Outlier detection algorithms for least squares time series regression

Q: *How big is an outlier? How to choose cut-off?*

A: *Choose cutoff indirectly from* ***gauge***:
how many false detections can you tolerate?

Asymptotic analysis of Forward Search

Outline

★ Model

★ Gauge $\widehat{\gamma}$

★ Forward Search Algorithm

★ Asymptotic theory for forward residual

★ Asymptotic theory for stopping Forward Search

★ Fulton Fish Example used throughout

Time series regression model

$$y_i = x_i'\beta + \varepsilon_i \qquad \text{for } i = 1, \ldots, n.$$

where

$$\varepsilon_i \ : \quad \text{independent } \mathsf{N}(0, \sigma^2)$$

$$x_i \ : \quad \begin{cases} \text{stationary} \\ \text{deterministic trends} \\ \text{random walk trends} \end{cases}$$

Thus: looking at outlier detection when
data generating process has no outliers

Gauge

Suppose we have estimators $\widehat{\beta}, \widehat{\sigma}$

Choose cut-off $c > 0$

Observation i is outlier if $|y_i - x_i'\widehat{\beta}| > c\widehat{\sigma}$

Sample gauge

$$\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} 1_{(|y_i - x_i'\widehat{\beta}| > c\widehat{\sigma})}$$

Population gauge

$$\gamma = \mathsf{E}\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{P}(|y_i - x_i'\widehat{\beta}| > c\widehat{\sigma})$$

Forward Search

1. Choose $m_0 < n$ and initial $\widehat{\beta}^{(m_0)}$. Let $m = m_0$.

2.1 Compute $\widehat{\xi}_i^{(m)} = |y_i - x_i'\widehat{\beta}^{(m)}|$ for $i = 1, \ldots, n$.

2.2 Find $(m+1)^{th}$ smallest order statistic $\widehat{\xi}_{(m+1)}^{(m)}$

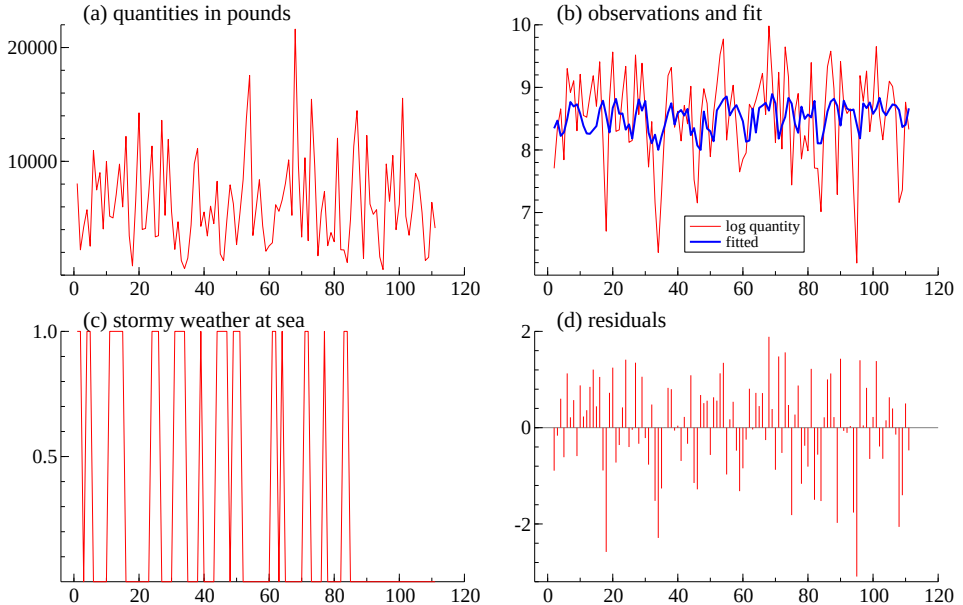
2.3 Define indicators

$$\widehat{v_i}^{(m)} = 1_{(\widehat{\xi}_i^{(m)} \leq \widehat{\xi}_{(m+1)}^{(m)})} = 1_{(|y_i - x_i'\widehat{\beta}^{(m)}| \leq \widehat{\xi}_{(m+1)}^{(m)})}$$

3. $\widehat{\beta}^{(m+1)}, (\widehat{\sigma}^{(m+1)})^2$ are LS for i so $\widehat{v_i}^{(m)} = 1$

