About Regression Estimators with High Breakdown Point

Vandev, D.L.
Inst. of Mathematics., Bulg. Acad. Sciences,
P.O.Box 373,1113 Sofia, Bulgaria
and Neykov, N.M. [1]
Inst. of Meteorology and Hydrology, Bulg. Acad.Sciences,
66 Tsarigradsko chaussee, 1784 Sofia, Bulgaria

Abstract

A generalisation of a theorem by Vandev (1993) concerning the finite sample breakdown point is given. Using this result the breakdown point properties of the LMS and LTS estimators of Rousseeuw (1984) and the rank-based regression estimator of Hössjer (1994) are studied. Moreover, the breakdown point properties of the weighted least trimmed estimators of order k in the case of grouped logistic regression are investigated, as well as linear regression with an exponential q-th power distribution.
1991 Mathematics Subject Classification: Primary 62J05; Secondary 62F35.
Keywords: Breakdown point, Least Median of Squares, Least Trimmed Squares, Modified maximum likelihood, R-estimators, Robust regression estimators, Logistic regression.

1 Introduction

It is well known that the Maximum Likelihood Estimator (MLE) can be very sensitive to some deviations from the assumptions, in particular to unexpected outliers in the data. To overcome this problem many robust alternatives of the MLE have been developed in the last decades. For detailed introduction, see Huber (1981), Hampel et al. (1986), and Rousseeuw and Leroy (1987).

Neykov and Neytchev (1990), following the definitions of the Least Median of Squares (LMS) and Least Trimmed Squares (LTS) estimator of Rousseeuw (1984), introduced two classes of estimators for the parameters of any unimodal distribution with regular density as an extension of the maximum likelihood principle. This modification considers the likelihood of individual observations as residuals and applies on them the basic idea of the LMS and LTS estimators. The corresponding estimators are called Least Median Estimator (LME(k)) and Least Trimmed Estimator (LTE(k)), where k is tuning constant which can be chosen by the user within some reasonable range of values, see Vandev and Neykov (1993).

In order to study the breakdown properties of such estimators Vandev (1993) developed a d–fullness technique. He proved that their breakdown point is not less than $(n-k)/n$ if k is within the range of values $(n+d)/2 \leq k \leq (n-d)$ for some constant d which depends upon the density considered. Vandev and Neykov (1993) determined the value of d for the set of log-density functions for the multivariate normal case.

In this paper a generalisation of the theorem by Vandev (1993) concerning the finite sample breakdown point is given. The concept of d–fullness is applied to a general class of estimators called Weighted Least Trimmed Estimator of order k (WLTE(k)). Thus in many cases we are able to find the exact value of the breakdown point. In particular:

[1]Correspondence to: N.Neykov, E-mail:neyko.neykov@meteo.bg

- in the linear regression setting with an arbitrary distribution of the errors it is shown that the breakdown point of the LMS and LTS regression estimators of Rousseeuw (1984) and the rank-based regression estimators of Hössjer (1994) is equal to $(n-k)/n$ if the observations $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position and k is within the bounds $(n+p+1)/2 \leq k \leq n-p-1$;

- further in the case of linear regression with q-th power exponential distribution for the errors it is proved that the breakdown point of the WLTE(k) estimators for simultaneous estimation of the regression and scale parameters is equal to $(n-k)/n$ if the observations $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position and k is within the bounds $(n+p+1)/2 \leq k \leq n-p-1$;

- in the case of grouped binary logistic regression it is shown that the breakdown point of the WLTE(k) estimators is $(m-k)/m$, where $(m+p+1)/2 \leq k \leq m-p-1$, m is the number of groups, $n = n_1 + n_2 + \ldots + n_m$ is the total number of the observations and y_i represents the number of responses that satisfy the inequality $0 < y_i < n_i$ for $i = 1, \ldots, m$.

Some of the estimators we introduce are too general to be used in practice. So an investigation of their asymptotic behaviour, consistency and equivariance properties is beyond the scope of the paper.

This paper is organised as follows. Section 2 defines the concepts of d–fullness and breakdown point. Section 3 considers the breakdown point properties of a general class of regression estimators in terms of d–fullness. Sections 4 and 5 establish the breakdown point of simultaneous estimation of the scale and parameters in the linear regression model with exponential q–th order distribution of error and the grouped binary logistic regression respectively. The proofs of all the theorems and lemmas (except Theorem 1) are given in Section 6.

2 Definition of breakdown point and d-fullness

We now define the replacement variant of the finite sample breakdown point given by Hampel et al. (1986), which is closely related to the one introduced by Donoho and Huber (1983). Let $\Omega = \{\omega_i \in R^p, \text{ for } i = 1, \ldots, n\}$ be a sample of size n.

Definition 1. The breakdown point of an estimator T at Ω is given by

$$\varepsilon_n^*(T) = \frac{1}{n} \max\{m : \sup_{\tilde{\Omega}_m} \|T(\tilde{\Omega}_m)\| < \infty\},$$

where $\tilde{\Omega}_m$ is any sample obtained from Ω by replacing any m of the points in Ω by arbitrary values.

Thus, there is a compact set such that the estimator T remains in it even if we replace any m elements of the sample Ω by arbitrary ones. The largest m/n for which this property holds is the breakdown point.

