Discussion Papers
Department of Economics
University of Copenhagen

No. 10-06

Discussion of **The Forward Search: Theory and Data Analysis,** by Anthony C. Atkinson, Marco Riani, and Andrea Ceroli

Søren Johansen, Bent Nielsen

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (online)

Discussion of **The Forward Search: Theory and Data Analysis** by Anthony C. Atkinson, Marco Riani, and Andrea Ceroli

Søren Johansen*
Department of Economics, University of Copenhagen
and CREATES, University of Aarhus
and
Bent Nielsen†
Department of Economics, University of Oxford

February 6, 2010

Abstract

The Forward Search Algorithm is a statistical algorithm for obtaining robust estimators of regression coefficients in the presence of outliers. The algorithm selects a succession of subsets of observations from which the parameters are estimated. The present note shows how the theory of empirical processes can contribute to the understanding of how the subsets are chosen and how the sequence of estimators is changing.

Keywords: Empirical processes, Huber's skip, least trimmed squares estimator, one–step estimator, outlier robustness.

JEL Classification: C2

*Address: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K. Denmark, Email: Soren.Johansen@econ.ku.dk. The author gratefully acknowledges support from Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation.

†Address: Nuffield College, Oxford OX1 1NF, UK, Email: bent.nielsen@nuffield.ox.ac.uk

1 Introduction

The paper by Atkinson, Riani and Ceroli, henceforth ARC, is concerned with detection of outliers and unsuspected structures which is rather important in practice. This is done through a Forward Search Algorithm. The statistical analysis of such algorithms poses many challenging problems, and we would like to contribute to the theory of the algorithm in this discussion.

We establish some results in a simple case for a single iteration of the algorithm using empirical process theory. This would then have to be extended to more models, and developed further to understand the properties of the full algorithm. The established results suggest that the heuristic results from ARC could be correct if the parameters were known, but not when the parameters are estimated.

A general reference for empirical process theory is the monograph by Koul (2002) which analyses weighted empirical processes. Some further developments are made in Johansen and Nielsen (2009) which we exploit here. For simplicity we only consider a simple location-scale problem but the results would generalize to regressions and time-series regressions.

The discussion is organized so that the algorithm is described in §2. Some potential results are described in §3. The analysis of a single step of the algorithm is then provided for the known parameter case in §4 and for the unknown parameter case in §5. We conclude in §6 and leave some proofs to an Appendix.

2 The algorithm

We consider the regression problem $y_i = \mu + \sigma\varepsilon_i, i = 1, \ldots, n$, where the i.i.d. ε_i follow a known distribution function F and symmetric density f, with mean zero and variance 1. The distribution is assumed to be continuous and to satisfy some regularity conditions for smoothness which are met by the standard normal distribution, see Johansen and Nielsen (2009, Assumption A). The absolute standardized error $|\varepsilon_i|$ has distribution function G satisfying $\mathsf{G}(u) = \mathsf{P}(|\varepsilon_i| \leq u) = \mathsf{F}(u) - \mathsf{F}(-u)$, and $u_\psi = \mathsf{G}^{-1}(\psi)$ is the ψ-quantile of G and its density is $\mathsf{g}(u) = 2\mathsf{f}(u)$.

2.1 Description of the algorithm

We start with some initial robust location estimator $\hat{\mu}$ and an initial observation set of size m_0. This set is constructed by calculating absolute residuals $\hat{r}_i = |y_i - \hat{\mu}|$, finding their order statistics $\hat{r}_{(i)}$, and defining the initial observation index set of size m_0 as the m_0 observations closest to $\hat{\mu}$, that is

$$S_*^{(m_0)} = \{i : |y_i - \hat{\mu}| \leq \hat{r}_{(m_0)}\}.$$

The algorithm then proceeds in the steps

1. Given an index set $S_*^{(m)}$ calculate estimators

$$\hat{\mu}^{(m)} = m^{-1}\sum_{i\in S_*^{(m)}} y_i, \qquad (\hat{\sigma}^{(m)})^2 = m^{-1}\sum_{i\in S_m}(y_i - \hat{\mu}^{(m)})^2,$$

residuals $\hat{r}_i^{(m)} = |y_i - \hat{\mu}^{(m)}|$, and their order statistics $\hat{r}_{(i)}^{(m)}$ for $i = 1, \ldots, n$.

2. Test whether the residual nearest to the observations in $S_*^{(m)}$ does not correspond to an outlier. In ARC the test is based upon

$$u_{test}^{(1)} = \min_{i\notin S_*^{(m)}} \hat{r}_i^{(m)}/\hat{\sigma}^{(m)},$$

but we suggest to use

$$u_{test}^{(2)} = \hat{r}_{(m+1)}^{(m)}/\hat{\sigma}^{(m)}.$$

3. In the next step we either continue with step 1 or stop the algorithm.

   (a) If the test based on the nearest residual does not reject, then define

$$S_*^{(m+1)} = \{i : \hat{r}_i^{(m)} \leq \hat{r}_{(m+1)}^{(m)}\}$$

   and return to 1.

   (b) If the test rejects, then set $\hat{m} = m$ and define the terminal estimators $\hat{\mu}_n = \hat{\mu}^{(\hat{m})}$, $\hat{\sigma}_n = \hat{\sigma}^{(\hat{m})}$ and the observation set $\hat{S}_n = S_*^{(\hat{m})}$.

An important problem is to determine distributions of test statistic and estimators in the case of a sample without outliers.