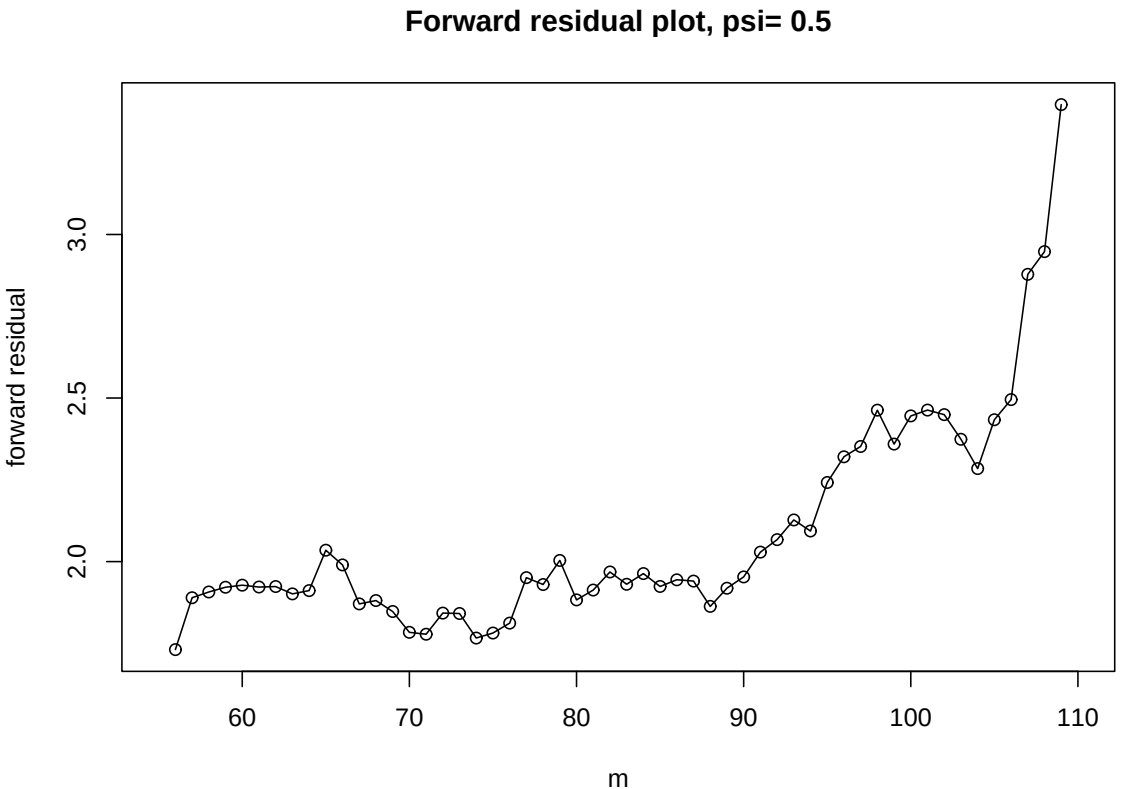
4. If $m < n$ & $\widehat{\xi}_{(m+1)}^{(m)}$ *"small"* repeat 2 and 3.

Fulton Fish Data



regress q_t on $1, q_{t-1}, S_t$

H&N outliers: $18, 34, 95$


Forward residual plot, psi= 0.5
forward residual
2.0
2.5
3.0
60
70
80
90
100
110
m

Asymptotic theory, forward residuals

Assume $N' \sum_{i=1}^{n} x_i x_i N \xrightarrow{\mathsf{D}} \Sigma > 0$

Assume $N^{-1}(\widehat{\beta}^{(m_0)} - \beta) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$

Let $m = n\psi$ and $\mathsf{P}(|\varepsilon_i| > c_\psi) = \psi$.

