Some asymptotics for elemental subsets in regression with applications

Keith Knight
Department of Statistical Sciences
University of Toronto
Toronto, Ont. M5G 1Z5

Abstract

In a linear regression model, elemental subsets consist of the minimum number of observations needed to estimate the regression parameter where the resulting estimators are called elemental estimators. Elemental estimators have a long history in statistics; many estimators are functions of elemental estimators and elemental estimators are used for outlier detection as well as for computation of highly robust estimators. In this paper, we consider some asymptotic theory for elemental subsets in linear regression. In particular, we derive the limiting distribution of the elemental subsets that produce "good" elemental estimators. A number of applications of this theory are also given, including a diagnostic for homoscedasticity, a heuristic for sampling elemental subsets in the computation of least median of squares estimates, and a diagnostic for the Powell estimator in a censored linear regression model.

Keywords: elemental subsets, homoscedasticity, least median of squares, Powell estimator

1 Introduction

Consider fitting the linear regression model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of length p of unknown parameters. An elemental estimator of $\boldsymbol{\beta}$ is based on p observations $(\boldsymbol{x}_{i_1}, Y_{i_1}), \cdots, (\boldsymbol{x}_{i_p}, Y_{i_p})$; defining $H = \{i_1, \cdots, i_p\}$ and

$$\begin{aligned} X_H^T &= \left(\boldsymbol{x}_{i_1} \cdots \boldsymbol{x}_{i_p}\right), \\ \boldsymbol{Y}_H^T &= \left(Y_{i_1} \cdots Y_{i_p}\right)^T, \end{aligned}$$

the elemental estimator $\widehat{\boldsymbol{\beta}}_H$ satisfies

$$\widehat{\boldsymbol{\beta}}_H = X_H^{-1}\boldsymbol{Y}_H$$

provided the inverse X_H^{-1} exists. The elemental estimators $\{\widehat{\boldsymbol{\beta}}_H\}$ are the foundation of a wide class of estimators of $\boldsymbol{\beta}$ in the model (1); roughly speaking, we can think of the set (or subset) of elemental estimators $\{\widehat{\boldsymbol{\beta}}_H\}$ as a set of multivariate observations (from some distribution) and use some measure of centrality of these data to define an estimator of $\boldsymbol{\beta}$ in (1). For example, the least squares estimator can be written as a weighted average of elemental estimators

$$\widehat{\boldsymbol{\beta}} = \frac{\sum\limits_H |X_H|^2 \widehat{\boldsymbol{\beta}}_H}{\sum\limits_H |X_H|^2}$$

where the sum extends over all subsets and $|X_H|$ is the absolute determinant of the matrix X_H.

Many other well-known estimators are either exactly elemental estimators or are determined by elemental estimators. For example, the L_1 (and other regression quantile) estimator (Koenker and Bassett, 1978) is an elemental estimator (for some subset H). Maire and Boscovich (1755) investigated the ellipticity of the earth using five measurements (at widely dispersed locations) of the arc length of one degree of latitude; their estimates were based on elemental estimates using pairs of observations. (More details on the work of Maire and Boscovich can be found in Stigler (1986) as well as Koenker and Portnoy (1987).) In the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the Theil-Sen (Theil, 1950; Sen, 1968) estimator of the slope β_1 is defined as the median of all elemental estimators of β_1 while Jurečková (1971) and Jaeckel (1972) propose rank estimators of β_1 that are weighted medians of elemental estimators of β_1. An extension of the Theil-Sen estimator to the multiple linear regression model (1) has been proposed by Dang *et al.* (2009); this estimator of $\boldsymbol{\beta}$ is defined to be a multivariate median of the elemental estimators $\{\widehat{\boldsymbol{\beta}}_H\}$. (Similar extensions have also been proposed by Oja and Niinimaa (1984) and by D'Esposito and Furno (1992a).) Rousseeuw and Bassett (1991) investigate using elemental estimators as approximations for computationally complex high breakdown estimators. Elemental estimation is also used in the detection of outliers; see, for example, Hawkins *at al.* (1984), D'Esposito and Furno (1992b) as well as Hadi and Simonoff (1993). Mayo and Gray (1997) give an excellent survey of elemental estimation.

In this paper, $\boldsymbol{\beta}$ (and $\{\varepsilon_i\}$) are somewhat arbitrary. For example, suppose that

$$Y_i = g(\boldsymbol{x}_i) + \nu_i \quad (i = 1, \cdots, n)$$

for some function g where $\{\nu_i\}$ are assumed to be independent random variables. Then for an arbitrary $\boldsymbol{\beta}$, we can write

$$Y_i \;=\; \boldsymbol{x}_i^T\boldsymbol{\beta} + \left\{g(\boldsymbol{x}_i) - \boldsymbol{x}_i^T\boldsymbol{\beta} + \nu_i\right\}$$

$$= \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$$

where ε_i now depends explicitly on $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$ unless $g(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

Asymptotic theory for elemental estimators has been considered by several authors, notably Hawkins (1993) as well as Olive and Hawkins (2007). Our main interest in this paper is the distribution of X_H for "good" elemental estimators $\widehat{\boldsymbol{\beta}}_H$, for example, those estimators that are close to "good" estimators of $\boldsymbol{\beta}$; this will be considered in section 2. In section 3, we will consider how this distribution can be used to check the extent to which the distribution of the errors $\{\varepsilon_i\}$ in (1) depends on $\{\boldsymbol{x}_i\}$. We will also apply the asymptotics to least median of squares estimation.

2 Asymptotics for X_H

For a bounded set $A \subset R^p$ with positive Lesbegue measure ($\lambda(A) > 0$) and some sequence $a_n \to \infty$ with $a_n/n \to 0$, define the conditional measure

$$\mathcal{P}_n(B|A) = \frac{\sum_H I\{X_H \in B, a_n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A\}}{\sum_H I\{a_n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A\}}. \tag{2}$$

We will assume (although it is not necessary to do so) that the summations in (2) are over the $\binom{n}{p}$ subsets of the form $H = \{i_1 < \cdots < i_p\}$ with $\boldsymbol{x}_1 \le \boldsymbol{x}_2 \le \cdots \le \boldsymbol{x}_n$ for some ordering on R^p. Thus we can restrict B to subsets of the order restricted space of matrices

$$\mathcal{O} = \{(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p) : \boldsymbol{t}_1 \le \boldsymbol{t}_2 \le \cdots \le \boldsymbol{t}_p\}.$$

$\mathcal{P}_n(\cdot|A)$ can be interpreted as the conditional distribution of X_H given that $a_n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A$.

The asymptotic behaviour of $\mathcal{P}_n$ can be studied by examining the behaviour of the numerator and denominator. Define

$$L_n(A, B) = \binom{n}{p}^{-1} \sum_H a_n^p I\{X_H \in B, a_n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A\}. \tag{3}$$

Suppose that $\{\varepsilon_i\}$ are i.i.d. with distribution function F with a density f with $f(0) > 0$; then for t close to 0, we have $F(t) - F(0) \approx t\, f(0)$. In this case,

$$E[L_n(A, B)] \approx \binom{n}{p}^{-1} f(0)^p \lambda(A) \sum_H I(X_H \in B)|X_H|$$

which suggests that

$$L_n(A, B) \xrightarrow{p} f(0)^p \lambda(A) \int_B |(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)| \, \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}$$

where μ is the limit (provided it exists) of the empirical measure of the covariates $\{\boldsymbol{x}_i\}$. This, in turn, would imply that

$$\mathcal{P}_n(B|A) \xrightarrow{p} \frac{\int_B |(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)| \, \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}}{\int_{\mathcal{O}} |(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)| \, \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}}$$

where limit is independent of A.

In fact, we can relax the i.i.d. assumption on $\{\varepsilon_i\}$ to allow the behaviour of the distribution functions near 0 of each ε_i to depend on $\boldsymbol{x}_i$.

