QÜESTIIÓ, vol. 21, 1 i 2, p. 9-36, 1997

ESTIMATING THE CONTAMINATION LEVEL OF DATA IN THE FRAMEWORK OF LINEAR REGRESSION ANALYSIS*

A. RUBIO*
Universidad de Extremadura
J.Á. VÍŠEK**
Academy of Sciences of the Czech Republic

An estimator of the contamination level of data is proposed in the framework of linear models and its asymptotic behavior is investigated. A numerical study illustrates its finite sample performance under an alternative.

Keywords: Huber's convex mixture model, contamination level estimation, linear model, asymptotic behavior, simulation study.

AMS Classification: Primary 62F35, secondary 62J05

* A. Rubio, Departamento de Matemáticas. Universidad de Extremadura. 100 71 Cáceres.

** J.Á. Víšek, Department of Macroeconomics and Finance. Institute of Economic Studies. Faculty of Social Sciences. Charles University. Staré město. Smetanovo nábřeží 6, 110 01 Prague 1 & Department of Stochastic Informatics, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic. Pod vodárenskou věži 4. 182 08 Prague 8.

⋆ This work was supported by Dirección General de Investigación Científica y Técnica (DGICYT), Ministery of Education and Sciences of Spanish Goverment

– Article rebut el maig de 1992.

– Acceptat l'octubre de 1996.

1. INTRODUCTION

It is well known that since the days of Sir Francis Galton (1886) the regression analysis has been used as an efficient tool of modelling the data. In the past it employed mostly the least squares (Legendre (1805), Gauss (1809)) although the method of the least absolute deviations has been proposed much earlier (Galilei (1632), Boscovisch (1757), Laplace (1793)). During the past quarter of this century a lot of other methods have appeared. Some of them produce in fact not a single estimator but the whole family of model estimators, in which every estimator corresponds to some value of the (tuning) parameter(s). Everything will be clear from the following example considering the family of the M-estimators of regression model corresponding to the family of Huber's ψ-functions, $\{\psi_k(z)\}_{k\in(0,\infty)}$ where

$$\psi_k(z) = \begin{cases} z & \text{for} \quad |z| \le k, \\ \\ k \cdot \operatorname{sign}(z) & \text{otherwise.} \end{cases} \tag{1}$$

The family of estimators $\{\hat{\beta}^{(n,k)}\}_{k\in(0,\infty)}$ is then given by

$$\hat{\beta}^{(n,k)} = \min_{\beta\in R^p} \left\{ \sum_{i=1}^{n} \rho_k(Y_i - X_i^T\beta) \right\}$$

where $\rho_k(z)$ is a criterial function with the derivative equal to $\psi_k(z)$ (for Y's and X's see (4) below). The value of the tuning constant k is in applications selected on the basis of experiences. We shall recall later (after introducing necessary notations) Huber's result which connects the optimal value of the tuning constant k with the mixture parameter ε in the Huber model of contamination. We shall also see that the «minimal» mixture parameter ε corresponds (in one-to-one way) to the contamination level $\varepsilon_{G,F}$ of data. It means that the optimal value k of the tuning constant depends on the contamination level $\varepsilon_{G,F}$. Now, if we select the value of k so that it «underestimates the contamination level» we may obtain wrong model of data. On the other hand «overestimating contamination level» leads to a needless loss of efficiency. Of course, a loss of efficiency is (typically) small and hence it is not important, see Víšek (1993).

There is however another problem. It is known that some robust methods focus themselves on a (too) restricted part of data considerably weighting down (or depressing completely) the influence of other part. Such behaviour may be observed especially for the methods with large breakdown point, i. e. for the methods assuming (extremely) high contamination level. So using methods designed for the high contamination level —as the least median of squares, the least trimmed squares or minimal biased estimators— we may obtain somewhat (or completely) misleading

estimate of model. Let us give (at least) one numerical example. First of all, let us recall definitions of the estimators which will be used. Let us put

$$r_i(\beta) = Y_i - \sum_{j=1}^{p} X_{ij}\beta_j, \quad i = 1,2,\dots,n \qquad\qquad h = \left[\frac{n}{2}\right] + \left[\frac{p+1}{2}\right] \tag{2}$$

and let $r^2_{(i:n)}(\beta)$ be the i-th order statistics among $r_i^2(\beta)$, $i = 1,2,\dots,n$. Further, let us recall that (with h given by (2))

$$\begin{aligned} \hat\beta^{\mathrm{LS}} &= \operatorname*{argmin}_{\beta\in R^p} \sum_{i=1}^{n} r_i^2(\beta), & \hat\beta^{\mathrm{LMS}} &= \operatorname*{argmin}_{\beta\in R^p} r^2_{(h:n)}(\beta), \\ \hat\beta^{\mathrm{LTS}} &= \operatorname*{argmin}_{\beta\in R^p} \sum_{i=1}^{h} r^2_{(i:n)}(\beta), & \hat\beta^{(\rho_k)} &= \operatorname*{argmin}_{\beta\in R^p} \sum_{i=1}^{n} \rho_k(r_i(\beta)) \end{aligned} \tag{3}$$

with ρ_1 – Huber's function (with $\psi_1(t) = t$ for $|t| < c$, $\psi_1(t) = c\cdot\operatorname{sign} t$ otherwise) and ρ_2 – Hampel's function (with $\psi_2(t) = \psi_1(t)$ for $|t| < 1.2c$, $\psi_2(t) = [c - \frac{5}{9}(t - 1.2c)]\cdot$ sign t for $1.2c < |t| < 3c$ and zero otherwise; both with the tuning constant $c = 1.2$). Finally, let

$$\hat\beta^{\mathrm{L_1}} = \operatorname*{argmin}_{\beta\in R^p} \sum_{i=1}^{n} |r_i(\beta)| \qquad \text{and} \qquad \hat\beta^{\mathrm{TLS}} = \operatorname*{argmin}_{\beta\in R^p} \sum_{i\in I_\alpha} r_i^2(\beta)$$

where I_α is the index-set of points obtained by the symmetric trimming according to α-regression quantiles of Koenker and Bassett (1978) (value of α was 0.1).

Example. Demographic Data (49 cases, Chatterjee and Hadi (1988)). Dependence of the gross national product per capita on: the infant death per thousand live births, the number of inhabitants per physician, the population per km^2, the population per 10^3 ha of agricultural land, the percentage of literate population over 15 years of age, the number of students enrolled in higher education per 10^5 population. (The software which was used for evaluation was either prepared by the experienced statisticians as Jaromír Antoch (1991, 1992), Roger Koenker (1978) or Alvio Marazzi (1992), and we are grateful of possibility to utilize it, or by our colleagues and it was tested in an extensive numerical studies, see e. g. Víšek (1996 a,1997).)

(Value of α for *TLS* was 0.1.) It is not even difficult to find artificial (one-dimensional) data for which *LMS* and *LTS* estimates are orthogonal each to other and *S*-estimate divides the angel between them on two halves, see Víšek (1994 a). (An objection may appear that perhaps in the example with Demographic data some regressors are insignificant. However, in the case of contaminated data, it is not simple task to say which regressors are significant and which not —for more arguments see

Víšek (1996 c).) Similarly, rather diverse results may be obtained when using different constants k when evaluating M-estimates.

Table 1. *Demographic Data*

Method	LS	LMS	LTS	TLS	L_1	Huber	Hampel
intercept	112.885	331.095	103.563	480.509	148.732	33.459	−146.586
Infant	−3.621	−2.774	−1.526	−6.764	−3.964	−3.025	−2.029
Inhabitants	0.009	−0.017	0.005	−0.013	0.032	0.015	0.032
Population	0.186	−1.024	0.009	−1.265	0.088	−0.052	0.242
Agriculture	0.003	0.072	−0.001	0.085	−0.005	0.000	−0.008
Literate	5.566	2.501	3.929	3.793	2.985	5.280	4.339
Higher	0.693	0.249	0.295	0.373	0.860	0.734	1.072

Table 2. *Demographic Data - Huber's estimator*

Tuning constant	0.6	0.8	1.0	1.2	1.5	2.0
intercept	110.136	69.369	42.526	33.459	67.058	113.059
Infant	−3.254	−3.191	−3.144	−3.025	−3.226	−3.545
Inhabitants	0.013	0.016	0.017	0.015	0.013	0.009
Population	0.018	0.013	−0.015	−0.052	−0.106	−0.165
Agriculture	−0.002	−0.001	−0.001	0.000	0.001	0.003
Literate	4.002	4.422	4.917	5.280	5.308	5.319
Higher	0.777	0.782	0.753	0.734	0.722	0.705

For other examples of such situations see Rubio *et al.* (1992) or Víšek (1994 a), (1994 b)).

So, one possible way how to begin any processing of data may be to estimate the contamination level of them and then to apply that method (from the family in question) which corresponds to the estimated contamination level. Of course, the problem with the selection of an adequate robust method is a complicated problem and we shall return to it briefly at the end of paper.

As we shall see later we will need for the estimation of the contamination level to estimate the density of data. It is simple to do it directly when we have at hand

a sample of i.i.d. r.v.'s, e.g. in the location problem, but naturally it cannot be applied directly on the response variable in the regression scheme. Nevertheless as we shall see below some preliminary considerations will open a straightforward way to a proposal of an estimator of contamination level in regression model, too.