We shall recall the basic definitions given by Vandev (1993), Vandev and Neykov (1993).

Definition 2. A real valued function $g(\theta)$ defined on a topological space Θ is called subcompact, if its Lebesgue sets $\{\theta : g(\theta) \leq C\}$ are compact for any constant C.

This notion is useful because the minima of any subcompact function exist, since the set of minima is an intersection of nonempty compacts.

Definition 3. A finite set F of n functions is called d–full, if for each subset of cardinality d of F, the supremum of this subset is a subcompact function.

Let the finite set $F = \{f_i(\theta) \geq 0, i = 1, \ldots, n, \text{ for } \theta \in \Theta\}$ be d–full and Θ is a topological space. Consider the function

$$D(k, \theta) = \sum_{i=1}^{k} w_i f_{\nu(i)}(\theta).$$

Here $f_{\nu(1)}(\theta) \leq f_{\nu(2)}(\theta) \leq \ldots \leq f_{\nu(n)}(\theta)$ are the ordered values of f_i at θ, $\nu = (\nu(1), \ldots, \nu(n))$ is the corresponding permutation of the indices, which may depend on θ. The weights $w_i \geq 0$ for $i = 1, \ldots, n$ are such that there exists an index $k = \max(i : w_i > 0)$.

The functions $S(k, \theta)$ and $M(k, \theta)$ of Vandev (1993) are obtained from $D(k, \theta)$ respectively if $w_i = 1$ for $i = 1, \ldots, k$ and $w_i = 0$ otherwise and if $w_k = 1$ and $w_i = 0$ for $i \neq k$.

The breakdown properties of the set of values of θ that minimize $D(k, \theta)$ is of primary interest in this paper. Let us denote this set by $R(k)$, in analogy to the sets $U(k)$ and $V(k)$ that correspond with $M(k, \theta)$ and $S(k, \theta)$. From a statistical point of view these sets can be considered as the sets of estimates if the functions $f_i(\theta)$ are appropriately chosen, e.g. depend on the observations. Vandev (1993) proves that the breakdown points of $U(k)$ and $V(k)$ are not less than $(n-k)/n$ if k is such that $(n+d)/2 \leq k \leq n-d$. The largest $(n-k)/n$ for which this property holds is related to the replacement version of the breakdown point of Hampel et al. (1986).

The following theorem characterizes the breakdown properties of the set $R(k)$ and is a generalization of the theorem of Vandev (1993).

Theorem 1. Let $F = \{f_i(\theta) \geq 0, \text{ for } i = 1, \ldots, n;\ \theta \in \Theta\}$ be d-full, the constants $w_i \geq 0$ for $i = 1, \ldots, n$, $w_k = 1$ for $k = \max\{i : w_i > 0\}$, $n \geq 3d$ and $(n + d)/2 \leq k \leq n - d$. Then the breakdown point of the set $R(k)$ is not less than $(n - k)/n$.

Proof. We shall follow the proof of Vandev (1993). We shall note that the set $R(k)$ is nonempty compact. If $k \geq d$ it can be represented as the intersection of finite subsets of compact sets of cardinality d as the set F is d–full. Let us fix $\theta_0 \in R(k)$ and then choose the constant

$$C = \sup_{\tau} \sum_{i=1}^{k} w_i f_{\tau(i)}(\theta_0),$$

where $\tau = (\tau(1), \tau(2), \ldots, \tau(n))$ is any permutation of the indices $1, 2, \ldots, n$.

Let $F' = \{f'_i(\theta) \geq 0, \text{ for } i = 1, \ldots, n;\ \theta \in \Theta\}$ be a perturbed version of F, which is obtained by replacing $(n - k)$ original elements of F with arbitrary nonnegative functions.

Obviously $D(k, \theta_0) \leq C$. Denote by $D'(k, \theta_0)$ the corresponding functional defined on F'. Also we have $D'(k, \theta_0) \leq C$ as there are k original elements belonging to F'.

Denote by $K(J, C) = \{\theta : f(\theta) \leq C, f \in J\}$, where $J \subset F$ with cardinality d. Denote by W the set of all subsets of F with cardinality d and by $K = \bigcup_{J \in W} K(J, C)$. As W is finite, obviously K is a compact set. Denote by I the set of the k lowest valued functions at the minima. The set K is nonempty and it contains $K(I, C) = \{\theta : f(\theta) \leq C, f \in I\}$ which contains $R(k)$, since $C \geq D(k, \theta) \geq w_k f_{\nu(k)}(\theta) = f_{\nu(k)}(\theta)$ is true for each $\theta \in R(k)$.

Let θ be a point outside K. If such a point does not exist the theorem is proved ($K \equiv \Theta$). First we note that for any d functions in F there exists at least one that has a value bigger than C at θ. Otherwise the point θ would belong to some $K(J, C)$, i.e. to K.

