Čížek, Pavel

Working Paper

Robust estimation in nonlinear regression and limited dependent variable models

SFB 373 Discussion Paper, No. 2001,100

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Čížek, Pavel (2001) : Robust estimation in nonlinear regression and limited dependent variable models, SFB 373 Discussion Paper, No. 2001,100, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10051178

This Version is available at:
https://hdl.handle.net/10419/62677

Robust Estimation in Nonlinear Regression and Limited Dependent Variable Models*

Pavel Čížek†

13th December 2001

Abstract

Classical parametric estimation methods applied to nonlinear regression and limited-dependent-variable models are very sensitive to misspecification and data errors. On the other hand, semiparametric and nonparametric methods, which are not restricted by parametric assumptions, require more data and are less efficient. A third possible estimation approach is based on the theory of robust statistics, which builds upon parametric specification, but provides a methodology for designing misspecification-proof estimators. However, this concept, developed in statistics, has so far been applied almost exclusively to linear regression models. Therefore, I adapt some robust methods, such as least trimmed squares, to nonlinear and limited-dependent-variable models. This paper presents the adapted robust estimators, proofs of their consistency, suitable computational methods, as well as examples of regression models which the proposed estimators can be applied to.

Keywords: least trimmed squares, limited-dependent-variable models, nonlinear regression, robust estimation

JEL: C13, C21, C24

* Finacial Support was received by the Deutsche Forschungsgemeinschaft, SFB 373 ('Quantifikation und Simulation Ökonomischer Prozesse), Humbold-Universität zu Berlin.

†Institute für Statistik und Ökonometrie, CASE, Humboldt-Universität zu Berlin, Spandauer Str. 1, D – 10178 Berlin, Germany. E-mail: cizek@wiwi.hu-berlin.de.

Contents

1 Introduction **3**

2 Definition of nonlinear trimmed estimators **7**
2.1 Least trimmed squares 7
2.2 Definition of maximum trimmed likelihood 8
2.3 Common framework for LTS and MTLE: general trimmed estimator 10

3 Consistency of GTE in nonlinear regression models **10**
3.1 Alternative definition of GTE, notation 11
3.2 Assumptions 12
3.3 Normal equations 16
3.4 Consistency of general trimmed estimator 17
3.5 Identification condition 26
3.5.1 Nonlinear least trimmed squares 28
3.5.2 Maximum trimmed likelihood 30

4 Consistency of GTE in limited-dependent-variable models **32**

5 Examples of trimmed estimators **34**
5.1 Nonlinear regression models 34
5.2 Limited-dependent-variable models 35
5.2.1 Truncated regression 35
5.2.2 Censored regression 38
5.3 Binary-choice models 39

6 Computation of trimmed estimators **42**
6.1 Subsample selection and estimation 42
6.2 Differential evolution 43

7 Conclusion **44**

A Proofs of lemmas and other auxiliary propositions **46**

1 Introduction

Various nonlinear and limited-dependent-variable models are quite natural in econometrics, both because a modeled relationship is of nonlinear nature or the dependent variable is truncated, censored, or discrete.[1] These models are typically estimated parametrically either by (nonlinear) least squares or by maximum likelihood and are, therefore, quite sensitive to misspecification. This sensitivity naturally depends on the type and extent of deviations from the parametric model used and on the model itself. Whereas the robustness of classical parametric estimators in nonlinear regression models is basically the same as in the case of linear regression (see Chatterjee and Hadi (1988) for an introduction to this topic), the robustness of parametric techniques is much lower in the case of limited-dependent-variable models because even non-normality of errors (see Hurd (1979) for truncated regression or Arabmazar and Schmidt (1982) for Tobit) or heteroscedasticity (see Arabmazar and Schmidt (1981) for Tobit) can lead to inconsistency. Similar results were demonstrated also in case of binary-choice models (Manski and Thompson (1986), Klein and Spady (1993) for probit). Consequently, most recently-developed estimation techniques for these models primarily address misspecification-sensitivity problems so that they lead to meaningful results even under relatively weak conditions.

