Robust Estimation with Discrete Explanatory Variables

Pavel Čížek*

Abstract

The least squares estimator is probably the most frequently used estimation method in regression analysis. Unfortunately, it is also quite sensitive to data contamination and model misspecification. Although there are several robust estimators designed for parametric regression models that can be used in place of least squares, these robust estimators cannot be easily applied to models containing binary and categorical explanatory variables. Therefore, I design a robust estimator that can be used for any linear regression model no matter what kind of explanatory variables the model contains. Additionally, I propose an adaptive procedure that maximizes the efficiency of the proposed estimator for a given data set while preserving its robustness.

JEL: C20, C24, C61

Keywords: discrete explanatory variables, linear regression, robust statistics, least trimmed squares

* Institut für Statistik und Ökonometrie, Wirtschaftwissenschaftliche Fakultät, Humboldt-Univerzität zu Berlin and Charles University, Center for Economic Research and Graduate Education, Prague

Abstrakt

Metoda nejmenších čtverců je jednou z nejčastěji používaných metod v regresní analýze. Její hlavní nevýhodou je velká citlivost vůči chybné specifikaci modelu a kontaminaci dat. Toto řeší tak zvané robustní odhady, kterých existuje pro lineární regresní model celá řada. Vetšinu robustních odhadů však není možno použít v případě, že model obsahuje diskretní proměnné. V tomto článku proto navrhuji robustní odhad, který je možno aplikovat v libovolném lineárním regresním modelu bez ohledu na charakter vysvětlujících proměnných. Nezanedbatelnou částí je i návrh a studium adaptivní metody, která maximalizuje účinnost odhadu při zachování jeho robustnosti.

Contents

List of Figures

List of Tables

1 Introduction

Estimation tasks that involve discrete dependent or discrete explanatory variables are quite natural in econometrics. The former is represented, for example, by any member of the wide class of discrete-response models. The latter is almost omnipresent in econometrics and occurs when we deal with various categorical variables that are used to represent non-continuous characteristics such as an individual's gender or education, or to characterize a general nonlinear relationship between regressors and the corresponding dependent variable. Thus, reliable and efficient estimation methods for models containing these kinds of variables are of considerable interest. In this paper, I concentrate on the second case, namely on the classical linear regression model with discrete explanatory variables.

Linear regression models are in most cases estimated using techniques based on the least squares principle. Although the least squares method is frequently used in regression analysis, mainly because of its simplicity and ease of use, it is quite sensitive to data contamination and model misspecification. Therefore, it is a bit surprising that some more reliable methods are not more widely spread, especially because it is not necessary to abandon a classical parametric model and its advantages in order to gain more robustness. The methods of robust statistics retain standard parametric assumptions but take into account possible misspecification and data contamination and their impact on estimation procedures in order to design misspecification- and data-contamination-proof estimators. For example, Orhan, Rousseeuw, and Zaman (2001) demonstrate the use of robust regression methods on three classical macroeconomics models estimated in the past by the least squares method. The main result is that the use of robust methods is highly recommended even in the case of a simple linear regression, because their use together with careful analysis of data sets lead to significantly different results than the least squares regression, at least in the case of the data sets analyzed by these authors.

On the other hand, although the asymptotic and robust properties of various robust estimators have been studied for several decades, at least in the case of regression with one explanatory variable, it is understandable from some points of view that robust estimation methods are not used more frequently in econometrics. There are several reasons for this and I will document them on the least trimmed squares (LTS) estimator (see Section 4.1 for more details), which was used by Orhan, Rousseeuw, and Zaman (2001). The first reason is computational: it is possible to compute LTS only approximately and even obtaining an approximation is relatively time consuming; moreover, a good approximation algorithm did not previously exist. However, the recent availability of a good and fast approximation algorithm (see, for example, Rousseeuw and Van Driessen (1999)), faster computers, and the presence of this algorithm in some widely-spread statistical packages[1] have made LTS more attractive.

The second reason is more troublesome: whereas discrete regressors do not cause any particular problems to standard estimation procedures (e.g., the least square or the maximum likelihood methods) if some regularity assumptions hold, the situation is completely different in the case of many robust regression methods. The main reason is that some robust methods completely reject a subset of observations. In other words, they completely ignore some observations and can consequently exclude a group of observations defined by categorical variables from regression estimation; this results in the problem of singular matrices, and consequently, some variables do not have to be identifiable. Given the significance

[1] For example, R, S-plus, TSP, and XploRe include procedures for the computation of LTS.

of discrete and categorical explanatory variables in econometric practice, this is a serious shortcoming that was already addressed by Hubert and Rousseeuw (1997), for instance. Nevertheless, the existing remedies do not represent an optimal solution—above all because they are limited only to a certain class of models (see Section 3)—and that is why I present here a new solution to this problem.

I essentially take the LTS estimator as the starting point and create a smoothed version of this estimator, removing thus the complete rejection of observations, the main cause of the problem. As we see later, this solution adds some further improvements to the LTS estimator, such as an increase of efficiency while preserving the robustness of LTS. The extent to which efficiency is improved and robustness is decreased depends heavily on the smoothing scheme used. Thus, I define first the smoothed LTS estimator in a general way and study its properties for a general smoothing scheme. Later, I propose a class of smoothing schemes and a rule that allows us, for a given data set, to adaptively find a smoothing scheme that maximizes the efficiency of the estimator while preserving its robustness properties. This is achieved by searching for an optimal choice among smoothing schemes defining smoothed LTS estimators ranging from the least trimmed squares ("most robust" option) to the least squares ("most efficient" option). Thus, given a data set, I try to come as close as possible to the least squares estimator without losing robustness of LTS, that is, without letting data anomalies significantly affect the estimate.

In the rest of this paper, I first describe basic concepts of robust statistics (Section 2). Later, I review the existing attempts at robust estimation in the presence of discrete and categorical explanatory variables (Section 3) and propose a smoothed version of the least trimmed squares estimator (Section 4). Next, the proofs of consistency and asymptotic normality are presented together with some elementary assertions that underlie one scheme for an adaptive choice of smoothing parameters (Section 5). Finally, the features of the proposed estimator are documented using Monte Carlo simulations (Section 6).

2 Robust statistics

Robust statistics aims to study the behavior of parametric estimators under deviations from the standard assumptions of parametric models and to develop estimators that behave well not only under correct parametric specification, but also in the presence of "small" deviations from the parametric assumptions. In other words, robust estimation methods are designed so that they are not easily endangered by the contamination of data. As a result, a subsequent analysis of regression residuals coming from a robust regression fit can hint at outlying observations. In addition, the use of a parametric model contributes efficiency, while features of these estimators ensure sufficient robustness. There are two main approaches to the formalization of robust statistics, namely Huber's *minimax approach* (Huber (1964), Huber (1981)) and Hampel's infinitesimal approach based on the *influence function* (Hampel et al. (1986)). Because of the advantages of the latter (see, for example, Hampel et al. (1986) and Peracchi (1990))[2], a more detailed description of robust statistics in the next section follows Hampel's approach.

2.1 Principles of robust statistics

Robust statistics formalizes certain desirable requirements for the behavior of various statistical procedures under deviations from parametric assumptions. This

[2] Most importantly, Hampel's approach can be generalized to any parametric model, while Huber's minimax strategy cannot.

is an important topic since the assumptions of parametric models are valid only "approximately" in many situations.[3] There are several reasons for this "approximate validity" of parametric models, which are summarized and exemplified in Hampel et al. (1986):

1. Nearly any data set contains some amount of gross errors, i.e., infrequent observations that are for some reason "wrong" (e.g., because of copying or computational errors; alternatively, these errors can be caused by an incorrect model specification or by a transitory phenomenon that affects only a few observations in the data set).

2. Frequently, data exhibit "small" deviations from the assumed parametric distribution, caused, for example, by rounding (data have always a limited accuracy), or by an approximate validity of asymptotic properties in finite samples.

3. The assumption of independence among observations is not "completely" satisfied.[4]

Unfortunately, these facts are usually not reflected by standard parametric methods, which is definitely one of the reasons for their sensitivity to even mild deviations from the model assumptions (see, for example, Tukey (1960)).[5] Therefore, the theory of robustness and robust procedures, that is, procedures that take into

[3] It is quite possible that observations in a given data sample do not precisely follow a given parametric specification, but, on the other hand, the behavior of the majority of observations is relatively close to this parametric specification. Therefore, it is often not necessary to completely reject the classic parameterization of a given model, even in case of nonlinear models (see Gerfin (1996)).

[4] See Hampel et al. (1986) for a discussion of the meaning of "wrong," "small," and "completely."

[5] Here, the sensitivity of classic parametric estimators does not cover only a possible inconsistency, but also a possible efficiency loss.

account not only standard parametric specifications but also possible deviations from them, has been developed. The main goals of the theory of robustness, as summarized in Hampel et al. (1986, page 11), are:

1. To describe the structure best fitting the bulk of the data.

2. To identify deviating data points (outliers) or deviating substructures for further treatment, if desired.

3. To identify and give a warning about highly influential data points ("leverage points").

4. To deal with unsuspected serial correlations, or more generally, with deviations from the assumed correlation structures.

A description of the main concepts used to formalize the above mentioned goals follows in Subsection 2.2. Examples of robust procedures are presented in Subsection 2.3.

2.2 Main concepts

Hampel et al. (1986) formalizes the aims of robust statistics by specifying a local measure of robustness—the *influence function*—and a global measure of robustness—the *breakdown point*. The influence function characterizes the sensitivity of an estimator T to infinitesimal contamination placed at a given point $x \in \mathbb{R}^p$: it is defined as a derivative of the estimator T taken as a functional on the space of distribution functions in the direction of x.[6] For example, one finite-sample measure, the sensitivity curve introduced by Tukey (1977), which

[6] A single point $x \in \mathbb{R}^p$ corresponds in the space of distribution functions to a degenerated distribution function.

in most cases converges to the asymptotically defined influence function, can be expressed as

$$SC_n(x) = n \cdot (T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1}))$$

for an estimator T_n evaluated at sample $x_1, \ldots, x_{n-1}$. There are also several other measures of robustness derived from the concept of the influence function, for example, the sensitivity to gross-errors, defined as the supremum of the influence function over all points $x \in \mathbb{R}^p$.

On the other hand, the global measure of robustness—the breakdown point—indicates how much contamination can make an estimate completely "useless".[7] This can be again illustrated using a finite-sample definition of the breakdown point for an estimator T_n at a sample $x_1, \ldots, x_n$ (Hampel et al. (1986)):

$$\varepsilon_n^* = \frac{1}{n} \max \left\{ m \,\middle|\, \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T_n(z_1, \ldots, z_n)| < +\infty \right\}, \tag{1}$$

where sample $z_1, \ldots, z_n$ is created from the original sample $x_1, \ldots, x_n$ by replacing observations $x_{i_1}, \ldots, x_{i_m}$ by values $y_1, \ldots, y_m$. The breakdown point usually does not depend on the sample $x_1, \ldots, x_n$. To give an example, it immediately follows from the definition that the finite-sample breakdown point of the arithmetic mean equals 0 in a one-dimensional location model, while for the median it is equal to 1/2. Actually, the breakdown point equal to 1/2 is the highest one that can be achieved at all; if the amount of contamination is higher, it is not possible to decide which part of the data is the correct one. Such a result is proven, for example, in Rousseeuw and Leroy (1987, Theorem 4, Chapter 3) for the case of

[7] For example, how much contamination can make the Euclidean norm of a given estimator higher than any given real constant.

regression equivariant estimators (the upper bound on ε_n^* is actually $([(n-p)/2]+1)/n$ in this case, where $[x]$ denotes the integer part of x).

These two concepts are of a different nature. The influence function, which is defined as a derivative of an estimator, characterizes the behavior of the estimator in a neighborhood of a given parametric model, in which the effect of contamination can be approximated by a linear function. On the contrary, the breakdown point specifies how far from the parametric model the estimator is still useful, in the sense that it produces usable results. In other words, while the influence function provides mainly an asymptotic tool that allows us to characterize and design, in some sense, asymptotically "optimal" estimators that exhibit certain robustness properties,[8] the breakdown point determines the robustness of the same estimators with respect to outliers and other deviations from the parametric model both asymptotically and when they are applied to real data.[9] As some kind of asymptotic optimality (e.g., asymptotic efficiency) of an estimator might be worthless if the robustness of the estimator is not high enough, a sufficiently high breakdown point is an important property of the estimator. Thus, the influence function and the breakdown point can be viewed as complementary characteristics.

2.3 Examples of robust estimators

The theory of robustness offers two main approaches for developing new estimators. First, estimators can be designed primarily to achieve a maximum possible

[8] The reason is that the influence function of an estimator does not characterize only one kind of robustness of the estimator, but is also related to the asymptotic variance of the estimator, see Hampel et al. (1986).

[9] Usually, the breakdown point ε_n^* is "quite close" to the limit $\lim_{n\to\infty}\varepsilon_n^*$ for any $n \in N$; for example, estimators that achieve the upper bound $([(n-p)/2]+1)/n$ have their breakdown point "quite close" to $1/2$.

breakdown point (which in general cannot be higher than 1/2). Second, the theory of robustness also supplies some "optimality" criteria that allow the construction of most efficient estimators with certain robustness qualities. One of them is *B-robustness*, which is defined as a finite sensitivity to gross errors. An optimally B-robust estimator is then defined as the most efficient estimator (according to its asymptotic variance) that has its sensitivity to gross errors lower or equal to some fixed upper bound. Examples of robust estimators of both classes, relevant for the presented research, are mentioned in this subsection.

Currently, there are many procedures with breakdown points close to 1/2, most of which are designed for the linear regression model. These high breakdown point estimators serve several purposes: (1) a reliable estimation of unknown parameters, which is possible because of their high breakdown point; (2) detection of outliers and leverage points (using the analysis of the residuals) so that they can be used as diagnostic tools; (3) a robust initial estimate for iterative estimation procedures. Examples of existing techniques designed for the linear regression model are the *least median of squares* (Rousseeuw (1984)), the *least trimmed squares* (Rousseeuw (1985)), and the *S-estimators* (Rousseeuw and Yohai (1984)). Recently, the least trimmed squares estimator became more preferred to the least median of squares because it features better asymptotic performance and a fast and reliable approximation algorithm (Rousseeuw and Van Driessen (1999)). All these estimators can withstand a high amount of contamination including *outliers* (observations that are distant in the direction of the dependent variable) and *leverage points* (observations outlying in the space of explanatory variables).[10] Unfortunately, they all have inherent problems with estimation which includes both continuous and categorical variables. Existing robust methods designed for

[10] If the meaning of terms "outliers" and "leverage point" are not intuitive or apparent enough, check, for example, the classification of outlying points in Rousseeuw (1997).

the estimation of such models are discussed in Section 3.

On the other hand, there is already a class of robust estimators that can be applied in models with categorical variables, which has been studied from the standpoint of the optimal B-robustness concept: the *M-estimators* (generalized maximum likelihood estimators). The main idea is to replace the maximum likelihood scores (or their integrals) with a general function of observed data points and unknown parameters. By choosing a suitable function, optimally B-robust M-estimators were developed in a rather general setup that covers many linear and nonlinear regression models. Unfortunately, they have an important disadvantage once they are applied in the multiple regression framework—a low breakdown point,[11] which cannot exceed $1/p$, where p is the dimension of the corresponding parameter space (Maronna, Bustos, and Yohai (1979)); moreover, they are often quite sensitive to leverage points. Thus, the global robustness of M-estimators becomes rather low when there are more parameters involved. Finally, M-estimators are usually not invariant with respect to scale (residuals have to be studentized in most cases), and hence, they are rather sensitive to its initial estimate. Altogether, this means that in most cases it is necessary to combine them with an estimator with a high breakdown point.

3 Existing approaches to robust estimation with discrete explanatory variables

There are several estimators that are robust in some way and can cope with discrete and categorical variables. The most obvious one is the least absolute

[11] Low compared to the breakdown point of the estimators discussed in the previous paragraph. Note that the breakdown point of MLE under the assumption of normally distributed errors is equal to zero.

deviation (L_1) estimator. However, it is not directly comparable with the high breakdown-point estimators discussed in Section 2.3, because, despite being resistant to outliers, it is not robust against leverage points. Therefore, new high breakdown point estimators for linear regression model with binary and categorical variables were designed—first for the special case of distributed intercept (see Hubert and Rousseeuw (1996)), later for a linear regression model with continuous and binary variables, where binary variables enter the regression equation only additively (Hubert and Rousseeuw (1997)). The best (from the viewpoint of robustness and the speed of convergence) from several proposed estimators is the so-called RDL_1 estimator (Robust Distance and L_1 regression). RDL_1 is a three stage procedure:

1. The *minimum volume ellipsoid* (MVE) estimator (Rousseeuw (1985)) of location and scatter matrix is applied on the set of all continuous explanatory variables, and based on it, robust distances are computed.

2. Using the robust distances, strictly positive weights w_i are defined in such a way that observations having a large distance from the center of data are down-weighted (distances are computed only in the space of continuous variables, because all categorical variables are encoded as dummy variables, which cannot be outlying by their nature). Then regression parameters are estimated by a weighted L_1 procedure with the constructed weights w_i.

3. The scale of residuals is estimated by the median absolute deviation (MAD) estimator applied on the vector of residuals coming from the L_1 regression in point 2.