One may learn from the text given below that the estimation of the contamination level is analogous to the estimation of the mixture parameter (see Rukhin(1994)). The difference is that in the case of the estimation of the contamination level we do not specify the «contaminating» distribution but only the «central» model. Of course, it implies that when we estimate the contamination level we may meet some additional (technical) difficulties in comparison with the estimation of the mixture parameter. So, taking into account that when looking for an estimator of the mixture parameter we are rarely able to prove more than the consistency, we cannot expect that for the estimator of the contamination level we will be able to give more than an asymptotic results about its behavior.

The results are accompanied by a simulation study. The reasons for the simulation study will be also discussed.

2. ESTIMATING CONTAMINATION LEVEL

Let N denote the set of all positive integers, R the real line and R^p the p−dimensional Euclidian space. We shall consider the linear model

$$Y = X \cdot \beta^0 + e \tag{4}$$

where for every $n \in N$ and some (fix) $p \in N$ we have $Y = (Y_1, Y_2, \dots, Y_n)^T$ (a response variable), $X = (x_{ij})_{i=1,2,\dots,n}^{j=1,2,\dots,p}$ (a design matrix), $\beta^0 = (\beta_1^0, \beta_2^0, \dots, \beta_p^0)^T$ (regression coefficients) and $e = (e_1, e_2, \dots, e_n)^T$ (random disturbances). We assume that the random variables in the sequence $\{e_i\}_{i=1}^{\infty}$ are i.i.d. and they are defined on a probability space $(\Omega, \mathcal{B}, P)$, and the carriers x_{ij}'s are fix and known while β^0 (the «true» value of the vector of regression coefficients) is unknown but also fix. (Let us remark here that in what follows all probabilistic assertions as «a. s.», «in probability» etc. will be understood with respect to P.) Assume moreover that there is a $K < \infty$ such that

$$\sup_{i \in N} \max_{j=1,2,\dots,p} |x_{ij}| < K \tag{5}$$

and

$$\lim_{n \to \infty} \frac{1}{n} X^T X = Q \tag{6}$$

where Q is a regular matrix. Sometimes we will use also an alternative notation $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, so that the expression $X_i^T\beta$ will stay instead of the sum $\sum_{j=1}^{p} x_{ij}\beta_j$. Finally, let ν be a σ-finite measure defined on $(R, \mathcal{B}(R))$ (where $\mathcal{B}(R)$ is the Borel σ-algebra) such that distribution function F of e_1 is absolutely continuous with respect to ν, and then let us denote by $\mathcal{F}_\nu$ the set of all distribution functions which are absolutely continuous with respect to ν.

In Huber (1964) the convex combination model

$$G(z) = (1-\varepsilon)F(z) + \varepsilon H(z) \tag{7}$$

$(\varepsilon \in [0,1], F, H \in \mathcal{F}_\nu)$ was used to describe the contamination of data. So, the first idea how to define the contamination level may start with this model. One finds immediately that for given G the decomposition (7) is given uniquely only in the case when

$$\int_{\{z:g(z)=0\}} \mathrm{d}F(z) > 0$$

where g is density of the distribution function G (with respect to ν). Then $\varepsilon = 1$ and $H(z) = G(z)$. If however (7) holds for some $\varepsilon \in [0,1)$ then also for any $\varepsilon^* \in [\varepsilon, 1]$ we may write

$$G(z) = (1-\varepsilon^*)F(z) + \varepsilon^* H^*(z)$$

where $H^*(z) = (\varepsilon^*)^{-1}\{(\varepsilon^* - \varepsilon)F(z) + \varepsilon H(z)\}$ and we evidently have $H^* \in \mathcal{F}_\nu$. It hints that the following definition of the contamination level, for the version using the densities, has to consider an essential minimum of possible values of ε's, essential with respect to the measure ν (notice that for our case the result is same if we take «ess inf» with respect to F because f is involved in the characterization of $\varepsilon_{G,F}$).

Definition 1 *For G and $F \in \mathcal{F}_\nu$ we shall define the contamination level as a value*

$$\varepsilon_{G,F} = \inf\{\varepsilon : G(z) = (1-\varepsilon)F(z) + \varepsilon H(z), H \in \mathcal{F}_\nu\}$$

or equivalently

$$\varepsilon_{G,F} = \inf\{\varepsilon : \nu\{z \ : \ g(z) \neq (1-\varepsilon)\, f(z) + \varepsilon h(z)\} = 0, H \in \mathcal{F}_\nu\}$$

where g, f and h are again the densities of G, F and of H with respect to ν, respectively.

Remark 1 The mixture

$$G(z) = (1-\varepsilon)F(z) + \varepsilon H(z)$$

expresses the fact that the bulk of data, the «proper data», are distributed according to F and some portion of data (this portion being considered as contamination) is distributed accordng to H. The distribution H is not usually fully specified while the

distribution function F is selected (at least as a type of distribution) by the statistician when processing data. So in other words, having at hand data, we assume that they were generated by some unknown G, nevertheless, we would like to describe them (or explain them) by F. The reason for selecting the distribution F instead of the «true» distribution G may be, e.g., the fact that the distribution function G is expected to be too «wild» while the distribution function F is easy (or at least easier) to work with, and at the same time being a reasonable approximation of the distribution G. However we are aware that it need not be precisely the distribution which has generated the data, i.e. that $G \neq F$, and hence we admit that there is a distribution H and $\varepsilon > 0$ so that $G = (1-\varepsilon)F + \varepsilon H$. It means that we select F and hope that the «true» distribution of data lies in a neighborhood of F.

□

Remark 2 In the last twenty years a lot of others contamination models have appeared being based on different types of distances in the space of probability measures (e.g. on the Kolmogorov-Smirnov or Prokhorov metrics (see Huber (1981)), on the combination of convex combinations and total variation (see Rieder (1977)), on 2-alternating capacities (see Huber, Strassen (1973)), or on divergences (see Vajda (1989)). We believe that a straightforward generalization of the notion of contamination level is possible in any of these contamination models. We also believe that for such generalization the modifications of the theory which will be presented below, will be also straightforward (something has been already done for the location problem, see Víšek (1989)).

□

Lemma 1 *(characterization of the contamination level $\varepsilon_{G,F}$) Let $G(z)$ and $F(z) \in \mathcal{F}_\nu$. Then*

$$\varepsilon_{G,F} = \left\| \frac{f(z) - g(z)}{f(z)} I_{\{z: f(z) > g(z)\}} \right\|_\infty$$

$f(z)$ and $g(z)$ being again the densities of the distibutions $F(z)$ and $G(z)$ with respect to ν.

Proof: At first notice that

$$\left\| \frac{f(z) - g(z)}{f(z)} I_{\{z: f(z) > g(z)\}} \right\|_\infty \geq 0$$

and the lower bound is attained only when $\nu\{z : f(z) \neq g(z)\} = 0$. Let us denote by $\mathcal{E}_{G,F}$ the set of ε's for which (7) holds, and let $\varepsilon \in \mathcal{E}_{G,F}$. It implies that for some density $h(z)$ we have $g(z) = (1-\varepsilon) f(z) + \varepsilon h(z)$, and hence $f(z) - g(z) \leq \varepsilon f(z)$, i.e. $\varepsilon \geq$

$\frac{f(z)-g(z)}{f(z)}$ for any $z \in R$ for which $f(z) > 0$. Then of course $\varepsilon \geq \| \frac{f(z)-g(z)}{f(z)} I_{\{z: f(z)>g(z)\}} \|_\infty$, and hence also $\varepsilon_{G,F} \geq \| \frac{f(z)-g(z)}{f(z)} I_{\{z: f(z)>g(z)\}} \|_\infty$.

On the other hand for any $\varepsilon \geq \| \frac{f(z)-g(z)}{f(z)} I_{\{z: f(z)>g(z)\}} \|_\infty$, $h_\varepsilon(z) = \frac{1}{\varepsilon}[g(z) - (1-\varepsilon) f(z)]$ is a density, and

$$g(z) = (1-\varepsilon) f(z) + \varepsilon h_\varepsilon(z).$$

It means that $\varepsilon \in \mathcal{E}_{G,F}$ and $\varepsilon_{G,F} \leq \| \frac{f(z)-g(z)}{f(z)} I_{\{z: f(z)>g(z)\}} \|_\infty$.