The strategy that is used most often for dealing with misspecification sensitivity relies on weakening the necessary regularity and identification conditions so that the probability of their violation is significantly decreased. This is the case of semiparametric and nonparametric methods. First, semiparametric procedures are typically built around a single-index model, the concept first proposed by Brillinger (1983), which encompasses, for example, binary-response models, censored Tobit regressions, and some duration models. Examples of existing estimators include the density-weighted average derivative estimator (Powell, Stock, and Stoker (1989)), a quasi-likelihood estimator for binary-choice models developed by Klein and Spady (1993), or the semiparametric least square method (SLS) and its weighted version (WSLS) designed by Ichimura (1993). Second, nonparametric methods do not assume any specific kind of a functional relationship and estimate directly the regression function using kernel estimation, see Härdle and Linton (1994) for more

[1] First, nonlinear regression models are represented, for example, by models with an exponential regression function and an additive error term. Next, a model is truncated if we cannot see observations with negative values of dependent variable, for instance. On the other hand, a regression model can be also censored, that is, all observations are visible, but we do not have the actual values of the dependent variable if it is negative, for example. Finally, it is possible that we are only able to recognize that a dependent variable is, let us say, negative or positive—then the true value is replaced by a discrete variable indicating what we observe.

information, including econometric applications.

Unfortunately, there are several problems related to the use of most semiparametric estimation methods. The first difficulty is the need for data: semiparametric and nonparametric methods generally require large samples compared to parametric methods in order to achieve an equally good approximation of the true parameter values, especially if some kind of nonparametric smoothing is used. Another problem lies in the fact that semiparametric estimators are still sensitive to misspecification of the regression function or other parametric components. On the other hand, if the regression function is not specified, it is almost impossible to obtain its derivatives with respect to the unknown parameters. Moreover, the limiting distribution is either non-normal or unknown in some cases. Finally, the sensitivity of the existing semiparametric methods to outliers and very influential observations, as well as the detection of outliers within the semiparametric framework, has not been explored yet.

On the other hand, there is another strategy, which retains standard parametric assumptions but takes into account possible misspecification and its impact on estimation procedures. This approach falls under the heading of so-called "robust statistics" (see Hampel et al. (1986)), which provides a methodology for developing estimators that behave well not only under correct parametric specification, but also in the presence of "small" deviations from the parametric assumptions. These deviations can be of almost any kind, but the use of a parametric model requires that at least part of data follows the model. Now, this strategy has not been used to design robust estimators for limited-dependent-variable models yet and my aim is to follow it and to design robust estimators that will provide reliable results without a high efficiency loss[2] in these models.

As a starting point, I use highly robust *least trimmed squares* (LTS), a statistical technique for the estimation of unknown parameters of the linear regression model. It was proposed by Rousseeuw (1985) as a robust alternative to the classical regression method based on minimizing the sum of squared residuals. Despite its advantages over classical parametric estimators (see Orhan, Rousseeuw, and Zaman (2001) for some econometric evidence), it is not applied in econometrics at all since it has several shortcomings concerning its applicability. First, LTS as well as many other highly robust estimators cannot be applied in regression models containing discrete explanatory variables, and only recently, their modifications appeared that allow such an application (see Hubert and Rousseeuw (1997) and Čížek (2001b)). Second, there are many classes of regression models in which

[2]The efficiency loss is meant in comparison with the efficiency of standard parametric methods providing that the assumptions of the appropriate parametric models are valid.

such a robust estimator cannot be used right now, and partly because its properties are not known in such models, partly because it is not adapted to suit such models.[3] This concerns the estimation of nonlinear models, limited-dependent-variable, and discrete-choice models.

This paper addresses the lack of robust techniques such as LTS in nonlinear and limited-dependent-variable regression and extends them so that they can be applied there as a robust alternative to classical parametric methods. Additionally, since some parametric estimators used in these models are based on principles different from the simple least squares minimization, I aim to generalize the idea of the LTS estimator first and create a concept of a general trimmed estimator. It should cover not only LTS and its eventual variants suitable for nonlinear regression models, but allow us to define trimmed versions of other estimators, such as a trimmed version of the maximum likelihood estimator (MLE) proposed and discussed here as well. The proposed trimmed estimators should be robust in the same sense as LTS and many other robust estimators: they can cope with any type of deviation from a parametric model provided that there is a core subset of data which follows the parametric model. Next, I make the first step concerning the use of trimmed estimators in nonlinear regression and limited-dependent-variable models. Namely, I prove the consistency of the proposed estimators and show that its order is equal to $\sqrt{n}$ (this extends the existing results regarding LTS even within the framework of standard nonlinear regression models). Last, but not least, I discuss specific examples of trimmed estimators encompassed by the framework of a general trimmed estimator and their applications.