This estimator achieves a high breakdown point, because the influence of leverage points is reduced by weights that are indirectly proportional to the robust distances of these points and the robustness against outliers is obtained by using the L_1 estimation method. On the other hand, the procedure has several disadvantages. One of them is the lack of efficiency in most usual cases caused by the use of the L_1 estimator; as a possible remedy, Hubert and Rousseeuw (1997) propose a four stage procedure that adds as the fourth step computing a weighted least squares estimator with weights based on studentized residuals from RDL_1 estimator. Another disadvantage is that this estimator, which is defined for linear regression models with dummy variables entering a model only additively, can hardly be generalized to more complicated models: for example, to general regression models with dummy and categorical variables (including cross-effects); to instrumental variable and similar models, for which results concerning least-squares-like estimators are readily available, but often missing for other types of estimators; or to nonlinear models, in which it is hard to predict the effect of large values of different variables, and thus, a simple down-weighting proportional to distances in space of explanatory variables does not make sense. Finally, RDL_1 can be relatively easily influenced by misspecification occurring in dummy and categorical variables simply because it does not treat dummy variables in any special way (this is documented in Section 7). Such an effect is naturally bounded so it does not affect the breakdown point as defined by (1), but it suffices to make the estimator inconsistent.

4 Smoothed least trimmed squares

Robust estimation of linear regression models with discrete and categorical explanatory variables has received some attention recently, but there is still vast area for improvement, as discussed in Section 3. In addition, the least trimmed squares estimator has been gaining more popularity because of its robustness and relatively high efficiency, but also there is a need for improvement, as I discuss below. Therefore, I define a smoothed version of the least trimmed squares estimator that should preserve the robustness of LTS, and at the same time, allow the estimation of general linear regression models with discrete explanatory variables and obtain better properties than the existing robust estimators in the area of efficiency. In this section, the smoothed LTS estimator is defined for a general smoothing scheme. An adaptive choice of smoothing, which should enable us to obtain as high efficiency as possible while preserving the robustness of the estimator, is discussed in more detail in Sections 6.2 and 7.

I first define the linear regression model used throughout this paper and describe the least trimmed squares estimator (LTS) introduced by Rousseeuw (1985) in Section 4.1. In Section 4.2 I define the smoothed version of LTS. Finally, I discuss the relation between the smoothed LTS and weighted least squares estimators in Section 4.3.

4.1 Linear regression model and least trimmed squares

LTS is a statistical technique for estimation of the unknown parameters of a linear regression model and provides a robust alternative to the classical regression methods based on minimizing the sum of squared residuals. Let us consider a linear regression model for a sample (y_i, x_i) with a response variable $y_i \in \mathbb{R}$ and a vector of explanatory variables $x_i \in \mathbb{R}^p$:

$$y_i = x_i^T \beta + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

The least trimmed squares estimator $\hat{\beta}_n^{(LTS)}$ is defined as

$$\hat{\beta}_n^{(LTS)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} r_{[i]}^2(\beta), \tag{3}$$

where $r_{[i]}^2(\beta)$ represents the ith order statistics of squared residuals $r_1^2(\beta), \ldots, r_n^2(\beta)$; $r_i(\beta) = y_i - x_i^T \beta$ and $\beta \in \mathbb{R}^p$ (p denotes the number of estimated parameters). The *trimming constant* h has to satisfy $\frac{n}{2} < h \leq n$. This constant determines the breakdown point of the LTS estimator since definition (3) implies that $n-h$ observations with the largest residuals do not affect the estimator (except for the fact that the squared residuals of excluded points have to be larger than the hth order statistics of the squared residuals). The maximum breakdown point is attained for $h = [n/2] + [(p+1)/2]$ (see Rousseeuw and Leroy (1987, Theorem 6)), whereas for $h = n$, which corresponds to the least squares estimator, the breakdown point is equal to 0. There is, of course, a trade-off: lower values of h, which are close to the optimal breakdown-point choice, lead to a higher breakdown point, while higher values of h improve efficiency (if the data are not too contaminated) since more (presumably correct) information in the data is utilized. The most robust choice of h is often employed when the LTS is used for diagnostic purposes. It may also be favored when LTS is used for a comparison with some less robust estimator, e.g., least squares, because a comparison of these two estimators can serve as a simple check of data and a model—if the estimates are not similar to each other, special care should be taken throughout the analysis. On the other hand, it may be sensible to evaluate LTS for a wide range of trimming-constant values and to observe how the estimate behaves with increasing h because this dependence can provide hints about the amount of contamination and possibly about suspicious structures in the studied data (for example, that the data actually contain a mixture of two different populations, see Benáček, Jarolím, and Víšek (1998)).

Before proceeding further, I will discuss the existence of the estimator and its statistical properties. First, the existence of the optimum in (3) under some reasonable assumptions can be justified in the following way: the minimization of the objective function in (3) can be viewed as a process in which we choose a subsample of h observations and find some β minimizing the sum of squared residuals for the selected subsample every time. Doing this for every subsample (there are $\binom{n}{h}$ of them), we get $\binom{n}{h}$ candidates for the LTS estimate and the one that commands the smallest value of the objective function is the final estimate. Therefore, the existence of the LTS estimator is basically equivalent to the existence of the least squares estimator for all subsamples of size h.

Finally, I discuss various finite-sample and asymptotic properties of LTS as well as some drawbacks related to the use of LTS. First, the least trimmed squares is regression, scale, and affine equivariant[12] (see, for example, Rousseeuw and Leroy (1987, Lemma 3, Chapter 3)). I also already remarked that the breakdown point of LTS reaches the upper bound $([(n-p)/2]+1)/n$ for regression equivariant estimators if the trimming constant h is equal to $[n/2]+[(p+1)/2]$ (p represents the number of unknown parameters). Furthermore, the $\sqrt{n}$-consistency and asymptotic normality of LTS can be proven for a general linear regression model with continuously distributed disturbances, see Víšek (1999a). Besides these important statistical properties, there are also some less practical aspects. One of them is the above mentioned complete rejection of observations. If discrete variables are present in the regression model, complete rejection can lead

[12] An estimator T as a function of data is equivariant with respect to a family of transformations F if and only if the estimator applied on any transformed data set is equal to the transformation of the estimator applied on the original data set: $(\forall f \in F)(\forall x_1,\ldots,x_n)(T(f(x_1),\ldots,f(x_n)) = f(T(x_1,\ldots,x_n)))$.

to a situation where the parameters of interest are not identified in a subsample of h observations included in the LTS objective function. Another disadvantage directly follows from the discontinuity[13] of the LTS objective function. Because of this, the sensitivity of the least trimmed squares estimator to a change in one or several observations might sometimes be rather high (Víšek (1999b)). This property, often referred as high subsample sensitivity, follows from the fact that high breakdown-point estimators search for a "core" subset of data that best fits a certain model (with all its assumptions) without taking into account the rest of the observations: a change in some observations may then lead to a large swing in the composition of this core subset. These problematic aspects of LTS will be fixed by the smoothed version of LTS proposed in the next section.

4.2 Definition of smoothed least trimmed squares

In this section, I define the *smoothed least trimmed squares* (SLTS) estimator. Let us consider a linear regression model (2) for a sample $(y_i, x_i), i = 1, \ldots, n$. Moreover, let $w = (w_1, \ldots, w_n)$ be a vector of weights such that $w_1 \geq w_2 \geq \ldots \geq w_n \geq 0$. Then the *smoothed least squares estimator* $\hat{\beta}_n^{(SLTS,w)}$ is defined by

$$\hat{\beta}_n^{(SLTS,w)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} w_i r_{[i]}^2(\beta), \tag{4}$$

where

- $\beta \in B \subseteq \mathbb{R}^p$ is a p-dimensional vector of unknown parameters and $B \subseteq \mathbb{R}^p$ is the corresponding parameter space,

- $r_{[i]}^2(\beta)$, $i = 1, \ldots, n$, represent the ordered sample of squared residuals

[13] The discontinuity of LTS refers here to the fact that residuals either enter the objective function or not at all.

$r_i^2(\beta) = (y_i - x_i^T \beta)^2$ for any $\beta \in B$, and

- w is a *weighting vector*: $w_1 \geq w_2 \geq \ldots \geq w_n \geq 0$.

The estimator is quite similar to the weighted least squares (WLS) estimator with one important difference: weights are assigned to the order statistics of squared residuals instead directly to the residual. Clearly, the behavior and properties of the SLTS estimator are given entirely by the choice of weights. Let me provide two simple and one complex examples:

1. $w_1 = \ldots = w_n = 1$: SLTS is equivalent to the least squares estimator;

2. $w_1 = \ldots = w_h = n/h$ for $\frac{n}{2} < h \leq n$ and $w_{h+1} = \ldots = w_n = 0$: SLTS is equivalent to the least trimmed squares estimator;

3. $w_i = f(\frac{i}{n}; \omega_1, \ldots, \omega_m)$ for all $i = 1, \ldots, n$, where $f(x; \omega_1, \ldots, \omega_m)$ is a real-valued function on $\langle 0, 1 \rangle$ parameterized by $\omega_1, \ldots, \omega_m \in \mathbb{R}^m$: in this case, weights follow a function $f(x; \omega_1, \ldots, \omega_m)$ and are actually given by the parameters $\omega_1, \ldots, \omega_m$. For example, such a function can be defined as

$$f(x; \omega) = \frac{1}{1 + e^{\omega(x - 1/2)}}$$

for all $x \in \langle 0, 1 \rangle$ and one parameter $\omega \in \langle 0, \infty)$. Then we have a smoothing scheme $w_i = f(\frac{i}{n}, \omega)$ for any given, but fixed value of ω, and moreover, we can choose among such smoothing schemes by selecting a suitable value of parameter ω. Note that this smoothing scheme converges to the one introduced in point 1 (least-squares weights) for $\omega \to 0$ (as then $w_i \to 1$) and also to the smoothing scheme in point 2 (LTS weights) for $w \to +\infty$ ($w_i \to 1$ for $i \leq h_n = \left[\frac{n}{2}\right]$ and $w_i \to 0$ for $i > h_n = \left[\frac{n}{2}\right]$).

Apparently, this estimator can share its robustness properties (namely a high breakdown point) with the already reviewed LTS, at least for choices of weights like in point 2 and in point 3 for $\omega \gg 1$. Additionally, once we restrict our attention only to strictly positive weights, i.e., $w_1 \geq w_2 \geq \ldots \geq w_n > 0$, we obtain an estimator that does not reject any observation completely. This means that all observations are included in the regression and binary and categorical variables do not cause problems anymore; moreover, removing the discontinuity of the objective function significantly reduces the sensitivity of SLTS to small changes of data. On the other hand, there are many similarities between LTS and SLTS. SLTS can still eliminate the effect of outliers and other data-contaminating observations in the same way as LTS does as long as weights are properly chosen, that is, if the effect of large residuals on the SLTS objective function is sufficiently reduced. Further, as I show later, the computation of SLTS could be done by using the weighted least squares (WLS) method with weights $w = (w_{P_1}, \ldots, w_{P_n})$ for each of $n!$ permutations $P = (P_1, \ldots, P_n)$ of $\{1, \ldots, n\}$ and taking as the final estimate the WLS estimate for the permutation that controls the minimum sum of squared residuals. Therefore, if the WLS estimator exists for all permutations of weight vectors, then SLTS also exists (it is the minimum of a finite number of values).

The crucial point is, of course, the choice of weights. There are several possibilities how weights can be chosen:

1. A fixed smoothing scheme, such as the least squares one ($w_1 = \ldots = w_n = 1$): the only advantage of this option is that we can use the resulting estimator in linear regression models with discrete explanatory variables if all weights are positive. However, in such a case, the robustness of the estimator suffers.

2. A data-dependent smoothing scheme: weights are based on data statistics. If we want to be on the safe side, the weights can be defined, for example, so that the smallest weights are inversely proportional to the distance of the point most distant from the center of data; or they can be based on some robust distances as in the case of the RDL_1 estimator.

3. An adaptive choice from a given class of smoothing schemes: given a class of smoothing schemes $f(x; \omega_1, \ldots, \omega_m)$ parameterized by $\omega_1, \ldots, \omega_m$ and requirements on robustness, we try to find an optimal choice of parameters $\omega_1, \ldots, \omega_m$ for a given data set.

There are certainly many possibilities how weight vectors can be defined. A fixed choice of a smoothing scheme (point 1) is neither robust, nor flexible. The strategy described in point 2 is also not suitable because we do not assign weights directly to residuals and because usual weight choices provide sufficient robustness only under some additional assumptions about a model. Therefore, the strategy that I would like to discuss in this paper is the adaptive choice of weighting scheme described in point 3. Consider, for example, such a weighting scheme defined by one parameter: $w_i = f(\frac{i}{n}, \omega)$, where $\omega \in \mathbb{R}$ and f is chosen so that the corresponding SLTS estimate converges to the least squares for some values of parameter ω (e.g., for $\omega \to 0$) and to the least trimmed squares for other ones (e.g., $\omega \to \infty$). Then we can by means of this single parameter ω choose how far or close the corresponding SLTS estimator is to LTS and LS. In other words, we control the balance between the robustness of the estimator and the amount of information it employs from data. See Section 5.2 and 6 for more information on this topic.

4.3 Relation between SLTS and WLS estimators

Now, I derive a lemma describing the relation between the SLTS and weighted least squares (WLS) estimator. This result will be useful not only for a better understanding of the behavior of SLTS, but also for computation of the SLTS estimator.

We observed in Section 4 that the SLTS estimator corresponds to a weighted least squares estimator with specially assigned weights. Let us make this assertion more precise.

Lemma 1 *Let $(y_i, x_i)_{i=1}^n$ be a fixed realization of random sequence $(y_i = x_i^T \beta^0 + \varepsilon_i, x_i)_{i=1}^n$ and $w = (w_1, \ldots, w_n)$ be a weighting vector, $w_{1n} \geq w_{2n} \geq \ldots \geq w_{nn} > 0$. Consider*

$$\hat{\beta}_n^{(SLTS,w)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} w_i r_{[i]}^2(\beta), \tag{5}$$

where $r_i(\beta) = y_i - x_i^T \beta$. Let $k_i(\beta) : \mathbb{R} \to \{1, \ldots, n\}$ be a function such that $k_i(\beta)$ is the index of the observation with the i*th largest squared residual at β, $r_{k_i(\beta)}^2(\beta) = r_{[i]}^2(\beta)$. Define now weights $v_{k_i(\hat{\beta}_n^{(SLTS,w)})} = w_i$ for all $i = 1, \ldots, n$. Then the weighted least squares estimator with weights $v_i, i = 1, \ldots, n$,*

$$\hat{\beta}_n^{(WLS,v)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} v_i r_i^2(\beta) = \arg\min_{\beta \in B} \sum_{i=1}^{n} v_i \left(y_i - x_i^T \beta\right)^2, \tag{6}$$

is equal to the smoothed least trimmed squares estimator: $\hat{\beta}_n^{(SLTS,w)} = \hat{\beta}_n^{(WLS,v)}$.

Proof: I prove the lemma by contradiction. Let $\hat{\beta}_n^{(SLTS,w)} \neq \hat{\beta}_n^{(WLS,v)}$. Then it follows from the definition of weights v and estimates $\hat{\beta}_n^{(SLTS,w)}$ and $\hat{\beta}_n^{(WLS,v)}$

(ordering of squared residuals $r_i^2(\beta)$ at $\hat{\beta}_n^{(SLTS,w)}$ is given) that

$$S_s\left(X_n, Y_n, w; \hat{\beta}_n^{(SLTS,w)}\right) = S_w\left(X_n, Y_n, v; \hat{\beta}_n^{(SLTS,w)}\right) > S_w\left(X_n, Y_n, v; \hat{\beta}_n^{(WLS,v)}\right). \tag{7}$$

Since the objective function of the weighted least squares estimator can be rewritten as $(\{k_1(\beta), \ldots, k_n(\beta)\} = \{1, \ldots, n\}$ for any $\beta)$

$$\begin{aligned} S_w\left(X_n, Y_n, v; \hat{\beta}_n^{(WLS,v)}\right) &= \sum_{i=1}^n v_i r_i^2\left(\hat{\beta}_n^{(WLS,v)}\right) \\ &= \sum_{i=1}^n v_{k_i(\hat{\beta}_n^{(WLS,v)})} r^2_{k_i(\hat{\beta}_n^{(WLS,v)})}\left(\hat{\beta}_n^{(WLS,v)}\right) \\ &= \sum_{i=1}^n v_{k_i(\hat{\beta}_n^{(WLS,v)})} r^2_{[i]}\left(\hat{\beta}_n^{(WLS,v)}\right) \end{aligned}$$

and the sets of weights $\{v_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$ are identical, it follows that

$$S_w\left(X_n, Y_n, v; \hat{\beta}_n^{(WLS,v)}\right) = \sum_{i=1}^n v_{k_i(\hat{\beta}_n^{(WLS,v)})} r^2_{[i]}\left(\hat{\beta}_n^{(WLS,v)}\right) \geq \sum_{i=1}^n w_i r^2_{[i]}\left(\hat{\beta}_n^{(WLS,v)}\right). \tag{8}$$

The argument behind this result is simple: if weights $v_{k_i(\hat{\beta}_n^{(WLS,v)})}$ are sorted in descending order, that is $v_{k_1(\hat{\beta}_n^{(WLS,v)})} \geq \ldots \geq v_{k_n(\hat{\beta}_n^{(WLS,v)})}$, then the sums in (8) are equal; otherwise, we just order weights $v_{k_i(\hat{\beta}_n^{(WLS,v)})}, i = 1, \ldots, n$, decreasingly to get vector w, and thus, put more weight on smaller squared residuals and less weight on larger squared residuals. Consequently, we get

$$\begin{aligned} S_s\left(X_n, Y_n, w; \hat{\beta}_n^{(SLTS,w)}\right) &> S_w\left(X_n, Y_n, v; \hat{\beta}_n^{(WLS,v)}\right) \\ &\geq \sum_{i=1}^n w_i r^2_{[i]}\left(\hat{\beta}_n^{(WLS,v)}\right) = S_s\left(X_n, Y_n, w; \hat{\beta}_n^{(WLS,v)}\right) \end{aligned}$$

and this is the contradiction: $\hat{\beta}_n^{(SLTS,w)}$ does not minimize $S_s\left(X_n, Y_n, v; \beta\right)$. □

Lemma 1 actually states that the SLTS estimator corresponds to a weighted least squares estimator with specially assigned weights. These weights are a permutation of the weight vector w defining SLTS. However, this permutation is specific to a given realization of random variables, so we get a different permutation of weights (and thus a different WLS estimator) for every sample $(y_i, x_i)_{i=1}^n$. Unfortunately, it is not possible to easily find out, which permutation defines a WLS estimator equivalent to SLTS in a given sample. Nevertheless, this lemma is very important for the rest of this paper in two ways: it helps us to understand the asymptotic results concerning SLTS and it provides a way (although not a straightforward one) to compute the SLTS estimator.