(A1) Assume the model (1) with independent $\{\varepsilon_i\}$ where the distribution function F_i of ε_i satisfies

$$F_i(t) - F_i(0) = \kappa(\boldsymbol{x}_i) t \, \{1 + r_i(t)\}$$

for some non-negative function κ such that

$$\lim_{t \to 0} \max_{1 \le i \le n} |r_i(t)| \to 0.$$

(A2) (a) The empirical measure of $\{\boldsymbol{x}_i\}$ converges weakly to some μ with

$$\int \kappa(\boldsymbol{x}) \|\boldsymbol{x}\| \, \mu(d\boldsymbol{x}) < \infty.$$

(b) In addition, for some sequence $a_n \to \infty$ with $a_n/n \to 0$ (as $n \to \infty$)

$$\begin{aligned} \max_{1 \le i \le n} \frac{\|\boldsymbol{x}_i\|}{a_n} &\to 0 \\ \max_{1 \le i \le n} \frac{\kappa(\boldsymbol{x}_i)\|\boldsymbol{x}_i\|}{n} &\to 0. \end{aligned}$$

The weak convergence of the empirical measure $\{\boldsymbol{x}_i\}$ in part (a) of condition (A2) implies the weak convergence of the empirical measure of $\{(\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_p}) : i_1 < \cdots < i_p\}$ to the product measure $\mu \times \cdots \times \mu$ on the restricted space $\mathcal{O}$ Clearly, part (b) of condition (A2) holds if $\{\boldsymbol{x}_i\}$ and $\{\kappa(\boldsymbol{x}_i)\}$ are bounded. More generally, part (b) guarantees the convergence of certain sums to their "intuitive" limits; for example, given part (a) of condition (A2) and the condition

$$\max_{1 \le i \le n} \frac{\kappa(\boldsymbol{x}_i)\|\boldsymbol{x}_i\|}{n} \to 0,$$

it follows that

$$\binom{n}{p}^{-1} \sum_{i_1 < \cdots < i_p} |(\kappa(\boldsymbol{x}_{i_1})\boldsymbol{x}_{i_1} \cdots \kappa(\boldsymbol{x}_{i_p})\boldsymbol{x}_{i_p})| \to \int_{\mathcal{O}} |(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1 \cdots \kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \, \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}.$$

THEOREM 1. Assume the model (1) and conditions (A1) and (A2) on $\{\varepsilon_i\}$, $\{\boldsymbol{x}_i\}$, $\{a_n\}$. If $\mathcal{P}_n(B|A)$ is defined as in (2) then

$$\mathcal{P}_n(B|A) \xrightarrow{p} \mathcal{P}(B) = \frac{\int_B |(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1 \cdots \kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}}{\int_{\mathcal{O}} |(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1 \cdots \kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}}.$$

Proof. Define $L_n(A, B)$ as in (3). It suffices to show that

$$\begin{aligned} E[L_n(A,B)] &\to \lambda(A) \int_B |(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)| \prod_{i=1}^{p} \kappa(\boldsymbol{t}_i) \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\} \\ &= \lambda(A) \int_B |(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1 \cdots \kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\} \qquad (4) \\ \mathrm{Var}[L_n(A,B)] &\to 0. \qquad (5) \end{aligned}$$

Define for $H = \{i_1, \cdots, i_p\}$, $\boldsymbol{\varepsilon}_H$ to be the vector with elements $\varepsilon_{i_1}, \cdots, \varepsilon_{i_p}$. Then we have

$$\begin{aligned} a_n^p E[I\{X_H \in B, a_n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A\}] &= a_n^p P(\boldsymbol{\varepsilon}_H \in a_n^{-1} X_H A) I(X_H \in B) \\ &= \left\{ \prod_{j=1}^{p} \kappa(\boldsymbol{x}_{i_j}) \right\} |X_H| I(X_H \in B) \lambda(A) + r_H(A) \end{aligned}$$

where $\sup_H |r_H(A)| \to 0$ for each A. Thus (4) follows. To show (5), note that for sets H_1 and H_2 with $c > 0$ common elements,

$$\mathrm{Cov}\left[I(\boldsymbol{\varepsilon}_{H_1} \in a_n^{-1} X_{H_1} A), I(\boldsymbol{\varepsilon}_{H_2} \in a_n^{-1} X_{H_2} A) \right] = O(a_n^{c-2p})$$

(with the covariance exactly 0 when $c = 0$) and there are $O(n^{2p-c})$ pairs of sets with c common elements. Thus by simple computation $\mathrm{Var}[L_n(A, B)] = O(a_n/n)$ and so (5) holds. □

What does Theorem 1 tell us? If we assume the model (1) where $\{\varepsilon_i\}$ are independent with densities $f_1, \cdots, f_n$ where $f_i(0) = \kappa(\boldsymbol{x}_i)$ then for "good" elemental estimators $\widehat{\boldsymbol{\beta}}_H = X_H^{-1} \boldsymbol{Y}_H$, X_H will have columns that come from biased sampling from $\{\boldsymbol{x}_i\}$ where the bias is proportional to

$$|(\kappa(\boldsymbol{x}_{i_1})\boldsymbol{x}_{i_1} \cdots \kappa(\boldsymbol{x}_{i_p})\boldsymbol{x}_{i_p})|.$$

Since the absolute determinant of a matrix is a measure of the dispersion of its columns, this means that for good elemental estimators, the dispersion of the vectors $\{\kappa(\boldsymbol{x}_{i_j})\boldsymbol{x}_{i_j} : i_j \in H\}$ is greater than if we simply sampled p vectors (without replacement) from $\{\boldsymbol{x}_i\}$. We can also obtain the marginal density (with respect to μ) of a single column of X_H producing a good elemental estimator; this density is proportional to

$$\int \cdots \int |(\kappa(\boldsymbol{x})\boldsymbol{x}\ \kappa(\boldsymbol{t}_2)\boldsymbol{t}_2 \cdots \kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \, \mu(d\boldsymbol{t}_2) \times \cdots \times \mu(d\boldsymbol{t}_p).$$

For example, suppose that $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $(i = 1, \cdots, n)$ where the limiting measure of $\{x_i\}$ is uniform on $[0, 1]$ and $\kappa(\boldsymbol{x}_i) = x_i^\gamma$ for some $\gamma \geq 0$ then this density (with respect to Lebesgue measure on $[0, 1]$) is

$$g_\gamma(x) = (2\gamma + 3)x^\gamma \left\{\gamma + 1 - (2 + \gamma)x + 2x^{\gamma+2}\right\}$$

(where the density g_γ integrates to 2 to reflect the number of vectors in X_H). When $\gamma = 0$ (so that $\kappa(\boldsymbol{x}_i)$ is constant), $g_\gamma(x)$ is symmetric around $x = 1/2$. As γ increases, relatively more weight is given to values of x close to 1 and as γ tends to infinity, this distribution becomes concentrated around 1; in particular, as $\gamma \to \infty$, we have

$$\int_{1-t/\gamma}^{1} g_\gamma(x)\, dx \to 2 - 2\exp(-t)\left\{t + \exp(-t)\right\}.$$

Note that Theorem 1 does not hold for $a_n = n$. Olive and Hawkins (2007) show (under mild regularity conditions on $\{\boldsymbol{x}_i\}$ and $\{\varepsilon_i\}$) that for any $\boldsymbol{\beta}$ there exists an elemental estimator $\widehat{\boldsymbol{\beta}}_H$ satisfying $\|\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}\| = O_p(n^{-1})$. In this case, we can define

$$M_n(A, B) = \sum_H I\{X_H \in B, n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A\}, \tag{6}$$

which has a Poisson limit under conditions (A1) and (A2).

THEOREM 2. Assume conditions (A1) and (A2) on the errors $\{\varepsilon_i\}$ and the covariates $\{\boldsymbol{x}_i\}$ in (1) for $a_n = n$. If M_n is defined as in (6) then

$$M_n(A, B) \xrightarrow{d} M(A, B)$$

where $M(A, B)$ has a Poisson distribution with mean

$$m(A, B) = \lambda(A) \int_B |(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1 \cdots \kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \{\mu(d\boldsymbol{t}_1) \times \cdots \times \mu(d\boldsymbol{t}_p)\}$$

Proof. This is a straightforward extension of the results of Silverman and Brown (1978) (see also Brown and Eagleson (1971) and Brown (1978)), which deals with Poisson convergence of U-statistics. Following the approach used to prove Theorem A of Silverman and Brown (1978), the result follows by noting that for H_1 and H_2 with $c > 0$ common elements

$$E\left[I(\boldsymbol{\varepsilon}_{H_1} \in n^{-1}X_{H_1}A)I(\boldsymbol{\varepsilon}_{H_2} \in n^{-1}X_{H_2}A)\right] = O(n^{c-2p})$$

where $\boldsymbol{\varepsilon}_H$ is defined as in the proof of Theorem 1. □

Theorem 2 can be easily extended to give the weak convergence of the point process M_n (Kallenberg, 1976) to a Poisson process M whose mean measure is given in Theorem 2.