Finally, let me give examples of several typical models which the proposed methods are intended for. First, intrinsically nonlinear regression models arise when we have to use, for instance, the Box-Cox transformation (see Box and Cox (1964)) or when a regression function is of exponential nature but an error term enters the regression equation additively; see Griffiths, Hill, and Judge (1993, Chapter 22) for more examples. Further, it is not uncommon in econometrics that a dependent variable cannot be fully observed (for example, above or below a threshold)—some of its values are either censored or truncated. Then we talk about limited-dependent-variable models. Finally, a nonlinear relationship can arise if we observe only a finite number of states instead of a continuous dependent variable—discrete-choice models are used then. Both limited-dependent-variable and discrete-choice models possess very specific error structures.

[3]The existing literature concentrates on behavior in standard regression models, assuming the independence of explanatory variables and error terms, and on computational issues; see for example, Chen, Stromberg, and Zhou (1997) and Čížek (2001a).

Let me now make the main aims of the paper more precise. Because the regression models mentioned above are estimated not only by least-squares estimators, but by maximum likelihood estimators as well, I propose first a general concept of trimmed estimators that unites the trimmed versions of commonly used estimation methods and allows me to deal with both kinds of estimators at the same time. Further in the paper, I study the behavior of trimmed estimators applied to the nonlinear regression model

$$y_i = h(x_i, \beta) + \varepsilon_i, \tag{1}$$

where y_i represents the dependent variable and $h(x, \beta)$ is a regression function ($i = 1, \ldots, n$). The vector of explanatory variables x_i and the error term ε_i are assumed to form sequences of independent and identically distributed random variables that possess an absolutely continuous distribution function.

For the next step, I continue with analysis of trimmed estimators in the limited-dependent-variable framework: here I assume that there is a structural model of the form

$$\tilde{y}_i = \tilde{h}(x_i, \beta) + \varepsilon_i, \tag{2}$$

where $\tilde{h}(x_i, \beta)$ is a known function of data x_i and a vector β. Moreover, $\tilde{y}_i$ is unobservable—we can observe only $y_i = \tau(\tilde{y}_i)$, where $\tau : \mathbb{R} \to \mathbb{R}$ is a known transformation. This leads to the definition of the corresponding reduced model, that is a model based on $y_i = \tau(\tilde{y}_i)$ as a dependent variable. It can be described by

$$y_i = h(x_i, \beta) + \nu_i, \tag{3}$$

where the dependent variable y_i equals $\tau(\tilde{y}_i)$, the regression function is defined by $h(x_i, \beta) = \mathsf{E}\,(y_i|x_i) = \mathsf{E}\,(\tau(\tilde{y}_i)|x_i)$, and the error term is given by $\nu_i = y_i - \mathsf{E}(y_i|x_i) = y_i - h(x_i, \beta)$. Notice that models (1) and (3) have the same form (although the relationships between the variables are different).

Finally, I consider several typical nonlinear models and examples of trimmed estimators that can be applied in these models. Additionally, I discuss possible computational methods for trimmed estimators.

In the rest of the paper, I first review important facts about LTS that are also related to other trimmed estimators, such as the maximum trimmed likelihood (Section 2). Later, I consider the nonlinear regression model (1) and discuss necessary assumptions for

the consistency of the proposed trimmed estimators (Section 3). Next, I deal with the limited-dependent-variable model (3) and again derive the consistency of trimmed estimators (Section 4). Finally, I choose several typical models from both classes and provide examples of trimmed estimators suitable for these models (Section 5) together with several computational procedures for trimmed estimators (Section 6).

2 Definition of nonlinear trimmed estimators

In this section, I describe first the least trimmed squares estimator (LTS), introduced for the linear regression model by Rousseeuw (1985), and its properties (Section 2.1). Next, I define a trimmed version of MLE (Section 2.2). Finally, I unite both trimmed estimators in the concept of a general trimmed estimator (Section 2.3).

2.1 Least trimmed squares

Consider a nonlinear regression model for a sample (y_i, x_i) with a dependent variable $y_i \in \mathbb{R}$ and a vector of explanatory variables $x_i \in \mathbb{R}^k$

$$y_i = h(x_i, \beta) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4}$$

where $h(x_i, \beta)$ is a known regression function of data x_i and a vector β of p unknown parameters[4] ($\beta \in B \subset \mathbb{R}^p$, where B is the corresponding parameter space). Such a regression model represents both (1) and (3).