5 Properties of smoothed least trimmed squares

In this section, I first introduce the assumptions necessary for proving consistency and asymptotic normality of the proposed estimator and then I derive these important asymptotic results in Section 5.1. Later, I discuss some elementary properties of the SLTS estimator, its objective function and corresponding regression residuals as functions of weights (Section 5.2). This will be useful for designing rules driving the proposed adaptive choice of smoothing schemes (Section 6.2).

Before doing so, let us introduce the assumptions and notation used in the theoretical part. Consider a linear regression model (2) for a sample (y_i, x_i) with a response variable y_i and a vector of explanatory variables x_i:

$$y_i = x_i^T \beta + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{9}$$

Let us denote $Y_n = (y_1, \ldots, y_n)^T$ and $X_n = (x_1, \ldots, x_n)^T$; similarly, $\mathrm{E}_n = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Moreover, let 1_n represent n-dimensional vector of ones, 0_n be n-dimensional vector of zeroes, and I_n be the $n \times n$ identity matrix of dimension n.

Further, let β^0 represent the true value of regression parameters and $\hat{\beta}_n^{(SLTS,w)}$ the SLTS estimator defined by

$$\hat{\beta}_n^{(SLTS,w)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} w_i r_{[i]}^2(\beta) \tag{10}$$

for weights $w = (w_1, \ldots, w_n)$. The objective function of SLTS at β is further referred to by $S_s(X_n, Y_n, w; \beta) = \sum_{i=1}^n w_i r_{[i]}^2(\beta)$; if it is written without weights, $w = 1_n$ is assumed, and thus, $S_s(X_n, Y_n; \beta) = \sum_{i=1}^n r_{[i]}^2(\beta) = \sum_{i=1}^n r_i^2(\beta)$ is the objective function of the least squares estimator at β. The objective function of the weighted least squares estimator at β is denoted $S_w(X_n, Y_n, w; \beta) = \sum_{i=1}^n w_i r_i^2(\beta)$ and again $S_w(X_n, Y_n; \beta) = \sum_{i=1}^n r_i^2(\beta)$.

Further, we discussed the possibility to define weights for the SLTS by means of a real function in Section 4.2. To make this concept more precise, let us consider a real-valued non-increasing function $f(\cdot; \omega_1, \ldots, \omega_m) \in L_1(\langle 0, 1 \rangle)$ parameterized by $\omega_1, \ldots, \omega_m \in \mathbb{R}^m$ ($L_1(C)$ represents the space of all absolutely integrable functions on C) such that $f(x; \omega_1, \ldots, \omega_m) \geq 0$ for all $x \in \langle 0, 1 \rangle$. For the given values of parameters $\omega_1, \ldots, \omega_m$, it is possible to define weights

$$w_i = f\left(\frac{2i-1}{2n}; \omega_1, \ldots, \omega_m\right)$$

for all $i = 1, \ldots, n$.[14] Then the function $f(\cdot; \omega_1, \ldots, \omega_m)$ is the generating function of the SLTS smoothing scheme parameterized by $\omega_1, \ldots, \omega_m$ and the weights are

[14] Fraction $\frac{2i-1}{2n}$ is used instead of the simple $\frac{i}{n}$ in order to obtain evenly spread values inside the open interval $(0, 1)$.

generated by the function f. In the following analysis, I focus only on strictly positive generating functions, which prevent a complete rejection of observation. Moreover, I discuss mainly the so-called stepwise generating functions:[15] $f(x)$ is a stepwise function on $\langle 0,1\rangle$ if there are $k_f \in \mathbb{N}$ and real constants $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_{k_f} = 1$ and $c_1, \ldots, c_{k_f} \in \mathbb{R}$ such that $f(x) = c_i$ for all $\alpha_{i-1} < x < \alpha_i$ and all $i = 1, \ldots, k_f$. Because we require that $w_1 \geq w_2 \geq \ldots \geq w_n > 0$ for a weighting vector $w = (w_1, \ldots, w_n)$, it has to hold $c_1 \geq c_2 \geq \ldots \geq c_{k_f} > 0$ for values of a stepwise generating function. Additionally, we can always assume without loss of generality that constants α_i and c_i are chosen such that $c_1 > c_2 > \ldots > c_{k_f} > 0$.

Finally, note that if we assume that weights $w = (w_1, \ldots, w_n)$ are generated by a stepwise function defined by constants k_f, $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_{k_f} = 1$, and $c_1 > c_2 > \ldots > c_{k_f} > 0$, we can rewrite the definition (10) of SLTS as[16]

$$\hat{\beta}_n^{(SLTS,w)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} r_i^2(\beta) \cdot \left(\sum_{j=1}^{k_f - 1} (c_j - c_{j+1})\, I\Big(r_i^2(\beta) \leq r_{[\alpha_j n]}^2(\beta)\Big) + c_{k_f} \right). \tag{11}$$

To obtain this formula, one has to realize that the $[\alpha_1 n]$ smallest residuals are assigned weight c_1, the $[\alpha_2 n]$ smallest residuals have weight $c_2 \leq c_1$, and so on. Moreover, for a given value of $\beta \in B$, the set of the $[\alpha_j n]$ smallest squared residuals corresponds to a set of those residuals that satisfy $r_i^2(\beta) \leq r_{[\alpha_j n]}^2(\beta)$.[17]

[15] This allows me to employ existing asymptotic results for LTS.

[16] By I(property describing a set A) we denote the indicator of the set A.

[17] In general, this definition is not equivalent to the original one. They are exactly equivalent if and only if all the residuals are different from each other. Under Assumption A stated below, this happens with zero probability and definitions (10) and (11) are equivalent almost surely as the cumulative distribution function of $r_i(\beta)$ is assumed to be absolutely continuous. Therefore, I use definition (11) for convenience.

For notational convenience, I denote

$$SI(i,\beta;\alpha,c)=\sum_{j=1}^{k_f-1}(c_j-c_{j+1})\,I\Big(r_i^2(\beta)\le r_{[\alpha_j n]}^2(\beta)\Big)+c_{k_f} \tag{12}$$

where $\alpha=(\alpha_1,\ldots,\alpha_{k_f})$ and $c=(c_1,\ldots,c_{k_f})$, so we can rewrite (11) as

$$\hat{\beta}_n^{(SLTS,w)}=\arg\min_{\beta\in B}\sum_{i=1}^{n}r_i^2(\beta)\cdot SI(i,\beta;\alpha,c),$$

and similarly, the objective function of SLTS at β is $S_s(X_n,Y_n,w;\beta)=\sum_{i=1}^{n}r_i^2(\beta)\cdot SI(i,\beta;\alpha,c)$. Additionally, I define an asymptotical equivalent of $SI(i,\beta;\alpha,c)$. I simply replace $r_{[\alpha_j n]}^2(\beta)$ in (12) by its probability limit:

$$SIT(i,\beta;\alpha,c)=\sum_{j=1}^{k_f-1}(c_j-c_{j+1})\,I\big(r_i^2(\beta)\le G_\beta^{-1}(\alpha_j)\big)+c_{k_f}, \tag{13}$$

where $G_\beta^{-1}(\alpha_j)$ represents the α_j-quantile of the distribution function of $r_{[\alpha_j n]}^2(\beta)$.

Now, let us finally specify the assumptions needed for the consistency and asymptotic normality of the SLTS estimator.

Assumption A**.**

A1 Let $W_n=(w_{in})_{i=1}^n$ be a sequence of weight vectors generated for all $n\in\mathbb{N}$ by a stepwise generating function $f_w(x):\langle 0,1\rangle\to\mathbb{R}_+$. We assume that there are constants $k_f\in\mathbb{N}$, $0=\alpha_0<\alpha_1<\ldots<\alpha_{k_f}=1$, and $+\infty>c_1>c_2>\ldots>c_{k_f}>0$ such that $f_w(x)=c_i$ for all $\alpha_{i-1}<x\le\alpha_i$ and all $i=1,\ldots,k_f$. Hence, $w_{1n}\ge w_{2n}\ge\ldots\ge w_{nn}>0$.

***Remark* 1** *As stated above, I derive consistency and asymptotic normality only for stepwise generating functions. However, this does not present a considerable* *restriction on the choice of smoothing schemes since every continuous function on $\langle 0, 1\rangle$ can be approximated with an arbitrary precision by a stepwise function. See Section 6.2 for more details.*

A2 Let $(x_i, \varepsilon_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1, \ldots, n$, be a sequence of independent identically distributed random vectors with finite fourth moments. Moreover,

$$n^{-1/4} \max_{i,j} |x_{ij}| = \mathcal{O}_p(1). \tag{14}$$

***Remark* 2** *The necessity to include restriction (14) is caused by the discontinuity of the objective function of LTS, which the SLTS objective function is composed of. A nonrandom version of this assumption was used for the first time by Jurečková (1984) and the presented version (14) was introduced by Víšek (1999a) and used by Čížek (2001). Apparently, this condition does not affect a random variable with a finite support at all. Moreover, Čížek (2001, Proposition 1) showed that equation (14) holds even for some distribution functions with polynomial tails, namely for those that have finite second moments. As the existence of finite second moments is almost always utilized, and moreover, is one of the necessary conditions here, assumption (14) should not pose a considerable restriction on the explanatory variables.*

A3 We assume

- $\mathsf{E}\left(x_1 x_1^T \cdot SIT(1, \beta; \alpha, c)\right) = Q(\beta)$, where $Q(\beta)$ is a nonsingular (positive definite) matrix for $\beta \in B$, where B is a compact parametric space,
- $\mathsf{E}\left(\varepsilon_1 \cdot SI(1, \beta^0; \alpha, c) | \, x_1\right) = 0$,
- $\mathsf{E}\left(\varepsilon_1^2 \cdot SI(1, \beta^0; \alpha, c) | \, x_1\right) = \sigma_T^2$, where $\sigma_T^2 \in (0, +\infty)$.

***Remark* 3** *These moment assumptions are nothing but a natural analogy to the usual orthogonality* $\mathsf{E}(\varepsilon|x) = 0$ *and spheriality* $\mathsf{E}(\varepsilon^2|x) = \sigma^2$ *conditions used for the least squares regression. They also closely resemble similar conditions used for LTS*

$$\mathsf{E}\left(\varepsilon_1 I\big(r_1^2(\beta^0) \leq r_{[\lambda n]}^2(\beta^0)\big)\,\middle|\, x_1\right) = 0, \quad \mathsf{E}\left(\varepsilon_1^2 I\big(r_1^2(\beta^0) \leq r_{[\lambda n]}^2(\beta^0)\big)\,\middle|\, x_1\right) = \sigma_T^2, \tag{15}$$

$\lambda \in \left\langle \frac{1}{2}, 1\right\rangle$*; see, for example Čížek (2001). Note that Assumption A2 is weaker than its counterparts (15) for LTS.*

The same applies to the regularity condition regarding explanatory variables— $\mathsf{E}\, x_1 x_1^T = Q$*, where* Q *is a nonsingular matrix, is a standard identification condition for the least squares estimator.*

A4 Further, let us denote $F_{\beta^0}(x)$ as the distribution function of ε_i and assume that $F_{\beta^0}(x)$ is absolutely continuous. Let f_{β^0} denote the probability density of F_{β^0}, which is assumed to be positive, bounded by $M_f > 0$ and differentiable on the whole support of the distribution function F_{β^0}.

***Remark* 4** *This assumption, which actually implies the continuity of the quantile function, is typical when trimmed order statistics of random variables are analyzed; see Víšek (1999a) and Čížek (2001), for instance.*

Let $G_{\beta^0}(z)$ represents the distribution function of $\varepsilon_i^2 \equiv r_i^2(\beta)$. It follows that $G_{\beta^0}(z) = F_{\beta^0}(\sqrt{z}) - F_{\beta^0}(-\sqrt{z})$ for $z > 0$, $G_{\beta^0}(z) = 0$ otherwise, and hence, it is also absolutely continuous. Therefore, we can define $g_{\beta^0}(z)$ to be the corresponding probability density function. Moreover, sometimes it is necessary to refer to the distribution function of $r_i(\beta)$ and $r_i^2(\beta)$; in such a case, F_β and G_β are used for the cumulative distribution functions and f_β and g_β for the corresponding probability densities.

A5 Finally, assume that for any $\varepsilon > 0$ and $U(\beta^0, \varepsilon)$ such that $B - U(\beta^0, \varepsilon)$ is compact, there exists $\alpha(\varepsilon) > 0$ such that it holds

$$\min_{\beta \in B - U(\beta^0, \varepsilon)} \mathsf{E}\left[r_i^2(\beta) \cdot SIT(1, \beta; \alpha, c)\right] - \mathsf{E}\left[r_i^2(\beta^0) \cdot SIT(1, \beta^0; \alpha, c)\right] > \alpha(\varepsilon).$$

Remark **5** *This is nothing but a standard identification condition—the expectation of the objective function is assumed to have asymptotically a unique global minimum at β^0. Compare, for example, to Čížek (2001) and White (1980).*

5.1 Consistency and asymptotic normality

Now, I derive the main asymptotic results, namely the consistency and asymptotic normality of SLTS.

Theorem 1 *Let Assumption A hold for a sequence $W_n = (w_{in})_{i=1}^n$ of weight vectors. Let $q_j = \sqrt{G_{\beta^0}^{-1}(\alpha_j)}, j = 1, \ldots, k_f$, and*

$$\sum_{i=1}^{k_j - 1} (c_j - c_{j+1}) \cdot \{\alpha_j - q_j \left[f_{\beta^0}(-q_j) + f_{\beta^0}(q_j)\right]\} + c_{k_f} \neq 0. \tag{16}$$

Then the smoothed least trimmed squares estimator $\hat{\beta}_n^{(SLTS, W_n)}$ is $\sqrt{n}$-consistent

$$\sqrt{n}\left(\hat{\beta}_n^{(SLTS, W_n)} - \beta^0\right) = \mathcal{O}_p(1) \tag{17}$$

and asymptotically normal,

$$\sqrt{n}\left(\hat{\beta}_n^{(SLTS, W_n)} - \beta^0\right) \xrightarrow{F} N(0, V) \tag{18}$$

as $n \to +\infty$, where

$$\begin{aligned} V &= \left\{ \sum_{i=1}^{k_j-1} (c_j - c_{j+1}) \cdot \{\alpha_j - q_j\, [f_{\beta^0}(-q_j) + f_{\beta^0}(q_j)]\} + c_{k_f} \right\}^{-2} \times \\ &\quad \times Q^{-1}(\beta^0) \operatorname{\mathsf{var}} \big(\varepsilon_1 x_1 \cdot SIT(1, \beta^0; \alpha, c)\big)\, Q^{-1}(\beta^0). \end{aligned} \tag{19}$$

Proof: First of all, the objective function

$$\begin{aligned} S_s(X_n, Y_n, W_n; \beta) &= \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 \cdot SI(i, \beta; \alpha, c) \qquad (20) \\ &= \sum_{j=1}^{k_f} (c_j - c_{j+1}) \cdot \left[\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 \cdot I\Big(r_i^2(\beta) \le r^2_{[\alpha_j n]}(\beta)\Big) \right] \\ &+ c_{k_f} \cdot \left[\sum_{i=1}^{n} (y_i - \beta^T x_i)^2 \right] \end{aligned}$$

is actually a sum of the objective functions of the LTS estimators (the sums in the square brackets are the mentioned LTS objective functions with trimming constants α_j). Because Assumption A covers all the assumptions relevant for the linear regression model used in Víšek (1999a) and Čížek (2001), I simply employ the existing results for LTS from these two papers by applying them to every element of sum (20).

Next, the SLTS estimator, minimizing its objective function $S_s(X_n, Y_n, W_n; \beta)$, can be also obtained from the normal equations $\frac{\partial S_s(X_n, Y_n, W_n; \beta)}{\partial \beta} = 0$. As derived by Víšek (1999a, page 6) and Čížek (2001, Section 3.3.1 and Lemma 1), the normal equations can almost surely be expressed as

$$\frac{\partial S_s(X_n, Y_n, W_n; \beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \beta^T x_i) x_i^T \cdot SI(i, \beta; \alpha, c) = 0. \tag{21}$$

The second derivative of the objective function $\frac{\partial^2 S_s(X_n,Y_n,W_n;\beta)}{\partial\beta^2}$ can be analogously expressed as

$$\frac{\partial^2 S_s(X_n,Y_n,W_n;\beta)}{\partial\beta^2} = \sum_{i=1}^{n} x_i x_i^T \cdot SI(i,\beta;\alpha,c).$$

Moreover, because of Assumption A, we can use the results from Čížek (2001, Corollary 5 and Lemma 7), which imply uniformly in β

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \cdot SI(i,\beta;\alpha,c) \to \mathsf{E}\left(x_i x_i^T \cdot SIT(i,\beta;\alpha,c)\right) = Q(\beta)$$

in probability for $n \to \infty$, where $Q(\beta)$ is a nonsingular positive definite matrix (see Assumption A3). Hence, for any $\varepsilon > 0$ it is possible to find $n_0 \in \mathbb{N}$ such that the matrix $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \cdot SI(i,\beta;\alpha,c)$ is positive definite for all β with a probability greater than $1-\varepsilon$. Consequently, the normal equations (21) have a unique solution with an arbitrarily high probability for a sufficiently high n.