Suppose now that $D'(k,\theta) \le C$ for some perturbed set F'. Denote by I' the subset of the k lowest-valued functions in F'. Note that there are at least $k-(n-k)$ functions in $F \cap I'$ and this number is not less than d since $(n+d)/2 \le k$. So there exists a function in this set with value exceeding C. This is a contradiction and hence we have that $D'(k,\theta) > C$ for any θ outside K for any perturbed set F'. Therefore the set

$$
\begin{aligned}
R'(k) &= \operatorname*{argmin}_{\theta \subset \Theta} D'(k,\theta) \subset \{\theta :\ D'(k,\theta) \le D'(k,\theta_0)\} \\
&\subset \{\theta :\ D'(k,\theta) \le C\} \subset K. \qquad \textbf{Q.E.D.}
\end{aligned}
$$

Remark 1. (i) The condition $n \ge 3d$ is sufficient for the existence of a positive integer k that satisfies the inequalities $(n+d)/2 \le k \le n-d$.
(ii) The requirement $w_k = 1$ is technical as the constants w_i, $i = 1,\dots,n$ can be normalized by dividing with the positive constant w_k. This will not change $R(k)$.
(iii) The proposed estimator $R(k)$ that combines both types of estimators $U(k)$ and $V(k)$ is defined for nonnegative functions. Therefore the result concerning the breakdown properties of the $R(k)$ estimator proposed in this paper could be considered as a generalization of the main result of Vandev (1993). Vandev's result concerning $U(k)$ estimators remains valid even for nonpositive functions.

3 Estimation of the parameters in the linear regression model

Consider the multiple regression model

$$
y_i = x_i^T \beta + \varepsilon_i \quad for \quad i = 1,\dots,n, \tag{1}
$$

where y_i is an observed response, x_i is a $p \times 1$–dimensional vector of explanatory variables and β is a $p \times 1$ vector of unknown parameters. Classically ε_i, $i = 1,\dots,n$ are assumed to be independent and identically distributed as $N(0,\sigma^2)$, for some $\sigma^2 > 0$.

Consider the class of regression estimators defined as

$$
\hat{\beta} = \operatorname*{argmin}_{\beta} \sum_{i=1}^{k} w_i \rho(|r(\beta)|_{(i)}), \tag{2}
$$

where $|r(\beta)|_{(1)} \le |r(\beta)|_{(2)} \le \dots \le |r(\beta)|_{(n)}$ are the ordered absolute values of the regression residuals $r_i = y_i - x_i^T\beta$, the weights w_i are the same as in the previous section and ρ is strictly increasing continuous function such that $\rho(0) = 0$.

Following the reasoning of Rousseeuw and Leroy (1987) it is easily verified that this class of estimators is regression, scale and affine equivariant. The following theorem characterizes their breakdown point.

Theorem 2. The breakdown point of the regression estimators (2) is equal to $(n-k)/n$ if $w_i \ge 0$ for $i = 1,2,\dots,n,\ w_k > 0$ for $k = \max\{i : w_i > 0\}$, the index k is within the bounds $(n+p+1)/2 \le k \le n-p-1,\ n \ge 3(p+1)$ and the data points $x_i \in R^p$ for $i = 1,\dots,n$ are in general position.

We shall remind that the observations $x_i \in R^p$ for $i = 1,\dots,n$ are in general position if any p of them are linearly independent. Geometrically this means that no $(p+1)$ observations lie in a subspace of dimension less than p.

The class of regression estimators (2) is closely related with a class of regression estimators proposed by Hössjer (1994), which contains the LTS regression estimator of Rousseeuw (1984) and his rank-based regression estimator defined by $w_i = a_n(i) = h^+(\frac{i}{n+1})$ and $\rho(|r|_{(i)}) = |r|_{(i)}$ for $i = 1, 2, \ldots, n$, where h^+ is a score generating real function mapping $(0, 1)$ into $[0, \infty)$ that satisfy the condition $\sup\{u : h^+(u) = \alpha\}$ with $0 < \alpha \leq 1$. Hössjer proved that the breakdown point of the proposed rank-based regression estimators is equal to $\min(n-k+1, k-p+1)/n$ if the index k satisfies the conditions $k \geq p$, $k = \max\{i : a_n(i) > 0\}$ and the observations $x_i \in R^p$ are in general position. He also showed the consistency and normality under various regularity conditions.

It is seen that the class of estimators (2) contains also: the Least Squares Estimators (LSE) if $\rho(|r|_{(i)}) = r^2_{(i)}$ and the Least Absolute Value Estimator (LAV) if $\rho(|r|_{(i)}) = |r|_{(i)}$ and $w_i \equiv 1$ for $i = 1, 2, \ldots, n$; the Chebishev minmax estimator if $\rho(|r|_{(n)}) = |r|_{(n)}$, $w_n = 1$ and $w_i = 0$ for $i = 1, 2, \ldots, n-1$; the LMS and LTS estimators of Rousseeuw (1984); the h-trimmed weighted L_q estimators of Müller (1995) if $\rho(|r|_{(i)}) = |r|^q_{(i)}$.

Coakley and Mili (1993) introduced a class of regression estimators called D-estimators. They characterized the breakdown point of D-estimators using the exact fit notion without the assumption that the observations are in general position and determined the optimal value of the quantile index of the LQS and LTS estimators in order to achieve the highest breakdown point.

Müller (1995) introduced a new definition of the breakdown point, the so called breakdown point for contamination free experimental condition in studying the breakdown point of the h–trimmed weighted L_q estimators and derived their breakdown point for situations with replications but without errors in the experimental conditions.