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

$$\hat{\beta}_n^{(LTS)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} r_{[i]}^2(\beta), \tag{5}$$

where $r_{[i]}^2(\beta)$ represents the ith order statistics of squared residuals $r_1^2(\beta), \ldots, r_n^2(\beta)$: $r_i^2(\beta) = (y_i - h(x_i, \beta))^2$. The trimming constant h has to satisfy $\frac{n}{2} < h \leq n$. This constant determines the robustness of the LTS estimator, since definition (5) implies that $n - h$ observations with the largest residuals do not have a direct influence on the estimator. The highest level of robustness is achieved for $h = [n/2] + [(p+1)/2]$ (Rousseeuw and Leroy (1987, Theorem 6), in the case of linear regression; see Stromberg (1993) for an analogous result in

[4] In general, the dimensions of x_i and β do not have to be the same, i.e., $k \neq p$.

nonlinear regression models). On the other hand, the robustness of LTS is lowest for $h = n$, which corresponds to the least squares estimator. There is, of course, a trade-off: lower values of h lead to a higher degree of robustness, 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 comparison with some less robust estimator, e.g., the least squares, because a comparison of these two estimators can serve as a simple check of data and the 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 specific structures in the studied data.

Before proceeding further, it is useful to discuss several issues, most importantly the existence of this estimator. The existence of the optimum in (5) under some reasonable assumptions can be justified in the following way: the minimization of the objective function in (5) can be viewed as a process in which we choose every time a subsample of h observations and find some β minimizing the sum of squared residuals for the selected subsample. Doing this for every subsample, 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 subsamples of size h. This is, of course, far from a usable computational procedure, see Section 6 for more information.

2.2 Definition of maximum trimmed likelihood

Although the least-squares-based approach is traditionally used in most (non)linear regression models, the maximum likelihood estimation is equally or even more important, especially if we want to estimate various limited-dependent-variable models. In the same way the least trimmed squares estimator is derived from the least squares, I propose analogously the maximum trimmed likelihood estimator for the model

$$y_i = h(x_i, \beta) + \varepsilon_i, \tag{6}$$

provided that the distribution of (x_i, ε_i), or equivalently (y_i, x_i), is known. The advantage of such an estimator against MLE is an increase in the robustness of the estimator because

only a part of the data has to follow a specified parametric model and only the data that follows closely the specified model are taken into account. So, let us assume that $(x_i, \varepsilon_i), i = 1, \ldots, n$, or equivalently $(x_i, y_i), i = 1, \ldots, n$, form sequences of independent and identically distributed random vectors and that the corresponding distribution functions are known. Moreover, let $l_i(\beta) = l(x_i, y_i; \beta)$ be the likelihood function of $\beta \in B$ associated with the observation (y_i, x_i). Then the *maximum trimmed likelihood estimator* (MTLE) $\hat{\beta}_n^{(MTLE,h)}$ is defined for regression model (6) by

$$\hat{\beta}_n^{(MTLE,h)} = \arg\max_{\beta \in B} \prod_{i=n-h+1}^{n} l_{[i]}(x_i, y_i; \beta), \tag{7}$$

where

- $\beta \in B \subset \mathbb{R}^p$ is a p-dimensional vector of unknown parameters and $B \subset \mathbb{R}^p$ is the corresponding parameter space,
- $l_{[i]}(x_i, y_i; \beta)$ represents the (ascendingly) ordered sample of likelihood functions $l_i(x_i, y_i; \beta) = l(x_i, y_i; \beta)$, $i = 1, \ldots, n$, for any $\beta \in B$,
- $h \in \left\{ \left[\frac{n+1}{2} \right], \ldots, n \right\}$ is the *trimming constant* (see Section 2.1).

This definition can also be rewritten as

$$\hat{\beta}_n^{(MTLE,h)} = \arg\max_{\beta \in B} \sum_{i=n-h+1}^{n} \ln l_{[i]}(x_i, y_i; \beta). \tag{8}$$

Apparently, the maximum trimmed likelihood estimator makes MLE robust in the same way as the trimming of least squares in case of LTS: the objective function of MTLE contains now only likelihoods for $n - h$ most probable observations at a given β. This means that all other observations, that is, the h least probable observations for a given parametric model, do not directly influence the objective function of the MTLE estimator. Thus, if a number of observations in a sample that do not follow the specified parametric model is smaller than h, the MTLE estimator is not influenced by these observations and estimates the model using only the (model-following) rest of the data.

