*Discussiones Mathematicae*
*Probability and Statistics 30 (2010) 203–219*

# ADAPTIVE TRIMMED LIKELIHOOD ESTIMATION IN REGRESSION

Tadeusz Bednarski

*Institute of Economic Sciences*
*Faculty of Law, Administration and Economy*
*Wrocław University, Uniwersytecka 22/26, 50–145 Wroclaw, Poland*

**e-mail:** t.bednarski@prawo.uni.wroc.pl

Brenton R. Clarke and Daniel Schubert

*Mathematics and Statistics, School of Chemical and Mathematical Sciences*
*Faculty of Minerals and Energy*
*Murdoch University, Murdoch, WA 6150, Australia*

**e-mail:** B.Clarke@murdoch.edu.au

**Summary**

In this paper we derive an asymptotic normality result for an adaptive trimmed likelihood estimator of regression starting from initial high breakdownpoint robust regression estimates. The approach leads to quickly and easily computed robust and efficient estimates for regression. A highlight of the method is that it tends automatically in one algorithm to expose the outliers and give least squares estimates with the outliers removed. The idea is to begin with a rapidly computed consistent robust estimator such as the least median of squares (LMS) or least trimmed squares (LTS) or for example the more recent MM estimators of Yohai. Such estimators are now standard in statistics computing packages, for example as in SPLUS or R. In addition to the asymptotics we provide data analyses supporting the new adaptive approach. This approach appears to work well on a number of data sets and is quicker than the related brute force adaptive regression approach described in Clarke (2000). This current approach builds on the work of Bednarski and Clarke (2002) which considered the asymptotics for the location estimator only.

**Key Words:** trimmed likelihood estimator, adaptive estimation, regression.

**2010 Mathematics Subject Classification:** 62F35, 62E20.

## 1. Introduction

In a relatively recent paper Bednarski and Clarke (2002) describe the asymptotic theory of an adaptive trimmed likelihood estimator of location, where the adaptive estimator chooses that solution which minimizes an estimated asymptotic variance of the trimmed likelihood estimator. This work built on initial asymptotic work detailed in Bednarski and Clarke (1993) where the idea of trimmed likelihood estimation was introduced, albeit at the same time Vandev and Neykov (1993) were investigating a similar proposal at least for the Gaussian case, and where the objective was in terms of high breakdown point estimators. Some recent history of related proposals can be found in Neykov and Müller (2003) and Müller and Neykov (2004).

Following up the trimmed likelihood approach for location an adaptive version of it was successfully investigated empirically in Clarke (1994) and for at least small to moderate sample sizes the adaptive estimator for regression in an algorithm called **ATLA** proved most effective in obtaining the outliers and only the outliers to be removed from the data before least squares is then implemented. This work defining and using **ATLA** is in Clarke (2000). The drawback of using **ATLA** is that in large samples the computing is intensive, and while one may desire a breakdown point one half (approximately) estimator the computing time soon becomes out of bounds. However there are estimators in statistical packages that can be implemented on large data sets, and for which estimates have large breakdown points. The approach of this paper is to use these consistent and robust estimates of regression to then use the adaptive estimate of location on the residuals to identify the outliers and then take least squares estimates with the identified outliers removed from the initial data. This allows one to use all the useful data in a robust and efficient approach to estimation. What we are proposing is a method synonymous with the approach of Rousseeuw (1984) and Rouseeuw and Leroy (1987), where one highlights the outliers. These may be of more interest than the actual regression in some examples. On the other hand we do not throw out half the data as indicated in those works before evaluating the estimator. The more efficient estimator of Yohai (1987) known as the MM-estimator while it leads to a high breakdown point and highly efficient estimator, fails to identify the outliers automatically. By applying our adaptive approach on the residuals from MM-estimator of regression we highlight the observations that are outliers in the regression analysis.

Following a similar idea starting from the LTS estimator Gamble (1999) showed empirically some good results in a thesis on the analysis of contaminated tidal data. This paper develops some of the asymptotics associated with such an approach.

The idea of adaptive estimation for location estimation and regression has been studied in related but different settings to the one countenanced in this paper in Dodge and Jurečková (1987, 1997), Jurečková, Koenker and Welsh (1994) and more recently in a monograph Dodge and Jurečková (2000).