■

Remark 3 It follows from the proof of Lemma 1 that the infimum in Definition 1 is attained so that we may write

$$G(z) = (1-\varepsilon_{G,F}) F(z) + \varepsilon_{G,F} H_{\varepsilon_{G,F}}(z)$$

for some $H_{\varepsilon_{G,F}}(z)$. It may be of interest that for a convex mixture of two normal distribution, say

$$(1-\varepsilon_H)(2\pi)^{-\frac{1}{2}} \exp(-\frac{z^2}{2}) + \varepsilon_H (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp(-\frac{z^2}{2\sigma^2}), \tag{8}$$

we obtain

$$\varepsilon_{G,F} = \varepsilon_H \qquad \text{for} \quad \sigma < 1 \tag{9}$$

and

$$\varepsilon_{G,F} = (1-\sigma^{-1})\varepsilon_H \qquad \text{for} \quad \sigma > 1. \tag{10}$$

Let us recall that Huber's result (proved in 1964, under condition that the logarithm of the density of central model is concave) says that the optimal selection of the ψ_k-function (see (1)) in the case that the data were generated by the mixture density (7) is given by $k = k(\varepsilon_H)$ where $k(\varepsilon_H)$ satisfies

$$(1-\varepsilon_H)^{-1} = \int_{t_0(\varepsilon_H)}^{t_1(\varepsilon_H)} g(t)\mathrm{d}t + \frac{g(t_0(\varepsilon_H)) + g(t_1(\varepsilon_H))}{k(\varepsilon_H)} \tag{11}$$

where $t_i(\varepsilon_H)$'s are given by

$$\frac{g'(t_i(\varepsilon_H))}{g(t_i(\varepsilon_H))} = (-1)^i \cdot k(\varepsilon_H), \quad i = 0, 1. \tag{12}$$

So an alternative notation for Huber's family of ψ-functions may be

$$\{\psi_{k(\varepsilon)}\}_{\varepsilon \in (0,1)}. \tag{13}$$

We shall need it in the discussion later.

□

Now we need to estimate $\|\frac{f(z)-g(z)}{f(z)} I_{\{z: f(z)>g(z)\}}\|_\infty$. Since the density $f(z)$ is known (see Remark 1) it is necessary to estimate the density $g(z)$. To be more precise, let $\{Z_k(\omega)\}_{k=1}^\infty$ be a sequence of i.i.d. random variables (defined on the probability space $(\Omega, \mathcal{B}, P)$). Assume that the corresponding distribution function $G_Z(z)$ belongs to $\mathcal{F}_\nu$ and denote by $g_Z(z)$ its density. We shall assume that $g_Z(z)$ vanishes outside an interval (c,d), $-\infty \le c < d \le \infty$. Moreover, following Csörgö and Révész (1981) and Rosenblatt (1971), let us assume:

Conditions A *The bounded integrable kernel $w : R \longrightarrow R$ vanishes outside an interval $(-a,a)$ with $-\infty \le -a \le c < d \le a \le \infty$ for some $a > 0$. Moreover, it is twice continuously differentiable with bounded first derivative $w'(z)$, say $\sup_{z\in R}|w'(z)| < L < \infty$ and with $(1+z^4)|w''(z)|$ also bounded. Finally, the kernel has a Fourier transform $\phi(t)$ with $(1+t^2)\cdot\phi(t)$ integrable and is symmetric, i. e. $w(z) = w(-z)$.*

Please notice that Conditions A imply that

$$\int_{-\infty}^{\infty} w'(z)\mathrm{d}z = 0 \qquad \text{(due to symmetry)} \tag{14}$$

and

$$\left|\int_{-\infty}^{\infty} z\cdot w'(z)\mathrm{d}z\right| < \infty. \tag{15}$$

Let us write $Z^T(\omega,n) = (Z_1(\omega), Z_2(\omega), \ldots, Z_n(\omega))$. Then define for any $z \in R$, $\omega \in \Omega$ the kernel estimator of density by

$$\hat{g}_n(z, Z(\omega,n)) = \frac{1}{nc_n}\sum_{i=1}^{n} w(c_n^{-1}(z - Z_i(\omega))) \tag{16}$$

and in what follows we shall assume that the sequence of the bandwidths $\{c_n\}_{n=1}^\infty \searrow 0$.

For the simplicity we shall assume that in the rest of paper ν is the Lebesgue measure.

Lemma 2 *(Csörgö, Révész). Let us denote*

$$\pi_n(g) = \sup_{z\in R}|\mathsf{E}\hat{g}_n(z, Z(\omega,n)) - g_Z(z)|.$$

Then

$$\sup_{z\in(-a,a)}|n(\hat{g}_n(z, Z(\omega,n)) - g_Z(z)) - \Gamma_n(z)| = O(c_n^{-1}\log^2 n + n\pi_n(g)) \qquad \text{a. s.}$$

where $\Gamma_n(z)$ is the following Gaussian process:

$$\Gamma_n(z) = \frac{1}{c_n}\int_0^1 \mathcal{K}(v,n)\mathrm{d}w(c_n^{-1}(z-G^{-1}(v))) \tag{17}$$

with $\mathcal{K}(v,n)$ being a suitable Kiefer process.

The proof is a specification of the proof of the Theorem 6.1.1 of Csörgö, Révész (1981), page 223.

Lemma 3 *Let $G(z)$ and $F(z)$ belong to $\mathcal{F}_\nu$. Further, let $\{S_n\}_{n=1}^{\infty}$ be a family of sets, $S_n \in \mathcal{B}(R)$, such that*

$$S_n \nearrow \{z : f(z) > 0\} \tag{18}$$

and

$$[\inf_{z\in S_n} f(z)]^{-1}\max\left\{\pi_n(g), n^{-1}c_n^{-1}\log^2 n, n^{-\frac{1}{2}}c_n^{-1}(\log\ \log n)^{\frac{1}{2}}\right\} = o(1). \tag{19}$$

Then

$$\sup_{z\in S_n}\frac{|\hat{g}_n(z,Z(\omega,n)) - g(z)|}{f(z)} = o(1) \qquad \text{a. s.} \qquad \text{as } n \longrightarrow \infty.$$

Proof: Let us denote by $M_n = [\inf_{z\in S_n} f(z)]^{-1}$. Now from the Lemma 2 we have for some $\Delta < \infty$

$$-\Delta M_n(\pi_n(g) + c_n^{-1}n^{-1}\log^2 n) \le \frac{\hat{g}_n(z,Z(\omega,n)) - g(z)}{f(z)} - \frac{\Gamma_n(z)}{nf(z)}$$

$$\le \Delta M_n(\pi_n(g) + c_n^{-1}n^{-1}\log^2 n).$$

Since the lower as well as the upper bound tends to zero it suffice to show that

$$\sup_{z\in S_n}\frac{|\Gamma_n(z)|}{nf(z)} = o(1). \tag{20}$$

However, due to (17) we have

$$c_n|\Gamma_n(z)| \le \sup_{z\in R}|\mathcal{K}(z,n)|\int_0^1 \mathrm{d}w(c_n^{-1}(z-G^{-1}(v))) \le \Delta\cdot\sup_{z\in R}|\mathcal{K}(z,n)|$$

where $\mathcal{K}(z,n)$ is again an appropriate Kiefer process. Now, due to (19) for any $\varepsilon > 0$ we may find $n_\varepsilon \in N$ so that for any $n > n_\varepsilon$

$$\sup_{z\in S_n}\frac{(\log\ \log n)^{\frac{1}{2}}}{c_n n^{\frac{1}{2}} f(z)} \le \varepsilon,$$

and we have

$$\sup_{z\in S_n}\frac{|\Gamma_n(z)|}{nf(z)}\le\frac{\Delta}{c_n n}\sup_{z\in R}\frac{|\mathcal{K}(z,n)|}{f(z)}\le n^{-\frac{1}{2}}\varepsilon\Delta\sup_{z\in R}|\mathcal{K}(z,n)|(\log\log n)^{-\frac{1}{2}}.$$

The proof of (20) then follows due to the law of iterated logarithm for the Kiefer process (see Corollary 1.15.1 of Csörgö, Révész (1981)).

■

As we have mentioned above we need to estimate $\|\frac{f(z)-g(z)}{f(z)}I_{\{x:f(z)>g(z)\}}\|_\infty$ and let us recall that

$$\|\frac{f(z)-g(z)}{f(z)}I_{\{z:f(z)>g(z)\}}\|_\infty=\inf\left\{a\ :\ \nu(\left\{\frac{z\ :}{f(z)@>a}\right\})=0\right\}.$$

A routine arguments yields that there are $\tilde{f}(z)$ an $\tilde{g}(z)$ such that $\nu(\{f(z)\neq\tilde{f}(z)\})=0$ and $\nu(\{g(z)\neq\tilde{g}(z)\})=0$, and

$$\|\frac{f(z)-g(z)}{f(z)}I_{\{z:f(z)>g(z)\}}\|_\infty=\sup_{z\in\left\{t:\tilde{f}(t)>0\right\}}\frac{\tilde{f}(z)-\tilde{g}(z)}{\tilde{f}(z)}$$

and so we may assume that $f(z)$ and $g(z)$ are such versions of densities of $F(z)$ and of $G(z)$ that

$$\|\frac{f(z)-g(z)}{f(z)}I_{\{z:f(z)>g(z)\}}\|_\infty=\sup_{z\in\{t:f(t)>0\}}\frac{f(z)-g(z)}{f(z)}.$$

Let us assue that we shall work in the rest of paper with such versions of densities. Then from the previous lemma it follows that we may estimate $\|\frac{f(z)-g(z)}{f(z)}I_{\{z:f(z)>g(z)\}}\|_\infty$ by

$$\sup_{z\in S_n}\frac{f(z)-\hat{g}_n(z,Z(\omega,n))}{f(z)} \tag{21}$$

due to the fact that

$$\sup_{z\in\{t:f(t)>0\}}\frac{f(z)-g(z)}{f(z)}-\sup_{z\in S_n}\frac{f(z)-\hat{g}_n(z,Z(\omega,n))}{f(z)}$$

$$\le\sup_{z\in S_n}\frac{\hat{g}_n(z,Z(\omega,n))-g(z)}{f(z)}=o(1)\qquad\text{a. s.}\qquad\text{as } n\longrightarrow\infty \tag{22}$$

(realize that we have assumed $S_n\nearrow\{z:\ f(z)>0\}$) and similarly

$$\sup_{z\in S_n}\frac{f(z)-\hat{g}_n(z,Z(\omega,n))}{f(z)}-\sup_{z\in\{t:f(t)>0\}}\frac{f(z)-g(z)}{f(z)}$$

(23) $$\leq \sup_{z\in S_n} \frac{g(z)-\hat{g}_n(z,Z(\omega,n))}{f(z)} = o(1) \qquad \text{a. s.} \qquad \text{as } n \longrightarrow \infty.$$

But let us consider the case when $g(z) = f(z)$, i.e. when there is no contamination. Then, due to the random fluctuations of $\hat{g}_n(z,Z(\omega,n))$, we obtain positive value at (21) for every $\omega \in \Omega$. That is the reason why we shall not propose in the following theorem the estimator of $\hat{\varepsilon}_{G,F}$ as $\sup_{z\in S_n} \frac{f(z)-\hat{g}_n(z,Z(\omega,n))}{f(z)}$ but in a little modified form.