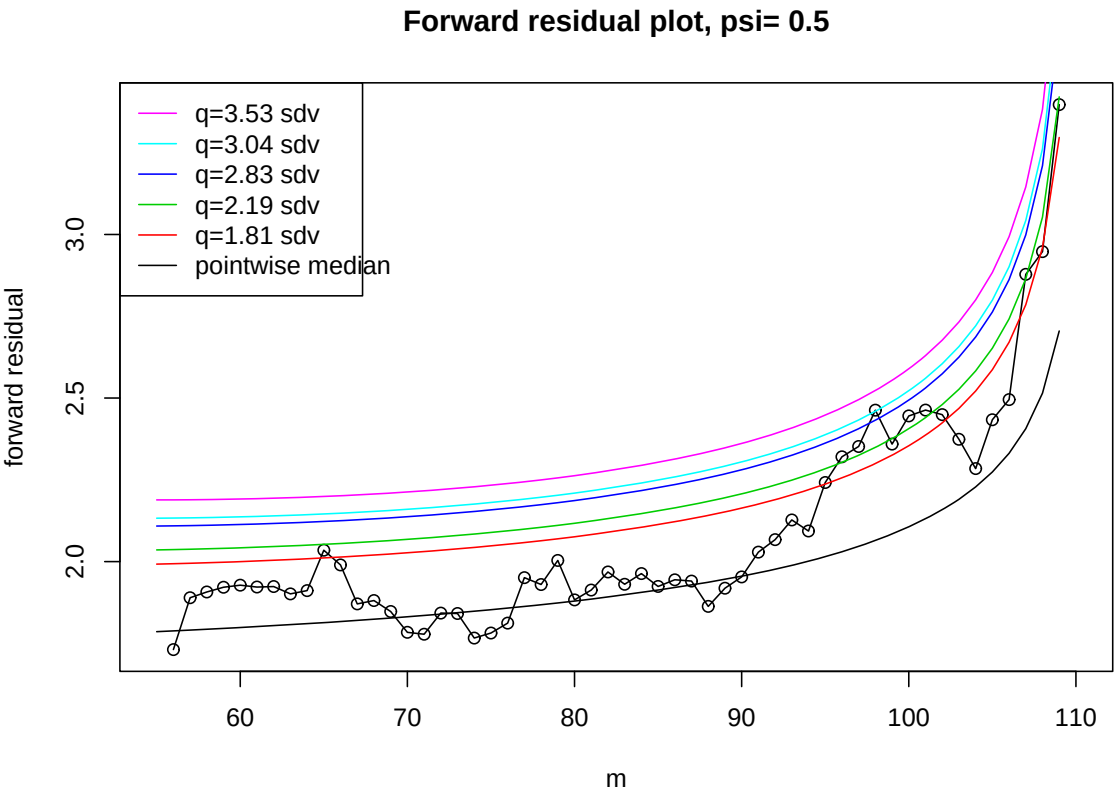
Forward residual $\widehat{z}^{(m)} = \widehat{\xi}^{(m)}_{(m+1)}$ satisfy

$$\sup_{0<\psi_0\le\psi\le n/(n+1)} \left| 2\mathsf{f}(c_\psi) \left(\frac{\widehat{z}_\psi}{\sigma} - c_\psi \right) + n^{-1/2} \sum_{i=1}^{n} \left\{ 1_{(|\varepsilon_i| \le \sigma c_\psi)} - \psi \right\} \right| = o_{\mathsf{P}}(1)$$

Similarly, scaled forward residuals satisfy

$$\sup_{0<\psi_0\le\psi\le n/(n+1)} \left| 2\mathsf{f}(c_\psi) \left(\frac{\widehat{z}_\psi}{\widehat{\sigma}_\psi} - c_\psi \right) + \mathbb{Z}_n(c_\psi) \right| = o_{\mathsf{P}}(1),$$

$$\mathbb{Z}_n \xrightarrow{\mathsf{D}} \mathbb{Z} \qquad \text{a Gaussian process}$$


Forward residual plot, psi= 0.5
q=3.53 sdv
q=3.04 sdv
q=2.83 sdv
q=2.19 sdv
q=1.81 sdv
pointwise median
forward residual
2.0
2.5
3.0
60
70
80
90
100
110
m

How to stop FS? Use gauge!

Choose $q > 0$

Stop first time forward residual exits "$+q$sdv".

This is a stopping time $\widehat{m}$

Number of outliers in then $n - \widehat{m}$

Sample gauge is $\widehat{\gamma} = \dfrac{n - \widehat{m}}{n}$

Population gauge is $\gamma = \mathsf{E}\widehat{\gamma}$

Can choose γ. This implies q.

Manipulate gauge

$$\begin{aligned}\widehat{\gamma} &= \frac{n-\widehat{m}}{n} \\ &= \frac{1}{n}\sum_{i=m_0}^{n-1}(n-m)1_{(\widehat{m}=m)} \\ &= \frac{1}{n}\sum_{j=m_0}^{n-1}1_{(\widehat{m}\leq j)}\end{aligned}$$

Thus

$$\gamma = \mathsf{E}\widehat{\gamma} = \frac{1}{n}\sum_{j=m_0}^{n-1}\mathsf{P}(\widehat{m}\leq j).$$