2.3 Common framework for LTS and MTLE: general trimmed estimator

In Section 3, I am going to derive important asymptotic properties for both the LTS and MTLE estimators. Therefore, it is advantageous to create a general framework that encompasses both estimators and allows us to derive their properties at the same time. Using regression models (1) or (3), we can define the *general trimmed estimator* (GTE) as

$$\hat{\beta}_n^{(GTE,h)} = \arg\min_{\beta \in B} \sum_{j=1}^{h} s_{[j]}(x_i, y_i; \beta), \tag{9}$$

where $s_{[j]}(x_i, y_i; \beta), j = 1, \ldots, n$, represent the (ascendingly) ordered sample of some general loss functions $s(x_i, y_i; \beta) = s(x_i, \varepsilon_i; \beta)$, $i = 1, \ldots, n$, for any $\beta \in B$.[5] The choice $s(x_i, y_i; \beta) = r_i^2(\beta) = (y_i - h(x_i, \beta))^2$ corresponds to the LTS estimator, $s(x_i, y_i; \beta) = -\ln l_i(x_i, y_i; \beta)$ represents the MTLE estimator. Naturally, it is necessary to restrict the possible choices of $s(x, y; \beta)$ in order to be able to derive reasonable results: thus, $s(x, y; \beta)$ should be, among others, a continuously differentiable function of β on B; see Section 3.2 for more details. In the following sections, I refer to this general trimmed estimator as GTE, and specifically, I understand under this name the previously defined LTS and MTLE. Moreover, I will refer to $s(x_i, y_i; \beta)$ as residuals or losses for the sake of simplicity.

***Note* 1** *For the loss function evaluated at an observation (x_i, y_i), we use notation $s(x_i, y_i; \beta)$ or $s_i(x_i, y_i; \beta)$. For the* j*th order statistics of $s(x_i, y_i; \beta), i = 1, \ldots, n$, we use symbol $s_{[j]}(x_i, y_i; \beta)$. In this case, index i inside the order statistics is just a formal notation and does not have any relationship to summation or other indices. It is to be understood so that x_i, y_i inside $s_{[j]}(x_i, y_i; \beta)$ indicate just the sample on which this order statistics is based (so correctly, one would have to write $s_{[j]}((x_i, y_i)_{i=1}^n; \beta)$).*

3 Consistency of GTE in nonlinear regression models

In this section, I present the main asymptotic result concerning GTE, namely, its asymptotic consistency in the nonlinear regression model (1). Before proving it, an alternative definition of GTE and some notational conventions used in the rest of the paper are introduced (Section 3.1) as well as the assumptions necessary for the asymptotic results

[5] The expressions $s(x_i, y_i; \beta)$ and $s(x_i, \varepsilon_i; \beta)$ represent the same loss function and are equivalent since $y_i = h(x_i, \beta) + \varepsilon_i$.

(Section 3.2).

3.1 Alternative definition of GTE, notation

Given a sample (y_i, x_i), the GTE estimator of an unknown parameter vector β is defined for models (1) and (3) by equation (9). The dependent variable is denoted $y_i \in \mathbb{R}$, the vector of explanatory variables is $x_i \in \mathbb{R}^k$, and ε_i represents the error term. In addition, Ω_y, Ω_x, and Ω_ε refer to probability spaces on which y_i, x_i, and ε_i are defined, so $\Omega = \Omega_y \times \Omega_x$ is the probability space of the random vector (y_i, x_i). The true underlying value of the vector β in regression models (1) and (3) will be referred to by β^0. The non-trimmed estimator corresponding to GTE, which naturally coincides with the nonlinear least squares or the maximum likelihood, is further denoted by

$$\hat{\beta}_n^{(GE)} = \arg\min_{\beta \in B} \sum_{j=1}^{n} s_{[j]}(x_i, y_i; \beta) = \arg\min_{\beta \in B} \sum_{i=1}^{n} s_i(x_i, y_i; \beta).$$

Here and in definition (9), $s_{[j]}(x_i, y_i; \beta)$ stands for the jth order statistics of $s_i(x_i, y_i; \beta)$. In other words, it holds that $s_{[1]}(x_i(\omega), y_i(\omega); \beta) \leq \cdots \leq s_{[n]}(x_i(\omega), y_i(\omega); \beta)$ for any $\beta \in B$ and $\omega \in \Omega$.