Now, I will find the solution to the normal equations (21). Because it is unique, it has to be equal to the SLTS estimate. Using the asymptotic linearity theorem for LTS (see Víšek (1999a, Theorem 1) and Čížek (2001, Theorem 1)) we can write that for any $M > 0$

$$\begin{aligned} \frac{\partial S_s(X_n,Y_n,W_n;\beta^0 - n^{-\frac{1}{2}}t)}{\partial\beta} &= \frac{\partial S_s(X_n,Y_n,W_n;\beta^0)}{\partial\beta} \\ &- n^{\frac{1}{2}} Q(\beta^0) t \cdot \left\{ \sum_{i=1}^{k_j-1} (c_j - c_{j+1}) \cdot C_j(\alpha) + c_{k_f} \right\} \\ &+ \mathcal{O}_p\left(n^{\frac{1}{4}}\right) \end{aligned} \tag{22}$$

uniformly for all $t \in T_M = \{t : \|t\| \leq M\}$, where

$$C_j(\alpha) = \alpha_j - q_j\left(f(-q_j) + f(q_j)\right)$$

(notation $q_j = \sqrt{G_{\beta^0}^{-1}(\alpha_j)}$ is used). We show that there is $t \in T_M$ such that $\frac{\partial S_s(X_n,Y_n,W_n;\beta^0-n^{-\frac{1}{2}}t)}{\partial\beta} = 0$ with an arbitrarily high probability. This means that $\beta = \beta^0 - n^{-\frac{1}{2}}t$ is then the only solution of normal equations. From equation (22), it follows that, for the solution of the normal equations,

$$\frac{\partial S_s(X_n,Y_n,W_n;\beta^0)}{\partial\beta} = n^{\frac{1}{2}}Q(\beta^0)t \cdot \left\{\sum_{i=1}^{k_j-1}(c_j - c_{j+1})\cdot C_j(\alpha) + c_{k_f}\right\} + \mathcal{O}_p\left(n^{\frac{1}{4}}\right)$$

and (remember, $\sum_{i=1}^{k_j-1}(c_j - c_{j+1})\cdot C_j(\alpha) + c_{k_f} \neq 0$ and $Q(\beta^0)$ is a nonsingular matrix)

$$t = Q^{-1}(\beta^0)\cdot\frac{1}{\sqrt{n}}\frac{\partial S_s(X_n,Y_n,W_n;\beta^0)}{\partial\beta}\cdot\left\{\sum_{i=1}^{k_j-1}(c_j - c_{j+1})\cdot C_j(\alpha) + c_{k_f}\right\}^{-1} + \mathcal{O}_p\left(n^{-\frac{1}{4}}\right) \tag{23}$$

as $n \to \infty$. Since the random variable

$$\frac{1}{\sqrt{n}}\frac{\partial S_s(X_n,Y_n,W_n;\beta^0)}{\partial\beta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i - x_i^T\beta^0)x_i^T \cdot SI(i,\beta^0;\alpha,c)$$

has asymptotically the normal distribution with zero expectation and variance

$$\mathsf{var}\left(\frac{1}{\sqrt{n}}\frac{\partial S_s(X_n,Y_n,W_n;\beta^0)}{\partial\beta}\right) = \mathsf{var}\left(\varepsilon_1 x_1 \cdot SIT(1,\beta^0;\alpha,c)\right)$$

(see Víšek (1999a, proof of Theorem 2) and Čížek (2001, proof of Theorem 4 and Lemma 6)), it is bounded in probability. Hence, t defined in (23) is bounded in probability as well and for any $\varepsilon > 0$ there is $M > 0$ such that term (22) equals zero for some $t \in T_M$ with probability higher than $1 - \varepsilon$. Then $\beta^0 - n^{-\frac{1}{2}}t$ is the unique solution of (21), and consequently, the SLTS estimate itself is $\hat{\beta}_n^{(SLTS,W_n)} = \beta^0 - n^{-\frac{1}{2}}t$. Apparently, it holds that $\sqrt{n}\left(\hat{\beta}_n^{(SLTS,W_n)} - \beta^0\right) = t$.

This finding has two important implications: the $\sqrt{n}$-consistency and asymptotic normality of SLTS. First, because we can find a compact set T_M and the solution to the normal equations $\hat{\beta}_n^{(SLTS,W_n)}$ such that $\left\|\sqrt{n}\left(\hat{\beta}_n^{(SLTS,W_n)} - \beta^0\right)\right\| = \|t\| \leq M$ with an arbitrarily high probability,

$$\sqrt{n}\left(\hat{\beta}_n^{(SLTS,W_n)} - \beta^0\right) = \mathcal{O}_p(1)$$

as $n \to +\infty$ (this is the $\sqrt{n}$-consistency of SLTS). Second, we found that the solution t of the normal equation (21) considered as a random variable equals

$$t = Q^{-1}(\beta^0) \cdot \left\{ \sum_{i=1}^{k_j-1} (c_j - c_{j+1}) \cdot C_j(\alpha) + c_{k_f} \right\}^{-1} \cdot Z + \mathcal{O}_p\left(n^{-\frac{1}{4}}\right)$$

(see (23)), where $Z = \frac{1}{\sqrt{n}} \frac{\partial S_s(X_n,Y_n,W_n;\beta^0)}{\partial \beta}$ is asymptotically normally distributed with zero expectation and variance $\mathsf{var}\, Z = \mathsf{var}\left(\varepsilon_1 x_1 \cdot SIT(1,\beta^0;\alpha,c)\right)$. Hence,

$$\sqrt{n}\left(\hat{\beta}_n^{(SLTS,W_n)} - \beta^0\right) = t \sim N\left(0,V\right)$$

for $n \to +\infty$, where

$$V = \left\{ \sum_{i=1}^{k_j-1} (c_j - c_{j+1}) \cdot C_j(\alpha) + c_{k_f} \right\}^{-2} \cdot Q^{-1}(\beta^0)\, \mathsf{var}\left(\varepsilon_1 x_1 \cdot SIT(1,\beta^0;\alpha,c)\right) Q^{-1}(\beta^0).$$

□

The proof of asymptotic normality is not useful just on its own, it gives us also an idea about the asymptotic variance of the SLTS estimator. This provides a comparison to the least squares estimator, and more importantly, an idea how a choice of weighting scheme used to define SLTS influences the asymptotic variance of the estimator. Nevertheless, remember that these results describe only the asymptotic behavior and we cannot use it efficiently without a prior assumption about the error distribution. To complement these asymptotic results, I study the finite sample performance behavior of SLTS using Monte Carlo simulations in Section 7.

5.2 Properties of the estimator as a function of weights

As I indicated in Section 4, the main focus of this paper is on the adaptive choice of weights, which should enable us by a choice of one or more parameters to control the balance between the robustness of the estimator and the amount of information it employs from data. As the first step in this direction, I derive some theoretical properties concerning the SLTS objective function $S_s(x, y, w; \beta)$ as a function of weights. In order to make the subsequent explanations and analysis tractable, I first restrict the choice of weights to a family of weighting schemes. Later, I discuss the principles of the adaptive weight choice and the corresponding theoretical results.

It is interesting to study SLTS not only for a fixed weighting scheme, but it is preferable to search for optimal weights from a class of weighting schemes. For this purpose, I introduced weights-generating functions that are parameterized by a vector of parameters. Whereas this concept requires a non-increasing function that is positive and integrable on $\langle 0, 1 \rangle$, the asymptotic properties of SLTS were proved only for stepwise functions. Both because the results derived in the rest of this paper can be proved generally for any generating function and because it is easier and more transparent to work with a general generating function, I assume from now on that a weights-generating function is a non-increasing continuous function that is positive and integrable on $\langle 0, 1 \rangle$. However, keeping in mind that only stepwise generating functions should be used for practical computation (asymptotic properties of SLTS are derived only for stepwise generating functions in Section 5.1), I assume that some fixed $1 > \varepsilon_p > 0$ and $n_p = [\varepsilon_p^{-1}] + 1$ are given, describing the precision of approximation by a stepwise function. This means that we use for practical computation a stepwise approximation $\bar{f}(x)$ instead of a general continuous function $f(x)$ on $\langle 0, 1\rangle$:

$$\bar{f}(x) = f\left(\frac{2i-1}{2n_p}\right) \quad \text{if} \quad \frac{i-1}{n_p} \leq f(x) \leq \frac{i}{n_p}$$

for all $i = 1, \ldots, n_p$.

Now, let us specify the restrictions regarding generating functions used in the rest of this section.

Assumption W.

Let $W_n(\omega) = (w_{in})_{i=1}^n$ be a sequence of weight vectors generated by function $f(x;\omega) : \langle 0, 1\rangle \to R$ parameterized by ω from an interval $(\omega_1, \omega_2) \subseteq \mathbb{R}$. Assume that for any $\omega \in (\omega_1, \omega_2)$

W1 $f(x;\omega)$ is a continuous, non-increasing, and everywhere positive function bounded on $\langle 0, 1\rangle$ by constant $K_w > 0$ uniformly for all $\omega \in (\omega_1, \omega_2)$

W2 $\int_0^1 f(x;\omega)\mathrm{d}x$ is independent of ω, and

W3 there is $\lambda \in \left\langle \frac{1}{2}, 1\right\rangle$ such that

- $f(x;\omega) \geq f(x;\omega')$ for any $\omega > \omega'$ and $x \leq \lambda$
- $f(x;\omega) \leq f(x;\omega')$ for any $\omega > \omega'$ and $x > \lambda$.

W4 Optionally, we can require that there are $\omega_1 < \omega_2 \in P_\omega$ such that

- for $\omega \to \omega_1$ it holds that $f(x;\omega) \to a_1 > 0$ for all $x \in \langle 0, 1\rangle$,

- for $\omega \to \omega_2$ it holds that $f(x;\omega) \to a_2 > 0$ for all $x \leq \lambda$ and $f(x;\omega) \to 0$ for all $x > \lambda$.

***Remark* 6** *Assumption W2 is just a normalization condition that allows us to compare the values of the SLTS objective function for weighting schemes corresponding to different ω. Assumption W3 formalizes the requirement that we put less weight on large residuals for some (here greater) values of parameter ω and vice versa. Optional assumption W4 states that the least squares and LTS estimators should be at least limiting cases within the class of smoothing schemes defined by $f(\cdot;\omega)$.*

A reasonable choice of weighting functions $f(x;\omega)$ might be, for example, functions of the form $1 - F(x;\omega)$, where $F(x;\omega)$ represents a cumulative density function from some suitable family of distributions. Let me give an example from Section 4, which actually corresponds to a generating function based on the logistic distribution function:

$$f_\lambda(x;\omega) = \frac{1}{1+e^{\omega(x-\lambda)}} \Bigg/ \int_0^1 \frac{1}{1+e^{\omega(x-\lambda)}} \mathrm{d}x \tag{24}$$

for $\omega \geq 0$ and $x \in \langle 0, 1\rangle$ ($\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ is a fixed number here).

Now, I would like to roughly describe the principle of the adaptive choice of weights defined by the weighting parameter ω. At the time of estimation, only a few characteristics of the estimate are readily available: the value of the objective function at the point of the current estimate $S_s\left(x, y, W_n(\omega); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right)$ and the corresponding regression residuals. So, if we want to find the best choice of the weighting parameter ω, we have to base our decision on some characteristics of regression residuals or on the behavior of $S_s\left(x, y, W_n(\omega); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right)$ for different values of ω. Let me give some examples of possible adaptive-choice procedures. One possible idea is based on the fact that the estimator minimizes the weighted sum of squared residuals and that the smaller sum represents a better fit. If there is contamination or a deviation from a regression model that makes the estimate for a given ω inconsistent, the value of the objective function will grow rapidly. This can, indeed, help to differentiate "good" and "bad" choices of the weighting parameter. Another possibility is to use regression residuals. Regression residuals have some mean value and variance, which indicate which residuals are acceptable or which residuals are suspicious. If there is contamination or a deviation from a regression model that makes the estimate for a given ω inconsistent, some regression residuals will be suspiciously large. This will again differentiate "good" and "bad" choices of the weighting parameter. Finally, knowing which values of the weighting parameter ω are acceptable ("good" ones), we choose the one providing the best efficiency. In order to find out which values of ω are acceptable and which are not, I now analyze some fundamental properties of the objective function as a function of ω. Later, I will discuss some theoretical results concerning regression residuals, again as a function of the weighting parameter ω.

So, let us analyze the behavior of $S_s\left(x, y, W_n(\omega); \hat{\beta}_n^{(SLTS,W_n(\omega))}\right)$, that is, of the objective function of SLTS at the optimum $\hat{\beta}_n^{(SLTS,W_n(\omega))}$, as a function of the parameter ω. We show first that this function is decreasing for all $\omega \in (\omega_1, \omega_2)$.

Proposition 1 *Let $(y_i, x_i)_{i=1}^n$ be a fixed realization of random sequence $(y_i = x_i^T \beta^0 + \varepsilon_i, x_i)_{i=1}^n$ and $W_n(\omega)$ be a sequence of weight vectors satisfying Assumption W. Consider*

$$\hat{\beta}_n^{(SLTS,W_n(\omega))} = \arg\min_{\beta \in B} \sum_{i=1}^n w_{in}(\omega) r_{[i]}^2(\beta). \tag{25}$$

Then

$$S_s\left(X_n, Y_n, W_n(\omega); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right) \geq S_s\left(X_n, Y_n, W_n(\omega'); \hat{\beta}_n^{(SLTS, W_n(\omega'))}\right)$$

holds for any $\omega < \omega'$ *from* (ω_1, ω_2).

Proof: Let $\omega < \omega'$. Assumption W implies that $f(x;\omega) \leq f(x;\omega')$ for $x \leq \lambda$ and $f(x;\omega) \geq f(x;\omega')$ for $x > \lambda$. In other words, a higher ω' causes bigger weights to be assigned to smaller residuals and smaller weights to larger residuals (compared with weights for ω). Hence,

$$S_s\left(X_n, Y_n, W_n(\omega); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right) \geq S_s\left(X_n, Y_n, W_n(\omega'); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right).$$

since the objective function S_s is evaluated at the same point $\hat{\beta}_n^{(SLTS, W_n(\omega))}$ on both sides of the inequality, so all residuals stay the same. Now, by the definition of SLTS,

$$S_s\left(X_n, Y_n, W_n(\omega'); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right) \geq S_s\left(X_n, Y_n, W_n(\omega'); \hat{\beta}_n^{(SLTS, W_n(\omega'))}\right),$$

and consequently, it follows that

$$S_s\left(X_n, Y_n, W_n(\omega); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right) \geq S_s\left(X_n, Y_n, W_n(\omega'); \hat{\beta}_n^{(SLTS, W_n(\omega'))}\right).$$

□

So, we know now that the objective function at optimum is decreasing in ω. Unfortunately, we can hardly analyze the shape of $S_s\left(X_n, Y_n, W_n(\omega); \hat{\beta}_n^{(SLTS, W_n(\omega))}\right)$ for the general weighting scheme introduced in Assumption W. On the other hand, the complete specification of weighting schemes in Assumption W provides another guideline: for small values of ω (close to ω_1), the SLTS estimates should converge to the least squares estimates; for large values of ω (close to ω_2), the SLTS estimates should converge to the least trimmed squares estimates. Thus, the lower and upper bound for the values of the SLTS objective function are given by the LS and LTS objective functions. Of course, these bounds cannot be estimated on a real data set because we do not know whether the (least squares) estimates are consistent. However, assuming a linear regression model with a known distribution of the error term, it is possible to compute the asymptotic ratio of the upper and lower bounds of the SLTS objective function. I compute this ratio in the case of the normal distribution in Proposition 2.

Proposition 2 *Let Assumption A hold and the error term be normally distributed $\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$. Consider two special choices of weight vectors: $W_n^1 = (w_{in}^1)$, $w_{in}^1 = 1$ (the least squares weights), and $W_n^2 = (w_{in}^2)$, $w_{in}^2 = \frac{n}{h_n} I(i \leq h_n)$, where $h_n = [\lambda n]$ and $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ (the least trimmed squares weights), for all $i = 1, \ldots, n$ and $n \in \mathbb{N}$. Then*

$$\frac{S_s\left(X_n, Y_n, W_n^1; \hat{\beta}_n^{(SLTS, W_n^1)}\right)}{S_s\left(X_n, Y_n, W_n^2; \hat{\beta}_n^{(SLTS, W_n^2)}\right)} = \frac{\sum_{i=1}^{n} r_{[i]}^2\left(\hat{\beta}_n^{(SLTS, W_n^1)}\right)}{\frac{n}{h_n}\sum_{i=1}^{h_n} r_{[i]}^2\left(\hat{\beta}_n^{(SLTS, W_n^2)}\right)} \to \frac{\lambda}{F_{\chi_3^2}\left(F_{\chi_1^2}^{-1}(\lambda)\right)}$$

as $n \to +\infty$, where $F_{\chi_d^2}$ represents the χ_d^2 cumulative distribution function with d degrees of freedom and $F_{\chi_d^2}^{-1}$ the quantile function of χ_d^2 distribution.