Asymptotic theory for gauge

From before, $m = \psi n$ and $\mathsf{P}(|\varepsilon_i| > \sigma c_\psi) = \psi$,

$$\widetilde{\mathbb{Z}}_n(c_\psi) = 2\mathsf{f}(c_\psi) n^{1/2} \left(\frac{\widehat{z}_\psi}{\widehat{\sigma}_\psi} - c_\psi \right)$$
$$= \mathbb{Z}_n(c_\psi) + \mathrm{o}_{\mathsf{P}}(1) \xrightarrow{\mathsf{P}} \mathbb{Z}(c_\psi)$$

Stopping time

$$\widehat{m} = \underset{m_0 \le m < n}{\arg\min} \left[\widetilde{\mathbb{Z}}_n(c_{m/n}) > q\mathsf{sdv} \left\{ \widetilde{\mathbb{Z}}_n(c_{m/n}) \right\} \right].$$

Look at events

$$(\widehat{m} \leq j) = \left[\max_{m_0 \leq m \leq j} \frac{\widetilde{\mathbb{Z}}_n(c_{m/n})}{\mathsf{sdv}\left\{\widetilde{\mathbb{Z}}_n(c_{m/n})\right\}} > q \right]$$

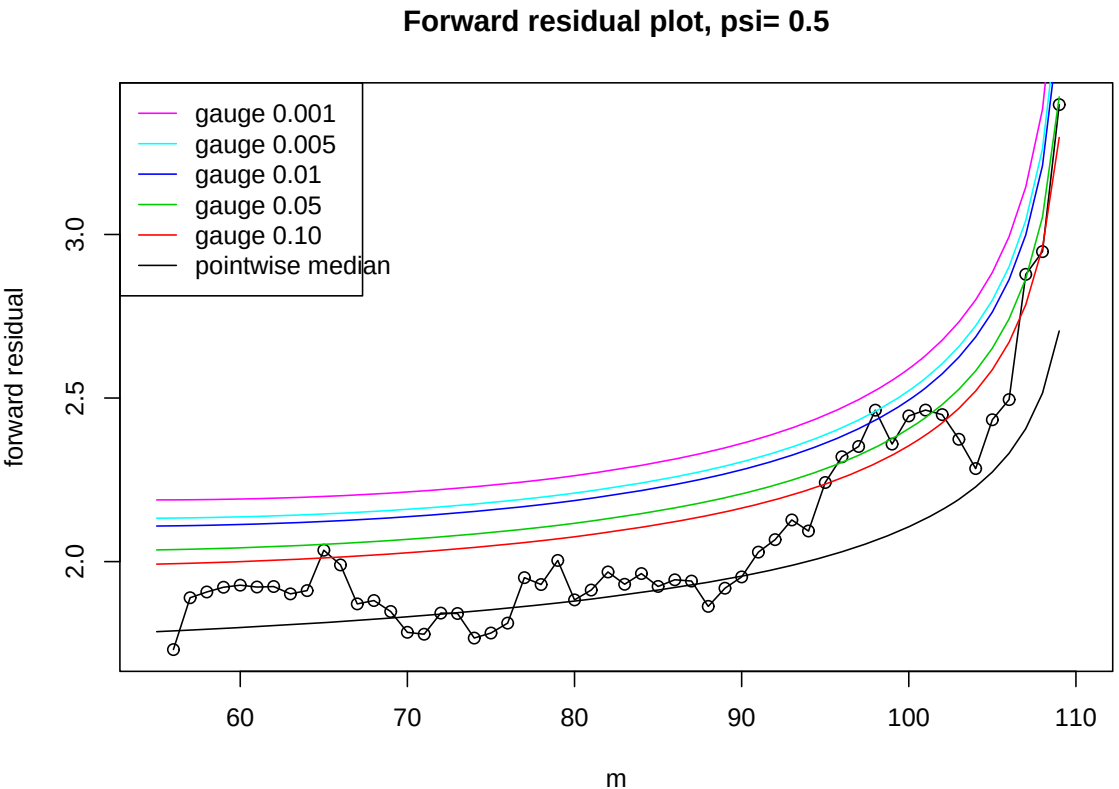
Thus

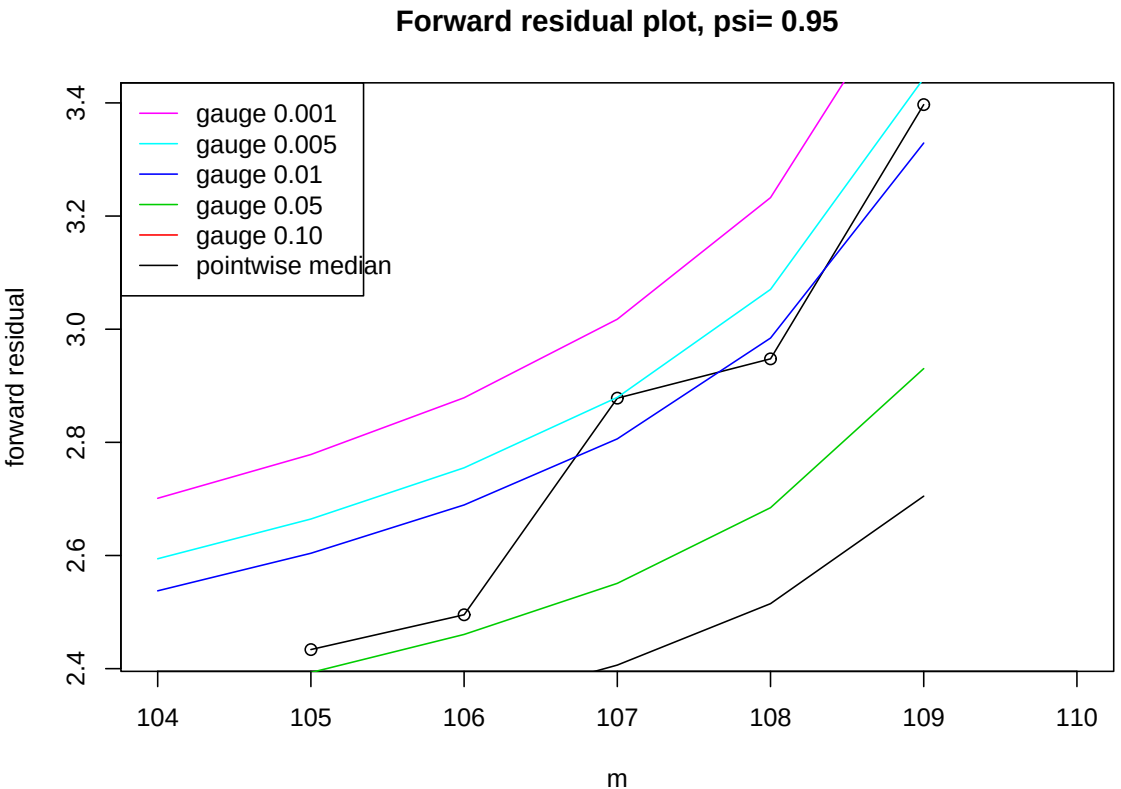
$$\gamma = \mathsf{E}\widehat{\gamma} = \frac{1}{n} \sum_{j=m_0}^{n-1} \mathsf{P}(\widehat{m} \leq j)$$

$$\rightarrow \gamma = \int_{\psi_0}^{1} \mathsf{P}\left[\sup_{\psi_0 \leq \psi \leq u} \frac{\mathbb{Z}(c_\psi)}{\mathsf{sdv}\left\{\mathbb{Z}(c_\psi)\right\}} > q \right] du.$$

