. Comput. Statist. Data Anal. 55, 1845–1866.

Davidian, M., Carroll, R.J., 1987. Variance function estimation. J. Amer. Statist. Assoc. 82, 1079–1091.

Economist, 2014. Uncontained. Economist, UK Edn 411 (8885), 59–60.

Fedorov, V.V., Leonov, S.L., 2014. Optimal Design for Nonlinear Response Models. Chapman and Hall/ CRC Press, Boca Raton.

Greene, W.H., 2002. Econometric Analysis, fifth ed. Macmillan, New York.

Greene, W.H., 2012. Econometric Analysis, seventh ed. Prentice Hall, Upper Saddle River, NJ.

Hampel, F.R., 1975. Beyond location parameters: robust concepts and methods. Bull. Int. Statist. Inst. 46, 375–382.

Harvey, A.C., 1976. Estimating regression models with multiplicative heteroscedasticity. Econometrica 44, 461–465.

Johansen, S., Nielsen, B., 2016a. Analysis of the Forward Search using some new results for martingales and empirical processes. Bernoulli 21, 1131–1183.

Johansen, S., Nielsen, B., 2016b. Asymptotic theory of outlier detection algorithms for linear time series regression models. Scand. J. Statist. 43, 321–348.

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

Neykov, N.M., Filzmoser, P., Neytchev, P.N., 2012. Robust joint modeling of mean and dispersion through trimming. Comput. Statist. Data Anal. 56, 34–48.

Riani, M., Atkinson, A.C., Cerioli, A., 2009. Finding an unknown number of multivariate outliers. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 447–466.

Riani, M., Atkinson, A.C., Perrotta, D., 2014a. A parametric framework for the comparison of methods of very robust regression. Statist. Sci. 29, 128–143.

Riani, M., Cerioli, A., Atkinson, A.C., Perrotta, D., 2014b. Monitoring robust regression. Electron. J. Stat. 8, 642–673.

Riani, M., Perrotta, D., Torti, F., 2012. FSDA: a MATLAB toolbox for robust analysis and interactive data exploration. Chemometr. Intell. Lab. Syst. 116, 17–32. http://dx.doi.org/10.1016/j.chemolab.2012.03.017.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Amer. Statist. Assoc. 79, 871–880.

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

Tallis, G.M., 1963. Elliptical and radial truncation in normal samples. Ann. Math. Statist. 34, 940–944.

Welsh, A.H., Carroll, R.J., Ruppert, D., 1994. Fitting heteroscedastic regression models. J. Amer. Statist. Assoc. 89, 100–116.

White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.