We can use Theorem 2 to derive, for example, the asymptotic distribution of the closest elemental estimator to $\boldsymbol{\beta}$ in the Euclidean norm. Define

$$\widetilde{\boldsymbol{\beta}}_n = \arg\min_H \|\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}\|_2.$$

Then $n(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$, where $\boldsymbol{U}$ has density function

$$g(\boldsymbol{u}) = c(\kappa,\mu)\exp\left\{-\frac{c(\kappa,\mu)\pi^{p/2}}{\Gamma(1+p/2)}\|\boldsymbol{u}\|_2^p\right\}$$

where

$$c(\kappa,\mu) = \int_{\mathcal{O}} |(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1\cdots\kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)|\{\mu(d\boldsymbol{t}_1)\times\cdots\times\mu(d\boldsymbol{t}_p)\}.$$

The asymptotic independence between the events $X_H \in B$ and $a_n(\widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta}) \in A$ in Theorems 1 and 2 is a consequence the linearity of the distribution functions F_i in a neighbourhood of 0 as outlined in assumption (A1). It is possible to relax (A1) by assuming that (for example) the distribution functions $\{F_i\}$ are regularly varying in a neighbourhood of 0, that is,

$$b_n\{F_i(t/a_n) - F_i(0)\} = \kappa(\boldsymbol{x}_i)\psi(t) + r_{ni}(t)$$

for some sequences $a_n, b_n \to \infty$ with $b_n/n \to 0$ where ψ is a strictly increasing function with derivative ψ' and $r_{ni}(t)$ tends to 0 sufficiently uniformly over t as $n \to \infty$. The function ψ is of the form

$$\psi(t) = \begin{cases} \lambda^+ t^\alpha & \text{for } t \geq 0 \\ -\lambda^-(-t)^\alpha & \text{for } t < 0. \end{cases}$$

where $\alpha > 0$ and $\lambda^+, \lambda^- \geq 0$ with $\lambda^+ + \lambda^- > 0$. Then (under additional regularity conditions), we will have

$$\mathcal{P}_n(B|A) \xrightarrow{p} \frac{\displaystyle\int_B \int_{(\mathbf{t}_1\cdots\mathbf{t}_p)A} \prod_{j=1}^p \{\kappa(\boldsymbol{t}_j)\psi'(s_j)\,ds_j\}\{\mu(d\boldsymbol{t}_1)\times\cdots\times\mu(d\boldsymbol{t}_p\}}{\displaystyle\int_{\mathcal{O}} \int_{(\mathbf{t}_1\cdots\mathbf{t}_p)A} \prod_{j=1}^p \{\kappa(\boldsymbol{t}_j)\psi'(s_j)\,ds_j\}\{\mu(d\boldsymbol{t}_1)\times\cdots\times\mu(d\boldsymbol{t}_p\}}$$

where the limit does depend on A when ψ is non-linear.

3 Some applications

Approximating $\mathcal{P}$

As we noted above, for each bounded set A, the limit of $\mathcal{P}_n(\cdot|A)$, $\mathcal{P}$ is independent of A when conditions (A1) and (A2) hold. In this case, the limiting distribution of $\{X_H : \widehat{\boldsymbol{\beta}}_H - \boldsymbol{\beta} \in a_n^{-1}A\}$ has a density (with respect to the product measure $\mu\times\cdots\times\mu$) proportional to

$$|(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1\cdots\kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)|.$$

Given the covariate vectors $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ whose empirical distribution is approximately μ, we can approximate $\mathcal{P}$ by $\mathcal{Q}_n$, a discrete distribution on the matrices $\{X_H\}$ with density

$$q_n(\boldsymbol{t}_1,\cdots,\boldsymbol{t}_p) = K(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)|(\kappa(\boldsymbol{t}_1)\boldsymbol{t}_1\cdots\kappa(\boldsymbol{t}_p)\boldsymbol{t}_p)| \quad \text{for } \boldsymbol{t}_1,\cdots,\boldsymbol{t}_p \in \{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}. \qquad (7)$$

(Alternatively, given μ, we can generate a random sample of vectors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ from μ to define q_n in (7).) We can easily generate samples of matrices from q_n using Gibbs sampling (Geman and Geman, 1984); to do so, it is convenient to dispense with the ordering and regard q_n as a density on the set of matrices whose columns lie in $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ with no restrictions. In this case, the conditional density of the j-th column given the remaining $p-1$ columns is

$$q_n(\boldsymbol{t}_j | \boldsymbol{t}_k \ k \neq j) = \frac{|D_j(\boldsymbol{t}_j | \boldsymbol{t}_k \ k \neq j)|}{\sum_{i=1}^{n} |D_j(\boldsymbol{x}_i | \boldsymbol{t}_k \ k \neq j)|} \quad \text{for } \boldsymbol{t}_j \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$$

where $D_j(\boldsymbol{x} | \boldsymbol{t}_k \ k \neq j)$ is the matrix whose j-th column is $\kappa(\boldsymbol{x})\boldsymbol{x}$ and whose k-th column (for $k \neq j$) is $\kappa(\boldsymbol{t}_k)\boldsymbol{t}_k$. For each j, the n determinants $|D_j(\boldsymbol{x}_1 | \boldsymbol{t}_k \ k \neq j)|, \cdots, |D_j(\boldsymbol{x}_n | \boldsymbol{t}_k \ k \neq j)|$ can each be evaluated in terms of a single set of p cofactors.

Assessing density homogeneity

Suppose that $\widehat{\boldsymbol{\beta}}_n$ is an estimator of $\boldsymbol{\beta}$ in the model (1). Homoscedasticity of the errors $\{\varepsilon_i\}$ in (1) can be assessed by looking at various residual plots, essentially looking for patterns that might indicate heteroscedasticity of the errors. Generally speaking, homoscedasticity is typically defined to mean that the variance (or some other measure of scale) of ε_i is constant (in particular, independent of $\boldsymbol{x}_i$). However, we can also define the notion of local homoscedasticity, which means that the densities $f_1, \cdots, f_n$ of $\varepsilon_1, \cdots, \varepsilon_n$ satisfy $f_1(0) = \cdots = f_n(0)$. In regression quantile estimation, the assumption of local homoscedasticity greatly facilitates asymptotic inference.

Theorem 1 suggests the following approach for assessing local homoscedasticity. Given $\widehat{\boldsymbol{\beta}}_n$, we can find elemental estimates $\widehat{\boldsymbol{\beta}}_{H_1}, \cdots, \widehat{\boldsymbol{\beta}}_{H_m}$ that are close to $\widehat{\boldsymbol{\beta}}_n$. Then under local homoscedasticity, the empirical distribution of the matrices $X_{H_1}, \cdots, X_{H_m}$ should look like the distribution $\mathcal{P}$ with $\kappa(\boldsymbol{x})$ constant; as above, we can approximate $\mathcal{P}$ by $\mathcal{Q}_n^{(0)}$ whose density is now

$$q_n^{(0)}(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_p) = K(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n) |(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)| \quad \text{for } \boldsymbol{t}_1, \cdots, \boldsymbol{t}_p \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}. \tag{8}$$

Given $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$, we can use Gibbs sampling to draw a sample from $q_n^{(0)}$, which can then be compared to the empirical distribution of $\{X_{H_j} : j = 1, \cdots, m\}$.

There are several practical issues to consider. First, since the space of matrices $R^{p \times p}$ is large, we need to choose appropriate real-valued functions ϕ of the columns of X_{H_j} for comparison to $\mathcal{Q}_n^{(0)}$ with density $q_n^{(0)}$ defined in (8). If q_n is as defined in (7) for some κ then for a given real-valued function ϕ, we can define the total variation distance between the distributions of $\phi(X_H)$ under q_n and $q_n^{(0)}$ by

$$\sup_A \left| \mathcal{Q}_n\{\phi(X_H) \in A\} - \mathcal{Q}_n^{(0)}\{\phi(X_H) \in A\} \right|$$