Theorem 1 *Let G and F belong to $\mathcal{F}_v$. Further let $\left\{\hat{\beta}_n\right\}_{n=1}^{\infty}$ be a sequence of $\sqrt{n}$-consistent estimators of β^0 and $\{V_n\}_{n=1}^{\infty}$ sequence of sets, $V_n \in \mathcal{B}(R)$, fulfilling the assumptions (18) and (19), and moreover let*

(24) $$[\inf_{z\in V_n} f(z)]^{-1} c_n^{-1} \|\hat{\beta}_n - \beta^0\| = o_p(1).$$

Finally, let $\{f_n(z,\omega)\}_{n=1}^{\infty}$ be a sequence of random processes such that

(25) $$\sup_{z\in V_n} |f_n(z,\omega) - f(z)| = o_p(\inf_{z\in V_n} f(z))$$

and put

$$\hat{\varepsilon}_{G,F} = \max\left\{0, \min\left\{\sup_{z\in V_n} \frac{f_n(z,\omega) - \hat{g}_n(z, r(\hat{\beta}_n)))}{f(z)}, 1\right\}\right\}$$

where we have denoted for any $\beta \in R^p$ by $r(\beta)$ the vector of residuals $(Y_1 - X_1^T\beta, Y_2 - X_2^T\beta, \ldots, Y_n - X_n^T\beta)^T$. Then

(26) $$\hat{\varepsilon}_{G,F} - \varepsilon_{G,F} = o_p(1).$$

Proof: If we had known the true value β^0 of the regression coefficients we may evaluate the «theoretical» residuals $Y_1 - X_1^T\beta^0, Y_2 - X_2^T\beta^0, \ldots, Y_n - X_n^T\beta^0$ and to plug them into the kernel estimator of density and the (26) would have followed directly from Lemma 1 and 3, and (22),(23) and (25). Nevertheless, due to the fact that for any $z \in V_n$ and any $\omega \in \Omega$ (see (16) and (24))

(27) $$\frac{|\hat{g}_n(z, r(\hat{\beta}_n)) - \hat{g}_n(z, r(\beta^0))|}{f(z)} \leq \frac{\sup_{t\in R} |w'(t)| \cdot p^{\frac{1}{2}} \cdot K \cdot \|\hat{\beta} - \beta^0\|}{c_n \cdot f(z)}$$

the (26) holds (for K see (5)).

■

In the next theorems we shall give an example of the «quantile» process $f_n(z)$, and some inequalities describing the asymptotic behavior of the estimator of the

contamination level. It will be clear from the proof of the inequalities that there is a hope that they are reasonably tight, of course, asymptotically. It means that the difference between the limit value and the lower bound is small. To be able to prove the assertions which were just mentioned, we shall need a result of Rosenblatt (1971). For the convenience of reader we are going to give it as a lemma.

Lemma 4 *(Rosenblatt (1971), p. 1828). Let Assumptions A be fulfilled and let a density $g(z)$ be continuously differentiable and bounded away from zero on $[0,1]$. Then if $n^{-\frac{1}{24}} = O(c_n)$ as $n \longrightarrow \infty$, it follows that*

$$P\left\{ \max_{0\le z\le 1} \left[\frac{c_n \cdot n}{\gamma g(z)}\right]^{\frac{1}{2}} \{\hat{g}_n(z, Z(\omega,n)) - \mathsf{E}\hat{g}_n(z, Z(\omega,n))\} \right.$$

$$\left. \le \left\{2\log\, c_n^{-1}\right\}^{\frac{1}{2}} + \frac{A+v}{(2\log\, c_n^{-1})^{\frac{1}{2}}} \right\} \longrightarrow \exp\{-\exp\{-v\}\}$$

for $n \longrightarrow \infty$ where G is the distribution function which corresponds to the density $g(z)$ and

$$\gamma = \int w^2(s)\mathrm{d}s$$

and

$$A = \log\frac{B^{\frac{1}{2}}}{2\pi}, \qquad B = -\frac{2}{\gamma}\frac{\mathrm{d}^2}{\mathrm{d}s^2}\left(\int w(u)\, w(u+s)\mathrm{d}u\right)\Bigg|_{s=0}.$$

Remark 4 The basic idea of the proof of Rosenblatt's lemma is as follows. At first, the process

$$\left[\frac{c_n \cdot n}{\gamma g(z)}\right]^{\frac{1}{2}} \{\hat{g}_n(z, Z(\omega,n)) - \mathsf{E}\hat{g}_n(z, Z(\omega,n))\}$$

is approximated by an appropriate Wiener process. Then an assertion about the distribution of the supremum of Wiener process (Rosenblatt (1971) uses Cramér, Leadbetter (1967)) is applied. It implies that due to symm etry of Wiener process we have also

$$P\left\{ \max_{0\le z\le 1} \left[\frac{c_n \cdot n}{\gamma g(z)}\right]^{\frac{1}{2}} \{\mathsf{E}\hat{g}_n(z, Z(\omega,n) - \hat{g}_n(z, Z(\omega,n))\} \right.$$

$$\left. \le \left\{2\log\, c_n^{-1}\right\}^{\frac{1}{2}} + \frac{A+v}{(2\log\, c_n^{-1})^{\frac{1}{2}}} \right\} \longrightarrow \exp\{-\exp\{-v\}\} \tag{28}$$

for $n \longrightarrow \infty$.

□

Remark 5 Rosenblat's result (28) is given in a somewhat unusual form. More usual form would be such which describes convergences of distribution functions, i. e.

$$P\left\{(2\log\, c_n^{-1})^{\frac{1}{2}} \max_{0\le z\le 1}\left[\frac{c_n\cdot n}{\gamma g(z)}\right]^{\frac{1}{2}}\{\mathsf{E}\hat{g}_n(z,Z(\omega,n)-\hat{g}_n(z,Z(\omega,n))\}\right.$$

$$\left.-2\log\, c_n^{-1}-A\le v\right\}\longrightarrow \exp\{-\exp\{-v\}\} \tag{29}$$

for $n\longrightarrow\infty$.

□

There are two things which we have to cope with to be able to use Rosenblat's result for our purposes. First of all, if we apply directly Rosenblat's result, there would be an inconvenient presence of $g^{\frac{1}{2}}(z)$ in the denominator of the above formula. Another difficulty is that the normalized difference contains $\mathsf{E}\hat{g}_n(z,Z(\omega,n))$ and does not give so a «distance» from $g(z)$. Moreover, the difference between $\mathsf{E}\hat{g}_n(z,Z(\omega,n))$ and $g(z)$ is proportional to a power of c_n (unfortunately not to power of n). A remedy for the all difficulties is a transformation of data (in our case the transformation of residuals). Let us assume that we have data $z_1,z_2,\ldots,z_n$ which are realization of a sequence of i.i.d. random variables, distributed according to a distribution function (d.f.) $G(z)$ and that we have selected some another d.f. $F(z)$ to explain them. The first step will be to estimate the asymptotic distribution of $\hat{\varepsilon}_{G,F}$ under the null hypothesis, i.e. under the hypothesis that $\varepsilon_{G,F}=0$ (realize that then $G(z)=F(z)$). Consider instead of $z_1,z_2,\ldots,z_n$ the data $u_1,u_2,\ldots,u_n$ such that $u_i=F(z_i)$ for $i=1,2,\ldots,n$. Then the density $f^*(u)$ of the transformed data is equal to 1 over the interval $[0,1]$ (and zero elsewhere). Moreover, in what follows let us assume that we have selected V_n (for V_n see Theorem 1) so that $F(V_n)\subset(h_n,1-h_n), h_n=c_n\cdot a$ (for a see Conditions A) where $F(V_n)=\{u: u=F(z), z\in V_n\}$ and let us compute the mean value of the kernel estimator for the transformed observations u_i's. We obtain for $u\in F(V_n)$

$$\mathsf{E}\hat{g}_n^*(u,U(\omega,n))=\int\{\frac{1}{nc_n}\sum_{i=1}^{n}w(c_n^{-1}(u-y))\,f^*(y)\}\mathrm{d}y$$

$$=\frac{1}{n}\sum_{i=1}^{n}\int_{-a}^{a}w(s)\,f^*(u-c_ns)\mathrm{d}s=1. \tag{30}$$