In conclusion, if $k = (n+p+1)/2$ we find the highest breakdown point that is derived by Rousseeuw and Leroy (1987) and Hössjer (1994) respectively about the LQS and LTS and rank-based regression estimators. The usefulness of d–fullness is evident: (i) The breakdown point can be exemplified by the range of values of k. This allows the statistician to choose the tuning parameter k according to the expected percentage of outliers in data. The corresponding estimator will possess a breakdown point less than the highest possible but it will be more efficient at the same time; (ii) The number of observations n in the linear regression setup must be at least $3(p+1)$; (iii) It is sufficient that only the observations $x_i \in R^p$ are in general position and not the pair (y_i, x_i^T) of observations for $i = 1, 2, \ldots, n$, as it is usually assumed.

4 Simultaneous estimation of the scale and parameters in the linear regression model with exponential q–th order distributions of the error

First we will propose a modification of the basic definition of Neykov and Neytchev (1990) concerning the robustified maximum likelihood. Let the observations $e_1, e_2, \ldots, e_n$ be generated by an arbitrary unimodal probability density function $\psi(e, \theta)$ with unknown vector parameter θ. Let the weights w_i for $i = 1, \ldots, n$ be the same as in the previous paragraph.
Definition 4. The Weighted Least Trimmed log likelihood Estimator (WLTE) of θ is defined

as

$$\mathrm{WLTE(k)}(e_1,\ldots,e_n) = \underset{\theta}{\mathrm{argmin}} \sum_{i=1}^{k} \{-w_i \log \psi(e_{\nu(i)},\theta)\},$$

where $\psi(e_{\nu(1)},\theta) \geq \psi(e_{\nu(2)},\theta) \geq \ldots \geq \psi(e_{\nu(n)},\theta)$ are the ordered density values and $\nu = (\nu(1),\ldots,\nu(n))$ is a permutation of the indices $1,\ldots,n$, which may depend on θ.

It is clear that the LME(k) estimator of Neykov and Neytchev (1990) is obtained if $w_i = 0$ for $i = 1,\ldots,k-1,k+1,\ldots,n$ and $w_k = 1$, and the LTE(k) is obtained if $w_i = 1$ for $i = 1,\ldots,k$ and $w_i = 0$ otherwise.

From the definition of the WLTE(k) estimator it follows that its minima are achieved over a subsample of size k, as the following equality may be shown to hold:

$$\min_{\theta} \sum_{i=1}^{k} \{-w_i \log \psi(e_{\nu(i)},\theta)\} = \min_{\tau} \min_{\theta} \sum_{i=1}^{k} \{-w_i \log \psi(e_{\tau(i)},\theta)\},$$

where $\tau = (\tau(1),\ldots,\tau(n))$ is any permutation of the indices $1,\ldots,n$.

Let us suppose that the errors ε_i of the regression model (1) are independent and identically distributed with q–th power exponential distribution, i.e. the density function of ε_i is given by

$$\phi(\varepsilon,\beta,\sigma) = \frac{q(1/2)^{(1+1/q)}}{\sigma\Gamma(1/2)} \exp\{-\frac{1}{2}|\frac{\varepsilon}{\sigma}|^q\},$$

where Γ is the gamma function. The Gaussian for $q = 2$, the Laplace for $q = 1$, the double exponential for $0 < q < 2$, the leptokurtic for $1 < q < 2$, the platikurtic for $q > 2$, the rectangular for $q \to \infty$ distributions are obtained as particular cases.

Now we are going to study the breakdown properties of the WLTE(k) regression estimators of β and error σ (when estimating simultaneously). We need to study the d–fullness of the set $f_i(\beta,\sigma),\ldots,f_n(\beta,\sigma)$, where

$$f_1(\beta,\sigma) = -\log(\phi(\varepsilon_i,\beta,\sigma)) = \frac{1}{2}|r_i/\sigma|^q + \log(\sigma) + C_1$$

and

$$C_1 = \log(\Gamma(1/2)) - \log(q(1/2)^{(1+1/q)}).$$

Theorem 3. The breakdown point of the WLTE(k) estimators of β and scale σ in the linear regression model with q–th order exponential distribution for any $q \in [1,\infty)$ is equal to $(n-k)/n$ if $n \geq 3(p+1)$, $x_i \in R^p$, $i = 1,\ldots,n$ are in general position, the weights $w_i \geq 0$ for $i = 1,2,\ldots,n$, $w_k > 0$ for $k = \max\{i : w_i > 0\}$, and the index k is within the bounds $(n+p+1)/2 \leq k \leq n-p-1$.

As far as we are acquainted with the literature concerning the breakdown point of the robust regression estimators there is no analogue of this Theorem.

If σ is known or treated as a nuisance parameter in $f_i(\beta,\sigma)$ the LME(k) (LTE(k)) estimators of β are equivalent to the LQS (LTS) regression estimators of Rousseeuw (1984).

It is easily seen also that some robust estimators such as Minimum Volume Ellipsoid and Minimum Covariance Determinant estimators of multivariate location and scatter introduced by Rousseeuw (1985) can be obtained if one chooses LME(k) or LTE(k) for the multivariate normal density function and if we specify the value of the parameter k; for details see Vandev and Neykov (1993).