Proof: Assumption A guarantees that both estimators $\hat{\beta}_n^{(SLTS, W_n^1)}$ and $\hat{\beta}_n^{(SLTS, W_n^2)}$ are consistent—they converge to the true parameter vector β^0 in probability. Hence, the residuals $r_i\left(\hat{\beta}_n^{(SLTS, W_n^l)}\right) = \varepsilon_i + x_i^T\left(\hat{\beta}_n^{(SLTS, W_n^l)} - \beta^0\right)$ converge in probability to ε_i for $l = 1, 2$ and $i = 1, \ldots, n$. Thus, the squared residuals divided by σ^2 are asymptotically distributed according to χ_1^2 distribution with one degree of freedom. Consequently, by the strong law of large numbers, $\frac{1}{n\sigma^2}\sum_{i=1}^n r_{[i]}^2\left(\hat{\beta}_n^{(SLTS,W_n^1)}\right) = \frac{1}{n\sigma^2}\sum_{i=1}^n r_i^2\left(\hat{\beta}_n^{(SLTS,W_n^1)}\right)$ converges almost surely to the expectation of the χ_1^2 distribution, that is

$$\frac{1}{n\sigma^2}\sum_{i=1}^n r_{[i]}^2\left(\hat{\beta}_n^{(SLTS,W_n^1)}\right) \to \int_0^\infty x \cdot f_{\chi_1^2}(x)dx = 1,$$

where $f_{\chi_1^2}(x)$ represents the probability density function of χ_1^2. Similarly,

$$\frac{1}{h_n\sigma^2}\sum_{i=1}^{h_n} r_{[i]}^2\left(\hat{\beta}_n^{(SLTS,W_n^2)}\right) \to \int_0^{F_{\chi_1^2}^{-1}(\lambda)} x\cdot\frac{f_{\chi_1^2}(x)}{F_{\chi_1^2}\left(F_{\chi_1^2}^{-1}(\lambda)\right)}dx = \frac{1}{\lambda}\int_0^{F_{\chi_1^2}^{-1}(\lambda)} x\cdot f_{\chi_1^2}(x)dx.$$

We can transform the integral in the following way:

$$\begin{aligned}\int_0^z x\cdot f_{\chi_1^2}(x)dx &= \int_0^z x\cdot\frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}x^{-\frac{1}{2}}e^{-\frac{x}{2}}dx\\ &= \int_0^z \frac{1}{2^{\frac{3}{2}}\Gamma(\frac{3}{2})}x^{\frac{1}{2}}e^{-\frac{x}{2}}dx\\ &= \int_0^z f_{\chi_3^2}(x)dx.\end{aligned}$$

This leads directly to

$$\frac{1}{h_n\sigma^2}\sum_{i=1}^{h_n} r_{[i]}^2\left(\hat{\beta}_n^{(SLTS,W_n^2)}\right) \to \frac{1}{\lambda}\int_0^{F_{\chi_1^2}^{-1}(\lambda)} f_{\chi_3^2}(x)dx = \frac{1}{\lambda}F_{\chi_3^2}\left(F_{\chi_1^2}^{-1}(\lambda)\right).$$

□

Such a result can be computed in a similar way also for other absolutely continuous distribution functions. See Section 6 for further discussion and the use of this result.

Besides the objective function $S_s\left(x, y, w; \hat{\beta}_n^{(SLTS,W_n(\omega))}\right)$, we have one more characteristic of an estimate available: the corresponding regression residuals. Like in Proposition 2 for the SLTS objective function, it is possible to asymptotically compare some statistics (e.g., variance) of regression residuals for the two limiting cases, LS ($\omega \to \omega_1$) and LTS ($\omega \to \omega_2$). Assuming a linear regression model with normally distributed errors, I compare regression residuals in these two cases in Proposition 3.

Proposition 3 *Let Assumption A hold and the error term be normally distributed $\varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$. Consider two special choices of weight vectors: $W_n^1 = (w_{in}^1)$, $w_{in}^1 = 1$ (the least squares weights), and $W_n^2 = (w_{in}^2)$, $w_{in}^2 = \frac{n}{h_n} I(i \leq h_n)$, where $h_n = [\lambda n]$ and $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ (the least trimmed squares weights), for all $i = 1, \ldots, n$ and $n \in \mathbb{N}$. Moreover, given a sample of regression residuals $r_i(\beta) = y_i - x_i^T \beta, i = 1, \ldots, n$, let $r_{\{i\}}(\beta)$ refer to the* i*th smallest residual in absolute value. Then*

$$\frac{1}{n} \sum_{i=1}^{n} r_{\{i\}} \left(\hat{\beta}_n^{(SLTS,W_n^1)} \right) = \frac{1}{n} \sum_{i=1}^{n} r_i \left(\hat{\beta}_n^{(SLTS,W_n^1)} \right) \to \mathsf{E}\, \varepsilon_i = 0 \tag{26}$$

and

$$\frac{1}{h_n} \sum_{i=1}^{h_n} r_{\{i\}} \left(\hat{\beta}_n^{(SLTS,W_n^2)} \right) \to \lim_{n \to \infty} \mathsf{E} \left(\varepsilon_i \cdot I\big(\varepsilon_i^2 \leq \varepsilon_{[h_n]}^2\big) \right) = 0 \tag{27}$$

as $n \to +\infty$. Similarly,

$$\frac{1}{n} \sum_{i=1}^{n} r_{\{i\}}^2 \left(\hat{\beta}_n^{(SLTS,W_n^1)} \right) = \frac{1}{n} \sum_{i=1}^{n} r_i^2 \left(\hat{\beta}_n^{(SLTS,W_n^1)} \right) \to \mathsf{var}\, \varepsilon_i = \sigma^2 \tag{28}$$

and

$$\begin{aligned}\frac{1}{h_n}\sum_{i=1}^{h_n} r^2_{\{i\}}\left(\hat\beta_n^{(SLTS,W_n^2)}\right) &\to \lim_{n\to\infty}\mathsf{E}\left(\varepsilon_i^2\cdot I\big(\varepsilon_i^2\le\varepsilon^2_{[h_n]}\big)\right)=\\ &= \frac{1}{\lambda}F_{\chi^2_3}\left(\Phi^{-1}\left(\frac{1+\lambda}{2}\right)\right)\cdot \mathsf{var}\,\varepsilon_i \qquad (29)\\ &= \frac{\sigma^2}{\lambda}\cdot F_{\chi^2_3}\left(\Phi^{-1}\left(\frac{1+\lambda}{2}\right)\right)\end{aligned}$$

as $n\to+\infty$, where $F_{\chi^2_d}$ represents the χ^2_d cumulative distribution function with d degrees of freedom and $F^{-1}_{\chi^2_d}$ the quantile function of χ^2_d distribution.

Proof: Assumption A guarantees that both estimators $\hat\beta_n^{(SLTS,W_n^1)}$ and $\hat\beta_n^{(SLTS,W_n^2)}$ are consistent—they converge to the true parameter vector β^0 in probability. Hence, the residuals $r_i\left(\hat\beta_n^{(SLTS,W_n^l)}\right)=\varepsilon_i+x_i^T\left(\hat\beta_n^{(SLTS,W_n^l)}-\beta^0\right)$ converge in probability to ε_i for $l=1,2$ and $i=1,\ldots,n$. Consequently, the first assertions (26) and (27) are an immediate result of the consistency of the LS and LTS estimators and of the strong law of large numbers (see Assumption A3). The same is true for (28) ($\mathsf{var}\,\varepsilon_i^2=\sigma^2$): by the strong law of large numbers, $\frac{1}{n\sigma^2}\sum_{i=1}^n r^2_{\{i\}}\left(\hat\beta_n^{(SLTS,W_n^1)}\right)=\frac{1}{n\sigma^2}\sum_{i=1}^n r^2_i\left(\hat\beta_n^{(SLTS,W_n^1)}\right)$ converges almost surely to the variance of the standard normal distribution $N(0,1)$,

$$\frac{1}{n\sigma^2}\sum_{i=1}^n r^2_{\{i\}}\left(\hat\beta_n^{(SLTS,W_n^1)}\right)=\frac{1}{n\sigma^2}\sum_{i=1}^n r^2_i\left(\hat\beta_n^{(SLTS,W_n^1)}\right)\to\int_{-\infty}^{\infty}x^2\cdot\phi(x)dx=1, \qquad (30)$$

where $\phi(x)$ represents the probability density function of $N(0,1)$. Thus, the only assertion to be proved is (29). Similarly to (30),

$$\begin{aligned}\frac{1}{h_n\sigma^2}\sum_{i=1}^{h_n} r^2_{\{i\}}\left(\hat{\beta}_n^{(SLTS,W_n^2)}\right) &\rightarrow \int_{\Phi^{-1}(\frac{1-\lambda}{2})}^{\Phi^{-1}(\frac{1+\lambda}{2})} x^2\cdot\frac{\phi(x)}{\Phi\left(\Phi^{-1}(\frac{1+\lambda}{2})\right)-\Phi\left(\Phi^{-1}(\frac{1-\lambda}{2})\right)}\mathrm{d}x\\ &= \frac{1}{\lambda}\int_{\Phi^{-1}(\frac{1-\lambda}{2})}^{\Phi^{-1}(\frac{1+\lambda}{2})} x^2\cdot\phi(x)\mathrm{d}x.\end{aligned}$$

We can transform the integral in the following way ($q_\lambda = \Phi^{-1}(\frac{1+\lambda}{2})$ is used for simplicity of notation):

$$\begin{aligned}\int_{\Phi^{-1}(\frac{1-\lambda}{2})}^{\Phi^{-1}(\frac{1+\lambda}{2})} x^2\cdot\phi(x)dx &= 2\int_0^{q_\lambda} x^2\cdot\frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}e^{-\frac{x^2}{2}}\mathrm{d}x\\ &= \int_0^{q_\lambda^2}\frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}t^{\frac{1}{2}}e^{-\frac{t}{2}}\mathrm{d}t\\ &= \int_0^{q_\lambda^2}\frac{1}{2^{\frac{3}{2}}\Gamma(\frac{3}{2})}t^{\frac{1}{2}}e^{-\frac{t}{2}}\mathrm{d}t\\ &= \int_0^{q_\lambda^2} f_{\chi_3^2}(x)\mathrm{d}x.\end{aligned}$$

This leads directly to

$$\frac{1}{h_n\sigma^2}\sum_{i=1}^{h_n} r^2_{\{i\}}\left(\hat{\beta}_n^{(SLTS,W_n^2)}\right) \rightarrow \frac{1}{\lambda}\int_0^{q_\lambda^2} f_{\chi_3^2}(x)\mathrm{d}x = \frac{1}{\lambda}F_{\chi_3^2}\left(\Phi^{-1}\left(\frac{1+\lambda}{2}\right)\right).$$

□

Proposition 3 describes the ratio between the variances of all regression residuals and the h_n smallest residuals (in absolute value), see (28) and (29). These h_n smallest residuals correspond to those observations that actually enter the objective function of the LTS estimator. The dependence of the ratio between the two variances is depicted in Figure 1.

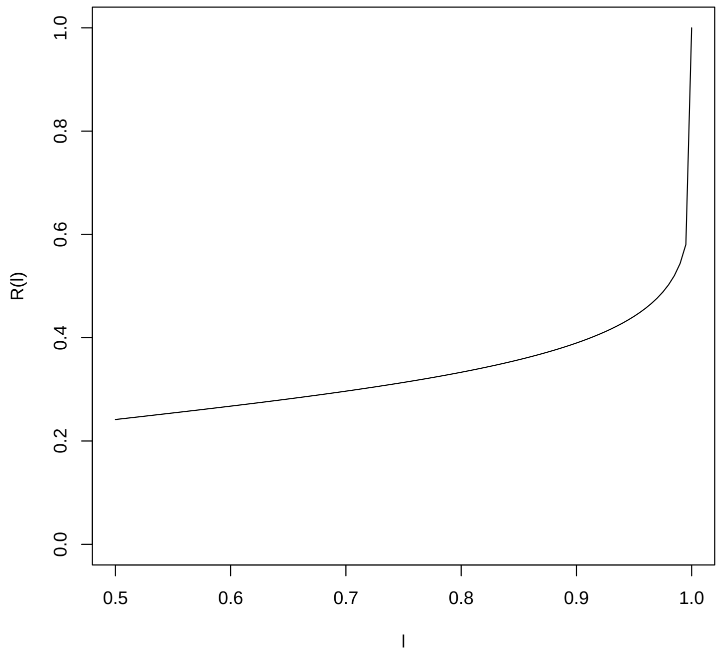

Figure 1: The ratio of variances (28) and (29) as a function of λ.

6 Computational aspects

Any practical computation of an estimate usually raises some further issues that need to be solved in additional to the theoretical problems. The choice of weights for SLTS can serve in our case as an important example. While we require only their positivity in the theory, the smallest weights in reality should not be chosen below $\varepsilon \cdot \frac{S_s(X_n,Y_n,w;\beta)}{\max_i |r_i(\beta)|}$, where $\varepsilon > 0$ is the smallest positive number such that $1 + \varepsilon > 1$ in a used computer representation—otherwise the residuals with such small weights cannot affect the minimized function. Nevertheless, most important is naturally the existence of an algorithm that computes the proposed SLTS estimate in an acceptable time and with an acceptable precision, see Section 6.1. Some specific choices of weights as well as possible schemes for adaptive choices of weights are discussed in Section 6.2.

6.1 Computation of SLTS for given weights

There are many ways to compute the SLTS estimates—one similar to the traditional way for computing LTS (see Rousseeuw and Van Driessen (1999)) and another one based on the so-called differential evolution. Both have their advantages, although the traditional way (Section 6.1.1) is more suitable for the computation within the classical linear regression model and is therefore described in more detail.

6.1.1 LTS-like approximation

First of all, let me briefly discuss the traditional strategy for determining the least trimmed squares estimates because it motivates the procedure I propose for computing SLTS. This strategy relies on the search through subsamples of size h and the consecutive least squares estimation: choose randomly an h-tuple of observations, apply the least squares method to it, and evaluate the residuals for all n observations given the estimated regression coefficients. Then select an h-tuple of data points with the smallest squared residuals and repeat the LS estimation for the selected h-tuple. If the sum of the h smallest squared residuals decreases, this step is repeated. When no further improvement can be found this way, a new subsample of h observations is randomly generated and the whole process is repeated. The search is stopped as soon as we get s times the same estimate or when we reach a pre-specified number of iterations. A more refined version of this algorithm suitable also for large data sets was proposed and described by Rousseeuw and Van Driessen (1999), who also provided theoretical arguments (the so-called **C**-step property) supporting the above outlined algorithm. The following lemma describes a similar property in the case of SLTS.

Lemma 2 *Let $(y_i, x_i)_{i=1}^n$ be a fixed realization of a random sample and $w = (w_1, \ldots, w_n)$ be a weighting vector, $w_{1n} \geq w_{2n} \geq \ldots \geq w_{nn} > 0$. Moreover, let $k_i(\beta) : \mathbb{R} \to \{1, \ldots, n\}$ be a function such that $k_i(\beta)$ is the index of the observation with the* i*th largest squared residual, $r^2_{k_i(\beta)}(\beta) = r^2_{[i]}(\beta)$ at β. Consider an arbitrary estimate $\hat{\beta}^0_n$ of the regression parameters and define weights $v_{k_i(\hat{\beta}^0_n)} = w_i$ for all $i = 1, \ldots, n$. Next, denote $\hat{\beta}^1_n$ as the weighted least squares estimator with weights $v_i, i = 1, \ldots, n$,*

$$\hat{\beta}^1_n = \hat{\beta}^{(WLS,v)}_n = \arg\min_{\beta \in B} \sum_{i=1}^n v_i r_i^2(\beta). \tag{31}$$

Then it holds for the SLTS objective function that

$$S_s(X_n, Y_n, w; \hat{\beta}^0_n) \geq S_s(X_n, Y_n, w; \hat{\beta}^1_n).$$

***Remark* 7** *The definition of weights v in Lemma 2 is the same as in Lemma 1.*

Proof: The property is almost trivial and is based on inequalities (7), (8), and (9) derived in the proof of Lemma 1:

$$S_s(X_n, Y_n, w; \hat{\beta}^0_n) = S_w(X_n, Y_n, v; \hat{\beta}^0_n) \geq S_w(X_n, Y_n, v; \hat{\beta}^1_n) \geq S_s(X_n, Y_n, w; \hat{\beta}^1_n).$$

□

Lemma 2 offers a way to improve the approximation of the SLTS estimate. Having an initial estimate $\hat{\beta}^0_n$, we can define weights v^1 as described in Lemma 2 and compute the weighted least squares estimate $\hat{\beta}^1_n$, which attains the same or a better value of the SLTS objective function than the initial $\hat{\beta}^0_n$. Next, we can use $\hat{\beta}^1_n$ in place of the initial estimate, define new weights v^2 and compute the WLS estimate $\hat{\beta}^2_n$, which again improves $S_s(X_n, Y_n, w; \beta)$. Repeating these steps yields an iterative process for the sequence $\hat{\beta}^0_n$, $\hat{\beta}^1_n$, $\hat{\beta}^2_n, \ldots$ such that $S_s(X_n, Y_n, w; \hat{\beta}^k_n) \geq$

$S_s(X_n, Y_n, w; \hat{\beta}_n^{k+1})$ for $k = 1, 2, \ldots$. The process stops when $S_s(X_n, Y_n, w; \hat{\beta}_n^k) = S_s(X_n, Y_n, w; \hat{\beta}_n^{k+1})$ for some $k = k^e$ (the sequence always converges and always has a minimum as it is a decreasing sequence of a finite number of nonnegative quantities). Unfortunately, this is not sufficient for $\hat{\beta}_n^{k^e}$ to be the global minimum of the SLTS objective function. Therefore, more such sequences are needed and the sequence that converges to the smallest value of S_s should be kept. This concept leads to the proposal of the following algorithm (we assume that data X_n, Y_n and weights W_n are given and $K_s \in \mathbb{N}$ is a fixed integer):

SLTS Algorithm:

1. Draw a random permutation $\Pi_n = (\pi_1, \ldots, \pi_n)$ of $\{1, \ldots, n\}$.

2. Define weights $v = (v_1, \ldots, v_n)$, $v_i = w_{\pi_i}$ for all $i = 1, \ldots, n$.

3. Compute the weighted least squares estimate $\hat{\beta}_n^0$ with weights v and set $k = 0$.

4. Sort the absolute values of residuals $r_i(\hat{\beta}_n^k)$, which give rise to a new permutation $\Pi_n = (\pi_1, \ldots, \pi_n)$ such that

$$\left| r_{\pi_1}\left(\hat{\beta}_n^k\right) \right| \leq \left| r_{\pi_2}\left(\hat{\beta}_n^k\right) \right| \leq \cdots \leq \left| r_{\pi_n}\left(\hat{\beta}_n^k\right) \right| .$$

5. Define weights $v = (v_1, \ldots, v_n)$, $v_i = w_{\pi_i}$ for all $i = 1, \ldots, n$.

6. Compute the weighted least squares estimate $\hat{\beta}_n^{k+1}$ with weights v.

7. If $S_s(X_n, Y_n, w; \hat{\beta}_n^k) > S_s(X_n, Y_n, w; \hat{\beta}_n^{k+1})$, set $k = k + 1$ and continue at point 4. Otherwise go to point 8.