## 2. Preliminaries

We consider the model

$$\mathbf{Y} = \boldsymbol{X}\,\boldsymbol{\beta}\, +\, \mathbf{e},$$

where $\mathbf{e} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_n})$, which is the multivariate normal distribution with mean zero and covariance $\sigma^2 \boldsymbol{I_n}$. Here $\boldsymbol{I_n}$ is the $n \times n$ identity matrix of order $n$; $\mathbf{X}$ is an $n \times p$ design matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameter coefficients. We shall assume the design matrix is of full rank $p$ and $n > p$. The Least Squares estimates of the parameter vector are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

and the sum of squared residuals is given by

$$\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \;=\; \boldsymbol{Y'Y} - \hat{\boldsymbol{\beta}}'(\boldsymbol{X'X})\hat{\boldsymbol{\beta}}\ .$$

Suppose we have some robust consistent estimator $\boldsymbol{\beta^*}$, for example an estimator such as LMS, LTS, MM, or an S-estimator, (which typically have a breakdown point of approximately one half). For instance, breakdown points are discussed in Rousseeuw and Leroy (1987). We set up the vector

$$\boldsymbol{r^*} = \boldsymbol{Y} - \boldsymbol{X\beta^*}$$

this being the vector of residuals gained from knowledge of the robust fit of the vector parameter $\boldsymbol{\beta^*}$. From $\boldsymbol{r^*}$ we calculate the quantity

$$V_n^*(g) = var(\tilde{\sigma}_*^2(g), g/n),$$

where

$$\tilde{\sigma}_*^2(g) = \frac{1}{n-p}\sum_{i=1}^{h}(r^*)^2_{i:n} \tag{2.1}$$

and $h = n - g$, while

$$var(\sigma^2, \alpha) = \frac{\sigma^2}{\{1-\alpha-\sqrt{\frac{2}{\pi}}z_{\alpha/2}e^{-z^2_{\alpha/2}}\}^2}$$

and $\alpha = g/n$. Note in a departure from Clarke (2000) we use the divisor $n-p$ rather than $h-p$ in equation (2.1) as then if the data have normal errors then the estimate $V_n^*(g)$, keeping the proportion $g/n = \alpha$ fixed, is such that when it is multiplied by $\boldsymbol{(X'X)^{-1}}$ tends toward the asymptotic variance of the trimmed likelihood estimator for regression. Above $z_{\alpha/2}$ is the critical point of the standard normal distribution, so that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ where $\Phi$ is the cumulative standard normal distribution. Also above we use the ordered squared residuals from the vector $\mathbf{r}^*$ where

$$(r^*)^2_{1:n} \leq (r^*)^2_{2:n} \leq ... \leq (r^*)^2_{n:n}. \tag{2.2}$$

The approach suggested in this paper is to minimize $V_n^*(g)$ over $0 \leq g \leq G^*(n)$, where $G^*(n) = n - [\frac{n}{2}] - [\frac{p+1}{2}]$ to determine

$$V_n^*(\tilde{g}) = min_{0\leq g\leq G^*(n)}V_n^*(g).$$

We denote $\tilde{S}_n$ to be the set of $\tilde{h} = n - \tilde{g}$ indices in $\mathbf{r}^*$ that give the smallest $(r^*)^2_{i:n}$. We then denote the modified adaptive trimmed likelihood estimator of the vector parameter $\boldsymbol{\beta}$ as

$$\boldsymbol{\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}},$$

where $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ are the original $(\mathbf{X}, \mathbf{Y})$ made up with indices in $\tilde{S}_n$, i.e. the least squares estimator with the "outliers" omitted.

For the initial estimators aforementioned there are consistency results in the sense that $\boldsymbol{\beta^*} \to \boldsymbol{\beta_0}$ where convergence is in probability. For a few of the estimators there exist under certain conditions an asymptotic normality result where $\sqrt{n}(\boldsymbol{\beta^*} - \boldsymbol{\beta_o})$ converges in distribution to a $N(\mathbf{0}, \sigma^2_\psi(\mathbf{X'X})^{-1}) = N(\mathbf{0}, \sigma^2_\psi\Lambda^{-1})$ distribution say.

It was conjectured in Bednarski and Clarke (2002) is that if the model put forward in the first equation in this section were to hold then $\tilde{\alpha} = \tilde{g}/n$ converges in probability to zero with large n and asymptotically the least squares estimates $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ agree. The question then comes, what if the actual density of the unobserved residuals in the first equation was in fact not $N(\mathbf{0}, \sigma^2 \boldsymbol{I_n})$ but the individual but independent errors $e_i$ had a distribution with a density as countenanced in that paper. That is the density is fatter tailed than normal, and then ideally we would find that

$$\tilde{\alpha} \xrightarrow{p} \alpha_A$$

and the regression modified trimmed likelihood estimator satisfies

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \xrightarrow{d} N(0, V(\alpha_A, F)(\mathbf{X'X})^{-\mathbf{1}}).$$

Here $V(\alpha_A, F)$ is as in the paper by Bednarski and Clarke (2002).