It means that the kernel estimator of the «transformed» density for any $u\in F(V_n)$ is unbiased. Let us assume that we have transformed the residuals using $F(z)$ and let us apply (28) on the transformed values. We obtain:

Theorem 2 *Let Conditions A and the assumptions of Theorem 1 and of Lemma 4 be fulfilled. For any* $p \in (0,1)$ *put* $v^* = -\log(-\log p)$ *and let*

$$b_n = \left(\frac{\gamma}{c_n}\right)^{\frac{1}{2}} \left[\left\{ 2\log c_n^{-1} \right\}^{\frac{1}{2}} + \frac{A + v^*}{(2\log c_n^{-1})^{\frac{1}{2}}} \right]. \tag{31}$$

Finally, let the density $f(z)$ *be bounded and let* $f_n(z) = f(z)(1 - b_n \cdot n^{-\frac{1}{2}})$. *Then under the assumption that* $\varepsilon_{G,F} = 0$ *we have*

$$\limsup_{n \longrightarrow \infty} P\{\hat{\varepsilon}_{G,F} = 0\} \geq p \qquad \text{for } n \longrightarrow \infty \tag{32}$$

and

$$\limsup_{n \longrightarrow \infty} P\left\{ \left(2\log c_n^{-1}\right)^{\frac{1}{2}} \left(\frac{c_n n}{\gamma}\right)^{\frac{1}{2}} \hat{\varepsilon}_{G,F} \leq v - v^* \right\} \geq \exp\{-\exp\{-v\}\} \tag{33}$$

for $(v - v^*)(2\log\, c_n^{-1})^{-\frac{1}{2}} [\frac{\gamma}{c_n n}]^{\frac{1}{2}}] \in (0,1)$.

Proof: First of all, let us say that in the proof some constants, say $C_1, C_2, \ldots$, will be used. Their definition will be assumed to hold only within the proof. We shall show that the proof follows immediately from (29). Let us consider the transformation $u = F(z)$ and let us denote the density of the transformed random variable by $g^*(u)$ and for any $\beta \in R^p$ put $r_u(\beta) = (r_{u1}(\beta), r_{u2}(\beta), \ldots, r_{un}(\beta))^T$, with

$$r_{ui}(\beta) = F(Y_i - X_i^T \beta). \tag{34}$$

Further, notice that the level of contamination is invariant with respect to the transformation. It is clear either from the heuristic background or from the formal expression. Really, the contamination level represents the percentage of the observations (among the data) which are not distributed according to the central model. So that if we transform data, the «earlier» central model is transformed into some «new» central model (in our case uniform distribution over $[0,1]$), and similarly, the «contaminating» distribution is transformed to some «new contaminating» distribution. So the percentage of «wrong» data is the same. On the other hand using the formal way, we see that the values of the fraction $\frac{f(z)-g(z)}{f(z)}$ are precisely the same as the values of $\frac{f(\text{inv}F(u))-g(\text{inv}F(u))}{f(\text{inv}F(u))}$ at the corresponding point $u = F(z)$. Multiplying by the Jacobian of the transformation both the numerator and the denominator, we do not change the value of the ratio. But then the density of the central model will become the density of the uniform distribution over $[0,1]$ and the density $g(z)$ is transformed on $g^*(u)$. So we obtain for the contamination level an equivalent expression

$$\sup_{u \in [0,1]} \{1 - g^*(u)\}.$$

Taking into account the specification of $f_n(z)$ given in the theorem which reads for the transformed residuals as

$$f_{n,U}(u) = 1 - b_n \cdot n^{-\frac{1}{2}} \tag{35}$$

we have for the estimator

$$\hat{\varepsilon}_{G,F} = \max\left\{0, \min\left\{\sup_{u\in F(V_n)}\left\{1 - b_n n^{-\frac{1}{2}} - \hat{g}_n^*(u, r_u(\hat{\beta}_n))\right\}, 1\right\}\right\} \tag{36}$$

where we have denoted the kernel density estimator based on the transforemed residuals by $\hat{g}_n^*(u, r_u(\hat{\beta}_n))$. Let us recall that for the a general $\beta \in R^p$ we have

$$\hat{g}_n^*(u, r_u(\beta)) = \frac{1}{nc_n}\sum_{i=1}^{n} w(c_n^{-1}(u - r_{ui}(\beta))).$$

Now we may write

$$\limsup_{n\longrightarrow\infty} P(\hat{\varepsilon}_{G,F} = 0) = \limsup_{n\longrightarrow\infty} P(\sup_{u\in F(V_n)}\{1 - b_n n^{-\frac{1}{2}} - \hat{g}_n^*(u, r_u(\hat{\beta}_n))\} \le 0)$$

$$= \limsup_{n\longrightarrow\infty} P(\sup_{u\in F(V_n)}\{1 - \hat{g}_n^*(u, r_u(\hat{\beta}_n))\} \le b_n n^{-\frac{1}{2}})$$

$$= \limsup_{n\longrightarrow\infty} P(\left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \sup_{u\in F(V_n)}[1 - \hat{g}_n^*(u, r_u(\hat{\beta}_n))] \le (2\log\, c_n^{-1})^{\frac{1}{2}} + \frac{A + v^*}{(2\log\, c_n^{-1})^{\frac{1}{2}}})$$

$$= \limsup_{n\longrightarrow\infty} P(\left(2\log\, c_n^{-1}\right)^{\frac{1}{2}}\left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \sup_{u\in F(V_n)}\{\mathsf{E}\hat{g}_n^*(u, r_u(\beta^0)) - \hat{g}_n^*(u, r_u(\beta^0))\}$$

$$+ \left(2\log\, c_n^{-1}\right)^{\frac{1}{2}}\left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \sup_{u\in F(V_n)}[\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - \hat{g}_n^*(u, r_u(\beta^0))] \le (2\log\, c_n^{-1}) + A + v^*)$$

$$\ge \limsup_{n\longrightarrow\infty} P(\left(2\log\, c_n^{-1}\right)^{\frac{1}{2}}\left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \max_{u\in[0,1]}\{\mathsf{E}\hat{g}_n^*(u, r_u(\beta^0)) - \hat{g}_n^*(u, r_u(\beta^0))\}$$

$$+ \left(2\log\, c_n^{-1}\right)^{\frac{1}{2}}\left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \max_{u\in[0,1]}[\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - \hat{g}_n^*(u, r_u(\beta^0))] \le 2\log\, c_n^{-1} + A + v^*.$$

So we shall need to estimate the difference

$$(2\log\, c_n^{-1})^{\frac{1}{2}}\left[\frac{c_n \cdot n}{\gamma}\right]^{\frac{1}{2}} \cdot \max_{u\in[0,1]}\left|\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - \hat{g}_n^*(u, r_u(\beta^0))\right|. \tag{37}$$

Taking into account (34), we may write

$$\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - \hat{g}_n^*(u, r_u(\beta^0)) = \frac{1}{nc_n} \sum_{i=1}^{n} \left[w(c_n^{-1}(u - r_{ui}(\hat{\beta}_n))) - w(c_n^{-1}(u - r_{ui}(\beta^0))) \right]$$

(38)
$$= \frac{1}{nc_n^2} \sum_{i=1}^{n} w'(\xi_i) \left[F(Y_i - X_i^T \beta^0) - F(Y_i - X_i^T \hat{\beta}_n) \right] = \frac{1}{nc_n^2} \sum_{i=1}^{n} w'(\xi_i) f(\eta_i) X_i^T \left(\hat{\beta}_n - \beta^0 \right)$$

where $\xi_i \in \left(c_n^{-1} \min \left\{ u - r_{ui}(\hat{\beta}_n), u - r_{ui}(\beta^0) \right\}, c_n^{-1} \max \left\{ u - r_{ui}(\hat{\beta}_n), u - r_{ui}(\beta^0) \right\} \right)$ and $\eta_i \in \left(\min \left\{ Y_i - X_i^T(\hat{\beta}_n), Y_i - X_i^T(\beta^0) \right\}, \max \left\{ Y_i - X_i^T(\hat{\beta}_n), Y_i - X_i^T(\beta^0) \right\} \right)$. Now, the expression (38) can be rewritten as

$$\frac{1}{nc_n^2} \sum_{i=1}^{n} \left[w'(\xi_i) - w'(c_n^{-1}(u - F(e_i))) \right] f(\eta_i) X_i^T \left(\hat{\beta}_n - \beta^0 \right) \tag{39}$$

$$+ \frac{1}{nc_n^2} \sum_{i=1}^{n} w'(c_n^{-1}(u - F(e_i))) \left[f(\eta_i) - f(e_i) \right] X_i^T \left(\hat{\beta}_n - \beta^0 \right) \tag{40}$$

$$+ \frac{1}{nc_n^2} \sum_{i=1}^{n} w'(c_n^{-1}(u - F(e_i))) f(e_i) X_i^T \left(\hat{\beta}_n - \beta^0 \right). \tag{41}$$