5 The grouped binary regression

The classical logistic and probit analyses are widely used in applied statistics. For a recent study see McCullagh and Nelder (1989). Copas (1988), Künsch, Stefanski and Carroll (1989), Carroll and Pederson (1993) and Wang and Carroll (1995) considered different types of robust M- and GM- estimators depending on the type of contamination in the binary regression models. The breakdown point of these estimators is not greater than $1/(p+1)$, where p is the number of covariates. A high breakdown point estimator for binary regression data is considered by Christmann (1994). The proposed estimator is based on LQS regression estimator of Rousseeuw (1984), but applied on appropriately weighted data.

In this section the breakdown properties of the WLTE(k) estimators for a grouped binary regression model are studied. The type of the data under consideration has the form (y_i, x_i^T) for $i = 1, \ldots, m$ where y_i is assumed to be binomially distributed, $b(y_i \mid n_i, \pi_i)$, where the group size is n_i, the probability of success is π_i, x_i is a $p \times 1$–dimensional vector of covariates (explanatory variables) and the total number of observations is $n = n_1 + n_2 + \cdots + n_m$.

We shall assume that π_i follows the linear logistic regression model

$$\pi_i = \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta),$$

where β is a $p \times 1$–dimensional vector of unknown parameters. Probit or other link functions could be used as well.

Hereafter we shall also assume that $0 < y_i < n_i$ for each i. Without this condition the subcompactness of the respective functions is not possible to be proved. This corresponds with the results concerning the existence, uniqueness and finiteness of the parameter estimate of β discussed by Haberman (1977). The case of ungrouped binary (Bernoulli) regression data thus cannot be considered here.

We need to study the d–fullness of the set $\{f_1(\beta), \ldots, f_n(\beta)\}$, where

$$f_i(\beta) = f_i(y_i, x_i^T, \beta) = -\log(b(y_i \mid n_i, \pi_i)) = -y_i(x_i^T \beta) + n_i \log(1 + exp(x_i^T \beta)) - \log \binom{n_i}{y_i}.$$

Theorem 4. The breakdown point of the WLTE(k) estimators in the grouped binary logistic linear regression model defined above is equal to $(m-k)/m$ if the data $x_i \in R^p$ for $i = 1, \ldots, m$ are in general position, the weights $w_i \geq 0$ for $i = 1, 2, \ldots, m$, $w_k > 0$ for $k = \max\{i : w_i > 0\}$, $m \geq 3(p+1)$ and the index k is within the bounds $(m+p+1)/2 \leq k \leq m - p - 1$.

Remark 2. Note that the meaning of breakdown point here is different from the one given by Definition 1 as we consider the triple (n_i, y_i, x_i^T) as one observation.

The above results also hold in the case of a homoscedastic linear regression model with replications. It is easily shown, by analogy of the grouped binary regression model, that the breakdown point of these estimators is equal to $(m-k)/m$ when the parameter k belongs to the range of values $(m+p+1)/2 \leq k \leq m-p-1$, since the respective set of functions is $p+1$ full if the observations $x_i \in R^p$ are in general position.

The results in this section show the inadequacy of both the breakdown point definition of Donoho and Huber (1983) and the d–fullness technique of Vandev (1993) when applied directly to the structured models with replications. The main reason is that the parameter k of the considered estimators belongs to the interval $(m+p+1)/2 \leq k \leq m-p-1$ in this case. These values of k may be too small in comparison with the total number of observations

$n = n_1 + n_2 + \cdots + n_m$. This means that a large number of regular observations would not be used, which contradicts the main goal of the robustness: the estimator should be based on the majority of the data. In order to escape from the dependence of the breakdown point on the sample, Coakley and Mili (1993) use the exact fit point in studying the robustness properties of their D-estimators, and in particular the LMS and LTS estimators of Rousseeuw (19984), in the case of linear regression with replications.

6 Appendix

The proof of Theorem 2 is based on the next lemmas and Theorem 1. We shall study the d-fullness of the set $\{|r_1|^q, \ldots, |r_n|^q\}$ where $q \in [1, \infty)$.
Lemma 1. For any $p+1$ points $u_1, \ldots, u_{p+1} \in R^p$ in general position, and arbitrary constants $v_1, \ldots, v_{p+1}$, the function

$$I(\beta) = \max_{i \in 1, \ldots, p+1} |v_i - u_i^T \beta|^q$$

is subcompact in β when $q \in [1, \infty)$.
Proof of Lemma 1. Denote by V and U the vector of the constants v_i and the matrix of the points u_i^T for $i = 1, \ldots, p+1$ respectively. We shall use the inequality:

$$C \geq I(\beta) \geq \frac{1}{p+1} \sum_{i=1}^{p+1} |v_i - u_i^T \beta|^q.$$

Since any two norms in the Euclidean space R^p are equivalent it follows after simple transformations that

$$C^{1/q} \geq \left(\frac{1}{p+1} \sum_{i=1}^{p+1} |v_i - u_i^T \beta|^q \right)^{1/q} \geq M \left(\frac{1}{p+1} \sum_{i=1}^{p+1} (v_i - u_i^T \beta)^2 \right)^{1/2},$$