## 3. Fisher consistency and principal regularity assumptions

We begin by letting $F^o$ be the model distribution for $Y$ and $\boldsymbol{X}$ in the regression model $Y = \boldsymbol{X'}\boldsymbol{\beta_o} + e$. Here $F_o$, the error distribution for $e$, is assumed to have a continuous positive and symmetric density function $f_o$ which is differentiable with bounded derivative. We let $G_o$ be the distribution of the covariate and suppose $e$ and $\boldsymbol{X}$ are independent random variables. For simplicity of subsequent derivations assume that the scale of the error distribution is 1 and that $G_o$ is a continuous distribution with bounded support. Occasionally we shall use the notation $F(\mathcal{A})$ to indicate the probability of event $\mathcal{A}$ under the law $F$. Then the Fisher consistency for the trimmed likelihood method amounts to checking if

$$\int \boldsymbol{x}(y - \boldsymbol{x'\beta}) J_\alpha[F^o\{(u, \boldsymbol{w}); f_o(u - \boldsymbol{w'\beta}) \geq f_o(y - \boldsymbol{x'\beta})\}] dF^o(y, \boldsymbol{x}) \tag{3.1}$$

is zero for $\boldsymbol{\beta} = \boldsymbol{\beta_o}$. Since the integral can be written as

$$\int_{|y - \boldsymbol{x'\beta_o}| \leq b} \boldsymbol{x}(y - \boldsymbol{x'\beta_o}) dF^o = \int \int_{|u| \leq b} \boldsymbol{x} u \; dF_o(u) dG_o(\boldsymbol{x})$$

where $F(|y-\boldsymbol{x'\beta_0}| \leq b) = 1-\alpha$ by symmetry of $f_o$ we get that the last integral is equal to zero.

Let $B$ be a bounded set in parameter space containing $\beta_o$. From now on we shall consider infinitesimal families of distributions $\mathcal{G}_n$ of $(Y, \boldsymbol{X})$, depending on the sample size $n$ so that the following conditions hold:

**a)**

$$\limsup_n \sup_{\beta\in B} \sup_{F\in\mathcal{G}_n} \sup_{y,x} \sqrt{n}|F(y,x) - F^o(y,x)| < \infty$$

and

$$\limsup_n \sup_{\beta\in B} \sup_{F\in\mathcal{G}_n} \sqrt{n} \sup_t |F(y-\boldsymbol{x'\beta} \leq t) - F^o(y-\boldsymbol{x'\beta} \leq t)| < \infty$$

**b)**

$$\lim_{\delta\to 0, n\to\infty} \sup_{F\in\mathcal{G}_n} \sup_{|t_0-t_1|<\delta} \sqrt{n}|F(t_0 < |y\boldsymbol{x'\beta_0}|< t_1)$$

$$-F^o(t_0< |y-\boldsymbol{x'\beta_0}|< t_1)|=0$$

for $\beta \in B$.

Conditions a) and b) are analogous to those given in Bednarski and Clarke (2002) (p.4). They will let us show that there exists a root-n consistent sequence of estimates for the families $\mathcal{G}_n$ and then build a uniformly valid expansion of the estimate for the empirical distribution functions, uniformly valid in $\alpha$ in closed subsets of $(0,1)$. The uniformity in $\alpha$, in turn, will let us show adaptivity of the method in the sense that the fixed $\alpha$ can be substituted by adaptively chosen $\alpha$ without the change of estimate's limiting distribution.

## 4. Root-n consistency of the trimmed likelihood estimator

The $\alpha$-trimmed likelihood functional $\boldsymbol{\beta}(F,\alpha)$ is defined to solve $L(F,\boldsymbol{\beta},b) = 0$, where

$$L(F,\boldsymbol{\beta},b) = \int_{|y-\boldsymbol{x'\beta}|\leq b} \boldsymbol{x}(y-\boldsymbol{x'\beta})dF$$

and $F(|y-\boldsymbol{x'\beta}| \leq b) = 1-\alpha$. More precisely, given $\boldsymbol{\beta}$, $b$ is defined as $inf\{b : F(|y-\boldsymbol{x'\beta}| \leq b) \geq 1-\alpha\}$ and it will further be denoted by $b(F,\boldsymbol{\beta})$.