ARC suggest that the initial estimator $\hat{\mu}$ could be chosen as the least trimmed squares estimator, see Rousseeuw (1984). This is constructed by choosing some $m > n/2$ and finding

$$\hat{\mu}^{(LTS,n,m)} = \arg\min_{\mu} \sum_{i=1}^{n} (r_i^{(\mu)})^2 1_{(r_i^{(\mu)} \leq r_{(m)}^{(\mu)})}, \qquad r_i^{(\mu)} = |y_i - \mu|. \tag{2.1}$$

2.2 Comments on the choice of test statistic

Comment 2.1 The motivation for the test statistic $\hat{\sigma}^{(m)} u_{test}^{(2)} = \hat{r}_{(m+1)}^{(m)}$ is that it will be the largest of the residuals with index in $S_*^{(m+1)}$. The rank of the observation $\hat{r}_{(m+1)}^{(m)}$ may, however, not enter $S_*^{(m)}$.

Comment 2.2 The index sets $S_*^{(m+1)}$ are constructed independently of the choice of test statistic, $u_{test}^{(1)}$ or $u_{test}^{(2)}$.

Comment 2.3 In general the test statistics $u_{test}^{(1)}$ and $u_{test}^{(2)}$ will be different. Three results follow.

1. the statistic $\hat{\sigma}^{(m)} u_{test}^{(1)} = \min_{i \notin S_*^{(m)}} \hat{r}_i^{(m)}$ suggested by ARC is not an order statistic of the residuals $r_i^{(m)}$, because $S_*^{(m)}$ is based on the previous set of residuals $\hat{r}_i^{(m-1)}$.

2. it holds $u_{test}^{(1)} \leq u_{test}^{(2)}$. Indeed, if $S_*^{(m)}$ is the ranks of $r_{(1)}^{(m)}, \ldots, r_{(m)}^{(m)}$ then these statistics are equal. If $S_*^{(m)}$ does not have this form then the complement of $S_*^{(m)}$ must include one of the ranks of $r_{(1)}^{(m)}, \ldots, r_{(m)}^{(m)}$.

3. The difference between the two test statistics relates to the distance $|\hat{\mu}^{(m-1)} - \hat{\mu}^{(m)}|$ and is therefore likely to be unimportant.

As a numerical example consider the data set

$$(-13, -8, -5, 5, 6, 7, 8).$$

The initial estimator is chosen as the sample average $\hat{\mu} = \bar{y} = 0$. The absolute residuals are then

$$(\hat{r}_i)_{i=1}^n = (13, 8, 5, 5, 6, 7, 8).$$

An initial index set of size 3 is then the ranks $S_*^{(3)} = (3, 4, 5)$ pointing at the observations $(-5, 5, 6)$. Now, in the first step of the algorithm compute the estimator and absolute residuals

$$\hat{\mu}^{(3)} = 2, \qquad (\hat{r}_i^{(3)})_{i=1}^n = (15, 10, 7, 3, 4, 5, 6).$$

Then $u_{test}^{(1)}$ is based on

$$\min_{i \notin S_*^{(3)}} \hat{r}_i^{(3)} = \min(15, 10, 5, 6) = 5,$$

whereas $u_{test}^{(2)}$ is based on the order statistic $\hat{r}_{(4)}^{(3)} = 6$. Regardless of the choice of test statistic the updated index set is $S_*^{(4)} = (4, 5, 6, 7)$ pointing at the observations $(5, 6, 7, 8)$. Note that

$$(3, 4, 5) = S_*^{(3)} \not\subset S_*^{(4)} = (4, 5, 6, 7),$$

so the sets $S_*^{(m)}$ are not in general increasing.

3 Some potential results

In order that the Forward Search Algorithm can be applied with confidence it is important to derive the distributions of the test statistics and the estimators. It would also be of interest to see if the sets $S_*^{(m)}$ are monotone in m.

3.1 Consistency

One would like to have a result as the following

Potential Theorem 3.1 *If the initial estimator $(\hat{\mu}, \hat{\sigma}^2)$ is consistent for $n \to \infty$, and if $m = \psi n + \mathrm{O}(1)$ for $0 < \psi < 1$, or if $m = \hat{m}$, then it holds that $\hat{\mu}^{(m)}$ and $\hat{\sigma}^{(m)}$ have a probability limit.*

Comment 3.1 We do not have a general proof of such a result, but some examples are given in §5. One can note here that the difference between order statistics from i.i.d. observations are of the order of $\mathrm{O_P}(n^{-1})$, so averaging order statistics which are that close to the initial (consistent estimator) will at most give a deviation from this of the order of $\mathrm{O_P}(n^{-1})$, and hence should not disturb consistency. If we could find the probability limit of $\hat{\mu}^{(m)}$ and $\hat{\sigma}^{(m)}$, we could correct the estimators to give consistent estimators, see ARC §3.3, and Theorem 5.3.

3.2 Monotonicity

The next result is alluded to in a number of places in ARC, even though it is realized that it does not hold without some conditions.

Potential Theorem 3.2 *The sets $S_*^{(m)}$ are monotone*

$$S_*^{(m)} \subset S_*^{(m+1)}, \ m = 2, \ldots, \hat{m}.$$

Comment 3.2 As it stands it is unfortunately false, as seen in the example in Comment 2.3. If in general we have found $S_*^{(m)}$ for some m, and want to add one point to $S_*^{(m)}$, call it y_{m+1}, then the new average becomes

$$\bar{y}^{(m+1)} = \bar{y}^{(m)} + \frac{1}{1+m}(y_{m+1} - \bar{y}^{(m)}).$$

Thus, if for instance $y_{m+1} > \bar{y}^{(m)}$, the average is moved up by $(y_{m+1} - \bar{y}^{(m)})/(m+1)$ which is of the order of $m^{-1} \approx n^{-1}$. The distance between order statistics is of the order of n^{-1}, so if one of the observations in $S^{(m)}$ is close to the lower boundary of the band defining $S_*^{(m)}$, then it could easily happen that it falls outside when the band is moved up by $(y_{m+1} - \bar{y}^{(m)})/(m+1)$. Thus a point can leave the band when another is added even for large m. This event may have small probability, however, so the following result could hold, but we have no proof.