Table: q as function of γ, ψ_0

γ vs ψ_0	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.10	2.28	2.14	1.99	1.81	1.60	1.31	0.82	-
0.05	2.58	2.46	2.33	2.19	2.02	1.79	1.45	0.69
0.01	3.14	3.04	2.94	2.83	2.71	2.55	2.33	1.91
0.005	3.35	3.26	3.15	3.04	2.95	2.81	2.62	2.26
0.001	3.77	3.69	3.62	3.53	3.43	3.32	3.18	2.92


Forward residual plot, psi= 0.5
gauge 0.001
gauge 0.005
gauge 0.01
gauge 0.05
gauge 0.10
pointwise median
forward residual
2.0
2.5
3.0
60
70
80
90
100
110
m


Forward residual plot, psi= 0.95
gauge 0.001
gauge 0.005
gauge 0.01
gauge 0.05
gauge 0.10
pointwise median
forward residual
2.4
2.6
2.8
3.0
3.2
3.4
104
105
106
107
108
109
110
m

References

J&N: Outlier detection algorithms for least squares time series regression.

To appear in *Scandinavian J Statistics* (with discussion)

J&N: Forward Search DP 2013

J&N: *Econometrics* 2013

Hoover & Perez *Econometrics J* 1999

Hendry & Doornik *Empirical Model Discovery* 2014

Hendry & Santos *Engle Festschrift* 2010

Castle, Doornik & Hendry *J Time Series Econometrics* 2011

Graddy *RAND J Economics* 1995

Hendry & N *Econometric Modelling* 2007