The root n consistency means that we can find a version of the functional $\beta(F,\alpha)$ so that $\sqrt{n}(\beta(F,\alpha)-\beta_{o})$ stays bounded uniformly in $F \in \mathcal{G}_n$. The following properties (1)–(3) which, as in Bednarski and Clarke (2002) imply root $n$ consistency of the estimator, follow from the model assumptions:

**(1)** $\sup_{\beta\in B} |b(F,\boldsymbol{\beta}) - b(F^o,\boldsymbol{\beta})| \le K/\sqrt{n}$ for some constant $K$ uniformly over $\alpha$ in any closed subinterval of $(0,1)$.

**(2)** There is $K>0$ such that for $\boldsymbol{\beta} \in B$

$$|L(F,\boldsymbol{\beta}) - L(F^o,\boldsymbol{\beta})| \le K||F - F^o||$$

uniformly over $\alpha$ in any closed subinterval of $(0,1)$.

**(3)** $L(F^o,\boldsymbol{\beta}) - L(F^o,\boldsymbol{\beta_o}) = (\boldsymbol{\beta}-\boldsymbol{\beta_o})\boldsymbol{A} + o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||)$ where

$$\boldsymbol{A} = -\int_{|y-\boldsymbol{x}'\boldsymbol{\beta_o}|\le b(F^o,\boldsymbol{\beta_o})} \boldsymbol{xx'}dF^o + 2b(F^o,\boldsymbol{\beta_o})f_o(b(F^o,\boldsymbol{\beta_o}))\int \boldsymbol{xx'}dG_o$$

uniformly over $\alpha$ in a closed subinterval of $(0,1)$.

Below we justify the validity of the above statements under the model assumptions:

***Proof of* (1):**
It follows from smoothness of $F^o$ and from a) that there are sequences $b_n^-$, $b_n^+$ so that

$$F(|y-\mathbf{x}'\boldsymbol{\beta}| \le b_n^+) \ge 1-\alpha$$

$$F(|y-\mathbf{x}'\boldsymbol{\beta}| \le b_n^-) \le 1-\alpha$$

$\sqrt{n}|b_n^+ - b_n^-| < M$ and $\sqrt{n}|b_n^+ - b(F^o,\beta)| < M$ for a constant $M>0$ uniformly in $\beta \in B$. We then have $b_n^- \le b(F,\beta) \le b_n^+$. In fact this convergence holds uniformly in $\alpha \in [\epsilon, \frac{1}{2}-\epsilon]$ for any $\epsilon > 0$.

***Proof of* (2):**
We have

$$L(F,\boldsymbol{\beta}) - L(F^o,\boldsymbol{\beta})$$

$$= \int_{|y-\boldsymbol{x'}\boldsymbol{\beta}|\le b(F,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})dF - \int_{|y-\boldsymbol{x'}\boldsymbol{\beta}|\le b(F_o,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})dF^o.$$

If $b(F^o,\boldsymbol{\beta}) \le b(F,\boldsymbol{\beta})$ the above difference equals

$$\int_{|y-\boldsymbol{x'}\boldsymbol{\beta}|\le b(F,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})d(F-F^o) + \int_{\substack{|y-\boldsymbol{x'}\boldsymbol{\beta}|> b(F^o,\boldsymbol{\beta})\\ |y-\boldsymbol{x'}\boldsymbol{\beta}|\le b(F,\boldsymbol{\beta})}} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})dF^o.$$

For the first integral the integration by parts gives the bound $K||F-F^o||$. For the second integral it is enough to use part (1). The bounds are by our assumption a) uniform in $\alpha$.

***Proof of* (3):**
We have

$$L(F^o,\boldsymbol{\beta}) - L(F^o,\boldsymbol{\beta_o})$$

$$= \int_{|y-\boldsymbol{x'}\boldsymbol{\beta}|\le b(F^o,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})dF^o - \int_{|y-\boldsymbol{x'}\boldsymbol{\beta_o}|\le b(F^o,\boldsymbol{\beta_o})} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta_o})dF^o$$

$$= -(\boldsymbol{\beta}-\boldsymbol{\beta_o})\int_{|y-\boldsymbol{x'}\boldsymbol{\beta_o}|\le b(F^o,\boldsymbol{\beta_o})} \boldsymbol{x}\boldsymbol{x'}dF^o + \int_{\substack{|y-\boldsymbol{x'}\boldsymbol{\beta_o}|> b(F^o,\boldsymbol{\beta_o})\\ |y-\boldsymbol{x'}\boldsymbol{\beta}|\le b(F^o,\boldsymbol{\beta})}} \boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})dF^o$$

$$= -(\boldsymbol{\beta}-\boldsymbol{\beta_o})(1-\alpha)\int \boldsymbol{x}\boldsymbol{x'}dG_o(\boldsymbol{x})$$

$$+\left(\int\int_{\boldsymbol{x'}\boldsymbol{\beta}-b(F^o,\boldsymbol{\beta})}^{\boldsymbol{x'}\boldsymbol{\beta_o}-b(F^o,\boldsymbol{\beta_o})} + \int\int_{\boldsymbol{x'}\boldsymbol{\beta_o}+b(F^o,\boldsymbol{\beta_o})}^{\boldsymbol{x'}\boldsymbol{\beta}+b(F^o,\boldsymbol{\beta})}\right)\boldsymbol{x}(y-\boldsymbol{x'}\boldsymbol{\beta})dF^o.$$