8. If $S_s(X_n, Y_n, w; \hat{\beta}_n^k) \leq S_s(X_n, Y_n, w; \hat{\beta}_n^{k+1})$, compare the value $S_s(X_n, Y_n, w; \hat{\beta}_n^k)$ with the values obtained from previously created sequences. If it is smaller, continue at point 1. Otherwise, check how many sequences have been tried without improving the global minimum of $S_s(X_n, Y_n, w; \beta)$. If less than K_S, continue at point 1; otherwise stop.

I implemented this algorithm in the S language, and as confirmed by many simulations, this algorithm converges fast enough for smaller data sets (no more than several thousands of observations). Its speed can be further improved in a similar way as proposed for LTS in Rousseeuw and Van Driessen (1999), but this is not the aim of this paper.

6.1.2 Differential evolution approach

Another method suitable for SLTS approximation is one of the global optimization methods, *differential evolution*, developed by Storn and Price (1995). Differential evolution is a direct search method that was recently found to be an efficient method for optimizing general real-valued functions (see Storn and Price (1996)). It uses a population of p-dimensional parameter vectors, which is initially randomly generated, and, in the simplest version, "generates new parameter vectors by adding the weighted difference between two population vectors to a third vector. If the resulting vector yields a lower objective function value than a predetermined population member, the newly generated vector replaces the vector, with which it was compared, in the next generation" (Storn (1996), page 1). There are many variants and refinements of this basic principle, but their discussion is outside of the scope of this paper. The main advantage of differential evolution is, besides its simplicity and generality (it does not require any special properties of the objective function), the parallel nature of the search (the algorithm works with a population of parameter vectors), because it suits the "combinatorial" nature of the (S)LTS objective function well.

The most important benefit of the differential-evolution algorithm is that it requires only evaluation of the objective function. Therefore, it is suitable for more complicated models (e.g., the application of SLTS in nonlinear models) or in the case of simultaneous optimization over the space of the regression parameters and the parameters controlling weight vectors. To check whether this method is really suitable for the computation of SLTS, I compared its performance in the case of the linear regression model with the algorithm described in Section 6.1.1 both for simulated and real data sets.[18] In all cases, the estimates obtained by the differential-evolution algorithm (schemes DE/rand/1 and DE/best/1, see Storn (1996)) are equal to those obtained by the other algorithm or even slightly better. On the other hand, the speed of the differential algorithm is lower when used in linear regression models, mostly two to three times than the SLTS algorithm described in Section 6.1.1.

6.2 Adaptive choice of weights

Having all the theoretical results and working computational procedures in hand, it is now possible to discuss the adaptive choice of weights for SLTS (for a fixed choice of weights, one can simply use the asymptotic results in Section 5.1 and the algorithms described in Section 6.1.1). I first describe the adaptive choice of weights theoretically (based on an abstract decision rule). Second, I propose two decision rules and combine them together into one final procedure for the adaptive choice of weighting schemes.

[18] The implementation of the variants of the differential-evolution algorithm is based on the source code written by the authors of the method—Storn and Price (1995).

The choice of weights for SLTS and the corresponding theoretical results derived to this point are limited only by Assumption A (Section 5) and Assumption W (Section 5.2). However, to exemplify the results and procedures discussed in this section, it is beneficial to demonstrate them on weights generated by functions from a specific class. For this purpose, I choose weighting schemes generated by logistic functions (they were introduced already in Section 5.2, equation (24), and they will be used in Section 7 as well):

$$w_i = f_\lambda(\frac{2i-1}{n};\omega), \text{ where } f_\lambda(x;\omega) = \frac{1}{1+e^{\omega(x-\lambda)}} \bigg/ \int_0^1 \frac{1}{1+e^{\omega(x-\lambda)}}\mathrm{d}x \qquad (32)$$

for $i = 1,\ldots,n$, $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ is a fixed trimming constant (equivalent to λ in Assumption W), and $\omega \geq 0$ is the parameter controlling the shape of the generating function $f_\lambda(x;\omega)$. Apparently, this weighting scheme satisfies Assumption

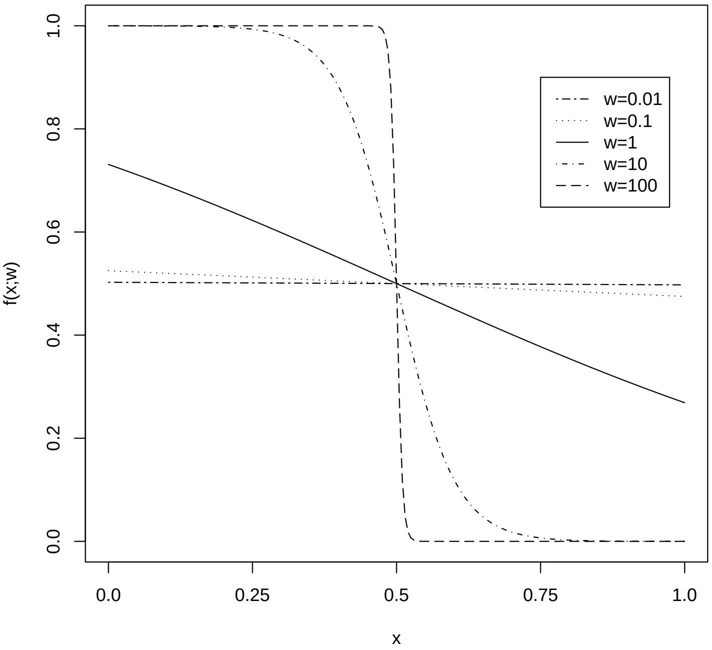

Figure 2: Logistic generating functions for $\omega = 0.01, 0.1, 1, 10, 100$.

W, including convergence to the least squares weights ($\omega \to 0$) and to the least trimmed squares weights ($\omega \to +\infty$). One can see the shape of function $f_\lambda(x;\omega)$ for different ω in Figure 2. The advantage of the presented logistic weights is that they satisfy Assumption W including the optional part and that they react quite sensitively to changes of the weighting parameter ω within a relatively small interval, but on the other hand, values outside of this interval produce only negligible changes in the estimates.

Let me describe now how an adaptive procedure for choosing a weight scheme works. For weighting schemes generated by a function $f(x;\omega)$ satisfying Assumption W, it holds that the corresponding SLTS estimator (see Figure 2)

- is more robust for $\omega \to \omega_2$ because the largest residuals are assigned very small weights (decreasing with ω approaching ω_2),
- is more efficient for $\omega \to \omega_1$ because all residuals have similar weights, none are extremely downweighted, and all observations influence significantly the SLTS objective function.

Altogether, decreasing the parameter ω increases efficiency and decreases the robustness of SLTS and vice versa. Therefore, an adaptive choice of weights can work in the following way: it starts with the highest possible ω (closest to ω_2) to obtain the most robust estimate. Given a data set, we do not know whether this maximum level of robustness is necessary at all, so the next step is to decrease ω. A decrease in ω improves the efficiency of the estimator (more information from data is used), but because it also decreases the robustness of SLTS, it is possible that the estimate is for lower values of ω already adversely affected by contamination or other data problems. Hence, we need a decision criterion that tells us how much we can decrease ω without threatening the robustness of the estimator. Having such a decision rule, the adaptive search for an optimal ω simply has to start with ω close to ω_2 and then to decrease ω toward ω_1 until the decision rule indicates that ω is already too low and the corresponding estimate not sufficiently robust. Thus, we obtain as low ω as possible, which means as efficient an estimator as possible. So, the aim of this section is to construct a decision criterion

$$D\left(\omega, S_s\left(X_n, Y_n, w(\omega); \hat{\beta}_n^{(SLTS,w)}\right), r_i\left(\hat{\beta}_n^{(SLTS,w)}\right)\right)$$

that indicates whether the current value of parameter ω is acceptable (i.e., does not lower robustness of the estimate too much) or not. Such a decision rule can be based either on the values of the objective function S_s or the regression residuals r_i and their statistics.

First, let us summarize what we know about the SLTS objective function as a function of weights: it is decreasing, it is bounded by $S_s(X_n, Y_n; \hat{\beta}_n^{(LS)})$ from above and by $S_s(X_n, Y_n, w_{LTS,h}; \hat{\beta}_n^{(LTS,h)})$ from below ($w_{LTS,h} = \frac{n}{h_n}(1_{h_n}, 0_{n-h_n})$, whereby $h_n = [\lambda n]$ and the multiplication by $\frac{n}{h_n}$ normalizes weights (see Assumption W and Lemma 3). Further, the ratio

$$R = \frac{S_s(X_n, Y_n; \hat{\beta}_n^{(LS)})}{S_s(X_n, Y_n, w_{LTS,h}; \hat{\beta}_n^{(LTS,h)})}$$

converges for normally distributed errors to

$$R_N = \frac{\lambda}{F_{\chi_3^2}\left(F_{\chi_1^2}^{-1}(\lambda)\right)}$$

(see Proposition 2), which for the choice $\lambda = 1/2$ results asymptotically in

$$R_N = \frac{1}{2} \Big/ F_{\chi_3^2}\left(F_{\chi_1^2}^{-1}(1/2)\right) \doteq 7.01.$$

Next, let us compare this outcome with some simulation results. Several estimates of ratio R are presented in Table 1. They come from a Monte Carlo simulation for the linear regression model $y_i = 0.3 + x_i + \varepsilon_i$, where $x_i \sim N(0, 100)$ and $\varepsilon_i \sim N(0, 4)$; the sample size is $n = 100$ and the results are based on 1000 simulations. Clearly, estimates for cases with normally distributed errors are a little bit higher than the asymptotically derived value. Nevertheless, most important is a drastic increase in R whenever outliers appear in the data[19] (the value for one percent of outliers is smaller mainly because this case represents only one randomly generated outlying observation ($n = 100$) which often does not outlie at all). It also seems that the value for the Student distribution t_3 is too large compared to the values for the normal distribution, but this is completely correct—if errors are distributed according to t_d with small degrees of freedom d,

Error distribution	Outliers (%)	$\hat{R}$	σ_R
$N(0,1)$	0	8.264	1.610
$N(0,4)$	0	8.251	1.782
$U\langle -1,1\rangle$	0	5.813	1.045
t_3^*	0	16.70	1.680
$N(0,1)^*$	1	9.875	1.212
$N(0,1)^*$	5	269.1	8.410
$N(0,1)^*$	15	1231.1	13.91

Table 1: Estimates of R: Simulation for $y_i = 0.3 + x_i + \varepsilon_i$ with various error distributions.

Entries in rows marked by * correspond to the median and the median absolute deviation, which were used instead of mean and standard deviation because of some extreme results in simulations concerning the least squares estimator.

[19] Although I used a simple linear regression, the results are the same for multiple regression models.

then the least squares estimator loses its efficiency and behaves as if the data were slightly contaminated (more information on this topic is presented in Section 7). Thus, we can conclude that the ratio R of the objective function of SLTS for $\omega \to 0$ and $\omega \to +\infty$ indicates quite well how much the data are contaminated, or in other words, how probable it is that the least squares estimator misbehaves.

Given these results, we can now propose the following decision rule (function $S_s(X_n, Y_n, w(\omega); \hat{\beta}_n^{(SLTS,w)})$ is further referred to by $S_s^*(\omega)$ for simplicity):

- start from a reasonably high ω_0[20] (e.g., $\omega_0 = 50$ for our logistic weights), estimate SLTS and remember the value of the objective function $S_s^*(\omega_0)$ at this point;

- gradually decrease the value of ω and stop when the estimated objective function $S_s^*(\omega)$ is greater than $M \cdot S_s^*(\omega_0)$, where $M = cR_N$ and R_N is the asymptotic value of the ratio R derived at the beginning of this section (cR_N with $c > 1$ can be used instead of R_N to allow for small sample deviations from the asymptotic value).

We showed that an increase in $S_s^*(\omega)$ indicates quite well whether data are contaminated. However, the described decision rule can work quite well in practice only for data that are not too contaminated. In general, it is possible that the estimate is already affected too much by contamination when we stop decreasing parameter ω (remember, $S_s^*(\omega) > MS_s^*(\omega_0)$, where $M \geq R_n > 7$). Therefore, the above rule should be complemented by another rule which is able to cope with highly contaminated data and will stop decreasing ω in time.

[20] By reasonably high ω_0 we understand ω_0 as close to ω_2 from Assumption W as possible, but such that it does not result in complete trimming numerically, that is, trimming caused by the limited computer precision (see Section 6).

Such a rule can be constructed based on regression residuals: we assume that the initial estimate corresponding to ω_0 is consistent and we know that the principle of most robust estimator is "to constrain the influence of observations with extremely large residuals on the estimate." Hence, we can construct estimates of location and scale for the consistently estimated residuals computed at ω_0 (most robust choice) and then compare them with the weighted residuals for a current ω to see whether some of them are already too large and thus have too big of an influence on the objective function and on the estimate itself. This decision rule can be summarized as follows:

- start from a reasonably high ω_0 (e.g., $\omega_0 = 50$ for our logistic weights), estimate SLTS and compute corresponding regression residuals along with robust estimates of their mean m_0 and variance v_0;
- gradually decrease the value of ω and stop when some weighted regression residuals $\sqrt{w_i} r_i(b)$ do not lie inside the interval $\langle m_0 - Cv_0, m + Cv_0 \rangle$ anymore (weighted regression residuals are used because they describe the effect of observations on the SLTS objective function).

The check for weighted residuals is based on the following principle. The mean value of residuals $r_i(b_0)$ (consistently estimated for ω_0) is m_0 and their variance is v_0. Hence, $\langle m_0 - Cv_0, m + Cv_0 \rangle$ represents a kind of confidence interval, and for a suitable choice of C, residuals should lie inside of this interval with a probability close to 1. It is, of course, possible that some residuals can lie outside of this interval, but such residuals should not have a bigger influence on the objective function of the SLTS estimate because they are most probably outliers.

Now, the crucial question is the choice of constant C for the confidence interval. Assuming normal distribution of the error term, it is tempting to choose

$C \in \langle 2.5, 3.0\rangle$ as this corresponds to 99%–99.9% confidence intervals. However, this would destroy the robustness of the SLTS estimator. Hence, in the same way as for LTS, we assume that only some fraction $\lambda \in \left\langle \frac{1}{2}, 1\right\rangle$ of observations closely follow a specified regression model. This is also reflected by Assumption W: the generating function is chosen so that weights for the $[\lambda n]$ smallest residuals increases for more robust choices of ω and the other weights converge to zero. Therefore, only these smallest residuals fully affect the SLTS objective function and all other observations are downweighted. Consequently, the adaptive decision rule should follow the same strategy: the $[\lambda n]$ smallest residuals can fully influence the SLTS objective function and the influence of all other residuals should be limited so that it will not be greater than the influence of these $[\lambda n]$ smallest residuals. This means that $\langle m_0 - Cv_0, m + Cv_0\rangle$ should represent the confidence interval for the $[\lambda n]$ smallest residuals and all greater residuals have to be downweighted so that they fall into this interval. Hence, assuming that the error term has normal distribution, constant C can be written as

$$C = D \cdot V_N(\lambda) = D \cdot \frac{1}{\lambda} \cdot F_{\chi_3^2}\left(\Phi^{-1}\left(\frac{1+\lambda}{2}\right)\right),$$

where $D \in \langle 2.5, 3.0\rangle$ is a constant we would use for the standard confidence interval of a normally distributed random variable and

$$V_N(\lambda) = \frac{1}{\lambda} \cdot F_{\chi_3^2}\left(\Phi^{-1}\left(\frac{1+\lambda}{2}\right)\right) \tag{33}$$

is the ratio of variances of the $[\lambda n]$ smallest residuals (in absolute value) and all residuals; this is derived in Lemma 3. For $\lambda = 1/2$, we get

$$V_N(\frac{1}{2}) = 2F_{\chi_3^2}\left(\Phi^{-1}\left(\frac{3}{4}\right)\right) \doteq 0.24.$$

Finally, let us combine both proposed decision rules with a general principle of the adaptive choice of SLTS weights. As a result, we obtain this adaptive-choice procedure (examples are always meant for the case of logistic generating function and weights, see (32)):

Adaptive choice 1 (one parameter)

1. Set the initial value of the weighting parameter ω to a reasonably high ω_0; for example, $\omega = \omega_0 = 50$.

2. Compute the SLTS estimate b_0 for ω_0, evaluate $S_s^*(\omega_0)$ and the characteristics of regression residuals: $m_0 = \text{med}_i \, r_i(b_0)$ and $v_0 = \text{MAD}_i \, r_i(b_0)$.

3. Decrease the weighting parameter, for example, $\omega = 0.8\omega$. If $\omega < \omega_1$, set $\omega = \omega_1$ and stop (ω_1 is the lower bound for ω).

4. Compute the SLTS estimate b for the new ω and evaluate $S_s^*(\omega)$.

5. If $S_s^*(\omega) > c \cdot R_N \cdot S_s^*(\omega_0)$, return to the previous value of ω and stop.

6. Compute the weighted regression residuals $\frac{\sqrt{w_i}}{\max_i \sqrt{w_i}} \cdot r_i(b)$ and check whether all of them are inside the interval $\langle m_0 - D \cdot V_N(\lambda) \cdot v_0, m + D \cdot V_N(\lambda) \cdot v_0 \rangle$. If not, return to the previous value of ω and stop. Otherwise continue at point 3.