Potential Theorem 3.3 *The sets $S_*^{(m)}$ are monotone with probability tending to one, that is for $m = \psi n + \mathrm{O}(1)$ as $n \to \infty$ then*

$$\mathsf{P}(S_*^{(m)} \subset S_*^{(m+1)}) \to 1,$$

or, perhaps,

$$\frac{1}{n - m_0 + 1} \sum_{m=m_0}^{n} 1_{(S_*^{(m)} \subset S_*^{(m+1)})} \xrightarrow{\mathsf{P}} 1.$$

3.3 The test statistics

In ARC it is suggested to find the distribution of the test statistics by simulation, and it is argued that it could be costly measured by computer time. An approximation is suggested in §3.3 of ARC, using the theory of order statistics, and in a previous paper (Atkinson and Riani, 2006) another approximation based on order statistics is suggested.

One would like to show that the test statistics are asymptotically normal.

Potential Theorem 3.4 *If the initial estimators* $(\hat{\mu}, \hat{\sigma})$ *are consistent for* $n \to \infty$, *and* $m = \psi n + \mathrm{O}(1)$, *or* $m = \hat{m}$, *and if* $(\hat{\mu}^{(m)}, \hat{\sigma}^{(m)})$ *is consistent, then the test statistic*

$$n^{1/2}\{(\hat{\sigma}^{(m)})^{-1}\hat{r}^{(m)}_{(m+1)} - u_\psi\}$$

is asymptotically normally distributed.

Comment We can prove this result under suitable conditions and we have collected these contributions in §5, where we outline a general strategy for finding these limit distributions.

4 The case of known location and scale

When the parameters are known, the residuals are $r_i = |y_i - \mu| = \sigma|\varepsilon_i|$, which we order as $r_{(1)} \le \cdots \le r_{(m)}$. Then

$$S^{(m)}_* = \{i : |u_i - \mu| \le r_{(m)}\}, \qquad m = 2, 3, \ldots, \hat{m}.$$

In this case we clearly have $S^{(m)}_* \subset S^{(m+1)}_*$ so Theorem 3.2 is correct.

The empirical distribution function of $|\varepsilon_i| = |y_i - \mu|/\sigma$ is denoted

$$\mathsf{G}_n(u) = n^{-1}\sum_{i=1}^{n} 1_{(|\varepsilon_i| \le u)}.$$

The order statistics $r_{(m)}$ have the well known relation

$$\sigma^{-1}r_{(m)} \le u \Leftrightarrow \frac{m}{n} \le \mathsf{G}_n(u) \text{ since } \mathsf{G}_n(\sigma^{-1}r_{(m)}) = \frac{m}{n}, \tag{4.1}$$

which transforms expressions in order statistics into expression involving the empirical distribution function.

Moreover, $u^{(1)}_{test} = u^{(2)}_{test} = \sigma^{-1}r_{(m)}$ and the distribution is given by the expression (5) in ARC by applying (4.1) as

$$\mathsf{P}(\sigma^{-1}r_{(m)} \le u) = \mathsf{P}\{\frac{m}{n} \le \mathsf{G}_n(u)\} = \textstyle\sum_{j=m}^{n}\binom{n}{j}\mathsf{G}(u)^j\{1 - \mathsf{G}(u)\}^{n-j}. \tag{4.2}$$

Thus, the distribution suggested as an approximation in ARC would in fact be the exact distribution if the parameters were known.

In Atkinson and Riani (2006) another approximation to the distribution of the test statistic is suggested. It is argued that if m is proportional to n, then we are essentially working in a truncated distribution where $100\psi\%$ have been included in the sample. Thus, it is suggested to approximate the distribution of the test statistic $u_{test}^{(j)} = \sigma^{-1} r_{(m)}$ by the distribution of the largest of m observations, z_i say, where z_i are drawn from the truncated distribution $\mathsf{G}(u)/\psi$ for $0 \leq u \leq u_\psi$. By the same methodology as in §3.3 of ARC we then get the approximate result

$$\mathsf{P}(\sigma^{-1} r_{(m)} \leq u) \approx \mathsf{P}(\max_{i \leq m} z_i \leq u) = \left\{\psi^{-1}\mathsf{G}(u)\right\}^m = \left[\left\{\psi^{-1}\mathsf{G}(u)\right\}^\psi\right]^n, \tag{4.3}$$

which is suggested as an approximation, which is clearly different from (4.2).

Yet another way of arguing is that when we have already used $r_{(1)}, \ldots, r_{(m)}$ to construct $S_*^{(m)}$ and the estimator $\hat{\mu}^{(m)}$, then significance of $r_{(m+1)}$ could be evaluated in the conditional distribution given $r_{(1)}, \ldots, r_{(m)}$, and that is given by

$$\mathsf{P}(\sigma^{-1} r_{(m+1)} \geq u | r_{(1)}, \ldots, r_{(m)}) = \left\{\frac{1 - \mathsf{G}(u)}{1 - \mathsf{G}(\sigma^{-1} r_{(m)})}\right\}^{n-m}, \qquad u \geq \sigma^{-1} r_{(m)}.$$

This distribution can be interpreted as the distribution of the smallest observation among $n - m$ observations $r_i = |y_i - \mu|$ with density $\sigma^{-1}\mathsf{g}(\sigma^{-1} r)/\{1 - \mathsf{G}(\sigma^{-1} r_{(m)})\}$ for $r \geq r_{(m)}$ and could be used for a conditional test of the next residual given what has already been used.