Suppose that $b(F^o,\boldsymbol{\beta}) < b(F^o,\boldsymbol{\beta_o})$. Applying the mean value theorem to the inner integral in the first of the two remaining integrals one obtains

$$[\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})+b(F^o,\boldsymbol{\beta})-b(F^o,\boldsymbol{\beta_o})]\boldsymbol{x}(\tilde{y}-\boldsymbol{x'\beta})f_o(\tilde{y})$$

where

$$\tilde{y}\in[\boldsymbol{x'\beta}-b(F^o,\boldsymbol{\beta}),\boldsymbol{x'\beta_o}-b(F^o,\boldsymbol{\beta_o})]$$

since the first factor tends to zero at the rate $||\boldsymbol{\beta}-\boldsymbol{\beta_0}||$ we can replace the above expression by

$$-[\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})+b(F^o,\boldsymbol{\beta})-b(F^o,\boldsymbol{\beta_o})]$$

$$\boldsymbol{x}\ b(F^o,\boldsymbol{\beta_o})f_o[-b(F^o,\boldsymbol{\beta_o})]+o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||)$$

where $f_o$ is the error distribution at the model.

The second inner integral gives similarly

$$[\boldsymbol{x'}(\boldsymbol{\beta}-\boldsymbol{\beta_o})+b(F^o,\boldsymbol{\beta})-b(F^o,\boldsymbol{\beta_o})]$$

$$\boldsymbol{x}\ b(F^o,\boldsymbol{\beta_o})\tilde{f}_o[b(F^o,\boldsymbol{\beta_o})]+o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||).$$

Their sum yields

$$2\boldsymbol{xx'}(\boldsymbol{\beta}-\boldsymbol{\beta_o})b(F^o,\boldsymbol{\beta_o})\tilde{f}_o[b(F^o,\boldsymbol{\beta_o})]+o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||),$$

whence we obtain

$$L(F^o,\boldsymbol{\beta})-L(F^o,\boldsymbol{\beta_o})$$

$$=-(\boldsymbol{\beta}-\boldsymbol{\beta_o})\left[\int_{|y-\boldsymbol{x'\beta}|\le b(F^o,\boldsymbol{\beta_o})}\boldsymbol{xx'}dG_o-2\int\boldsymbol{xx'}dG_o.b(F^o,\boldsymbol{\beta_o})\tilde{f}_o[b(F^o,\boldsymbol{\beta_o})]\right]$$

$$+o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||).$$

In fact, the expressions given above hold uniformly in $\alpha \in [\epsilon, \frac{1}{2} - \epsilon]$.

## 5. Expansion - Differentiability

The aim of this section is to show that for $F \in \mathcal{G}_n$, $\beta = \beta(F, \alpha)$, $\beta_o = \beta(F^o, \alpha)$ the following expansion holds

$$(\boldsymbol{\beta} - \boldsymbol{\beta_o}) = \int_{|y - \boldsymbol{x'}\boldsymbol{\beta_o}| \le b(F^o, \boldsymbol{\beta_o})} \boldsymbol{x}(y - \boldsymbol{x'}\boldsymbol{\beta_o}) d(F - F^o)\boldsymbol{A^{-1}}$$

$$+ o(||F - F^o||)$$

where $\boldsymbol{A} = [1 - \alpha - 2b(F^o, \boldsymbol{\beta_o})\tilde{f}_o[b(F^o, \boldsymbol{\beta_o})]] \int \boldsymbol{x}\boldsymbol{x'} dG_o$ and the expansion is uniform in $\alpha \in [\epsilon, \frac{1}{2} - \epsilon]$ for any $\epsilon > 0$.