Taking into account that

$$|w'(c_n^{-1}\xi_i) - w'(c_n^{-1}(u - F(e_i))| \leq \sup_{z \in R} |w''(z)| c_n^{-1} K p^{\frac{1}{2}} \|\hat{\beta}_n - \beta^0\|,$$

and also the fact that we have assumed that $\sqrt{n}\|\hat{\beta}_n - \beta^0\| = O_p(1)$, we have

$$\sup_{u \in R} \left\{ \left(2 \log c_n^{-1} \right)^{\frac{1}{2}} \left[\frac{c_n n}{\gamma} \right]^{\frac{1}{2}} \frac{1}{nc_n^2} \left| \sum_{i=1}^{n} \left[w'(\xi_i) - w'(c_n^{-1}(u - F(e_i))) \right] f(\eta_i) X_i^T \left(\hat{\beta}_n - \beta^0 \right) \right| \right\}$$

$$\leq C_1 \cdot c_n^{-\frac{5}{2}} n^{-\frac{1}{2}} \left(\log c_n^{-1} \right)^{\frac{1}{2}}$$

where C_1 is a positive (and finite) constant and hence the expression (39) converges to zero as $n \longrightarrow \infty$. Similarly

$$\sup_{u \in R} \left\{ \left(2 \log c_n^{-1} \right)^{\frac{1}{2}} \left[\frac{c_n n}{\gamma} \right]^{\frac{1}{2}} \left| \frac{1}{nc_n^2} \sum_{i=1}^{n} w'(c_n^{-1}(u - F(e_i))) \left[f(\eta_i) - f(e_i) \right] X_i^T \left(\hat{\beta}_n - \beta^0 \right) \right| \right\}$$

$$\leq C_2 \cdot c_n^{-\frac{5}{2}} n^{-\frac{1}{2}} \left(\log\, c_n^{-1}\right)^{\frac{1}{2}}$$

where again C_2 is a positive (and finite) constant and hence the expression (40) converges also to zero as $n \longrightarrow \infty$. It remains to cope with supremum of the expression (41) which may be bounded by

$$\frac{1}{n^{\frac{1}{2}} c_n^2} \sup_{u \in R} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \left[w'(c_n^{-1}(z - F(e_i))) f(e_i) \right.\right.$$

$$\left.\left. - \mathsf{E}\left\{ w'(c_n^{-1}(u - F(e_i))) f(e_i) \right\} \right] X_i^T \left(\hat{\beta}_n - \beta^0 \right) \right\} \tag{42}$$

(43)

$$+ \frac{1}{n^{\frac{1}{2}} c_n^2} \sup_{u \in R} \left\{ \mathsf{E}\left\{ w'(c_n^{-1}(u - F(e_1))) f(e_1) \right\} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} X_i^T \right\} \left(\hat{\beta}_n - \beta^0 \right) \right\}.$$

Taking into account once again that $\sqrt{n}\|\hat{\beta}_n - \beta^0\| = O_p(1)$ and following Csörgö and Révész (1981), theorem 6.1.1, we find that the expression (42) is of order $c_n^{-2} n^{-1}$ in probability, and hence after multiplication by the factor $\left(2\log\, c_n^{-1}\right)^{\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}}$ we obtain order of this expression equal to $c_n^{-3} n^{-\frac{1}{2}} \log n \cdot \left(\log\, c_n^{-1}\right)^{\frac{1}{2}}$. Now we may calculate

$$\mathsf{E}\left\{ w'(c_n^{-1}(u - F(e_1))) f(e_1) \right\} = \int w'(c_n^{-1}(u - F(t))) f^2(t) \mathrm{d}t$$

$$= \int w'(s) f^2(\mathrm{inv}F(u - c_n s)) c_n \mathrm{d}s = \int w'(s) f^2(\mathrm{inv}F(u)) c_n \mathrm{d}s - 2 \int w'(s) f(\zeta) f'(\zeta) c_n^2 s \mathrm{d}s$$

where ζ is again an appropriately selected point. Taking into account (6), (14) and (15), we come to the conclusion that the term in (43) (after multiplication by $\left(\log\, c_n^{-1}\right)^{\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}}$) converges also to zero. So denoting

$$\kappa_n = \left(2\log\, c_n^{-1}\right)^{\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \sup_{u \in F(V_n)} \left\{ \left[\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - \hat{g}_n^*(u, r_u(\beta^0)) \right] \right\}$$

using (29) and taking into account that $v^* = -\log(-\log p)$ we have

$$\limsup_{n \longrightarrow \infty} P(\hat{\varepsilon}_{G,F} = 0) =$$

$$= \limsup_{n \longrightarrow \infty} P\left(\left(2\log\, c_n^{-1}\right)^{\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \sup_{u \in F(V_n)} \left\{ \mathsf{E}\hat{g}_n^*(u, r_u(\beta^0)) - \hat{g}_n^*(u, r_u(\beta^0)) \right\} + \kappa_n \right.$$

$$\leq 2\log\, c_n^{-1} + A + v^*) \longrightarrow \exp\{-\exp\{-(v^* + \kappa_n)\}\} \longrightarrow p.$$

It concludes the proof of (32). Repeating the same steps we arrive at

$$\limsup_{n \longrightarrow \infty} P\left\{ (2\log\, c_n^{-1})^{\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \hat{\varepsilon}_{G,F} \leq v - v^* \right\}$$

$$\geq \lim_{n \longrightarrow \infty} P\left\{ (2\log\, c_n^{-1})^{\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \max_{u\in[0,1]} \left\{ \mathsf{E}\hat{g}_n(u, r_u(\hat{\beta}_n)) - \hat{g}_n(u, r_u(\hat{\beta}_n)) \right\} \right.$$
$$\left. + \kappa_n \leq 2\log\, c_n^{-1} + A + v \right\} \longrightarrow \exp\{-\exp\{-v\}\}$$

which concludes the proof.

■

Remark 6 It follows from the proof of the previous theorem that the lower bound in (33) is tight.

□

Similarly for a sequence of local alternatives we may obtain:

Theorem 3 *Let the assumptions of Theorem 2 be fulfilled. Put for any* $p \in (0,1)$ *again* $v^* = -\log(-\log\, p)$ *and let* b_n *be given by (31). Finally, let* $f_n(z) = f(z)(1 - b_n \cdot n^{-\frac{1}{2}})$ *and* $H(x)$ *any distribution such that* $\varepsilon_{H,F} \neq 0$ *and* $\max_{u\in[0,1]} h(\mathrm{inv}F(u) < \infty$. *Then for any* $\varepsilon \in (0,1)$ *and the sequence of the local alternative* $\{G_n(x)\}_{n=1}^{\infty}, G_n(x) = (1 - \varepsilon n^{-\frac{1}{2}})F(x) + \varepsilon n^{-\frac{1}{2}} H(x)$ *we have*

$$\limsup_{n \longrightarrow \infty} P\left\{ \left(2\log c_n^{-1}\right)^{\frac{1}{2}} \left(\frac{c_n n}{\gamma}\right)^{\frac{1}{2}} (\hat{\varepsilon}_{G_n,F} - \varepsilon_{G_n,F}) \leq v - v^* \right\} \leq \exp\{-\exp\{-v\}\}$$

for $(v - v^*)(2\log c_n^{-1})^{-\frac{1}{2}} \left[\frac{c_n n}{\gamma}\right]^{\frac{1}{2}} \in (0,1)$.

Proof: mimics the proof of Theorem 2. First of all, we shall make an idea about $\varepsilon_{G_n,F}$. Let again $f^*(u)$, $h^*(u)$ and $g^*(u)$ denote the transformed densities. Taking into account that $f^*(u) = 1$ and $h^*(u) \geq 0$, we have

$$\varepsilon_{G_n,F} = \sup_{u\in[0,1]} 1 - g^*(u) \leq 1 - 1 + \varepsilon n^{-\frac{1}{2}} = \varepsilon n^{-\frac{1}{2}}.$$

We will need also to estimate $\mathsf{E}\hat{g}_n^*(u, r_u(\hat{\beta}_n))$. Similarly as in (30) we utilize the fact that $f^*(u) = 1$ but now we need also to employ the assumption that $h^*(u)$ is bounded. Then we obtain

$$|\mathsf{E}\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - 1| \leq \int \left|\hat{g}_n^*(u, r_u(\hat{\beta}_n)) - 1\right| \left((1 - \varepsilon n^{-\frac{1}{2}}) + \varepsilon n^{-\frac{1}{2}} h^*(u)\right) \mathrm{d}u \leq C_1 n^{-\frac{1}{2}}$$

where C_1 is a finte constant. Now we may write

$$P\left\{ \left(2\log c_n^{-1}\right)^{\frac{1}{2}} \left(\frac{c_n n}{\gamma}\right)^{\frac{1}{2}} (\hat{\varepsilon}_{G_n,F} - \varepsilon_{G_n,F}) \leq v - v^* \right\}$$

$$\geq P\left\{ \left(2\log c_n^{-1}\right)^{\frac{1}{2}} \left(\frac{c_n n}{\gamma}\right)^{\frac{1}{2}} (\mathsf{E}\hat{\varepsilon}_{G_n,F} - \hat{\varepsilon}_{G_n,F}) - 2\log\, c_n^{-1} - A \leq v - C_2 n^{-\frac{1}{2}} \right\}$$

where again C_2 is a finite constant and the assertion of the theorem follows.