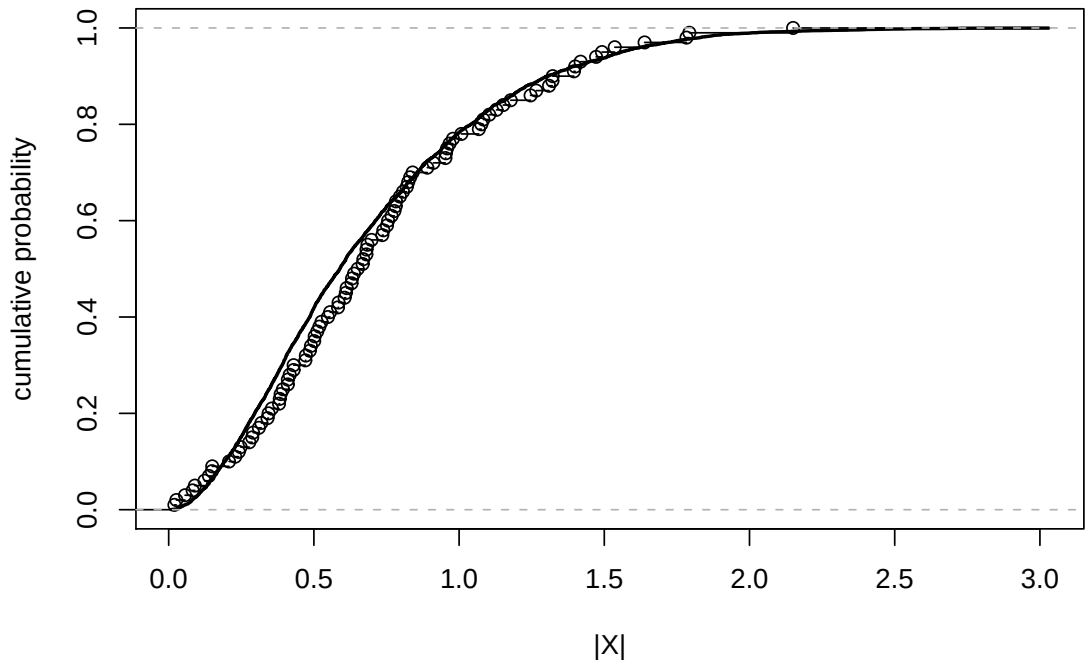

Figure 1: Distribution of $|X_{H_j}|$ for $j = 1, \cdots, 100$ compared to distribution from $\mathcal{Q}_n$ for homoscedastic errors.

and this is maximized for

$$\phi(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_p) = \prod_{j=1}^{p} \kappa(\boldsymbol{t}_j).$$

To use this in practice, however, we would need to hypothesize (or guess) the form of κ. Another possibility is to assume that the density of X_{H_j} has the local alternative form

$$p_n(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_p) = K_\delta(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)|(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)|^{1+\delta} \quad \text{for } \boldsymbol{t}_1, \cdots, \boldsymbol{t}_p \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$$

for some $\delta \neq 0$ (a "local" approximation to $q_n^{(0)}$), which leads to $\phi(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_p) = |(\boldsymbol{t}_1 \cdots \boldsymbol{t}_p)|$ maximizing the total variation distance; in the absence of any specific knowledge of κ, this seems to be a good default choice. Alternatively, we could also use the trace or the largest eigenvalue of $X_{H_j}^T X_{H_j}$. Second, we need to define what we mean by $\widehat{\boldsymbol{\beta}}_H$ being close of $\widehat{\boldsymbol{\beta}}_n$. The obvious definition is based on some distance measure (for example, Euclidean distance) between $\widehat{\boldsymbol{\beta}}_H$ and $\widehat{\boldsymbol{\beta}}_n$. Alternative, if $\widehat{\boldsymbol{\beta}}_n$ minimizes some objective function g, we could define distance in terms of $g(\widehat{\boldsymbol{\beta}}_H) - g(\widehat{\boldsymbol{\beta}}_n)$. Third is the issue of how to choose m, the number of matrices to be compared to $\mathcal{Q}_n$. Suppose that we are able to compute N elemental estimates, either via complete enumeration (in which case $N = O(n^p)$) or by sampling the p columns of X_H without replacement from $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$. Then we should take $m = N^\tau$ for some $\tau \in (0, 1)$; ideally τ should be sufficient large so that $n^{1-\tau}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = o_p(1)$.

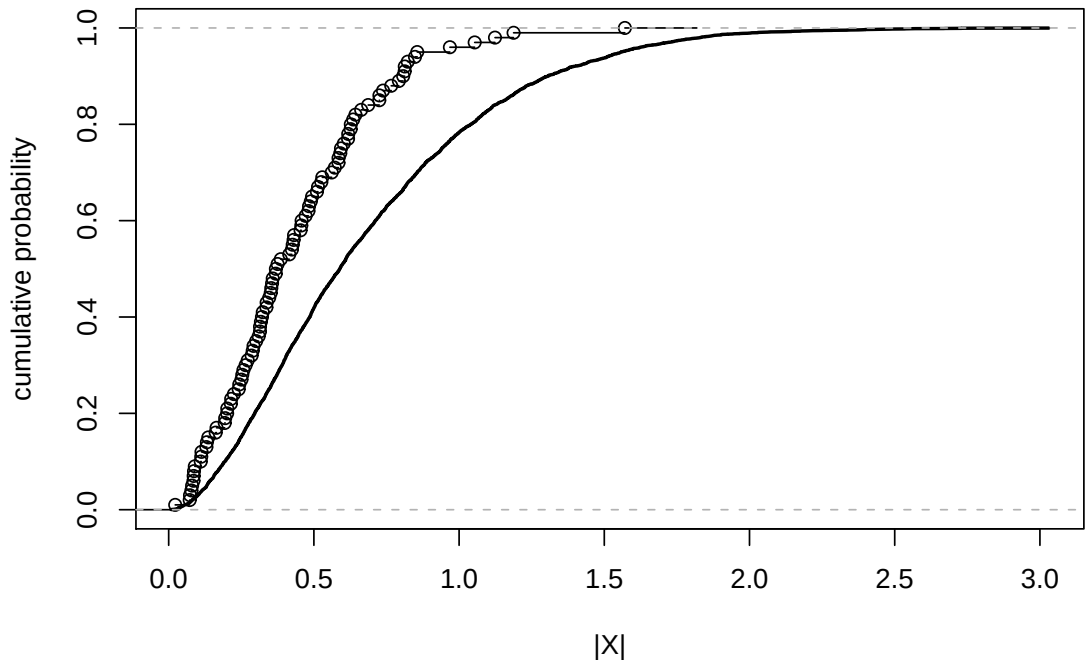

Figure 2: Distribution of $|X_{H_j}|$ for $j = 1, \cdots, 100$ compared to distribution from $\mathcal{Q}_n$ for heteroscedastic errors.

To illustrate, we consider fitting the model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Values of the two covariates $\{(x_{1i}, x_{2i})\}$ for $i = 1, \cdots, 100$ were drawn from a zero mean bivariate normal distribution with covariance matrix

$$C = \begin{pmatrix} 1 & 1 \\ 1 & 1.04 \end{pmatrix}.$$

The two covariates are highly correlated; the sample correlation is 0.975 while the theoretical correlation is 0.962. We will consider two scenarios for the errors $\{\varepsilon_i\}$:

(a) $\{\varepsilon_i\}$ are i.i.d. $\mathcal{N}(0, 1)$;

(b) $\{\varepsilon_i\}$ are independent with $\varepsilon_i \sim \mathcal{N}(0, |x_{1i} - x_{2i}|^2)$.

The heteroscedasticity given by (b) is not immediately apparent looking at (bivariate) scatterplots of the response versus each of the covariates. In both cases, $\boldsymbol{\beta}$ is estimated using least squares and we take a sample of 1000 elemental subsets and look at the distribution of

$|X_H|$ for the best 100 elemental estimates using residual sum of squares as a criterion. Figures 1 and 2 show the empirical distributions of $\{|X_{H_j}| : j = 1, \cdots, n\}$ for scenarios (a) and (b) compared to an estimate of the distribution from $\mathcal{Q}_n$ computed from 5000 observations using Gibbs sampling. Note that the empirical distribution for the i.i.d. errors agrees very well with $\mathcal{Q}_n$ while, on the other hand, the absolute determinants for the heteroscedastic errors are clearly stochastically smaller than those from $\mathcal{Q}_n$.

Least median of squares

Least median of squares (LMS) estimation was introduced by Rousseeuw (1984) as a highly robust (high breakdown point) estimator of $\boldsymbol{\beta}$ in (1). The LMS estimator minimizes the objective function

$$g(\boldsymbol{\phi}) = \text{median}_{1 \le i \le n} |Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}|;$$

the objective function g is very non-smooth and cannot be minimized using classical optimization methods. However, the LMS estimator can be shown to be a Chebyshev (or L_∞) estimator of $\boldsymbol{\beta}$ for some subset of $n/2$ observations; see, for example, Stromberg (1993). A Chebyshev estimator is itself an elemental estimator based on $p + 1$ observations from the augmented model

$$Y_i = \gamma + \boldsymbol{x}_i^T \boldsymbol{\beta} + (\varepsilon_i - \gamma) \tag{9}$$

$$Y_i = -\gamma + \boldsymbol{x}_i^T \boldsymbol{\beta} + (\varepsilon_i + \gamma) \tag{10}$$

for $i = 1, \cdots, n$. An elemental estimator $(\widehat{\gamma}_H, \widehat{\boldsymbol{\beta}}_H)$ satisfies for some subset $H = \{i_1, \cdots, i_{p+1}\}$

$$Y_{i_j} = \pm \widehat{\gamma}_H + \boldsymbol{x}_{i_j}^T \widehat{\boldsymbol{\beta}}_H \quad \text{for } j = 1, \cdots, p+1. \tag{11}$$

Theorem 1 (with appropriate modifications) will hold for elemental estimators $(\widehat{\gamma}_H, \widehat{\boldsymbol{\beta}}_H)$ (in the augmented model (9) and (10)) satisfying (11).