Next, an alternative definition of GTE employed in the theoretical part of this paper instead of (9) is given by[6]

$$\hat{\beta}_n^{(GTE,h)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)\big). \tag{10}$$

To obtain this formula, one has to realize that for a given value of $\beta \in B$, the minimization of the h smallest elements $s(x_i, y_i; \beta)$ means that we include in the objective function only those elements that satisfy $s(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)$.[7] One additional note concerns the trimming constant: whenever asymptotic properties of GTE are studied, that is $n \to +\infty$, we have to work with a sequence of trimming constants h_n (for every sample size n, there has to be a corresponding choice of h). As this constant determines the robustness properties of GTE, we want to asymptotically prescribe a fixed fraction λ of observations that are

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

[7] 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 the assumptions given in Section 3.2, this happens with zero probability and definitions (9) and (10) are equivalent almost surely as the cumulative distribution function of $s_i(x_i, y_i; \beta)$ is assumed to be absolutely continuous. Therefore, we further use definition (10) for convenience.

considered to be correct, $\frac{1}{2} < \lambda \leq 1$, or alternatively, a fraction $1 - \lambda$ of observations that are excluded from the objective function of GTE ($0 \leq 1 - \lambda < \frac{1}{2}$). The trimming constant can then be defined for a given $n \in \mathbb{N}$ by $h_n = [\lambda n]$, where $[x]$ represents the integer part of x; hence, $h_n/n \to \lambda$. From now on, we assume that there is such a number $\lambda \in (\frac{1}{2}, 1\rangle$ for a sequence h_n of trimming constants defining the general trimmed estimator.

To close this section, we introduce the remaining mathematical notation. As observations and parameters considered here always belong to a Euclidean space $\mathbb{R}^l$, we shall need to define a neighborhood of a point $x \in \mathbb{R}^l$: an open neighborhood (open ball) $U(x, \delta) = \{z \in \mathbb{R}^l : \|z - x\| < \delta\}$ and a closed neighborhood (closed ball) $\bar{U}(x, \delta) = \{z \in \mathbb{R}^l : \|z - x\| \leq \delta\}$. Moreover, let us denote a convex span of $x_1, \ldots, x_m \in \mathbb{R}^l$ by $[x_1, \ldots, x_m]_{\varkappa}$. Finally, several symbols from linear algebra are introduced: 1_n represents an n-dimensional vector of ones, $b_1, \ldots, b_n$ are standard basis vectors in $\mathbb{R}^n$, i.e., $b_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, and $\mathcal{I}_n$ is the identity matrix of dimension n.

3.2 Assumptions

The assumptions necessary to prove the asymptotic consistency of GTE form three groups: distributional Assumption D for random variables in (1), Assumption H concerning properties of the loss function $s(x, y; \beta)$, and Assumptions NC and NN needed for the uniform law of large numbers. The latter set of assumptions is presented separately in Section 3.4.

First of all, let me discuss the distributional assumptions dealing with the random variables used in model (1). Most of these conditions are either equivalent to standard assumptions used in (nonlinear) regression models or additional assumptions needed to derive any reasonable results for order statistics. Moreover, we argue in a number of remarks that the following assumptions do not restrict us in any way in real applications.

Assumption D.

D1 Let $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^k$, $i = 1, \ldots, n$, be a sequence of independent identically distributed random vectors with finite fourth moments. Moreover,

$$n^{-1/6} \max_{i,j} \{|y_i|, |x_{ij}|, |\varepsilon_i|\} = \mathcal{O}_p(1). \tag{11}$$

***Remark* 1** *The necessity to include restriction (11) is caused by the discontinuity of the objective function of GTE (the discontinuity has to be understood from the inclusion-of-observations point of view: every observation either fully enters the objective function or*

does not enter it at all). A similar assumption was used for the first time by Jurečková (1984). Using Proposition 1 (see below this remark), we can say that equation (11) holds even for some distribution functions with polynomial tails, namely for those that have finite fourth moments. This becomes apparent once we realize that a distribution with tails behaving like one over a polynomial of the fifth (or lower) order does not have finite fourth moments. As the existence of finite fourth moments is one of the necessary conditions here, assumption (11) should not pose a considerable restriction on the explanatory variables. One can also notice that a random variable with a finite support is not restrained by this assumption in any way.