■

3. SIMULATION STUDY

All results derived in the previous section are of the asymptotic type. Moreover, the results were obtained in the asymptotic framework in which several parameters changed simultaneously. Except of the number of observations which increased to infinity also the width of window, quantile process and the set over which the supremum had been taken, converged to the corresponding limits. Anybody who had sometimes tried to use such results to approximate the corresponding probabilities for the finite samples, has find out that «an adjustment» of the parameters (the width of window, quantile process, etc.) needs some simulation studies. Sometimes we may even meet with the standpoint that such asymptotic results should be interpreted only as a guarantee of the consistency (or coherence, if you want) of our approach with the general «structure of mathematics». That is the reason why the behavior of the statistics which was proposed above should be studied for the finite samples by simulations. In this section we shall offer a very first experience in the case when the data are contaminated, i.e. for $\varepsilon_{G,F} \neq 0$.

For the numerical study we have simulated data in the following way: Considering the regression model

$$Y_i = 2 \cdot X_{i1} + 3 \cdot X_{i2} + 4 \cdot X_{i3} + e_i, \qquad i = 1, 2, \ldots, 30$$

we have generated 30 three-dimensional vectors uniformly distributed over $[0, 10]$ (used as the carriers $X_i = (X_{i1}, X_{i2}, X_{i3})^T$, $i = 1, 2, \ldots, 30$) and 30 random numbers distributed according to the standard normal distribution (used as the errors $e_1, e_2, \ldots, e_{30}$).

The normality was checked by chi-square and Kolmogorov-Smirnov tests accompanied by the test of skewness and of kurtosis (see Shapiro, Wilk (1965)). Then we have randomly selected from $\{e_i\}_{i=1}^{30}$ four numbers, say $e_{i_1}, e_{i_2}, e_{i_3}, e_{i_4}$, multiplied them by 4 (they have represented the contamination). According to the Remark 3 we have then

$$\varepsilon_{G,F} = \frac{4}{30}(1-\frac{1}{4}) = 0.1. \tag{44}$$

Then the least trimmed square (LTS) algorithm was used to estimate the coefficients of the regression model, i.e.

$$\hat{\beta}^{(LTS)} = \underset{\beta \in R^p}{\operatorname{argmin}} \sum_{i=1}^{h} r^2_{[i:n]}(\beta)$$

where $r^2_{[i:n]}$'s represent ordered squared residuals $r_i^2(\beta) = [Y_i - \sum_{j=1}^3 x_{ij}\beta_j]^2$, i.e. $r^2_{[1:n]} \leq r^2_{[2:n]} \leq \cdots \leq r^2_{[n:n]}$, and $h = 17$ was selected to reach the maximal possible break-down point of the estimator (see Rousseeuw,Leroy (1987)). Finally the obtained residuals $r_i(\hat{\beta}^{(LTS)}) = Y_i - \sum_{j=1}^3 x_{ij}\hat{\beta}_j^{(LTS)}$, $i = 1,2,\ldots,30$ were transformed (see Theorem 2). The whole procedure was 30 times repeated. In what follows let $(\hat{\varepsilon}_{G,F})_k$ denote the estimate of the contamination level for the k-th sample.

In such a way set of 30 samples of transformed residuals with the contamination level equal to 0.1 (see (44)) was obtained (each sample contained 30 residuals). This collection of samples was used as a training set. As follows from (36) the value of $\hat{\varepsilon}_{G,F}$ for any fix sample of data is a nonincreasing function of b_n, $\hat{\varepsilon}_{G,F}(b_n) : [0,\infty) \longrightarrow [0, 1 - \min_{u \in F(V_n)} \hat{g}_n(u, r_u(\hat{\beta}_n))]$. Since in our case we had $\min_{u \in F(V_n)} \hat{g}_n(u, r_u(\hat{\beta}_n)) < 0.9$, it was possible to find b^*_{30} so that

$$\frac{1}{30}\sum_{k=1}^{30}(\hat{\varepsilon}_{G,F}(b^*_{30})_k) = 0.1$$

(see (44) once again). We have obtained $b^*_{30} = 1.80049$ (or in other words, we have learnt that the quantile process $f_{30,U}(u)$ (see (35)) for this type of data, this kernel etc. should be (approximately) equal to $1 - b^*_n n^{-\frac{1}{2}} = 1 - \frac{1.80047}{\sqrt{30}} = 0.67128$). The results of estimating regression coefficients and of the values of the estimates $\hat{\varepsilon}_{G,F}$ (after assigning the value 1.80049 to b^*_{30}) have been collected in the Table 3.

$$\frac{1}{30}\sum_{k=1}^{30}(\hat{\varepsilon}_{G,F})_k = 0.1 \qquad \hat{\operatorname{var}}(\hat{\varepsilon}_{G,F}) = 0.02961$$

Table 3. *Results of estimation of regression coefficients and level of contamination for the training set of samples*

case	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\varepsilon}_{G,F}$
1	1.878	3.116	4.027	0.0000
2	2.113	3.139	3.731	0.0000
3	1.898	3.028	4.100	0.1048
4	2.022	2.688	4.282	0.0073
5	2.172	2.837	4.046	0.1679
6	1.887	2.975	4.120	0.2185
7	2.086	3.154	3.796	0.0590
8	2.059	2.857	4.063	0.1817
9	1.924	3.090	4.009	0.0000
10	2.161	2.964	3.898	0.0000
11	2.003	3.129	3.896	0.0894
12	1.834	2.983	4.201	0.0000
13	1.936	2.893	4.150	0.2644
14	1.948	3.104	3.949	0.0000
15	2.297	3.016	3.766	0.1416
16	1.911	3.072	4.006	0.0967
17	1.836	3.087	4.110	0.0734
18	1.978	2.994	3.997	0.1159
19	2.007	2.871	4.174	0.1060
20	2.079	3.082	3.857	0.0000
21	2.075	2.856	4.039	0.3431
22	1.959	3.030	3.995	0.0555
23	1.945	3.082	4.017	0.1684
24	2.470	2.686	3.813	0.0835
25	1.969	3.091	3.973	0.0000
26	1.782	3.167	3.983	0.3292
27	1.910	3.134	3.980	0.0000
28	1.905	3.121	3.959	0.1230
29	1.871	2.941	4.183	0.0975
30	1.979	2.943	4.093	0.1882

Table 4. *Results of estimation of regression coefficients and level of contamination for the testing set of samples*

case	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\varepsilon}_{G,F}$
1	2.092	3.012	3.972	0.1610
2	2.003	2.992	4.029	0.2534
2 3	2.054	2.972	3.982	0.0000
4	1.969	3.134	3.914	0.0000
4 5	1.959	3.008	3.999	0.1306
6	2.014	3.037	3.917	0.0000
6 7	2.013	2.919	4.043	0.0138
8	2.198	2.943	3.803	0.1033
9	1.858	2.916	4.207	0.1440
10	1.862	3.035	4.071	0.1477
11	1.851	3.177	4.007	0.0000
12	2.086	2.889	4.024	0.2434
13	1.959	2.993	4.062	0.1582
14	1.979	3.024	4.032	0.2457
15	2.042	2.728	4.200	0.0000
16	1.802	2.986	4.192	0.0791
17	2.240	2.815	3.937	0.0412
18	2.063	3.040	3.888	0.3725
19	1.915	2.921	4.241	0.0977
20	1.928	2.981	4.073	0.1879
21	1.719	3.103	4.174	0.0000
22	2.237	2.913	3.844	0.0000
23	1.914	3.034	4.033	0.1315
24	1.975	2.969	4.081	0.2305
25	1.976	3.049	3.951	0.0000
26	1.845	2.963	4.172	0.1968
27	2.114	3.005	3.888	0.0000
28	1.987	2.882	4.173	0.0742
29	1.952	3.162	3.863	0.0000
30	1.979	3.120	3.918	0.0000

In the same way as the training set was prepared we have generated a testing set consisting again 30 samples, each of them containing 30 observations. Again LTS were applied and the corresponding results of the estimation of regression coefficients together with the results of estimation of contamination level of the residuals are given in the Table 4.

$$\frac{1}{30}\sum_{k=1}^{30}(\hat{\varepsilon}_{G,F})_k = 0.10042 \qquad \hat{\mathrm{var}}(\hat{\varepsilon}_{G,F}) = 0.03047$$

4. CONCLUSIONS

The paper brings a (theoretical) background for the selection of some free parameters of the robust methods (of the linear regression analysis) via estimating the contamination level. As it was already discussed the asymptotic results, may be except of the consistency of the estimator, have mainly a theoretical importance of some coherence of our approach with the general principles of mathematics. For the practical applications we should rely (mainly) on the results of a simulation studies.

In more details, we may procede as follows.

At first we estimate contamination level.

Of course, to be able to do it we need to adjust some value to the quantile process, to the width of window, to select appropriately the set V_n etc. It may be done on the base of experiences with the data of the same or similar character, or using the results of «reasonably» organized simulation study. It is clear that the type of distribution of the random errors is relevant, and so we have to employ our ideas about the character of these disturbances.