As mentioned above, the LMS estimator $\widehat{\boldsymbol{\beta}}_n$ satisfies (11) for some H; under i.i.d. errors $\{\varepsilon_i\}$ with a unimodal density f, the corresponding estimator $\widehat{\gamma}_n = \widehat{\gamma}_H$ (satisfying (11)) is a consistent estimator of γ_0 satisfying $f(-\gamma_0) = f(\gamma_0)$ and $P(|\varepsilon_i| \le \gamma_0) = 0.5$.

Algorithms for computing LMS estimates typically involve some sort of randomized selection of elemental subsets in the case where the computation of all elemental estimators is prohibitive; see, for example, Hawkins and Olive (2002). It may be possible that sampling elemental subsets from the density q_n in (7) for some κ (whose value may vary from iteration to iteration) might be useful in practice.

As an example, we consider the data from Rousseeuw and Leroy (1987) consisting of measurements from 47 stars in the stellar cluster CYGOB1; we have $\{(x_i, y_i) : i = 1, \cdots, 47\}$ where x_i is the logarithm of temperature of the surface of the star and y_i is the logarithm of the light intensity. The data are shown in Figure 3 with the LMS and least squares fits.

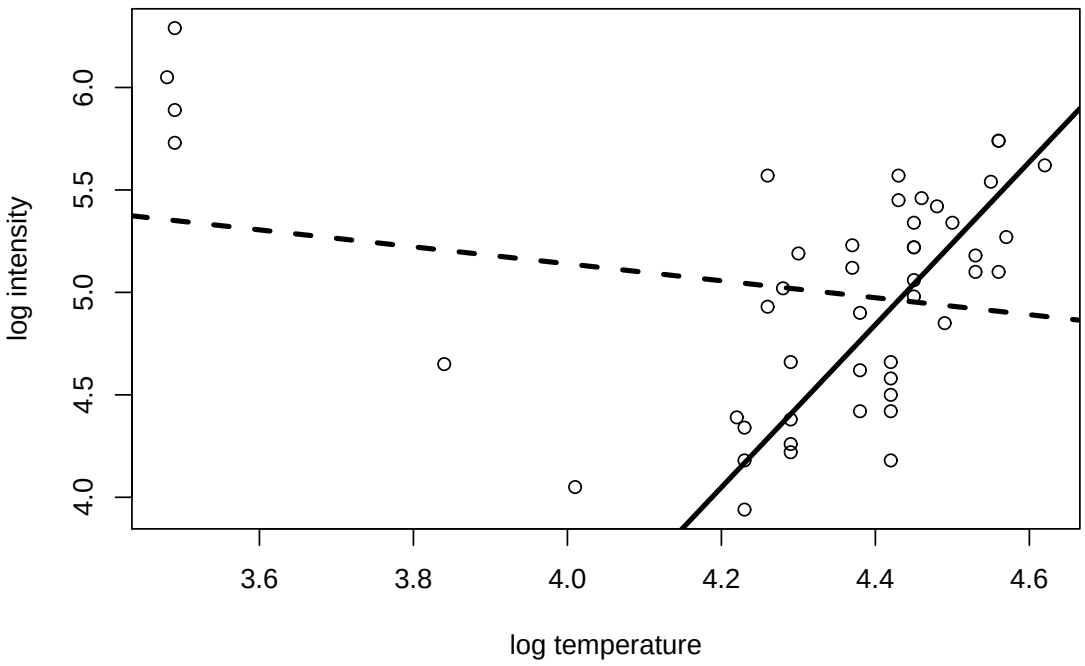

Figure 3: CYGOB1 data (Rousseeuw and Leroy, 1987) with LMS (solid line) and least squares (dashed line) fits.

From the LMS fit, there appear to be four stars (with the smallest x_i) that are anomalous; we would expect that the best elemental estimates (with respect to median absolute residual) would ignore these observations. To investigate this, we look at the distribution of the best 500 (out of all possible) elemental LMS fits. Figure 4 shows the empirical distribution of $|X_H|$ (from the augmented model (9) and (10)) compared to the distribution of the absolute determinants from the densities

$$q^{(1)}(\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3) = K(x_1, \cdots, x_{47})|(\boldsymbol{t}_1\ \boldsymbol{t}_2\ \boldsymbol{t}_3)| \quad \text{for } \boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3 \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{47}\} \tag{12}$$

$$q^{(2)}(\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3) = K(x_1, \cdots, x_{43})|(\boldsymbol{t}_1\ \boldsymbol{t}_2\ \boldsymbol{t}_3)| \quad \text{for } \boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3 \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{43}\} \tag{13}$$

$$q^{(3)}(\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3) = K(x_1, \cdots, x_{42})|(\boldsymbol{t}_1\ \boldsymbol{t}_2\ \boldsymbol{t}_3)| \quad \text{for } \boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3 \in \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{42}\} \tag{14}$$

where $x_1 \geq x_2 \geq \cdots \geq x_{47}$ and the vector $\boldsymbol{x}_i = (\pm 1, 1, x_i)^T$. Clearly, the empirical distribution is closest to $q^{(3)}$, which deletes the four clearly anomalous observations plus the observation with the next smallest value of x_i; in fact, none of the best 500 elemental estimates uses these five observations. (Other other observation is similarly excluded from the best 500 elemental estimates.)

Finally, we compare the frequency of each observation in the best 500 and best 100 elemental estimates to its "expected" frequency under the distribution $q^{(3)}$ defined in (14).

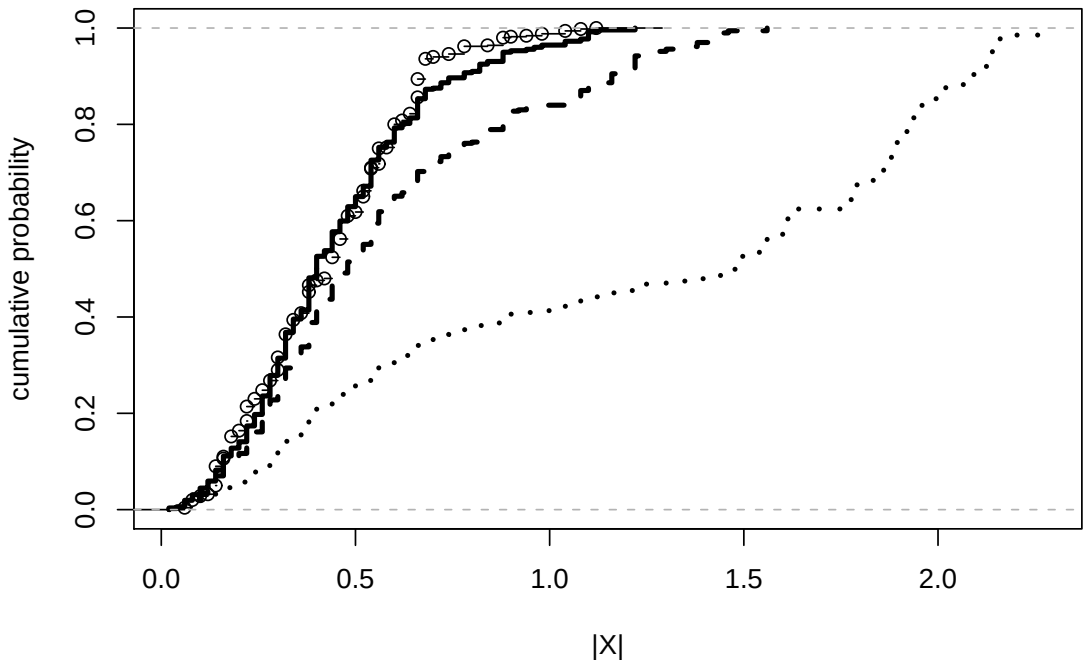

Figure 4: Distribution of $|X_H|$ for the best 500 LMS fits compared to distribution from $q^{(1)}$ (dotted lines), $q^{(2)}$ (dashed lines), and $q^{(3)}$ (solid lines) as defined in (12), (13), and (14).