Suppose the distributions $F$ satisfy the assumptions (a), (b) and fix $\alpha \in [\epsilon, 1/2 - \epsilon]$. Consider the difference $L(F, \boldsymbol{\beta}) - L(F^o, \boldsymbol{\beta_o})$ which is zero and can also be written as

$$L(F, \boldsymbol{\beta}) - L(F^o, \boldsymbol{\beta}) + L(F^o, \boldsymbol{\beta}) - L(F^o, \boldsymbol{\beta_o}).$$

By the property (3) we can write

$$L(F^o, \boldsymbol{\beta}) - L(F^o, \boldsymbol{\beta_o}) = (\boldsymbol{\beta} - \boldsymbol{\beta_o})\boldsymbol{A} + o\left(\frac{1}{\sqrt{n}}\right).$$

What remains to be shown is that

$$L(F, \boldsymbol{\beta}) - L(F^o, \boldsymbol{\beta})$$

$$= \int_{|y - \boldsymbol{x'}\boldsymbol{\beta_o}| \le b(F^o, \boldsymbol{\beta_o})} \boldsymbol{x}(y - \boldsymbol{x'}\boldsymbol{\beta_o}) d(F - F^o) + o(||F - F^o||).$$

This difference on the left hand side of the above equation equals

$$\int_{|y-\boldsymbol{x'\beta}|\leq b(F,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta})dF - \int_{|y-\boldsymbol{x'\beta}|\leq b(F^o,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta})dF^o$$

and it can be written as

$$\begin{aligned} &\int_{|y-\boldsymbol{x'\beta}|\leq b(F,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta})d(F-F^o) \\ &- \int_{b(F,\boldsymbol{\beta})<|y-\boldsymbol{x'\beta}|\leq b(F^o,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta})dF^o. \end{aligned} \tag{5.1}$$

Notice that by integration by parts the first integral differs from

$$\int_{|y-\boldsymbol{x'\beta}|\leq b(F,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta_o})d(F-F^o)$$

by term of order smaller than $||F-F^o||$. Moreover since the integration region $|y-\boldsymbol{x'\beta}|\leq b(F,\boldsymbol{\beta})$ can be written as

$$-x'(\beta_o-\beta)-b(F,\beta)\leq y-x'\beta_o\leq -x'(\beta_o-\beta)+b(F,\beta)$$

we can deduce from assumption b) that

$$\int_{|y-\boldsymbol{x'\beta}|\leq b(F,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta_o})d(F-F^o)$$

$$- \int_{|y-\boldsymbol{x'\beta_o}|\leq b(F^o,\boldsymbol{\beta_o})} \boldsymbol{x}(y-\boldsymbol{x'\beta_o})d(F-F^o)$$

is of order smaller then $||F-F^o||$.

We now show that the second term in the sum of integrals (6.1) is of order $o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||)$.

$$\int_{b(F,\boldsymbol{\beta})\le|y-\boldsymbol{x'\beta}|\le b(F^o,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta})dF^o$$

$$= \int_{b(F,\boldsymbol{\beta})<|y-\boldsymbol{x'\beta_o}+\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})|\le b(F^o,\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta_o})dF^o + \ o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||)$$

$$= \int_{b(F,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})<y-\boldsymbol{x'\beta_o}<b(F^o,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta_o})dF^o$$

$$+\int_{-b(F^o,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\beta)<y-\boldsymbol{x'\beta_o}<-b(F,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})} \boldsymbol{x}(y-\boldsymbol{x'\beta_0})dF^o+o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||)$$

$$= \int\int_{b(F,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})<u\le b(F^o,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})} \boldsymbol{x}u dF_o dG_o$$

$$+ \int\int_{-b(F^o,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})<u\le -b(F,\boldsymbol{\beta})-\boldsymbol{x'}(\boldsymbol{\beta_o}-\boldsymbol{\beta})} \boldsymbol{x}u dF_o dG_o \ + \ o(||\boldsymbol{\beta}-\boldsymbol{\beta_o}||).$$

The first of the above two integrals, after applying the mean value theorem, gives

$$\int \boldsymbol{x}[b(F^o,\boldsymbol{\beta})-b(F,\boldsymbol{\beta})]f_o(\tilde{u}(\boldsymbol{x}))b(F^o,\boldsymbol{\beta})dG_o(\boldsymbol{x})$$

while the second gives

$$-\int \boldsymbol{x}[b(F^o,\boldsymbol{\beta})-b(F,\boldsymbol{\beta})]f_o(\tilde{u}(\boldsymbol{x}))b(F^o,\boldsymbol{\beta})dG_o(\boldsymbol{x})$$

and so they cancel out in the limit. The expansion is uniform in $\alpha$ for any closed subinterval of $(0,1)$ by properties 1) – 3).