5 The case of unknown location and scale

We assume we have $n^{1/2}$-consistent estimators $(\hat{\mu}, \hat{\sigma}^2) = (\mu, \sigma^2) + \mathrm{O}_\mathsf{P}(n^{-1/2})$ and define $\hat{r}_i = |y_i - \hat{\mu}|$, with order statistics $\hat{r}_{(i)}$. We define $\tilde{r} = \hat{r}_{(m)}$ and

$$\tilde{\mu} = m^{-1} \sum_{i=1}^n y_i 1_{(|y_i - \hat{\mu}| \leq \tilde{r})}, \tag{5.1}$$

$$\tilde{\sigma}^2 = m^{-1} \sum_{i=1}^n (y_i - \tilde{\mu})^2 1_{(|y_i - \hat{\mu}| \leq \tilde{r})}. \tag{5.2}$$

We find stochastic expansions and limit distributions of $\tilde{r}$ and the one-step estimators $(\tilde{\mu}, \tilde{\sigma}^2)$, which are based upon the m observations closest to the initial estimator $\hat{\mu}$.

When applied to the first iteration of the algorithm then $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{r}_{(m)}$ represent the initial estimator $\hat{\mu}$ and residual $\hat{r}_{(m)}$ along with some suitable variance estimator so $\tilde{\mu} = \hat{\mu}^{(m_0)}$ and $\tilde{\sigma} = \hat{\sigma}^{(m_0)}$.

When applied to a later iteration of the algorithm then $\hat{\mu}$, $\hat{\sigma}$, $\hat{r}_{(m)}$ represent $\hat{\mu}^{(m-1)}$, $\hat{\sigma}^{(m-1)}$, $\hat{r}_{(m)}^{(m-1)}$ so $\tilde{\mu} = \hat{\mu}^{(m)}$ and $\tilde{\sigma} = \hat{\sigma}^{(m)}$.

5.1 The asymptotic expansions

We first find an expansion of the test statistic $u_{test}^{(2)} = \hat{\sigma}^{-1}\tilde{r}$ which can be applied to find the asymptotic distribution of the test for different choices of estimators $(\hat{\mu}, \hat{\sigma}^2)$, and then give expansions of the one-step estimators $\tilde{\mu}$ and $\tilde{\sigma}^2$, which show how the estimator is changed from initial estimators $(\hat{\mu}, \hat{\sigma}^2)$ to one-step estimators $(\tilde{\mu}, \tilde{\sigma}^2)$. The proofs are given in the Appendix.

Theorem 5.1 *If* $m = \psi n + \mathrm{O}(1)$ *and if* $(\hat{\mu} - \mu, \hat{\sigma}^2 - \sigma^2) \in \mathrm{O}_{\mathsf{P}}(n^{-1/2})$ *then*

$$n^{1/2}(\hat{\sigma}^{-1}\tilde{r} - u_\psi) = -\frac{1}{2\mathsf{f}(u_\psi)} n^{-1/2} \textstyle\sum_{i=1}^{n} \{1_{(|\varepsilon_i| \le u_\psi)} - \psi\} \\ - \frac{u_\psi}{2} n^{1/2}(\sigma^{-2}\hat{\sigma}^2 - 1) + \mathrm{o}_{\mathsf{P}}(1).$$

Theorem 5.2 *If* $m = \psi n + \mathrm{O}(1)$ *and if* $(\hat{\mu} - \mu, \hat{\sigma}^2 - \sigma^2) \in \mathrm{O}_{\mathsf{P}}(n^{-1/2})$ *then the estimator* $\tilde{\mu}$ *defined in (5.1) satisfies*

$$n^{1/2}(\tilde{\mu} - \mu) = \frac{\sigma}{\psi} n^{-1/2} \textstyle\sum_{i=1}^{n} \varepsilon_i 1_{(|\varepsilon_i| < u_\psi)} + \frac{2u_\psi \mathsf{f}(u_\psi)}{\psi} n^{1/2}(\hat{\mu} - \mu) + \mathrm{o}_{\mathsf{P}}(1).$$

Theorem 5.3 *If* $m = \psi n + \mathrm{O}(1)$ *and if* $(\hat{\mu} - \mu, \hat{\sigma}^2 - \sigma^2) \in \mathrm{O}_{\mathsf{P}}(n^{-1/2})$ *then the estimator* $\tilde{\sigma}^2$ *defined in (5.2) satisfies* $\tilde{\sigma}^2 \xrightarrow{\mathsf{P}} \psi^{-1}\tau_2^{u_\psi}$, *where* $\tau_2^{u_\psi} = \int_{-u_\psi}^{u_\psi} u^2 \mathsf{f}(u) du$.
We therefore define $\tilde{\sigma}^2_{corrected} = \psi\tilde{\sigma}^2/\tau_2^{u_\psi}$. *It holds*

$$n^{1/2}\{\sigma^{-2}\tilde{\sigma}^2_{corrected} - 1\} = (\tau_2^{u_\psi})^{-1} n^{-1/2} \textstyle\sum_{i=1}^{n} (\varepsilon_i^2 - \psi^{-1}\tau_2^{u_\psi}) 1_{(|\varepsilon_i| < u_\psi)} \\ - (\tau_2^{u_\psi})^{-1}(u_\psi^4 - \psi^{-1}\tau_2^{u_\psi}) n^{-1/2} \textstyle\sum_{i=1}^{n} \{1_{(|\varepsilon_i| \le u_\psi)} - \psi\} + \mathrm{o}_{\mathsf{P}}(1).$$

Comment 5.1 Note that these results are all derived for symmetric distributions. If this assumption is dropped, a bias term will appear in some asymptotic distributions and terms including $n^{1/2}(\hat{\mu} - \mu)$ will appear in Theorems 5.1, 5.3.

5.2 Examples

We illustrate these results by finding the asymptotic distribution of the test statistic for different choices of initial estimators $(\hat{\mu}, \hat{\sigma}^2)$.

First, for comparison we give the result for the test statistic for known parameters, where we re-discover a classical result on the asymptotic distribution of order statistics, see for instance David (1981, Theorem 9.2, p. 255).