Figures 5 and 6 give the ratio of the raw frequency $f(x_i)$ of each observation to its expectation $f_0(x_i)$ under $q^{(3)}$ (estimated using 50000 samples) for the best 500 and best 100 elemental estimates, respectively; the areas of the circles increase with the ratio $f(x_i)/f_0(x_i)$. In both cases, the average of $f(x_i)/f_0(x_i)$ for x_i in a neighbourhood of a given x is approximately 1 with this ratio being largest for x_i lying close to the lines

$$y = -12.63 \pm 0.26 + 3.97x$$

where -12.63 and 3.97 are the LMS estimates of β_0 and β_1, respectively, and $\widehat{\gamma}_H = 0.26$ is the best elemental estimate of γ in the augmented model (9) and (10).

The Powell estimator

Consider the censored linear regression model

$$Y_i = \max\{\boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i, 0\} \quad (i = 1, \cdots, n) \tag{15}$$

where the response variables $\{Y_i\}$ are (left-) censored at 0, and assume that $\{\varepsilon_i\}$ are independent random variables with median 0; $\boldsymbol{x}^T\boldsymbol{\beta}$ is the conditional median of the uncensored

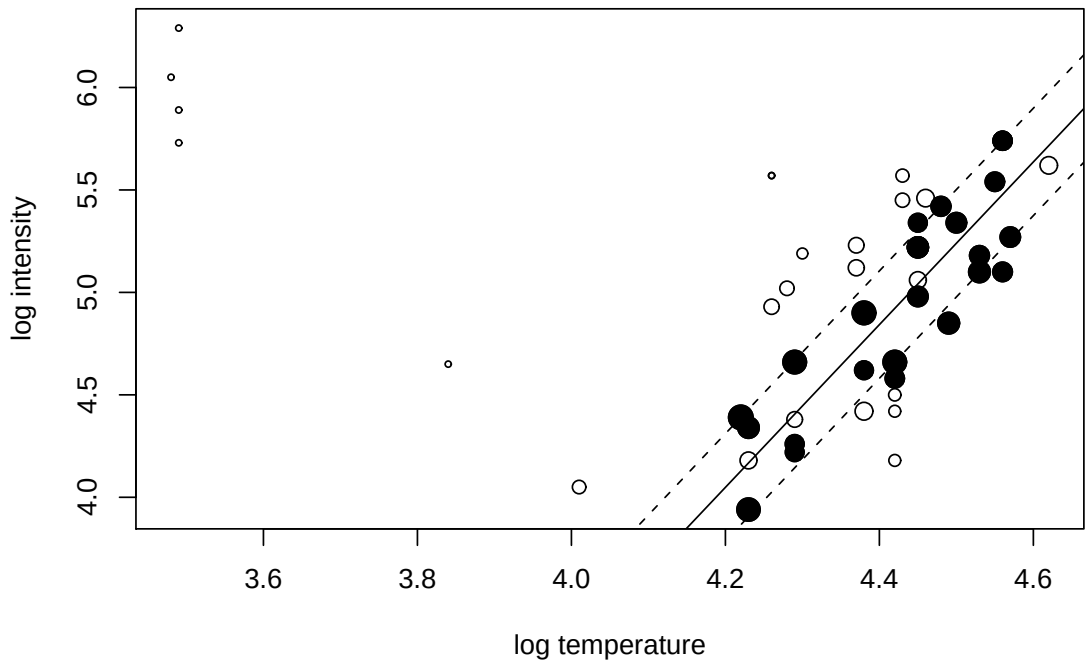

Figure 5: Frequency ratio $f(x_i)/f_0(x_i)$ for the best 500 elemental estimates for the CYGOB1 data; the area of the circles increase with $f(x_i)/f_0(x_i)$ and observations whose ratio is greater than 1 are shaded.

response at the predictor $\boldsymbol{x}$. Powell (1984) proposed to estimate $\boldsymbol{\beta}$ by minimizing the objective function

$$g(\boldsymbol{\phi}) = \sum_{i=1}^{n} |Y_i - \max\{\boldsymbol{x}_i^T \boldsymbol{\phi}, 0\}|. \tag{16}$$

(Powell (1986) extends this to censored regression quantile estimation by replacing the absolute value function by the "check" function $\rho_\tau(x) = x\{\tau - I(x \leq 0)\}$.) The Powell estimator $\widehat{\boldsymbol{\beta}}$ is an elemental estimator although efficient computation of it is complicated by the fact that this objective function is not convex, and may have multiple local minima. A number of methods for computing the Powell estimator have been proposed; see, for example, Fitzenberger and Winker (2007) and Hosseinkouchack (2011). However, the difficulty of computation has been used as a motivation for finding easier-to-compute alternatives to the Powell estimator with the same asymptotic properties.

Under standard regularity conditions, the Powell estimator is asymptotically equivalent to minimizing

$$\sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T \boldsymbol{\phi}| I(\boldsymbol{x}_i^T \boldsymbol{\beta} > 0) \tag{17}$$

where $\boldsymbol{\beta}$ is the unknown parameter in the model (15). However, Koenker (2008) notes that

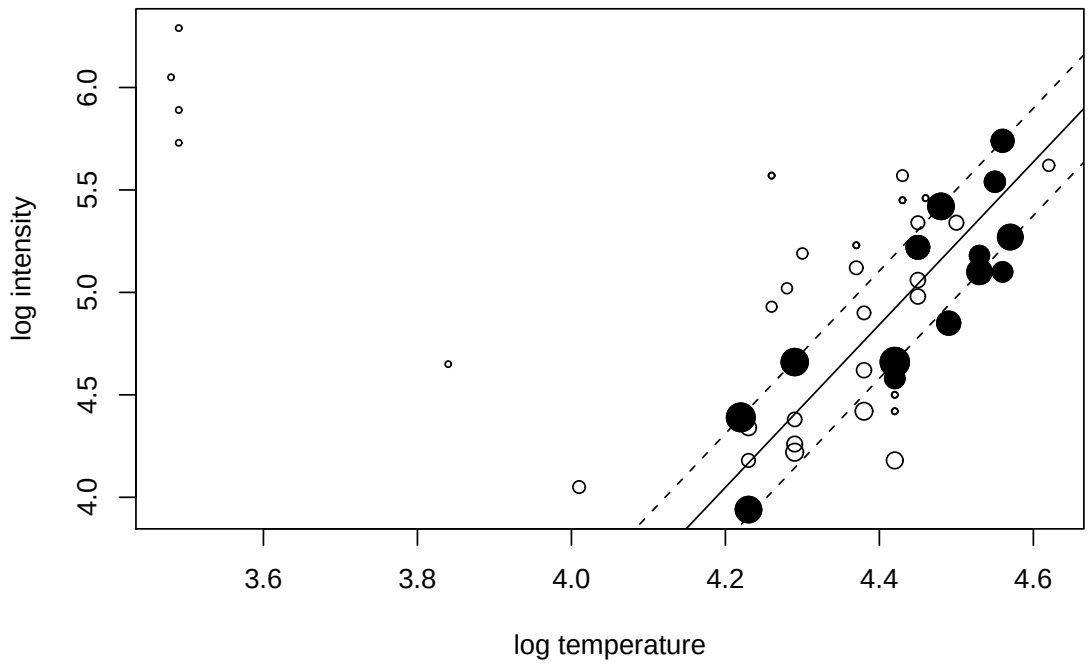

Figure 6: Frequency ratio $f(x_i)/f_0(x_i)$ for the best 100 elemental estimates for the CYGOB1 data; the area of the circles increase with $f(x_i)/f_0(x_i)$ and observations whose ratio is greater than 1 are shaded.

the global minimizer of (16) is not necessarily a good estimator of $\boldsymbol{\beta}$ in (15), particularly when there is a large amount of censoring, and that estimators based on sub-optimal solutions are often more sensible; one possible explanation is that for observations with $Y_i = 0$, the contribution to the objective function is 0 whenever $\boldsymbol{x}_i^T\boldsymbol{\phi} \leq 0$ (and fixed otherwise), which may contribute to non-trivial finite sample bias. In other words, if we define

$$M(\boldsymbol{\phi}) = \sum_{i=1}^{n} I(\boldsymbol{x}_i^T\boldsymbol{\phi} > 0) \tag{18}$$

then $M(\widehat{\boldsymbol{\beta}})$ may tend to underestimate $M(\boldsymbol{\beta})$, particularly if $M(\boldsymbol{\beta})$ is close to n and the model (15) holds. Chernozhukov and Hong (2002) as well as Tang *et al.* (2012) propose multi-step estimation procedures whose estimators ultimately minimize of proxy of (17),

$$\sum_{i=1}^{n} |Y_i - \boldsymbol{x}_i^T\boldsymbol{\phi}|\widehat{w}(\boldsymbol{x}_i)$$

where $\widehat{w}(\boldsymbol{x}_i) = 1$ if the (estimated) probability that the response at $\boldsymbol{x}_i$ is non-censored exceeds some threshold with $\widehat{w}(\boldsymbol{x}_i) = 0$ otherwise. If $\{\widehat{w}(x_i)\}$ are estimated appropriately