## 6. Asymptotic normality

The limiting distribution of the estimator can be easily concluded form the uniform expansion (compact differentiability). Since the empirical distribution function of $(Y_1,X_1),....,(Y_n,X_n)$ satisfies the conditions a) and b) in probability, under the model distribution $F^o$, we infer that for the trimmed likelihood estimator

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_o}) = \sqrt{n}\boldsymbol{A^{-1}} \int_{|y-\boldsymbol{x'}\boldsymbol{\beta_o}| \le b(F^o,\boldsymbol{\beta_o})} d(F_n - F^o) + o_p(||F_n - F||).$$

Therefore

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta_o}) \sim N(0, V(\alpha, F^o)),$$

where

$$V(\alpha, F^o) = \int \boldsymbol{A^{-1}} I_{|y-\boldsymbol{x'}\boldsymbol{\beta_o}| \le b(F^o,\boldsymbol{\beta_o})} \boldsymbol{x}(y - \boldsymbol{x'}\boldsymbol{\beta_o})(y - \boldsymbol{x'}\boldsymbol{\beta_o})\boldsymbol{x'}(\boldsymbol{A^{-1}})' dF^o$$

$$= \frac{\sigma_\alpha^2 \left\{ \int \boldsymbol{x}\boldsymbol{x'} dG_o \right\}^{-1}}{\{1 - \alpha - 2b(F^o, \boldsymbol{\beta_o})\tilde{f}_o[b(F^o, \boldsymbol{\beta_o})]\}^2},$$

while

$$\sigma_\alpha^2 = \int_{|y-\boldsymbol{x'}\boldsymbol{\beta_o}| \le b(F^o,\boldsymbol{\beta_o})} (y - \boldsymbol{x'}\boldsymbol{\beta_o})^2 dF^o$$

depends on $\alpha$ only and it equals to

$$\int_{-u_\alpha}^{u_\alpha} u^2 dF_o(u),$$

where $u_\alpha = F_o^{-1}(1 - \alpha/2)$.

Note that this quantity depends on both $\alpha$ and $F_o$ (the error distribution)

Since the expansion is uniform in $\alpha$ in any closed subinterval of $(0, 1)$ we can use $\alpha$ adaptively, as long as it converges to some $\alpha_o$ in probability. That limiting value will give the ultimate asymptotic limit.

## 7. Some tabulated empirical analysis

We illustrate the performance of the estimator post MM estimation, although one could just as easily implement the estimator post LTS estimation (where one trims approximately 50% of the data before implementing the adaptive estimate). For each $g$, the quantity $V_n^*(g)$ is calculated and $J(g)$ corresponds to those $g$ observations that lead to the largest ordered squared residuals post MM-estimation as indicated in (2.2). The data sets below are well known and are described in Clarke (2000). The one data set which continues to defy the adaptive approach in this case is the Stack Loss Data, where the method only highlights 2 potential outliers, observations 21 and 4., whereas there are several authors who advocate 4 potential outliers, observations 1,3,4, and 21. It is noted however that these data without the four potential outliers do not appear to have normal errors and that the data may need to be transformed. See Clarke (2000) and the references therein. It can be noted that otherwise for the adaptive approach post MM-estimation leads to exactly the outliers and only the outliers as indicated by several other authors in the data sets below. Since the speed of this algorithm is only determined essentially by the calculation of the MM-estimator and the ordering of a set of resulting squared residuals from which certain "variances" are calculated the present approach is considerably faster than the computationally intensive approach of **ATLA** described in the earlier paper of Clarke (2000).

Table 1. Hill data post MM linear regression as implemented by R.

| | $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ | $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ |
|---|---|---|---|---|---|---|
| | 0 | 277.37457 | - | 8 | 101.92590 | 30 |
| | 1 | 226.19626 | 18 | 9 | 114.99803 | 24 |
| | 2 | 89.13534 | 7 | 10 | 133.47432 | 16 |
| $\tilde{g} =$ | **3** | **63.81509** | 33 | 11 | 154.57125 | 26 |
| | 4 | 69.88278 | 19 | 12 | 182.44779 | 15 |
| | 5 | 78.22866 | 6 | 13 | 214.61558 | 12 |
| | 6 | 86.35742 | 8 | 14 | 254.08608 | 13 |
| | 7 | 93.51284 | 14 | 15 | 300.95174 | 3 |

Table 2. Stack Loss Data post MM linear regression as implemented by R.

| | $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ | $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ |
|---|---|---|---|---|---|---|
| | 0 | 13.10590 | - | 5 | 14.74116 | 1 |
| | 1 | 11.80638 | 21 | 6 | 17.79392 | 2 |
| $\tilde{g} =$ | **2** | **9.70778** | 4 | 7 | 21.08204 | 15 |
| | 3 | 11.01214 | 3 | 8 | 23.67754 | 6 |
| | 4 | 12.82279 | 13 | 9 | 29.72438 | 20 |