As a result, we obtain some ω, which define the optimal SLTS estimator within the used class of smoothing schemes for a given data set. In the following text, we refer to SLTS used with a smoothing scheme chosen by means of "Adaptive choice 1" as SLTS-AC1.

***Remark* 8** *Constants c and D determine the maximum accepted increase of $S_s^*(\omega)$ and the width of the confidence interval for the $[\lambda n]$ smallest residuals,* *respectively. Reasonable values are $c \in \langle 1, 2\rangle$ and $D \in \langle 2, 3\rangle$, as discussed above. The effects of the choice of D are also studied in Section 7.1.*

***Remark* 9** *There is one more important issue to be discussed. The algorithm for the adaptive choice of a weighting scheme, which I propose in this section, is based on theoretical results derived for normally distributed errors. Although this might seem to be non-robust, it is in fact robust. The least squares estimators generally perform best under errors having the normal distribution, and moreover, they are easily affected by observations with large residuals. Therefore, the decision rules discussed above are designed so that they are optimized for normal errors and they stop too early if the error term has a distribution with heavier tails or outliers are present. This implies that ω stays closer to ω_2 (more robust choice) and the SLTS-AC1 estimator "prefers" more robust, although less efficient estimates to efficient, but rather non-robust ones.*

On the other hand, this implies that an undersmoothing can occur (actually from two reasons: either the optimal smoothing is not reached—the adaptive procedure stops too early, or there is a better smoothing in a family of smoothing schemes not taken into account). To find the optimal decision rule and smoothing class, it is necessary to study the behavior of SLTS not only at a given distribution function, but also in its neighborhood. Unfortunately, SLTS under such distributional assumptions is hard to study because the asymptotic results concerning LTS that I used throughout the analysis of SLTS are not readily available under these assumptions. However, I will argue that the proposed adaptive procedures, although sub-optimal in this sense, are superior to the existing solutions in many aspects, see simulations in Section 7.

The proposed adaptive choice of SLTS weights describes a situation when a weighting scheme is controlled only by one parameter. This is not always optimal. We can consider, for instance, the logistic weighting scheme used throughout this section: it is generated by functions $f_\lambda(x;\omega)$, where the trimming constant λ is a fixed number, $\lambda \in \left\langle \frac{1}{2}, 1 \right\rangle$ (it is depicted in Figure 2). For $\omega \to 0$, it gives (almost) the same weight to all observations; for $\omega \to \infty$, the weights assigned to $[(1-\lambda)n]$ largest residuals converges to zero. Now, from the shape of the function, it is obvious that if more observations have to be significantly downweighted (let us say more than 1–5%), then all $[(1-\lambda)n]$ observations with largest residuals are significantly downweighted as well. This means that most of the information of all $[(1-\lambda)n]$ observations with largest residuals is not used in the presence of any contamination, which in turn leads to a loss of efficiency. Apparently, this inefficiency can be fixed when it is possible to adjust the parameter λ as well. Then, adaptively choose two parameters—λ and ω—and the logistic generating functions have to be considered as a function of these two parameters:

$$f(x;\lambda,\omega) = f_\lambda(x;\omega) = \frac{1}{1+e^{\omega(x-\lambda)}} \Bigg/ \int_0^1 \frac{1}{1+e^{\omega(x-\lambda)}} \, .$$

The adaptive choice of two parameters λ and ω can be done relatively easily using the same decision rules that were used for the adaptive choice of one parameter ω. Start again with the most robust choice: $\lambda = \lambda_0 = \frac{1}{2}$ and $\omega = \omega_0$. As the next step, find the optimal value for λ (i.e., the amount of observations that does not have to be downweighted at all) without changing ω—increase λ and stop when the decision rules indicate to do so. Finally, fix λ and start to search for the optimal value of ω in the same way as in Adaptive choice 1. The complete adaptive procedure for the two parameters can be summarized as follows (examples are again provided for the logistic generating functions):

Adaptive choice 2 (two parameters)

1. Set the initial value of the weighting parameters $\lambda = \lambda_0 = \frac{1}{2}$ and ω to a reasonably high ω_0; for example, $\omega = \omega_0 = 50$.

2. Compute the SLTS estimate b_0 for λ_0, ω_0, evaluate $S_s^*(\lambda_0, \omega_0)$ and the characteristics of regression residuals: $m_0 = \mathrm{med}_i\, r_i(b_0)$ and $v_0 = \mathrm{MAD}_i\, r_i(b_0)$.

3. Increase the trimming constant λ and keep parameter ω fixed (for example, $\lambda = \lambda + 0.05$). If $\lambda > 1$, set $\lambda = 1$ and stop (1 is the upper bound for λ).

4. Compute the SLTS estimate b for the new λ and ω and evaluate $S_s^*(\lambda, \omega)$.

5. If $S_s^*(\lambda, \omega) > c \cdot R_N \cdot S_s^*(\lambda_0, \omega_0)$, return to the previous value of λ and continue at point 7.

6. Compute the weighted regression residuals $\frac{\sqrt{w_i}}{\max_i \sqrt{w_i}} \cdot r_i(b)$ and check whether all of them are inside the interval $\langle m_0 - D \cdot V_N(\lambda) \cdot v_0, m + D \cdot V_N(\lambda) \cdot v_0 \rangle$. If not, return to the previous value of λ and continue at point 7. Otherwise continue at point 3.

7. Decrease the weighting parameter ω (λ is already fixed at its optimal level); for example, $\omega = 0.8\omega$. If $\omega < \omega_1$, set $\omega = \omega_1$ and stop (ω_1 is the lower bound for ω).

8. Compute the SLTS estimate b for the new λ and ω and evaluate $S_s^*(\lambda, \omega)$.

9. If $S_s^*(\lambda, \omega) > c \cdot R_N \cdot S_s^*(\lambda_0, \omega_0)$, return to the previous value of ω and stop.

10. Compute the weighted regression residuals $\frac{\sqrt{w_i}}{\max_i \sqrt{w_i}} \cdot r_i(b)$ and check whether all of them are inside the interval $\langle m_0 - D \cdot V_N(\lambda) \cdot v_0, m + D \cdot V_N(\lambda) \cdot v_0 \rangle$. If not, return to the previous value of ω and stop. Otherwise continue at point 7.

At the end of this algorithm for the adaptive choice of two parameters λ and ω, we obtain two values, ω and λ, which define the optimal SLTS estimator within the used class of smoothing schemes for a given data set. The main difference to SLTS-AC1 is that we have extended the class of smoothing schemes from $\{f_\lambda(\cdot;\omega) : \omega \in \mathbb{R}_+\}$ (λ fixed) to $\left\{f(\cdot;\lambda,\omega) : \lambda \in \left\langle\frac{1}{2},1\right\rangle, \omega \in \mathbb{R}_+\right\}$. Once again, we refer to SLTS using a weighting scheme found via "Adaptive choice 2" as SLTS-AC2. The simulations using the described adaptive procedures are presented in Section 7.

7 Simulations

In Section 6.2, we constructed adaptive choice procedures for the SLTS estimator, which allow us to select an optimal set of weights from a family of weighting schemes parameterized by one or two real parameters. As an example, we used weights generated by standardized logistic functions (32). In this section, I would like to demonstrate finite sample properties of the SLTS estimator with weights generated by logistic functions with one adaptively chosen parameter in Section 7.1 (SLTS-AC1) and with two parameters in Section 7.2 (SLTS-AC2). Please note that, despite the limitation to only one smoothing scheme, the qualitative results presented later in this section are valid also for some other weighting schemes (e.g., one generated by the cumulative distribution function with polynomial tails). Finally, I examine the effect of misspecification of categorical variables on the LS, RDL_1, and SLTS estimators in Section 7.3.

Before discussing the simulation results, let me describe the models used in Monte Carlo simulations. First, for most simulations, I use the linear regression model

$$y_i = 0.3 + x_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{34}$$

where x_i is a continuously distributed random variable, $x_i \sim N(0, 10)$; the error term ε_i has a continuous distribution, for example, normal, Student, or exponential. Continuous random variables are used in many cases so that it is possible to compare SLTS and LTS. Second, for simulations involving both continuous and discrete variables, I use

$$y_i = 0.3 + x_i - 1.5d_i + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{35}$$

where $x_i \sim N(0, 10)$ and $d_i \sim Bi(0.5, 1)$. Both models (34) and (35) are sufficiently simple, enable a comparison of SLTS with other existing estimators, and most importantly, the simulation results are qualitatively the same as for more complicated models. Finally, some simulations study the effects of contamination on the estimators. In these cases, contamination is simulated as a uniform random noise. This is actually one of the most favorable cases for the RDL$_1$ estimator because it treats observations only according to their robust distance from the center of the data cloud. On the other hand, LTS and SLTS treat any type of observations and any kind of contamination in the same way, so it does not matter so much for the simulations, which type of contamination we simulate.

7.1 Adaptive choice with one parameter

The simulation results presented in this section are for models (34) and (35). The results are in all cases based on 1000 simulations and samples consisting of 100 observations. Nevertheless, I obtained the same qualitative results for sample sizes ranging from 50 to 500 observations. Further, I present here results for the least squares, LTS with trimming constant $h = [n/2] + [(p+1)/2]$, SLTS-AC1 with logistic weights (see Section 6.2), and RDL$_1$ estimators—first, under different error distributions, later, under contamination.

Now, the use of the adaptive-choice algorithm deserves one additional note. One of the decision rules discussed in Section 6.2 checks whether all weighted residuals belong to a confidence interval $\langle m_0 - Cv_0, m_0 + Cv_0 \rangle$. For example, for normally distributed errors, we obtain the 99% confidence interval for $C = 2.58$. However, we argued that it is necessary to construct this confidence interval only for the $[\lambda n]$ smallest residuals in order to preserve robustness of the SLTS estimator. Therefore, we should set $C = D \cdot V_N(\lambda)$. However, to see the effect of such a choice, simulations are performed for a range of values—from 3.0 to $0.72 = 3.0 \cdot V_N(0.5)$.

Estimator	Parameter	Coefficient	$\varepsilon \sim N(0,1)$		$\varepsilon \sim t_3$		$\varepsilon \sim Exp(1)$	
	C		Mean	Var	Mean	Var	Mean	Var
LS		Intercept	0.290	0.099	0.294	0.178	0.305	0.139
LS		Slope	0.998	0.033	1.002	0.057	1.001	0.045
SLTS	2.5	Intercept	0.290	0.099	0.297	0.135	0.303	0.117
SLTS	2.5	Slope	0.998	0.033	1.002	0.043	1.001	0.038
SLTS	1.5	Intercept	0.290	0.102	0.299	0.139	0.303	0.112
SLTS	1.5	Slope	0.999	0.035	1.002	0.043	1.001	0.035
SLTS	1.0	Intercept	0.289	0.110	0.299	0.149	1.001	0.116
SLTS	1.0	Slope	0.999	0.038	1.002	0.046	0.303	0.036
SLTS	0.75	Intercept	0.289	0.119	0.299	0.157	1.001	0.120
SLTS	0.75	Slope	0.999	0.042	1.002	0.049	0.302	0.038
LTS		Intercept	0.286	0.278	0.292	0.248	0.302	0.176
LTS		Slope	0.994	0.089	1.001	0.079	1.003	0.057
RDL$_1$		Intercept	0.290	0.129	0.300	0.150	0.303	0.119
RDL$_1$		Slope	0.997	0.049	1.000	0.051	1.000	0.043

Table 2: Simulations for clear data sets of size $n = 100$ and SLTS-AC1.

Entries in column "Parameter" indicate which confidence interval for residuals was used for the decision rule within the algorithm Adaptive choice 1: $\langle m_0 - C \cdot v_0, m + C \cdot v_0 \rangle$, where $m_0 = \text{med}_i\, r_i(b_0)$, $v_0 = \text{MAD}_i\, r_i(b_0)$, and b_0 is the initial (most robust) estimate.

The first set of simulations studies the behavior of the estimators for a clean data set (no contamination) and model (34) under different error distributions, namely, the standard normal distribution $N(0,1)$, the Student distribution t_3 with 3 degrees of freedom, and the exponential distribution with parameter 1. The simulation results are presented in Table 2 (I obtained very similar results also for distributions $N(0,\sigma^2), \sigma^2 \in \left(\frac{1}{4}, 4\right)$ and $t_d, d \in \{1,\ldots,10\}$). In all cases and for all distributions, the estimators provide consistent results. For normal distribution $N(0,1)$, the least square method is the most efficient one (measured by variance of the estimate) with SLTS closely following it. For SLTS, a more strict decision rule (i.e., lower C) leads to higher robustness and lower efficiency. The performance of RDL_1 is a bit weaker, but still much better than that of the LTS estimator. For the Student distribution t_3, the final picture is quite similar with one exception: the variance of the least squares estimator increases in such a way that LS performs worse than all robust estimators except for LTS. This documents that robust estimators can provide more efficient estimates than the least squares in situations when the least squares estimator is consistent, but the error distribution has heavier tails than the normal distribution. Finally, the exponential distribution is presented as well, because it represents the optimal case for estimators minimizing the sum of absolute values of residuals. In this last case, the least squares estimator is (besides LTS) the least efficient estimator. Moreover, SLTS exhibits about the same or even better efficiency than RDL_1.

The second simulation repeats the first one for the case of normally distributed errors, but a dummy variable is included in model (35). The results are summarized in Table 3 and they are quite similar to those described in the last paragraph. The main conclusion is that the simulation confirms that SLTS can cope with discrete explanatory variables as well as with continuous ones.

Estimator	Parameter	Coefficient	$\varepsilon \sim N(0,1)$	
	C		Mean	Var
LS		Intercept	0.295	0.142
LS		Slope	0.999	0.033
LS		Dummy	-1.494	0.203
SLTS	2.5	Intercept	0.294	0.142
SLTS	2.5	Slope	0.999	0.034
SLTS	2.5	Dummy	-1.494	0.203
SLTS	1.0	Intercept	0.291	0.162
SLTS	1.0	Slope	0.999	0.039
SLTS	1.0	Dummy	-1.490	0.233
SLTS	0.75	Intercept	0.290	0.180
SLTS	0.75	Slope	0.999	0.043
SLTS	0.75	Dummy	-1.489	0.258
RDL$_1$		Intercept	0.290	0.184
RDL$_1$		Slope	1.000	0.050
RDL$_1$		Dummy	-1.489	0.262

Table 3: Simulations with one dummy variable for clear data sets of size $n = 100$ and SLTS-AC1.

Entries in column "Parameter" indicate which confidence interval for residuals was used for the decision rule within the algorithm Adaptive choice 1: $\langle m_0 - C \cdot v_0, m + C \cdot v_0 \rangle$, where $m_0 = \text{med}_i\, r_i(b_0)$, $v_0 = \text{MAD}_i\, r_i(b_0)$, and b_0 is the initial (most robust) estimate.

The third set of simulations studies the behavior of the estimators again using model (34) and normally distributed errors, but a positive amount of contamination is present in this case. Three cases presented in Table 4 correspond to contamination levels 1%, 10%, and 40% (this means that the respective amount of observations is replaced by random noise). To indicate which estimates are significantly biased, I test the one-sided hypothesis that the slope parameter equals its true value (all estimators have asymptotically normal distribution). The one-sided test is used since the simulated contamination leads to a bias towards zero. The estimates for which we reject this hypothesis are marked. First, the least squares estimator seems to be biased a bit already for 1% contamination and it does not provide any reasonable results for higher levels of contamination (the

Estimator	Parameter	Coefficient	Cont. 1%		Cont. 10%		Cont. 40%	
	C		Mean	Var	Mean	Var	Mean	Var
LS		Intercept	0.294	0.214	0.242	0.552	0.317	0.958
LS		Slope	0.849	0.200	0.292^c	0.234	0.077^c	0.142
SLTS	2.5	Intercept	0.295	0.159	0.311	0.205	0.333	0.379
SLTS	2.5	Slope	0.981	0.059	0.881^a	0.097	0.589^a	0.258
SLTS	1.5	Intercept	0.296	0.178	0.316	0.208	0.338	0.321
SLTS	1.5	Slope	0.990	0.059	0.935	0.082	0.649^a	0.257
SLTS	1.0	Intercept	0.297	0.192	0.317	0.217	0.323	0.265
SLTS	1.0	Slope	0.993	0.061	0.958	0.078	0.756	0.242
SLTS	0.75	Intercept	0.298	0.198	0.318	0.219	0.314	0.243
SLTS	0.75	Slope	0.995	0.063	0.964	0.077	0.819	0.218
LTS		Intercept	0.296	0.279	0.322	0.271	0.299	0.208
LTS		Slope	1.003	0.087	1.000	0.080	0.993	0.080
RDL_1		Intercept	0.297	0.134	0.307	0.138	0.313	0.195
RDL_1		Slope	0.999	0.047	0.987	0.048	0.903^a	0.075

Table 4: Simulations for contaminated data sets of size $n = 100$ and SLTS-AC1.

Entries in column "Parameter" indicate which confidence interval for residuals was used for the decision rule within the algorithm Adaptive choice 1: $\langle m_0 - C \cdot v_0, m + C \cdot v_0 \rangle$, where $m_0 = \mathrm{med}_i\, r_i(b_0)$, $v_0 = \mathrm{MAD}_i\, r_i(b_0)$, and b_0 is the initial (most robust) estimate. Constant C actually corresponds to $D \cdot V_N(\lambda)$.