Proposition 1 *Let $x_1, x_2, \ldots$ be a sequence of independent identically distributed random variables with a distribution function $F(x)$. Let $b(x)$ be a lower bound for $F(x)$ in a neighborhood U_1 of $+\infty$. If $b(x)$ can be chosen as $1 - \frac{1}{P_6(x)}$, where $P_6(x)$ is a polynomial of the fourth order, then it holds that $n^{-\frac{1}{6}} \max_{i=1,\ldots,n} x_i = \mathcal{O}_p(1)$ as $n \to +\infty$. Analogously, let $c(x)$ be an upper bound for $F(x)$ in a neighborhood U_2 of $-\infty$. If $c(x)$ can be chosen as $\frac{1}{P_6(x)}$, where $P_6(x)$ is a polynomial of the fourth order, then it holds that $n^{-\frac{1}{6}} \min_{i=1,\ldots,n} x_i = \mathcal{O}_p(1)$ as $n \to +\infty$.*

***Remark* 2** *Assumption D1 can be weakened to (x_i, y_i) having finite second moments and*

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

if we assume more about the functional form of the loss function $s(x, y; \beta)$—for instance, if we assume least-squares based function, $s(x, y; \beta) = (y - h(x, \beta))^2$, see Čížek (2001a).

D2 We assume $\mathsf{E}\, x_i x_i^T = Q$, where Q is a nonsingular matrix.

D3 Let $G(x)$ represent the distribution function of $s(x_i, y_i; \beta^0)$, $G(x)$ be absolutely continuous, and $g(x)$ be the corresponding probability density function. Moreover, $g(x)$ is assumed to be positive and bounded by constant $M_g > 0$ on the whole support of the distribution function G.

***Remark* 3** *Note that Assumption D3 implies the following property of the distribution function $G(x)$ and its density $g(x)$: for any $0 < \alpha < 1$ we can find $\varepsilon > 0$ such that $\inf_{x \in (G^{-1}(\alpha)-\varepsilon, G^{-1}(\alpha)+\varepsilon)} \min\{G(x), g(x)\} > 0$.*

Moreover, we have to add a piece of notation. Sometimes it is necessary to refer to the distribution function of $s(x_i, y_i; \beta)$ for any $\beta \in B$. In such a case, G_β is used for the cumulative distribution functions and g_β for the corresponding probability density function. It follows that $G = G_{\beta^0}$, and similarly, $g = g_{\beta^0}$.

D4 Assume that

$$m_{gg} = \inf_{\beta \in B} \inf_{z \in (-\delta, \delta)} g_\beta \left(G_\beta^{-1}(\lambda) + z \right) > 0$$

and

$$M_{gg} = \sup_{\beta \in B} \sup_{z \in \mathbb{R}} g_\beta(z) < +\infty,$$

where G_β and g_β are the cumulative distribution function and the probability density function of $s_i(x_i, y_i; \beta)$.

***Remark* 4** *Although Assumption D4 might look unfamiliar at first sight, it just guarantees that the distribution functions of random variables $s_i(x_i, y_i; \beta)$ do not converge to some extreme cases for some $\beta \in B$. Namely, these conditions exclude cases when the expectation or variance of $s_i(x_i, y_i; \beta)$ converge to infinite values for some $\beta \in B$ or when the distribution function G_β converges to a discrete distribution function for some $\beta \in B$. This does not restrict us in commonly used regression models, because the parametric space B is compact. See also remark 3 to Assumption D3.*

As I aim to apply GTE to nonlinear models, several conditions on the loss function $s(x, \varepsilon; \beta)$ have to be specified (it is in many cases closely related to the regression function $h(x, \beta)$). Most of them are just regularity conditions that are employed in almost any work concerning nonlinear regression models. Because the assumptions stated below rely on the value of β and because I do not have to require their validity over the whole parametric space, I restrict β to a neighborhood $U(\beta^0, \delta)$ in these cases and suppose that there exists a positive constant δ such that all the assumptions are valid.

Assumption H.