Table 3. Belgium telephone data post MM linear regression as implemented on R.

| $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ | | $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ |
|---|---|---|---|---|---|---|
| 0 | 52.86791 | - | | 6 | 1.64680 | 15 |
| 1 | 61.88021 | 20 | | 7 | 0.20243 | 21 |
| 2 | 62.33533 | 19 | $\tilde{g} =$ | **8** | **0.12525** | 14 |
| 3 | 58.43027 | 18 | | 9 | 0.14019 | 1 |
| 4 | 48.75151 | 17 | | 10 | 0.14792 | 22 |
| 5 | 33.51871 | 16 | | 11 | 0.15462 | 8 |

Table 4. Wood Specific Gravity post MM linear regression as implemented by R.

| $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ | | $g$ | $V_n^*(g)$ | $J(g)\backslash J(g-1)$ |
|---|---|---|---|---|---|---|
| 0 | 0.01399 | - | $\tilde{g} =$ | **4** | **0.00036** | 4 |
| 1 | 0.01849 | 19 | | 5 | 0.00042 | 5 |
| 2 | 0.01985 | 6 | | 6 | 0.00055 | 18 |
| 3 | 0.01447 | 8 | | 7 | 0.00077 | 1 |

**Acknowledgements**

This paper is dedicated to the memory of Dr. Daniel Schubert, who died in a motorbike accident in 2007.

## References

[1] T. Bednarski and B.R. Clarke, *Trimmed likelihood estimation of location and scale of the normal distribution*, Austral. J. Statist. **35** (1993), 141–153.

[2] T. Bednarski and B.R. Clarke, *Asymptotics for an adaptive trimmed likelihood location estimator*, Statistics **36** (2002), 1–8.

[3] P. Billingsley, *Convergence of Probability Measures*. New York: John Wiley 1968.

[4] R.W. Butler, *Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule*, Ann. Statist. **10** (1982), 197–204

[5] B.R. Clarke, *Empirical evidence for adaptive confidence intervals and identification of outliers using methods of trimming*, Austral. J. Statist. **36** (1994), 45–58.

[6] B.R. Clarke, *An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes*, Discussiones Mathematicae, Probability and Statistics **20** (2000), 25–50.

[7] Y. Dodge and J. Jurečková, *Adaptive combination of least squares and least absolute deviation estimators*, in: Statistical Data Analysis Based on the $L_1$-Norm and Related Methods ed. Y. Dodge, North Holland, Elsevier Science Publishers 1987.

[8] Y. Dodge and J. Jurečková, *Adaptive choice of trimming proportion in trimmed least-squares estimation*, Statistics & Probability Letters **33** (1997), 167–176.

[9] Y. Dodge and J. Jurečková, *Adaptive Regression*, New York, Springer-Verlag 2000.

[10] D. Gamble, The Analysis of Contaminated Tidal Data, PhD Thesis, Murdoch University, Western Australia 1999.

[11] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, New York, Wiley 1986.

[12] P.J. Huber, *The behaviour of maximum likelihood estimates under nonstandard conditions*, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability **1** (1967), 73–101.

[13] P.J. Huber, *Robust Statistical Procedures*, 2nd ed., CBMS-NSF, Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia 1996.

[14] L.A. Jaeckel, *Some flexible estimates of location*, Ann. Math. Statist. **42** (1971), 1540–1552.

[15] J. Jurečková, R. Koenker and A.H. Welsh, *Adaptive choice of trimming proportions*, Ann. Inst. Statist. Math., **46** (1994), 737–755.

[16] N.M. Neykov and C.H. Müller, Breakdown point and computation of trimmed likelihood estimators in generalized linear models, Developments in robust statistics (Vorau 2001), Heidelberg, Physica 2003.

[17] C.H. Müller and N. Neykov, *Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models*, J. Statist. Plann. Inference **116** (2004), 503–519.

[18] P.J. Rousseeuw, *Multivariate estimation with high breakdown point*, pp. 283–297 in: Mathematical Statistics and Applications. Vol.B, (1985), eds. Grossman, Pflug, Vincze and Wertz, Dordrecht:Reidel Publishing Co 1983.

[19] P.J. Rousseeuw, *Least Median of Squares Regression*, Journal of the American Statistical Association **79** (1984), 871–880.

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

[21] D.L. Vandev and N.M. Neykov, *Robust maximum likelihood in the Gaussian case*, p. 259–264 in: New Directions in Data Analysis and Robustness, Morgenthaler, S., Ronchetti, E. and Stahel, W.A. (eds.), Birkhäuser Verlag, Basel 1993.

Received 9 April 2010