Corollary 5.1 *If* $m = \psi n + \mathrm{O}(1)$ *and if* $\hat{\mu} = \mu, \hat{\sigma} = \sigma$ *then*

$$n^{1/2}(\sigma^{-1}\tilde{r} - u_\psi) \xrightarrow{\mathsf{D}} \mathsf{N}[0, \frac{\psi(1-\psi)}{\{2\mathsf{f}(u_\psi)\}^2}].$$

Proof of Corollary 5.1. In the expansion of Theorem 5.1 the second term drops out since $\hat{\sigma} = \sigma$. Then apply the Central Limit Theorem to the first term. ■

Alternatively, initial estimators could be chosen as full sample estimators $\hat{\mu} = n^{-1}\sum_{i=1}^{n} y_i$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. This changes the limit distribution.

Corollary 5.2 *If $m = \psi n + \mathrm{O}(1)$ and if $(\hat{\mu}, \hat{\sigma}^2)$ are the full sample estimators then*

$$n^{1/2}(\hat{\sigma}^{-1}\tilde{r} - u_\psi) \xrightarrow{\mathsf{D}} \mathsf{N}[0, \frac{\psi(1-\psi)}{\{2\mathsf{f}(u_\psi)\}^2} + \frac{u_\psi^2}{2}\{1 + \frac{\tau_2^{u_\psi} - \psi}{u_\psi \mathsf{f}(u_\psi)}\}].$$

Proof of Corollary 5.2. Apply Theorem 5.1 by inserting the expansion

$$n^{1/2}(\sigma^{-2}\hat{\sigma}^2 - 1) = n^{-1/2}\textstyle\sum_{i=1}^{n}(\varepsilon_i^2 - 1) + \mathsf{o}_\mathsf{P}(1)$$

to get that $n^{1/2}(\hat{\sigma}^{-1}\tilde{r} - u_\psi)$ equals

$$-\frac{1}{2\mathsf{f}(u_\psi)} n^{-1/2}\textstyle\sum_{i=1}^{n}\{1_{(|\varepsilon_i| \le u_\psi)} - \psi\} - \frac{u_\psi}{2} n^{-1/2}\sum_{i=1}^{n}(\varepsilon_i^2 - 1) + \mathsf{o}_\mathsf{P}(1).$$

This is asymptotically normal with a variance as indicated. ■

Comment 5.2 In general we get a different limit distribution for the test statistic when the variance is estimated. Note the curious result that for the standard normal distribution we find

$$\tau_2^{u_\psi} = \int_{-u_\psi}^{u_\psi} \varepsilon^2 \mathsf{f}(\varepsilon) d\varepsilon = \int_{-u_\psi}^{u_\psi} \mathsf{f}(\varepsilon) d\varepsilon - [\varepsilon \mathsf{f}(\varepsilon)]_{\varepsilon=-u_\psi}^{\varepsilon=u_\psi} = \psi - 2u_\psi \mathsf{f}(u_\psi).$$

so the asymptotic variance in Corollary 5.2 becomes $\{2\mathsf{f}(u_\psi)\}^{-2}\psi(1-\psi) - u_\psi^2/2$, which is less than the variance we get for known parameters.

Finally, we shall see what happens when we choose $\hat{\mu}$ as the least trimmed squares estimator $\hat{\mu}^{(LTS,n,m)}$ defined in (2.1). A stochastic expansion of $\hat{\mu}^{(LTS,n,m)}$ is given by Víšek (2006, Theorem 1, p. 215) as

$$n^{1/2}(\hat{\mu}^{(LTS,n,m)} - \mu) = \frac{\sigma}{\psi - 2u_\psi \mathsf{f}(u_\psi)} n^{-1/2}\textstyle\sum_{i=1}^{n}\varepsilon_i 1_{(|\varepsilon_i| \le u_\psi)} + \mathsf{o}_\mathsf{P}(1). \qquad (5.3)$$

Corollary 5.3 *If $m = \psi n + \mathrm{O}(1)$ and $\hat{\mu} = \hat{\mu}^{(LTS,n,m)}$ we find the expansion*

$$n^{1/2}(\tilde{\mu} - \mu) = \frac{\sigma}{\psi - 2u_\psi \mathsf{f}(u_\psi)} n^{-1/2}\textstyle\sum_{i=1}^{n}\varepsilon_i 1_{(|\varepsilon_i| \le u_\psi)} + \mathsf{o}_\mathsf{P}(1),$$

so that the limit distribution is

$$n^{1/2}(\tilde{\mu} - \mu) \xrightarrow{\mathsf{D}} \mathsf{N}[0, \frac{\sigma^2 \tau_2^{u_\psi}}{\{\psi - 2u_\psi \mathsf{f}(u_\psi)\}^2}].$$

If F is standard normal then $\tau_2^{u_\psi} = \psi - 2u_\psi \mathsf{f}(u_\psi)$ so the variance is $\sigma^2/\tau_2^{u_\psi}$.

Proof of Corollary 5.3. Insert Víšek's expansion (5.3) in Theorem 5.2 to get

$$n^{1/2}(\tilde{\mu}-\mu) = \frac{\sigma}{\psi}\{1+\frac{2u_\psi \mathsf{f}(u_\psi)}{\psi-2u_\psi \mathsf{f}(u_\psi)}\}n^{-1/2}\textstyle\sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i|\le u_\psi)} + \mathsf{o_P}(1)$$
$$= \frac{\sigma}{\psi-2u_\psi \mathsf{f}(u_\psi)}n^{-1/2}\textstyle\sum_{i=1}^n \varepsilon_i 1_{(|\varepsilon_i|\le u_\psi)} + \mathsf{o_P}(1).$$

This converges to a normal distribution with the variance as indicated. ■

Comment 5.3 Thus, if the initial estimator $\hat{\mu}$ has the expansion of the least trimmed squares estimator, then so does the one-step estimator $\tilde{\mu}$. In this sense the least trimmed squares estimator is a "fixed point" in the mapping from the initial estimator to the one-step estimator. ARC do indeed suggest, possibly for other beneficial reasons, to start with the least trimmed squares estimator. A similar result holds if the initial estimator is the Huber skip, which has the same expansion and limit distribution as the least trimmed squares estimator, see Johansen and Nielsen (2009).