where M is a positive constant. Therefore the following inequality holds

$$\begin{aligned} (1/M)^2 C^{2/q} \geq \frac{1}{p+1} \sum_{i=1}^{p+1} (v_i - u_i^T \beta)^2 &= \frac{1}{p+1} (V - U\beta)^T (V - U\beta) \\ &= \frac{1}{p+1} (\hat{r} - U(\beta - \hat{\beta})^T (\hat{r} - U(\beta - \hat{\beta}), \end{aligned}$$

where $\hat{\beta} = (U^T U)^{-1} U^T V$ is the LSE and $\hat{r}$ is the corresponding vector of residuals. The LSE $\hat{\beta}$ exists since the observations u_i are in general position. By $U^T \hat{r} = 0$ it follows that

$$(1/M)^2 C^{2/q} - \frac{\hat{r}^T \hat{r}}{p+1} \geq \frac{1}{p+1} (\beta - \hat{\beta})^T U^T U (\beta - \hat{\beta}).$$

Since $det(U^T U)$ is nonzero, because the observations u_i are in general position, it follows that β belongs to a bounded set, which is compact as the function $I(\beta)$ is continuous. Therefore the function $I(\beta)$ is subcompact. **Q.E.D**.
Lemma 2. For any $q \in [1, \infty)$ the set $\{|r_1|^q, \ldots, |r_n|^q\}$ is $(p+1)$–full if the points $x_i \in R^p$, $i = 1, \ldots, n$ are in general position.
Proof of Lemma 2. The proof follows directly from Lemma 1 and definition 3.

Proof of Theorem 2. From the inequalities

$$C \geq \max_{i\in\{1,\dots,p+1\}} \rho(|r_i(\beta)|) \geq \rho(\max_{i\in\{1,\dots,p+1\}} |r_i(\beta)|)$$

it follows that

$$\rho^{-1}(C) \geq \max_{i\in\{1,\dots,p+1\}} |r_i(\beta)|,$$

as the function ρ is strictly increasing and continuous. According to Lemma 2 the set of functions $\{\rho(|r_i(\beta)|)$ for $i = 1, 2, \dots, n\}$ is $(p+1)$–full since the points $x_i \in R^p$ for $i = 1, \dots, n$ are in general position. Therefore the breakdown point of this class of estimators is not less than $(n-k)/n$ according to Theorem 1.

The opposite inequality is proved easily. Let us increase the number of the contaminated observations by one, i.e., their number becomes $n-k+1$. Since the estimator is defined over a subsample of k observations, it follows that one of them is a contaminated one. So the estimator can obtain arbitrary large values, and therefore its breakdown point is less than $(n-k)/n$. The equality follows from both inequalities. **Q.E.D.**

The proof of Theorem 3 is based on the next lemmas and Theorem 1.

Lemma 3. For any $p+1$ points $u_1, \dots, u_{p+1} \in R^p$ in general position, and arbitrary constants $v_1, \dots, v_{p+1}$, the function

$$I(\beta, \sigma) = \max_{i\in\{1,\dots,p+1\}} f_i(\beta, \sigma) = \max_{i\in\{1,\dots,p+1\}} \frac{1}{2}|v_i - u_i^T\beta|^q + \log(\sigma) + C_1$$

is subcompact in β and σ when $q \in [1, \infty)$.

Proof of Lemma 3. Denote by V and U the vector of the constants v_i and the matrix of the points u_i^T for $i = 1, \dots, p+1$ correspondingly. Denote by A $= \max_i |v_i - u_i^T\beta|$ for $i \in \{1, \dots, p+1\}$. Note that A is not zero since the points u_i for $i = 1, \dots, p+1$ are in general position. The function $I(\beta, \sigma)$ for fixed β has a local minimum at $\tilde{\sigma} = A(q/2)^{1/q}$. We will use the following inequalities:

$$C \geq 2I(\beta, \sigma) \geq 2I(\beta, \tilde{\sigma}) = A^q/\tilde{\sigma^q} + 2\log(\tilde{\sigma}) + 2C_1 = 2/q + (2/q)\log(q/2) + 2\log A + 2C_1.$$

Denote $C_2 = C - 2C_1 - 2/q - (2/q)\log(q/2)$. From the inequalities

$$\exp(C_2) \geq A^2 = \left(\max_{i\in\{1,\dots,p+1\}} |v_i - u_i^T\beta|\right)^2 = \max_{i\in\{1,\dots,p+1\}} (v_i - u_i^T\beta)^2 \geq \frac{1}{p+1}\sum_{i=1}^{p+1}(v_i - u_i^T\beta)^2$$

and by analogy with the proof of Lemma 1 it follows that β belongs to a compact set.