Secondly, we select the «tuning» parameter(s) of the corresponding family of robust methods.

(As an example of such family may serve the family of Huber's ψ-functions $\{\psi_{k(\varepsilon)}\}_{\varepsilon\in(0,1)}$, see (13).) Such selection may be performed either according to a known formula, connecting the contamination level with the «tuning» parameter(s), or by means of some theoretical tool of the type of efficiency rate or the local deficiency. Let us give an example.

After evaluating the estimate $\hat{\varepsilon}_{G,F}$ of the contamination level $\varepsilon_{G,F}$ we may calculate the estimate $\hat{\varepsilon}_H$ of the appropriate Huber mixture parameter ε_H by means of (9) or of (10). Then using the relation (11) and (12) we assign the value of the tuning constant $k(\hat{\varepsilon}_H)$ for the Huber's ψ-function .

In the case of another type of robust procedure we may use e. g. the efficiency rate and/or the local deficiency to select the «tuning» constant corresponding to the estimated contamination level. An example of using the efficiency rate for the selection of proper α for α-estimators is given in Rubio, Víšek (1992). Let us be again more explicit.

Let $\{T_\alpha\}_{\alpha\in(0,1)}$ be a family of α-estimators of a parameter θ. (For the simplicity let us assume that θ is scalar.) Let us recall that the α-estimators are defined as the minimal distance estimators minimizing the α-divergence between the empirical d. f. and a d. f. from a projection family of d. f.'s (see Vajda (1989)). Having evaluated the efficiency rate of T_α in the corresponding model of contamination, we find the optimal selection of α for the estimated contamination level $\hat\varepsilon_{G,F}$, say $\alpha(\hat\varepsilon_{G,F})$. (Let us recall that the efficiency rate was defined in Rubio, Víšek (1992) as the derivative of the supremum of the (asymptotic) variances of the estimators; supremum is taken over the given model of contamination, usually over some neighborhood of a central model; derivative is evaluated with respect to the parameter of the family of estimators, in our example derivative with respect to α. Of course if an objective function would be other than supremum of the variances, we should evaluate derivative of this function in the model of contamination.) Then we use for the estimation of θ the estimator $T_{\alpha(\hat\varepsilon_{G,F})}$ which then minimizes the supremum of the asymptotic variances for given contamination level.

We are aware that the selection of tuning constant on the basis of an estimate of contamination level is the selection within the limits of one type of estimators (e. g. Huber's ones). Selection among different types of estimators should be based on some general principles (e. g. homogeneity of residuals over factor space, see Rubio *et al.* (1993) or Rubio and Víšek (1994), or subsample stability of the estimates, Víšek (1996 b)), and/or on some model oriented rules (explicitly) formulated by the expert who has collected data (see Víšek (1995)).

The results of simulation study presented in Table 4 showed that the estimates of the contamination level are scattered quite near around the «true» value (see the estimate of the variance of the estimator). It supports a hope that the proposed estimator of the contamination level may work well. On the other hand, it is clear that the method belongs among computationally intensive ones.

5. ACKNOWLEDGEMENT

We would like to thank the anonymous referee for a very detailed suggestions of corrections of all omisions. His/her proposals led to the considerable improvement of the text.

6. REFERENCES

[1] **Antoch, J.** and **Víšek, J.Á.** (1991). «Robust estimation in linear models and its computational aspects». *Contributions to Statistics: Computational Aspects of Model Choice*. Springer Verlag, (1992), ed. J. Antoch, 39–104.

[2] **Antoch, J.** and **Vorlíčková, D.** (1992). *Selected methods of statistical analysis.* Academia, Praha.

[3] **Boscovisch, R.J.** (1757). «De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura eius ex exemplaria etiam sensorum impressa». *Bononiensi Scientiarum et Artium Instituto Atque Academia Commentarii*, **4**, 353–396.

[4] **Chatterjee, S.** and **Hadi, A. S.** (1988). *Sensitivity Analysis in Linear Regression.* New York: J. Wiley & Sons.

[5] **Cramér, H.** and **Leadbetter, M.R.** (1967). *Stationary and Related Processes.* Wiley, New York.

[6] **Csörgö, M.** and **Révész, P.** (1981). *Strong Approximation in Probability and Statistics.* Akademia Kiadó, Budapest.

[7] **Galilei, G.** (1632). *Dialogo dei masimi sistemi.*

[8] **Galton F.** (1886). «Regression towards mediocrity in hereditary stature». *Journal of the Antropological Institute*, **15**, 246–263.

[9] **Gauss F.C.** (1809). «Theoria molus corporum celestium». *Hamburg, Perthes et Besser.*

[10] **Hewitt, E.** and **Stromberg, K.** (1965). *Real and Abstract Analysis.* Springer - Verlag, Berlin.

[11] **Huber, P.J.** (1964). «Robust estimation of a location parameter». *Ann. Math. Statist.*, **36**, 1753–1758.

[12] **Huber, P.J.** (1981). *Robust Statistics.* J. Wiley & Sons, New York.

[13] **Huber, P.J.** and **Strassen, V.** (1973). «Minimax tests and Neyman-Pearson lemma for capacities». *Ann. Statist.*, **1**, 251–263.

[14] **Koenker,R.** and **Bassett, G.** (1978). «Regression quantiles». *Econometrica*, **46**, 33–50.

[15] **Laplace, P.S.** (1793). «Sur quelques points du systeme du mode». *Memoires de l'Academic Royale des Sciences de Paris*, 1–87.

[16] **Legendre A.M.** (1805). «Nouvelles méthodes pour la détermination des orbites des comètes». *Paris, Courcier.*

[17] **Marazzi, A.** (1992). *Algorithms, Routines and S Functions for Robust Statistics.* Wadsworth & Brooks/Cole Publishing Company, Belmont, California, 1992.

[18] **Rosenblatt, M.** (1971). «Curve estimates». *Ann. Math. Statist.*, **42**, 1815–1842.

[19] **Rousseeuw, P.J.** and **Leroy, A.M.** (1987). *Robust Regression and Outlier Detection*. J.Wiley & Sons, New York.

[20] **Rubio, A.M., Aguilar, L.** and **Víšek, J.Á.** (1992). «Testing for difference between models». *Computational Statistics*, **8**, 57–70.

[21] **Rubio, A. M., Quintana, F.** and **Víšek, J.Á.** (1993). «Diagnostics of non-linear regression». *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague*, 1994, 203–206.

[22] **Rubio, A. M.** and **Víšek J.Á.** (1992). «Efficiency rate and local deficiency of Huber's estimator of location and of estimators». *Trabajos de Estadistica*, **6**, 67–85.

[23] **Rubio, A. M.** and **Víšek, J.Á.** (1994). «Diagnostics of regression model: Test of goodness of fit». *Proceedings of the Fifth Prague Symposium on Asymptotic Statistics*, eds. M. Hušková & P. Mandl, Physica Verlag, 423–432.

[24] **Rukhin, A.L.** (1994). «Optimal estimation of the mixture parameter». *Transactions of the Twelfth Prague Conference on Information Theory*, Statistical Decision Functions and Random Processes, Prague, 1994, 214–216.

[25] **Shapiro, S.S.** and **Wilk, M.B.** (1968). «A comparative study of various tests for normality». *J. Amer. Statist. Assoc.*, **63**, 1343–1372.

[26] **Vajda, I.** (1989). *Theory of Statistical Inference and Information*. Kluwer Academic Publ., Dordrecht.

[27] **Víšek, J.Á.** (1989). «Estimation of contamination level in the model of contamination with general neighbourhoods». *Kybernetika*, **25**, 278–297.

[28] **Víšek, J.Á.** (1993). «On the role of contamination level and the least favorable behaviour of gross-error sensitivity». *Probability and Mathematical Statistics*, **14, 2**, 173–187.

[29] **Víšek, J.Á.** (1994 a). «A cautionary note on the method of Least Median of Squares reconsidered». *Transactions of the Twelfth Prague Conference on Information Theory*, Statistical Decision Functions and Random Processes, Prague, 1994, 254–259.

[30] **Víšek, J.Á.** (1994 b). «High robustness and an illusion of truth». *Transactions of ROBUST'94*, J const CMF (Union of Czech mathematicians), eds. J. Antoch & G. Dohnal, ISBN 80-7015-492-6, 172–185.

[31] **Víšek, J.Á.** (1995). «On the heuristics of statistical results». *Proceedings of 'PROBASTAT'94', Tatra Mountains Publ.*, **7**, 349–357, Bratislava

[32] **Víšek, J.Á.** (1996 a). «n high breakdown point estimation». *Computational Statistics*, **11**, 137–146, Berlin.

[33] **Víšek, J.Á.** (1996 b). «Sensitivity analysis of *M*-estimates». *Annals of the Institute of Statistical Mathematics*, **48**, 469–495.

[34] **Víšek, J.Á.** (1996 c). «On the coefficient of determination: Simple but...». *Bulletin of the Czech Econometric Society*, **5**, 117–124.

[35] **Víšek, J.Á.** (1997). «On the diversity of estimates». Submitted to *Computational Statistics and Data Analysis*.