6 Some final comments

6.1 The simulation method

The above theoretical results indicate that the idea of judging significance using the exact theory of order statistics, seems to be fine if parameters are known. But if parameters are estimated the (asymptotic) distributions change, depending on the choice of initial estimator.

Thus it would be very helpful with some simulations of the algorithm, as it is used, to check if asymptotic distributions can describe the variation of estimators and tests. We have seen that different initial estimators give different (limit) distributions. There may also be a problem for very large ψ, where a different asymptotic theory may be needed.

6.2 Generalizations

The results of Johansen and Nielsen (2009) cover models with general fixed or random regressors, as well as time series regressions, both stationary and non-stationary. Tools to study the general theory of empirical processes for residuals of such models are outlined in Engle and Nielsen (2009). So it is possible to extend the theory of the Forward Search Algorithm much beyond what we have indicated in this discussion.

6.3 Algorithms for time series

In §7 of ARC it is suggested "However, many things remain to be developed in the application of the Forward Search in the time series context, such as ... the construction of an algorithm which can automatically distinguish among the different types of outliers and level shifts".

While it could prove very useful to develop extensions of the Forward Search to time series one should bear in mind that the algorithm *Autometrics* by Doornik (2009) building on the work of Hoover and Perez (1999) and the PcGets algorithm of Hendry and Krolzig (2005) has been developed for this purpose.

In any case, it is a bit surprising to see in ARC that the results suggested for regression analysis be applied to ozone data where the residuals are likely to be autocorrelated.

A Proofs of main theorems

The proofs exploit Theorem 1.17 of Johansen and Nielsen (2009); see also their equation 1.46. For $(\hat{a}, \hat{b}) = \mathrm{O}_{\mathsf{P}}(n^{-1/2})$ and $\ell = 0, 1, 2$, it was shown under regularity conditions that

$$n^{-1/2}\sum_{i=1}^{n}\varepsilon_i^{\ell}\{1_{(|\varepsilon_i-\hat{b}|\leq a+\hat{a})} - 1_{(|\varepsilon_i|\leq a)}\} = a^{\ell-1}n^{1/2}(\hat{a}\xi_{\ell+1}^{u_\psi} + \hat{b}\xi_{\ell}^{u_\psi}) + \mathrm{o}_{\mathsf{P}}(1), \tag{A.1}$$

where, for a symmetric density $\xi_{2j}^{u_\psi} = 0$ and $\xi_{2j+1}^{u_\psi} = 2u_\psi^{2j+1}\mathsf{f}(u_\psi)$ for $j = 0, 1, \ldots$

Proof of Theorem 5.1. In order to find the asymptotic distribution of the test statistic $\hat{\sigma}^{-1}\tilde{r}$ we expand as

$$n^{1/2}(\frac{\tilde{r}}{\hat{\sigma}} - u_\psi) = n^{1/2}\frac{\sigma}{\hat{\sigma}}\{(\frac{\tilde{r}}{\sigma} - u_\psi) - u_\psi(\frac{\hat{\sigma}}{\sigma} - 1)\}.$$

Since $\hat{\sigma}^2 - \sigma^2 = \mathrm{O}_{\mathsf{P}}(n^{-1/2})$ then $\sigma^{-1}\hat{\sigma} - 1 = (\sigma^{-2}\hat{\sigma}^2 - 1)/2 + \mathrm{o}_{\mathsf{P}}(n^{-1/2})$ and $\hat{\sigma}^{-1}\sigma = 1 + \mathrm{o}_{\mathsf{P}}(1)$. It follows that

$$n^{1/2}(\frac{\tilde{r}}{\hat{\sigma}} - u_\psi) = n^{1/2}\{(\frac{\tilde{r}}{\sigma} - u_\psi) - \frac{u_\psi}{2}(\frac{\hat{\sigma}^2}{\sigma^2} - 1)\}\{1 + \mathrm{o}_{\mathsf{P}}(1)\} + \mathrm{o}_{\mathsf{P}}(1). \tag{A.2}$$

The first term in (A.2), is now shown to be asymptotically normal

$$n^{1/2}(\frac{\tilde{r}}{\sigma} - u_\psi) \xrightarrow{\mathsf{D}} \mathsf{N}\{0, \frac{\psi(1-\psi)}{2\mathsf{f}(u_\psi)}\}. \tag{A.3}$$

The quantile $\tilde{r}$ satisfies $\widehat{\mathsf{G}}_n(\tilde{r}/\sigma) = m/n$ where $\widehat{\mathsf{G}}_n(r) = n^{-1}\sum_{i=1}^{n} 1_{(|\varepsilon_i - \sigma^{-1}(\hat{\mu}-\mu)|<r)}$ is an empirical distribution function; see (4.1). Thus it holds

$$\begin{aligned}\mathcal{P}_n \stackrel{\text{def}}{=} \mathsf{P}\{n^{1/2}(\frac{\tilde{r}}{\sigma} - u_\psi) \leq z\} &= \mathsf{P}\{\frac{\tilde{r}}{\sigma} \leq u_\psi + n^{-1/2}z\} \\ &= \mathsf{P}\{\frac{m}{n} \leq \widehat{\mathsf{G}}_n(u_\psi + n^{-1/2}z)\} = \mathsf{P}(0 \leq \mathcal{G}_n),\end{aligned}$$

where

$$\mathcal{G}_n = n^{1/2}\{\widehat{\mathsf{G}}_n(u_\psi + n^{-1/2}z) - \frac{m}{n}\} = n^{-1/2}\sum_{i=1}^{n}[1_{(|\varepsilon_i - \sigma^{-1}(\hat{\mu}-\mu)|<u_\psi+n^{-1/2}z)} - \frac{m}{n}].$$