Let $\tilde{\beta}$ is the minimal value of the function A $= \max_i |v_i - u_i^T\beta|$ for $i \in \{1, \dots, p+1\}$. Then the inequalities hold

$$C \geq 2I(\beta, \sigma) \geq 2I(\tilde{\beta}, \sigma) \geq \tilde{A}^q/\sigma^q + 2\log(\sigma) + 2C_1.$$

By simple transformation after addition and subtraction of the term $2\log(\tilde{A})$ on the right side of the last inequality we obtain that

$$C \geq \tilde{A}^q/\sigma^q - (2/q)\log(\tilde{A}^q/\sigma^q) + 2\log(\tilde{A}) + 2C_1.$$

Denote $H = C - 2C_1 - 2log(\tilde{A})$. Then

$$H \geq \ \tilde{A}^q/\sigma^q - (2/q)\log(\tilde{A}^q/\sigma^q).$$

Let us consider the function $s(\sigma) = \tilde{A}^q/\sigma^q - (2/q)\log(\tilde{A}^q/\sigma^q)$. It reaches its minimum at $\hat{\sigma} = \tilde{A}(q/2)^{1/q}$, and the corresponding minimum equals $\hat{\sigma} = 2/q - 2/q\log(2/q)$, which is nonnegative when $1 \leq q < \infty$. Therefore for any q in the interval $1 \leq q < \infty$

$$H \geq 2/q - (2/q)\log(2/q) > 0,$$

which implies

$$e^{-H} < \tilde{A}^q/\sigma^q \leq \frac{Hq}{qe-2}.$$

The left inequality is obvious. The right one follows easily from the inequality

$$\tilde{A}^q/\sigma^q \leq H + (2/q)\log(\tilde{A}^q/\sigma^q) \leq H + (2/q)(\tilde{A}^q/\sigma^q)/e,$$

as $\log(x) \leq x/e$ for any $x > 0$. By the definition of A and the last inequality it follows that σ (which is positive) also belongs to a compact set. Therefore $I(\beta, \sigma)$ is subcompact in β and σ. **Q.E.D.**

Lemma 4. For any $q \in [1, \infty)$ the set of functions $\{f_1(\beta, \sigma), \ldots, f_n(\beta, \sigma)\}$ is $(p+1)$–full if the points $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position.

Proof of Lemma 4. According to Lemma 3 the function $I(\beta, \sigma)$ is subcompact in β and σ for any $(p+1)$ points, since the points $x_i \in R^p$ for $i = 1, \ldots, n$ are in general position. Therefore the set of functions $\{f_1(\beta, \sigma), \ldots, f_n(\beta, \sigma)\}$ is $(p+1)$–full. **Q.E.D.**

Proof of Theorem 3. According to Theorem 1 it follows that the breakdown point of the WLTE(k) estimator is not less than $(n-k)/n$ since the set $\{f_1(\beta, \sigma), \ldots, f_n(\beta, \sigma)\}$ is $(p+1)$–full by Lemma 4. The opposite inequality is proved in the same way as in the second part of the proof of Theorem 2. **Q.E.D.**

Remark 3. A large constant must be added to $f_i(\beta, \sigma)$ for $i = 1, \ldots, n$ in order to ensure their nonnegativeness. Clearly this constant will not affect the estimator but only the value of the corresponding loss function. It is seen that the choice of this constant depends on the value of σ which should be bound from below by a small positive constant say δ. Therefore the WLTE(k) estimate of σ must be obtained under the constraint $\sigma > \delta$, which can be done as the observations are in general position. As the terms $|(y_i - x_i^T\beta)/\sigma|^q$ are nonnegative for any i such a constant $C_\delta = \max_i C_{\delta i}$ can be determined for a fixed δ from the inequalities

$$C_{\delta i} > \log\left(q(1/2)^{(1+1/q)}\right) - \log(\Gamma(1/q)) - \log(\delta) - |(y_i - x_i^T\beta)/\delta|^q, \ i = 1, \ldots, n.$$

The proof of Theorem 4 is based on the following lemmas and Theorem 1.

Lemma 5. The function

$$I(\beta) = \sup_{i \in \{1, \ldots, p+1\}} \left\{ -y_i(u_i^T\beta) + n_i\log(1 + exp(u_i^T\beta)) - \log\binom{n_i}{k_i} \right\}$$

is subcompact in β if $0 < y_i < n_i$ for $i = 1, \ldots, p+1$ and the points $u_1, \ldots, u_{p+1} \in R^p$ are in general position.

Proof of Lemma 5. In order to show that the set $\{\beta : I(\beta) \leq C\}$ is compact for any constant C we shall use the inequality

$$C \geq I(\beta) = \max_{i \in \{1, \ldots, p+1\}} \left\{ -y_i(u_i^T\beta) + n_i\log(1 + exp(u_i^T\beta)) - \log\binom{n_i}{k_i} \right\}.$$

Consider the function $g(\lambda) = -y\lambda + n\log(1+exp(\lambda))$ under the restriction $0 < y < n$. It reaches its minimum at $\tilde{\lambda} = \log(\frac{y/n}{1-y/n})$. It is seen that $g(\lambda) \to +\infty$ when $\lambda \to \pm\infty$ and

$$g(\lambda) \geq \min(y, n-y)\{|\lambda - \tilde{\lambda}| - |\tilde{\lambda}|\} \geq |\lambda - \tilde{\lambda}| - |\tilde{\lambda}|$$

.