^{abc} For these estimates, the one-sided test of the hypothesis that the parameter is equal to its true value is rejected at 10% (^{a}), 5% (^{b}), or 1% (^{c}) levels, respectively. The one-sided test is used since the simulated contamination biases slope estimates towards zero.

intercept is estimated consistently by LS, but it is just because the random noise simulating contamination is symmetric around zero). Second, the robust estimators LTS and RDL_1 can cope with contamination quite well. RDL_1 is most efficient at lower levels of contamination, but it is biased at high levels of contamination. On the other hand, LTS, which is the least efficient estimator in most cases, provides the best and most efficient estimates for the 40% level of contamination. Finally, let us discuss SLTS. For non-robust choices of the decision rule ($C > 1$), lower levels of contamination do not affect the estimates too much (except for $C = 2.5$), but extreme 40% contamination destroys them completely. A quite robust choice $C = 0.75$ can cope relatively well with contamination, although it seems to be biased for the 40% contamination level. If necessary, it is possible to use an even more robust choice $C = 0.5$. Nevertheless, these results show that

- it is necessary to stick to robust decision rules ($C = 1$) even though it might decrease efficiency in the ideal case of normally distributed errors and a clean data set,
- the efficiency of SLTS is not very good in presence of contamination (it is certainly worse than that of RDL_1). The reason for this was already discussed in Section 6.2: SLTS-AC1 with logistic weights has to downweight almost half of all observations if contamination is present.

Altogether, we can conclude that SLTS performs quite well for clean data sets regardless of the error distribution. It provides robust estimates under contamination, but loses efficiency already under moderate contamination. These deficiencies are addressed by the proposed SLTS-AC2 and we examine its behavior in Section 7.2.

7.2 Adaptive choice with two parameters

The simulation results presented in this section are for model (34) and they correspond to the simulations in Section 7.1. The results are again based on 1000 simulations and sample size $n = 100$. The main difference is that SLTS-AC2 (see Section 6.2) is added and compared with all other estimators. This second adaptive SLTS estimator optimizes not only the parameter ω, controlling the shape of smoothing, but also the trimming constant λ, see (32). Moreover, the decision rule is now based only on robust confidence intervals $\langle m_0 - Cv_0, m + Cv_0 \rangle$, that is, $C = D \cdot V_N(\lambda)$, where $D = 3$ or $D = 4$. For the fixed choice of $\lambda = 0.5$, these two cases, $D = 3$ and $D = 4$, correspond to SLTS-AC1 with constants $C = 1$ and $C = 0.75$ presented in Section 7.1.

Estimator	Parameter nP: D	Coefficient	$\varepsilon \sim N(0,1)$		$\varepsilon \sim t_3$		$\varepsilon \sim Exp(1)$	
			Mean	Var	Mean	Var	Mean	Var
LS		Intercept	0.297	0.101	0.304	0.187	0.303	0.139
LS		Slope	1.002	0.033	1.000	0.057	0.998	0.045
SLTS	1P: 4.0	Intercept	0.298	0.117	0.305	0.147	0.300	0.115
SLTS	1P: 4.0	Slope	1.003	0.039	1.000	0.046	0.997	0.039
SLTS	1P: 3.0	Intercept	0.297	0.128	0.305	0.156	0.300	0.112
SLTS	1P: 3.0	Slope	1.003	0.043	1.001	0.048	0.997	0.038
SLTS	2P: 4.0	Intercept	0.296	0.103	0.303	0.130	0.301	0.117
SLTS	2P: 4.0	Slope	1.002	0.034	0.999	0.041	0.997	0.040
SLTS	2P: 3.0	Intercept	0.298	0.119	0.300	0.138	0.301	0.119
SLTS	2P: 3.0	Slope	1.003	0.040	1.000	0.043	0.997	0.039
LTS		Intercept	0.295	0.280	0.309	0.251	0.294	0.173
LTS		Slope	1.009	0.086	1.003	0.079	0.999	0.058
RDL$_1$		Intercept	0.296	0.136	0.307	0.149	0.299	0.117
RDL$_1$		Slope	1.002	0.049	1.001	0.052	0.998	0.044

Table 5: Simulations for clear data sets of size $n = 100$, SLTS-AC1 and SLTS-AC2.

Entries in column "Parameter" indicate: (a) which adaptive-choice algorithm is used for SLTS ("1P" means Adaptive choice 1 (SLTS-AC2), "2P" represents Adaptive choice 2 (SLTS-AC2), see Section 6.2); (b) which confidence interval for residuals was used for the decision rule within the algorithms Adaptive choice 1 and 2: $\langle m_0 - D \cdot V_n(\lambda) \cdot v_0, m + C \cdot V_n(\lambda) \cdot v_0 \rangle$, where $m_0 = \text{med}_i\, r_i(b_0)$, $v_0 = \text{MAD}_i\, r_i(b_0)$, and b_0 is the initial (most robust) estimate.

The first set of simulations concentrates again on the behavior of the estimators for a clean data set (no contamination) and model (34) under different error distributions. The simulation results are presented in Table 5. The results concerning LS, LTS, SLTS-AC1, and RDL$_1$ are naturally the same as in Section 7.1, so I pay attention mainly to SLTS-AC2. First, it is consistent, and additionally, it is more efficient than the corresponding SLTS-AC1 in the case of normal and Student distributions (for the exponential distribution, it is a bit worse). More interestingly, SLTS-AC2 with $D = 4$ reaches the efficiency of the least squares for normally distributed errors and overtakes least squares in the other cases. SLTS-AC2 also performs better then RDL_1 in all cases.

Estimator	Parameter nP: D	Coefficient	Cont. 1%		Cont. 10%		Cont. 40%	
			Mean	Var	Mean	Var	Mean	Var
LS		Intercept	0.294	0.214	0.276	0.530	0.164	0.946
LS		Slope	0.849	0.200	0.304^c	0.238	0.064^c	0.151
SLTS	1P: 4.0	Intercept	0.297	0.192	0.294	0.218	0.294	0.261
SLTS	1P: 4.0	Slope	0.993	0.061	0.957	0.075	0.767	0.253
SLTS	1P: 3.0	Intercept	0.298	0.198	0.300	0.224	0.299	0.238
SLTS	1P: 3.0	Slope	0.995	0.063	0.963	0.074	0.819	0.238
SLTS	2P: 4.0	Intercept	0.298	0.119	0.300	0.123	0.298	0.231
SLTS	2P: 4.0	Slope	0.998	0.046	0.983	0.048	0.842	0.244
SLTS	2P: 3.0	Intercept	0.299	0.136	0.300	0.141	0.294	0.200
SLTS	2P: 3.0	Slope	0.998	0.049	0.983	0.054	0.885	0.205
LTS		Intercept	0.296	0.279	0.296	0.272	0.298	0.207
LTS		Slope	1.003	0.087	0.996	0.086	0.993	0.076
RDL_1		Intercept	0.297	0.134	0.295	0.138	0.298	0.183
RDL_1		Slope	0.999	0.047	0.990	0.049	0.906^a	0.067

Table 6: Simulations for contaminated data sets of size $n = 100$, SLTS-AC1 and SLTS-AC2.

Entries in column "Parameter" indicate: (a) which adaptive-choice algorithm is used for SLTS ("1P" means Adaptive choice 1 (SLTS-AC1), "2P" represents Adaptive choice 2 (SLTS-AC2), see Section 6.2); (b) which confidence interval for residuals was used for the decision rule within the algorithms Adaptive choice 1 and 2: $\langle m_0 - D \cdot V_n(\lambda) \cdot v_0, m + C \cdot V_n(\lambda) \cdot v_0 \rangle$, where $m_0 = \mathrm{med}_i \, r_i(b_0)$, $v_0 = \mathrm{MAD}_i \, r_i(b_0)$, and b_0 is the initial (most robust) estimate.

^{abc} For these estimates, the one-sided test of the hypothesis that the parameter is equal to its true value is rejected at 10% (^{a}), 5% (^{b}), or 1% (^{c}) levels, respectively. The one-sided test is used since the simulated contamination biases slope estimates towards zero.

Now, let us analyze the results for all the estimators under contamination. The three cases presented in Table 6 correspond to contamination levels 1%, 10%, and 40%. Again, I test the one-sided hypothesis that the slope parameter equals its true value. Results concerning LS, LTS, SLTS-AC1, and RDL_1 correspond again to those in Section 7.1, so let us concentrate on SLTS-AC2. First of all, its estimates are less affected by contamination than the SLTS-AC1 estimates, especially under very high contamination (40%). Moreover, the adaptive search over two parameters considerably improves the efficiency of SLTS, especially for a moderate amount of contamination. Consequently, if the contamination level is not extremely high, it performs as good as RDL_1 or even better.

***Remark* 10** *Due to space consideration, it is not possible to present all the available numerical results. Therefore, I have chosen two main levels of contamination—10% and 40% levels. Whenever I speak about "moderate" amount of contamination, I mean lower levels of contamination. Simulations show that under the moderate level of contamination it is possible to understand contamination levels up to 30% in the sense that SLTS behaves in a similar way as for 10% contamination. Other cases (contamination levels higher than 30%) are referred to as high or extreme contamination. This threshold can be increased, indeed, because the robustness of SLTS can be further improved by using a smaller D (and thus smaller confidence intervals) for decision rules: until now, $D \geq 3$, which corresponds to at least 99.9% confidence intervals under normally distributed errors, but we can use also $D = 2.5$, which corresponds to the 99% confidence interval.*

The simulation results discussed in this section clearly indicate that the SLTS-AC2 estimator is superior to SLTS-AC1 both in robustness and efficiency. In almost all cases, it performed as good as or better than all other estimators including RDL_1. The only exception is estimation with highly contaminated data, because then SLTS loses efficiency and it is not so stable as the original LTS estimator.

7.3 Misspecification of categorical variables

To this point, RDL$_1$ has performed very well, even compared to SLTS (but remember, we have chosen a quite favorable type of contamination for RDL$_1$). On the other hand, RDL$_1$ is designed for a simple additive model (it is difficult to generalize it if cross-effects are to be included), and moreover, it only takes care of continuous variables. This does not effect its breakdown point (categorical variables are always bounded and cannot therefore bring the estimator out of any bounds), but, as we demonstrate in this section, makes RDL$_1$ vulnerable to misspecification in categorical variables.

Estimator	Parameter	Coefficient	$\varepsilon \sim N(0,1)$	
	nP: D		Mean	Var
LS		Intercept	2.162^c	0.041
LS		Slope	-1.001	0.037
LS		Dummy	2.836^c	0.081
SLTS	2P: 3.0	Intercept	1.026	0.060
SLTS	2P: 3.0	Slope	-1.000	0.017
SLTS	2P: 3.0	Dummy	3.978	0.106
RDL$_1$		Intercept	1.268^c	0.070
RDL$_1$		Slope	-1.000	0.023
RDL$_1$		Dummy	3.733^b	0.116

Table 7: Simulations with one misspecified dummy variable for data sets of size $n = 100$ and SLTS-AC2.

Entries in column "Parameter" indicate: (a) which adaptive-choice algorithm is used for SLTS ("1P" means Adaptive choice 1 (SLTS-AC1), "2P" represents Adaptive choice 2 (SLTS-AC2), see Section 6.2); (b) which confidence interval for residuals was used for the decision rule within the algorithm Adaptive choice 2: $\langle m_0 - D \cdot V_n(\lambda) \cdot v_0, m + C \cdot V_n(\lambda) \cdot v_0 \rangle$, where $m_0 = \text{med}_i\, r_i(b_0)$, $v_0 = \text{MAD}_i\, r_i(b_0)$, and b_0 is the initial (most robust) estimate.

^{abc} For these estimates, the two-sided test of the hypothesis that the parameter is equal to its true value is rejected at 10% (^{a}), 5% (^{b}), or 1% (^{c}) levels, respectively.

The misspecification sensitivity is again exemplified using a Monte Carlo simulation. I consider the model $y_i = 1 - x_i + 4d_i + \varepsilon_i$, where $i = 1, \ldots, n$, $\varepsilon_i \sim N(0,1)$, and $d_i \in \{0,1\}$. Further, assume that 20% percent of the observations have a misspecified binary variable d_i (it can correspond, for example, to wrong entries about the sex of individuals in a sample). In other words, d_i contains a wrong value for 20 percent of the sample. The results obtained for sample size $n = 200$ and 1000 simulations are summarized in Table 7.3. To indicate which estimates are significantly biased, I tested the two-side hypothesis that the intercept and slope parameters equal their true values. Apparently, both LS and RDL$_1$ estimates are inconsistent. Notice that the slope coefficient is estimated correctly, but the intercept and the effect of the dummy variable are wrong. On the contrary, the SLTS estimate provides consistent results, which are not affected by the misspecification of the dummy variable.

8 Conclusion

In this paper, I introduced the smoothed least trimmed squares estimator and derived its asymptotic properties. Thus, I extended applicability of the LTS procedure is extended to general regression models that involve categorical explanatory variables. The resulting estimator is currently the only robust estimator with a high breakdown point that can be applied in general regression models with categorical variables. Equally important is the improvement in efficiency compared to the LTS estimator and also to the RDL$_1$ estimator, which represented until now the only solid robust estimator for linear regression models involving binary covariables. The only exception concerning the efficiency improvement is highly contaminated data (40% contamination and more), because especially LTS performs better than SLTS for such data. This inefficiency of SLTS can probably be reduced by a better choice of smoothing, but one does not currently exist. I constructed a procedure that adaptively chooses weighting schemes for

SLTS and thus controls the balance between the robustness and the efficiency of the estimator. The adaptive procedure actually starts from an estimate close to LTS (most robust) and decides how far it can go towards the least squares (improvement in efficiency) without endangering the robustness of SLTS.

On the other hand, I studied behavior of the adaptive choice of a smoothing scheme only for one possible class of generating functions, which is quite suitable, but it does have to be the optimal one. Hence, finding an optimal smoothing class with respect to the asymptotic variance of SLTS would be a very valuable improvement of SLTS and it is one of the main issues for further research. Another unresolved issue closely related to the adaptive choice of smoothing is the construction of a distribution-free decision rule.

References

[1] Amemiya, T. (1985): *Advanced econometrics.* Harvard University Press, Massachusetts.

[2] Benáček, V., Jarolím, M. and Víšek, J. Á. (1998): Supply-side characteristics and the industrial structure of Czech foreign trade. *Proceedings of the conference Business and economic development in central and eastern Europe: Implications for economic integration into wider Europe, ISBN 80-214-1202-X*, Technical university in Brno together with University of Wisconsin, Whitewaters, and the Nottingham Trent university, 51–68.

[3] Čížek, P. (2001): Robust estimation in nonlinear regression models. *SFB Discussion paper,* 25/2001, Humboldt University, Berlin.

[4] Gerfin, M. (1996): Parametric and semi-parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometrics*, Vol. 11 / 3, 321–339.

[5] Hampel, F. R. (1971): A general qualitative definition of robustness. *Annals of Mathematical Statistics*, Vol. 42, 1887–1896.

[6] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986): *Robust statistics, The approach based on influence function.* Wiley, New York.

[7] Huber, P. J. (1964): Robust estimation of a location parameter. *Annals of Mathematical Statistics*, Vol. 35, 73–101.

[8] Hubert, M., Rousseeuw, P. J. (1997): Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference*, Vol. 57, 153–163.

[9] Jurečková, J. (1984): Regression quantiles and trimmed least squares estimator under a general design. *Kybernetika*, Vol. 20, 345–357.

[10] Jurečková, J., and Sen, P. K. (1989): Uniform second order asymptotic linearity of M-statistics. *Statistics & Decisions*, Vol. 7, 263–276.

[11] Marrona, R. A., Bustos, O. H., and Yohai, V. J. (1979): Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In T. Gasser, M. Rossenblatt eds., *Smoothing techniques for curve estimation*, Lecture Notes in Mathematics 757, Springer, Berlin, 91–116.

[12] Orhan, M., Rousseeuw, P. J., and Zaman, A. (2001): Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, Vol. 71, 1–8.

[13] Peracchi, F. (1990): Robust M-estimators. *Economic Reviews*, 1990, 1–30.

[14] Rousseeuw, P. J. (1984): Least median of squares regression. *Journal of American Statistical Association*, Vol. 79, 871–880.

[15] Rousseeuw, P. J. (1985): Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze, and W. Wertz eds., *Mathematical statistics and applications, Vol. B*, Reidel, Dordrecht, Netherlands, 283–297.

[16] Rousseeuw, P. J. (1997): Introduction to positive-breakdown methods. In Maddala, G. S., and Rao, C. R., eds., *Handbook of statistics, Vol. 15: Robust inference*, Elsevier, Amsterdam, 101–121.

[17] Rousseeuw, P. J., and Leroy, A. M. (1987): *Robust regression and outlier detection.* Wiley, New York.

[18] Rousseeuw, P. J. and Yohai, V. J. (1984): Robust regression by means of S-estimators. In J. Franke, W. Härdle, R. D. Martin eds., *Robust and nonlinear time series analysis*, Lecture notes in statistics, Vol. 26, Springer, New York, 256–272.

[19] Rousseeuw, P. J., and Van Driessen, K. (1999): Computing LTS regression for large data sets. *Technical report, University of Antwerp*, submitted.

[20] Storn, R. (1996): On the usage of differential evolution for function optimization. *NAFIPS 1996, Berkeley*, 519–523.

[21] Storn, R., and Price, K. (1995): Differential evolution—a simple and efficient adaptive scheme for global optimization over continuous spaces. *Technical report TR-95-012, ICSI.*

[22] Tukey, J. W. (1960): A survey of sampling from contaminated distributions. In I. Olkin ed., *Contributions to probability and statistics,* Stanford University Press, Stanford, 448–485.

[23] Tukey, J. W. (1977): *Exploratory data analysis.* Addison-Wesley, Reading.

[24] Víšek, J. Á (1996a): Sensitivity analysis of M-estimators. *Annals of the Institute of statistical mathematics*, Vol. 48, 469–495.

[25] Víšek, J. Á (1996b): On high breakdown point estimation. *Computational Statistics*, Vol. 11, 137–146.

[26] Víšek, J. Á (1999a): The least trimmed squares—random carriers. *Bulletin of the Czech econometric society*, Vol. 10/1999, 1–30.

[27] Víšek, J. Á (1999b): On the diversity of estimates. To appear in *Computational Statistics and Data Analysis.*

[28] White, H. (1980): Nonlinear regression on cross-section data. *Econometrica*, Vol. 48 / 3, 721–746.