This can be expanded as

$$\begin{aligned}\mathcal{G}_n &= n^{-1/2}\textstyle\sum_{i=1}^n\{1_{(|\varepsilon_i|<u_\psi)} - \psi\} \\ &\quad + n^{-1/2}\textstyle\sum_{i=1}^n[1_{(|\varepsilon_i-\sigma^{-1}(\hat\mu-\mu)|<u_\psi+n^{-1/2}z)} - 1_{(|\varepsilon_i|<u_\psi)}] + n^{1/2}(\psi - \frac{m}{n}). \qquad \text{(A.4)}\end{aligned}$$

Here, the first term is asymptotically $\mathsf{N}\{0, \psi(1-\psi)\}$. Applying (A.1) with $\ell = 0$, $a = u_\psi$, $\hat a = n^{-1/2}z$, $\hat b = \sigma^{-1}(\hat\mu - \mu)$ the second term is $2\mathsf{f}(u_\psi)z + \mathsf{o_P}(1)$. The third term vanishes by assumption. Thus, $\mathcal{G}_n$ is asymptotically $\mathsf{N}\{2\mathsf{f}(u_\psi)z, \psi(1-\psi)\}$. In particular, it follows that

$$\mathcal{P}_n \stackrel{\text{def}}{=} \mathsf{P}\{n^{1/2}(\frac{\tilde r}{\sigma} - u_\psi) \le z\} \to \Phi\{\frac{2\mathsf{f}(u_\psi)z}{\sqrt{\psi(1-\psi)}}\}.$$

In turn, $\tilde r$ is asymptotically normal as stated in (A.3).

Next, an expansion for $\tilde r$ is derived. Since $\widehat{\mathsf{G}}_n(\tilde r/\sigma) = m/n$ then

$$n^{1/2}\{\widehat{\mathsf{G}}_n(\frac{\tilde r}{\sigma}) - \psi\} = n^{1/2}(\frac{m}{n} - \psi) = \mathsf{o_P}(1),$$

by the assumption to m. Expanding the left hand term as in (A.4) then gives

$$n^{-1/2}\textstyle\sum_{i=1}^n\{1_{(|\varepsilon_i|<u_\psi)} - \psi\} + n^{-1/2}\sum_{i=1}^n[1_{(|\varepsilon_i-\sigma^{-1}(\hat\mu-\mu)|<\sigma^{-1}\hat r_{(m)})} - 1_{(|\varepsilon_i|<u_\psi)}] = \mathsf{o_P}(1).$$

Applying (A.1) with $\ell = 0$, $a = u_\psi$, $\hat a = \sigma^{-1}\tilde r - u_\psi$, $\hat b = \sigma^{-1}(\hat\mu - \mu)$ then shows

$$n^{-1/2}\textstyle\sum_{i=1}^n\{1_{(|\varepsilon_i|<u_\psi)} - \psi\} + 2\mathsf{f}(u_\psi)n^{1/2}(\sigma^{-1}\tilde r - u_\psi) = \mathsf{o_P}(1).$$

Solving this equation for $n^{1/2}(\sigma^{-1}\tilde r - u_\psi)$ gives the desired expansion when inserted into (A.2). ■

Proof of Theorem 5.2. The estimator $\tilde\mu$ satisfies

$$\tilde\mu - \mu = m^{-1}\textstyle\sum_{i=1}^n(y_i - \mu)1_{(|y_i-\hat\mu|\le\tilde r)} = \frac{n}{m}n^{-1}\sum_{i=1}^n(y_i-\mu)1_{(|\varepsilon_i-\sigma^{-1}(\hat\mu-\mu)|<\sigma^{-1}\tilde r)}.$$

Adding and subtracting $1_{(|\varepsilon_i|<u_\psi)}$ and using $y_i = \mu + \sigma\varepsilon_i$ it holds

$$\begin{aligned}\frac{m}{n}\sigma^{-1}n^{1/2}(\tilde\mu - \mu) &= n^{-1/2}\textstyle\sum_{i=1}^n\varepsilon_i 1_{(|\varepsilon_i|<u_\psi)} \\ &\quad + n^{-1/2}\textstyle\sum_{i=1}^n\varepsilon_i\{1_{(|\varepsilon_i-\sigma^{-1}(\hat\mu-\mu)|<\sigma^{-1}\tilde r)} - 1_{(|\varepsilon_i|<u_\psi)}\}.\end{aligned}$$

Here, $m^{-1}n \to \psi$ by assumption. The first term on the right converges in distribution by the central limit theorem. Applying (A.1) with $\ell = 1$, $a = u_\psi$, $\hat a = \sigma^{-1}\tilde r - u_\psi$, $\hat b = \sigma^{-1}(\hat\mu - \mu)$ the second term is $2u_\psi\mathsf{f}(u_\psi)\sigma^{-1}n^{1/2}(\hat\mu - \mu) + \mathsf{o_P}(1)$. Inserting these results gives the desired expansion. ■

Proof of Theorem 5.3. The estimator $\tilde{\sigma}^2$ satisfies

$$\sigma^{-2}\tilde{\sigma}^2 - \psi^{-1}\tau_2^{u_\psi} = m^{-1}\sum_{i=1}^n \{\sigma^{-2}(y_i - \hat{\mu})^2 - \psi^{-1}\tau_2^{u_\psi}\} 1_{(|y_i - \hat{\mu}| \le \tilde{r})}.$$