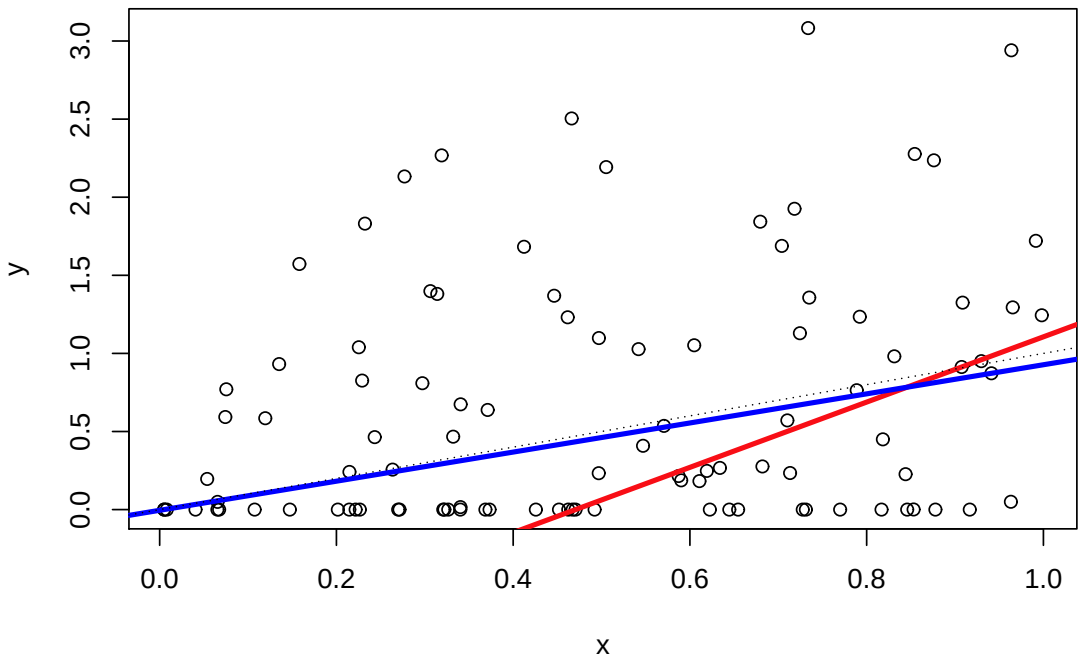

Figure 7: Scatterplot of simulated data from the model (19) using $\beta_0 = 0$ and $\beta_1 = 1$ with estimated conditional medians — the red line uses the Powell estimates and the blue line is the best elemental estimate $\widehat{\boldsymbol{\beta}}_H$ with $M(\widehat{\boldsymbol{\beta}}_H) = 100$. The dotted line is the true conditional median function.

then we should have

$$\sum_{i=1}^{n} \widehat{w}(\boldsymbol{x}_i) \approx M(\boldsymbol{\beta}).$$

The results in section 2 suggest that we can find "good" estimators of $\boldsymbol{\beta}$ by looking at elemental estimators $\widehat{\boldsymbol{\beta}}_H$ that nearly minimize the objective function g in (16); if we suspect that bias in the Powell estimator may be an issue, we may want to consider $\widehat{\boldsymbol{\beta}}_H$ such that $M(\widehat{\boldsymbol{\beta}}_H) > M(\widehat{\boldsymbol{\beta}})$ for $M(\cdot)$ defined in (18). For example, consider data generated from the simple model

$$Y_i = \max\{\beta_0 + \beta_1 x_i + \varepsilon_i, 0\} \tag{19}$$

where $\{x_i\}$ are uniformly distributed on $[0, 1]$ and $\{\varepsilon_i\}$ are i.i.d. $\mathcal{N}(0, 1)$ random variables. Figure 7 shows a scatterplot of 100 observations generated using $\beta_0 = 0$ and $\beta_1 = 1$ (so that $\beta_0 + \beta_1 x_i > 0$ for all i); the two lines are based on the Powell estimates of β_0 and β_1 (red line) and the best elemental estimates of β_0 and β_1 (blue line) subject to the constraint $M(\widehat{\boldsymbol{\beta}}_H) = 100$. For these data, the Powell estimates are particularly poor ($M(\widehat{\boldsymbol{\beta}}) = 49$) while the elemental estimates with $M(\widehat{\boldsymbol{\beta}}_H) = 100$ are quite good. (A simulation of this

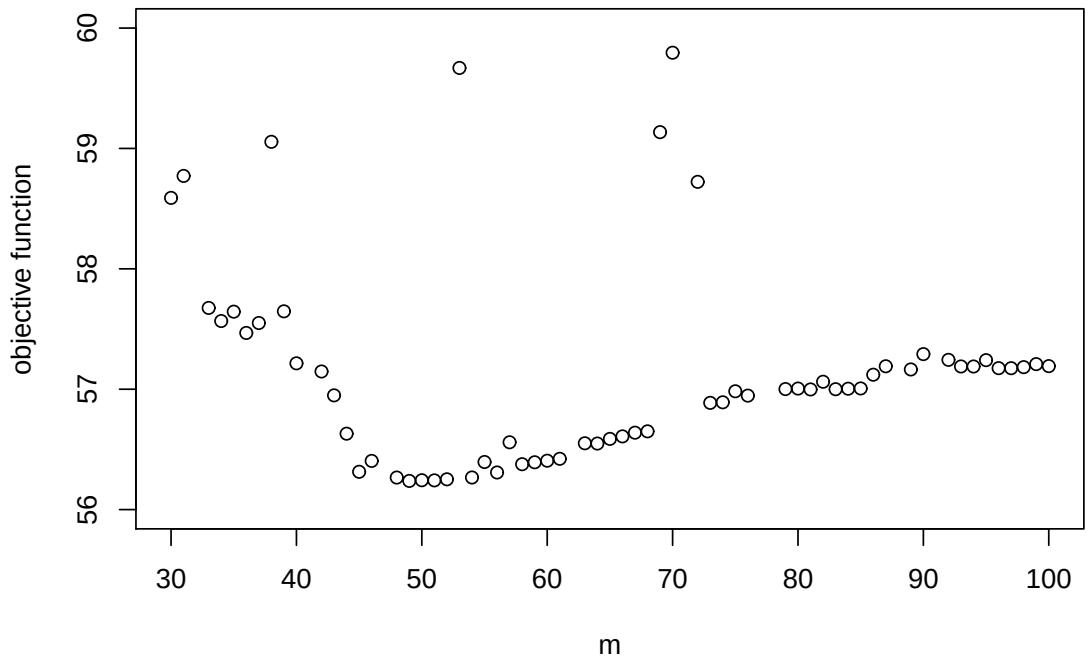

Figure 8: Plot of $g(\widehat{\boldsymbol{\beta}}_H)$ for the objective function g defined in (16) for elemental estimates $\widehat{\boldsymbol{\beta}}_H$ minimizing g with $M(\widehat{\boldsymbol{\beta}}_H) = m$ for $m = 30, \cdots, 100$; the data are 100 observations are generated from the model (19) using $\beta_0 = 0$ and $\beta_1 = 1$ so that $M(\boldsymbol{\beta}) = 100$.

model with 100 observations shows that $M(\widehat{\boldsymbol{\beta}}) = 100$ with (approximate) probability 0.46 and $M(\widehat{\boldsymbol{\beta}}) \leq 80$ with probability 0.24.) Of course, this observation is biased by our knowledge of the model generating the data; however, the value of the objective function (16) for the Powell estimates $\widehat{\boldsymbol{\beta}} = (-0.983, 2.089)^T$ is 56.24 compared to 57.19 for the elemental estimates $\widehat{\boldsymbol{\beta}}_H = (-0.0046, 0.932)^T$, a small difference given the variability of the noise in the model. Figure 8 shows a plot of the (minimized) objective function (16) for the best elemental estimates satisfying $M(\widehat{\boldsymbol{\beta}}_H) = m$ where $m = 30, \cdots, 100$.