Denote by $\lambda_i = u_i^T\beta$ and $C_3 = \max_{i \in 1,\ldots,p+1} \log\binom{n_i}{k_i}$. Then from the properties of the function $g(\lambda_i)$ it follows that

$$\begin{aligned} C + C_3 &\geq \max_{i\in\{1,\ldots,p+1\}} \left\{ -y_i(u_i^T\beta) + n_i \log(1+exp(u_i^T\beta)) - \log\binom{n_i}{k_i} \right\} \\ &= \max_{i\in\{1,\ldots,p+1\}} g(\lambda_i) \geq \max_{i\in\{1,\ldots,p+1\}} \{|\lambda_i - \tilde{\lambda}_i| - |\tilde{\lambda}_i|\}, \end{aligned}$$

which is equivalent to

$$C + C_3 + \max_{i\in\{1,\ldots,p+1\}} |\tilde{\lambda}_i| \geq \max_{i\in\{1,\ldots,p+1\}} |\lambda_i - \tilde{\lambda}_i| \geq$$

$$\frac{1}{p+1}\sum_{i=1}^{p+1} |\lambda_i - \tilde{\lambda}_i| = \frac{1}{p+1}\sum_{i=1}^{p+1} |u_i^T\beta - \log(\frac{y_i/n_i}{1-y_i/n_i})|.$$

By analogy with the proof of Lemma 1 it follows that β belongs to a compact set. **Q.E.D.**
Lemma 6. The set of functions $\{f_1(\beta),\ldots,f_m(\beta)\}$ is $(p+1)$– full if the points $x_i \in R^p$ for $i = 1,\ldots,m$ are in general position and $0 < y_i < n_i$.
Proof of Lemma 6. As any p+1 points of the sample are in general position, the function $I(\beta)$ is subcompact in β by Lemma 5. Therefore the set $\{f_1(\beta),\ldots,f_m(\beta)\}$ is $(p+1)$–full. **Q.E.D.**
Proof of Theorem 4. According to Theorem 1 the breakdown point of WLTE(k) estimators is not less than $(m-k)/m$ since the set of functions $\{f_1(\beta),\ldots,f_m(\beta)\}$ is $(p+1)$–full according Lemma 6 while the opposite inequality is proved in the same way as in the second part of the proof of Theorem 2. **Q.E.D.**

Acknowledgements We are grateful to Prof. Peter J. Rousseeuw for his continuing encouragement. We would also like to thank the referees for helpful comments on an earlier version of the paper. Part of this work was funded by contracts from the Ministry of Education, Science and Technologies of Republic of Bulgaria (M-440/94 and NZ-627/96).

References

[1] Carroll, R. J. and Pederson, S. (1993). On Robustness in the Logistic Regression Model. *J. R. Statist. Soc.* B, 55, pp. 693-706.

[2] Coakley, C. W. and Mili, L. (1993). Exact fit points under simple regression with reeplication. *Statistics and Probability Letters*, 17, pp. 265-271.

[3] Copas, J. B. (1988). Binary Regression Models for Contaminated Data (with Discussion). *J. R. Statist. Soc.* B, 50, pp. 225-265.

[4] Christmann, A. (1994). Least Median of Weighted Squares in Logistic Regression with Large Strata, *Biometrika*, 81, pp. 413-417.

[5] Donoho, D. L. and Huber, P. J. (1983). The Notion of Breakdown Point, In: *A Festschrift for Eric Lehmann*, P.J.Bickel, K.A.Doksum, and J.L.Hodges, (eds.), (Belmont, CA: Wadsworth), pp. 157-184.

[6] Haberman, S. J. (1977). Maximum Likelihood estimates in exponential response models. *Ann. Statist.* 5, pp. 815-841.

[7] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on influence Functions.* (John Wiley & Sons, New York).

[8] Huber, P. (1981). Robust Statistics, (John Wiley & Sons, New York).

[9] Hössjer, O. (1994). Rank-Based Estimates in the Linear Model with High Breakdown Point. *J. Amer. Statist. Assoc*, 89, pp. 149-158.

[10] Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989). Conditionally Unbiased Bounded Influence Estimation in General Regression Models, with Applications to Generalized Linear Models. *J. Amer. Statist. Assoc.*, 84, pp. 460-466.

[11] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* (Chapman & Hall, London).

[12] Müller, C. H. (1995). Breakdown Point for Designed Experiments. *J. Statist. Plan. Inference*, 45, pp. 413-427.

[13] Neykov, N. M. and Neytchev, P. N. (1990). A Robust Alternative of the ML Estimators. *COMPSTAT'90, Short communications*, Dubrovnik, Yugoslavia, pp. 99-100.

[14] Rousseeuw, P. J. (1984). Least Median of Squares Regression. *J. Amer. Statist. Assoc.*, 79, pp. 871-880.

[15] Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point, In: *Mathematical Statistics and Applications*, vol. B, W. Grossman, G. Pflug, I. Vincze and W. Wertz (eds.), Reidel, Dordrecht, pp. 283-297.

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

[17] Vandev, D. L. (1993). A Note on Breakdown Point of the Least Median of Squares and Least Trimmed Estimators. *Statistics and Probability Letters*, 16, pp. 117-119.

[18] Vandev, D. L. and Neykov, N. M. (1993). Robust Maximum Likelihood in the Gaussian Case. In: *New Directions in Statistical Data Analysis and Robustness*, S. Morgenthaler, E. Ronchetti, and W. A. Stahel (eds.), (Basel: Birkhauser Verlag) pp. 257-264.

[19] Wang,C.Y. and Carroll,R.J. (1995). On Robust Logistic Case-Control studies with Response-Dependent Weights. *J. Statist. Plan. Inference*, 43, pp. 331-340.