Using $y_i - \hat{\mu} = \sigma\varepsilon_i - (\hat{\mu} - \mu)$ then

$$\sigma^{-2}(y_i - \hat{\mu})^2 - \psi^{-1}\tau_2^{u_\psi} = (\varepsilon_i^2 - \psi^{-1}\tau_2^{u_\psi}) + \sigma^{-2}(\hat{\mu} - \mu)^2 - 2\sigma^{-1}(\hat{\mu} - \mu)\varepsilon_i,$$

while $1_{(|y_i - \hat{\mu}| \le \tilde{r})} = 1_{(|\varepsilon_i - \sigma^{-1}(\hat{\mu} - \mu)| < \sigma^{-1}\tilde{r})}$ as before. We define the functions $h_2(u) = u^2 - \psi^{-1}\tau_2^{u_\psi}$, $h_1(u) = \varepsilon$, and $h_0(u) = 1$, and, for $\ell = 0, 1, 2$,

$$\begin{aligned}
\mathcal{S}_\ell &= n^{-1/2}\textstyle\sum_{i=1}^n h_\ell(\varepsilon_i) 1_{(|\varepsilon_i - \sigma^{-1}(\hat{\mu} - \mu)| < \sigma^{-1}\tilde{r})} = n^{-1/2}\sum_{i=1}^n h_\ell(\varepsilon_i) 1_{(|\varepsilon_i| < u_\psi)} + \mathcal{R}_\ell, \\
\mathcal{R}_\ell &= n^{-1/2}\textstyle\sum_{i=1}^n h_\ell(\varepsilon_i)[1_{(|\varepsilon_i - \sigma^{-1}(\hat{\mu} - \mu)| < \sigma^{-1}\tilde{r})} - 1_{(|\varepsilon_i| < u_\psi)}],
\end{aligned}$$

so that

$$\frac{m}{n} n^{1/2}\{\sigma^{-2}\tilde{\sigma}^2 - \psi^{-1}\tau_2^{u_\psi}\} = \mathcal{S}_2 + \sigma^{-2}(\hat{\mu} - \mu)^2\mathcal{S}_0 - 2\sigma^{-1}(\hat{\mu} - \mu)\mathcal{S}_1. \tag{A.5}$$

For $\mathcal{S}_2$ the first term converges in distribution by the central limit theorem. Applying (A.1) with $\ell = 2$ and $\ell = 0$, $a = u_\psi$, $\hat{a} = \sigma^{-1}\tilde{r} - u_\psi, \hat{b} = \sigma^{-1}(\hat{\mu} - \mu)$, the second term is $\mathcal{R}_2 = 2\mathsf{f}(u_\psi)(u_\psi^4 - \psi^{-1}\tau_2^{u_\psi}) n^{1/2}(\sigma^{-1}\tilde{r} - u_\psi) + \mathrm{o}_\mathsf{P}(1)$. It follows that

$$\mathcal{S}_2 = n^{-1/2}\textstyle\sum_{i=1}^n (\varepsilon_i^2 - \dfrac{\tau_2^{u_\psi}}{\psi}) 1_{(|\varepsilon_i| < u_\psi)} + 2\mathsf{f}(u_\psi)(u_\psi^4 - \dfrac{\tau_2^{u_\psi}}{\psi}) n^{1/2}(\sigma^{-1}\tilde{r} - u_\psi) + \mathrm{o}_\mathsf{P}(1).$$

Inserting the expression for $\tilde{r}$ from Theorem 5.1 with $\hat{\sigma} = \sigma$ then shows

$$\mathcal{S}_2 = n^{-1/2}\textstyle\sum_{i=1}^n (\varepsilon_i^2 - \dfrac{\tau_2^{u_\psi}}{\psi}) 1_{(|\varepsilon_i| < u_\psi)} - (u_\psi^4 - \dfrac{\tau_2^{u_\psi}}{\psi}) n^{-1/2}\sum_{i=1}^n \{1_{(|\varepsilon_i| \le u_\psi)} - \psi\} + \mathrm{o}_\mathsf{P}(1).$$

The terms $\mathcal{S}_0$ and $\mathcal{S}_1$ are $\mathrm{O}_\mathsf{P}(1)$ by a similar argument. These terms, are however, pre-multiplied by vanishing terms in that $\mu - \mu = \mathrm{O}_\mathsf{P}(n^{-1/2})$ by assumption. Inserting these results in (A.5) noting that $m^{-1}n \to \psi$ by assumption shows

$$\psi n^{1/2}\{\sigma^{-2}\tilde{\sigma}^2 - \psi^{-1}\tau_2^{u_\psi}\} = \mathcal{S}_2 + \mathrm{o}_\mathsf{P}(1).$$

With $\tilde{\sigma}^2_{corrected} = \psi\tilde{\sigma}^2/\tau_2^{u_\psi}$ the left hand side becomes

$$\tau_2^{u_\psi} n^{1/2}\{\sigma^{-2}\tilde{\sigma}^2_{corrected} - 1\}$$

and the desired result follows. ■

References

Atkinson, A.C. and Riani, M. (2006) Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics* 15, 460–476.

Atkinson, A.C., Riani, M. and Ceroli, A. (2009) The forward search: Theory and data analysis. Discussion paper *Journal of Korean Statistical Society.*

David, H.A. (1981) *Order Statistics*. 2nd ed. New York: Wiley.

Doornik, J. (2009) Autometrics. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pp. 88–121. Oxford: Oxford University Press.

Engler, E. and Nielsen, B. (2009) The empirical process of autoregressive residuals. *Econometrics Journal* 12, 367–381.

Hendry, D.F. and Krolzig, H.-M. (2005) The properties of automatic Gets modelling. Economic Journal 115, C32–C61.

Hoover, K.D. and Perez, S. J. (1999) Data mining reconsidered: Encompassing and the general to specific approach to specification search. *Econometric Journal* 2, 167–191.

Johansen, S. and Nielsen, B. (2009) An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pp. 1–36. Oxford: Oxford University Press.

Koul, H.L. (2002) *Weighted Empirical Processes in Dynamic Nonlinear Models*, 2nd edition. New York: Springer.

Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association* 79, 871–880.

Víšek, J.Á. (2006) The least trimmed squares. Part III: Asymptotic normality. *Kybernetika* 42, 203–224.