The plot shown in Figure 8 can be used as a diagnostic for assessing possible bias in the Powell estimator. Figure 9 shows the same plot for 100 observations generated from the model (19) using $\beta_0 = -1$ and $\beta_1 = 2$; for these data, $M(\boldsymbol{\beta}) = 44$ (on average $M(\boldsymbol{\beta}) \approx 50$ since $2x - 1 > 0$ for $x > 1/2$) and $M(\widehat{\boldsymbol{\beta}}) = 49$. (Simulations for this model show that the $M(\widehat{\boldsymbol{\beta}}) < 50$ with (approximate) probability 0.55 and $M(\widehat{\boldsymbol{\beta}}) > 50$ with probability 0.43, reflecting the same downward bias as before albeit to a lesser extent.) In contrast to Figure 8 (where $g(\widehat{\boldsymbol{\beta}}_H)$ is relatively flat for m between 49 and 100), $g(\widehat{\boldsymbol{\beta}}_H)$ is relatively flat for m between 30 and 60 and increases substantially for $m > 60$.

In the examples presented here, it is simple to evaluate all 4950 elemental estimates.

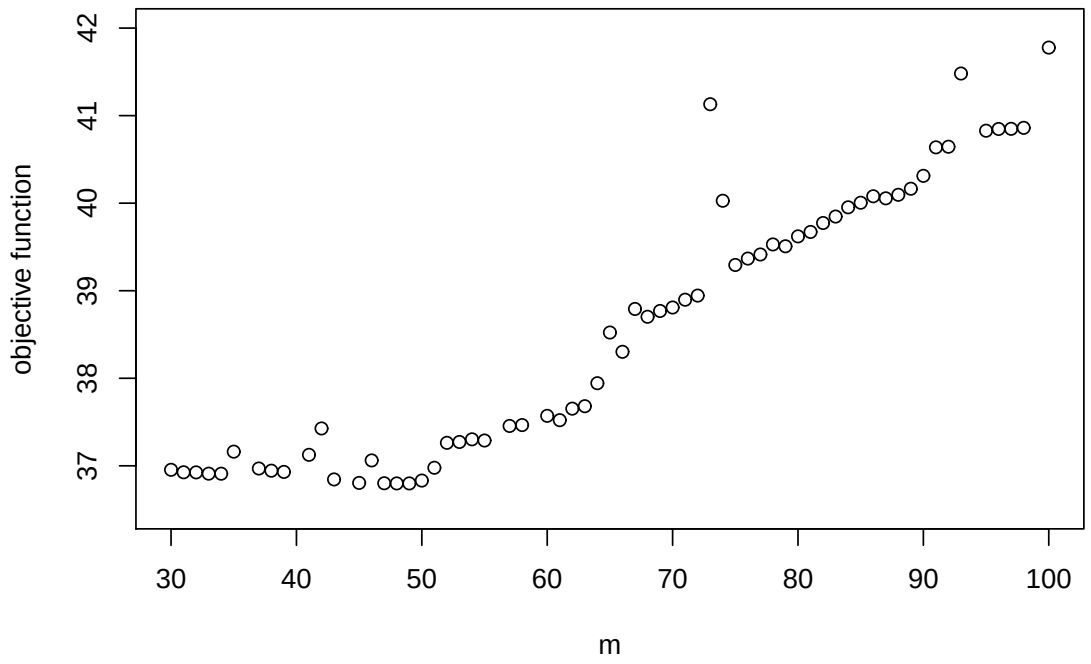

Figure 9: Plot of $g(\widehat{\boldsymbol{\beta}}_H)$ for the objective function g defined in (16) for elemental estimates $\widehat{\boldsymbol{\beta}}_H$ minimizing g with $M(\widehat{\boldsymbol{\beta}}_H) = m$ for $m = 30, \cdots, 100$; the data are 100 observations are generated from the model (19) using $\beta_0 = -1$ and $\beta_1 = 2$ with $M(\boldsymbol{\beta}) = 44$.

However, evaluating all elemental estimates for larger n and p is not necessarily feasible and in practice, it will typically be necessary to randomly sample subsets in order to produce the plots described above. While the random sampling will introduce some additional uncertainty, it should not reduce the utility of these plots.

References

Brown, B.M. and Eagleson, G.K. (1971) Martingale convergence to infinitely divisible laws with finite variance. *Transactions of the American Mathematical Society.* **162**, 449-453.

Brown, T.C. (1978) A martingale approach to the Poisson convergence of point processes. *Annals of Probability.* **6**, 615-628.

Chernozhukov, V. and Hong, H. (2002) Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association.* **97**, 872-882.

Dang, X., Peng, H., Wang, X. and Zhang, H. (2009) Theil-Sen estimators in a multiple linear regression model. (unpublished manuscript)

D'Esposito, M.R. and Furno, M. (1992a) Robust estimation in multiple regression model

using p-dimensional subsets. *Metron.* **50**, 115-136.

D'Esposito, M.R. and Furno, M. (1992b) Location of outliers in multiple regression using resampled values. *Computer Science in Economics and Management.* **5**, 171-182.

Fitzenberger, B. and Winker, P. (2007) Improving the computation of censored quantile regressions. *Computational Statistics and Data Analysis.* **52**, 88-108.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **6**, 721-741.

Hadi, A.S. and Simonoff, J.S. (1993) Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association.* **88**, 1264-1272.

Hawkins, D.M. (1993) The accuracy of elemental set approximations for regression. *Journal of the American Statistical Association.* **88**, 580-589.

Hawkins, D.M., Bradu, D. and Kass, G.V. (1984) Location of several outliers in multiple-regression using elemental sets. *Technometrics.* **26**, 197-208.

Hawkins, D.M. and Olive, D.J. (2002) Inconsistency of resampling algorithms for high-breakdown regression estimators and a new algorithm (with discussion). *Journal of the American Statistical Association.* **97**, 136-159.

Hosseinkouchack, M. (2011) Further improvements in the calculation of censored quantile regressions. *Journal of Computational and Applied Mathematics.* **235**, 1429-1445.

Jaeckel, L.A. (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics.* **43**, 1449-1458.

Jurečkovä, J. Nonparametric estimates of regression coefficients. *Annals of Mathematical Statistics.* **42**, 1328-1338.

Kallenberg, O. (1976) *Random Measures.* Berlin: Akademie Verlag.

Koenker, R. (2008) Censored quantile regression redux. *Journal of Statistical Software* **27**, 1-24.

Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica.* **46**, 33-50.

Koenker, R. and Portnoy, S. (1987) L-estimation for linear models. *Journal of the American Statistical Association.* **82**, 851-857.

Maire, C. and Boscovich, R.J. (1755) *De Litteraria Expeditione per Pontificiam Ditionem ad Dimetiendos Duos Meridiani Gradus et Corrigendum Mappam Geographicam.* Rome: Palladis.

Mayo, M.S. and Gray, J.B. (1997) Elemental subsets: the building blocks of regression. *American Statistician.* **51**, 122-129.

Oja, H. and Niinimaa, A. (1984) On robust estimation of regression coefficients. (unpublished manuscript)

Olive, D.J. and Hawkins, D.M. (2007) Behavior of elemental sets in regression. *Statistics and Probability Letters.* **77**, 621-624.

Powell, J.L. (1984) Least absolute deviations estimation for the censored regression model. *Journal of Econometrics.* **25**, 303-325.

Powell, J.L. (1986) Censored regression quantiles. *Journal of Econometrics.* **32**, 143-155.

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

Rousseeuw, P.J. and Bassett, G.W. (1991) Robustness of the p-subset algorithm for regression with high breakdown point. In *Directions in Robust Statistics and Diagnostics, Part II*, editors W. Stahel and S. Weisberg. 185-194.

Rouuseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection.* New York: Wiley.

Sen, P.K. (1968) Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association.* **63**, 1379-1389.

Silverman, B.W. and Brown, T.C. (1978) Short distances, flat triangles and Poisson limits. *Journal of Applied Probability.* **15**, 815-825.

Stigler, S. (1986) *History of Statistics.* Cambridge: Harvard University Press.

Stromberg, A.J. (1993) Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. *SIAM Journal of Scientific Computing.* **14**, 1289-1299.

Tang, Y., Wang, H.J., He, X., and Zhu, Z. (2012) An informative subset-based estimator for censored quantile regression. *Test.* **21**, 635-655.

Theil, H. (1950) A rank-invariant method of linear and polynomial regression analysis, I-III. *Koninklijke Nederlandse Akademie van Wetenschappen, Series A.* **53**, 386-402, 521-525, 1397-1412.