H1 Let $s(x, y; \beta)$ be a continuous (uniformly over any compact subset of the support of (x, y)) in $\beta \in B$ and twice differentiable function in β on $U(\beta^0, \delta)$ almost surely:

$$\left(\forall \beta \in U(\beta^0, \delta); \forall x \in A_x \subseteq \mathbb{R}^k, P(A_x) = 1; j, k \in \{1, \ldots, p\}\right) \left(\exists \frac{\partial s(x, y; \beta)}{\partial \beta_j}, \frac{\partial^2 s(x, y; \beta)}{\partial \beta_j \partial \beta_k} \right).$$

The first derivative is continuous in $\beta \in U(\beta^0, \delta)$ almost surely.

H2 Furthermore, let us assume that the second derivatives $s''_{\beta_j\beta_k}(x, y; \beta)$ satisfy locally the Lipschitz property in a neighborhood of β^0, i.e., for any compact subset of $\operatorname{supp} x \times \operatorname{supp} y$ there exists a constant $L_p > 0$ such that for all $\beta, \beta' \in U(\beta^0, \delta)$, and $j, k = 1, \ldots, p$

$$\left| s''_{\beta_j\beta_k}(x, y; \beta) - s''_{\beta_j\beta_k}(x, y; \beta') \right| \leq L_p \cdot \|\beta - \beta'\| .$$

H3 Let

$$n^{-1/3} \max_{1\leq i\leq n} \max_{1\leq j\leq p} \left\| s'_{\beta_j}(x_i, y_i; \beta) \right\| = \mathcal{O}_p(1) \tag{12}$$

as $n \to +\infty$ uniformly over $\beta \in B$.

***Remark* 5** *This assumption depicts another regularity condition that is going to be fulfilled in most cases. It is a nonlinear equivalent of Assumption D1, equation (11). For example, for a function of the form $s(x, y; \beta) = (y - h(x^T\beta))^2$, where h is a twice differentiable function, we can immediately observe that*

$$\begin{aligned} s'_{\beta_j}(x, y; \beta) &= 2(y - h(x^T\beta))h'(x^T\beta)x_j = 2(\varepsilon + h(x^T\beta^0) - h(x^T\beta))^T h'(x^T\beta)x_j \\ &= 2(\varepsilon + h'(x^T\xi))h'(x^T\beta)x_j = h'(x^T\beta)\varepsilon x_j + h'(x^T\xi)h'(x^T\beta)x_j \\ &= s'_{\beta_j}(x, \varepsilon; \beta), \end{aligned}$$

where $\xi \in [\beta^0, \beta]_{\varkappa}$. Hence, assumption (12) is a direct consequence of (11) as long as the first derivative of $h(\cdot)$ is bounded on any compact subset of its domain.

H4 To proceed further, I have to postulate some assumptions about the following expectations:

- Let $\mathsf{E}\left[s(x_i, \varepsilon_i; \beta)\right]^m$ exist and be finite for $m = 1, 2$ and any $\beta \in B$.
- Let $\mathsf{E}\left[s'_{\beta_j}(x_i, y_i; \beta^0)\right]^m$ and $\mathsf{E}\left[s''_{\beta_j\beta_k}(x_i, y_i; \beta^0)\right]$ exist and be finite for $m = 1, 2$, and for all $j, k = 1, \ldots, p$.
- Moreover, I assume that $\mathsf{E}\left[s''_{\beta\beta}(x_i, y_i; \beta^0) \cdot I(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda))\right] = Q_h$, where Q_h is a nonsingular positive definite matrix.

***Remark* 6** *It is important to remember that these assumptions correspond in our nonlinear model to the existence of finite fourth moments (see Assumption D1). Moreover, the second part of Assumption H4 is a natural analogy to Assumption D2 in the linear regression model.*

The presented Assumptions D and H can be divided into several groups. First, some of them are standard in nonlinear regression, for example, D1, D2, H1, H2, and H4. Second, there are several assumptions that are needed to prove almost any result for order statistics (D3) and to analyze a trimmed objective function (D1, D4, H3). Finally, note that some assumptions can be weakened if one wants to derive only consistency instead of $\sqrt{n}$-consistency (for example, one would not then require the existence of the derivatives of the objective function).

3.3 Normal equations

In order to analyze the behavior of the GTE estimator (especially to prove $\sqrt{n}$-consistency), we use normal equations as the starting point, i.e., instead of minimizing the objective function

$$\rho(\beta) = \sum_{i=1}^{n} s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \le s_{[h]}(x_i, y_i; \beta)\big)$$

over all $\beta \in B$, we consider a solution of $\frac{\partial \rho(\beta)}{\partial \beta} = 0$. The normal equations (for $\beta \in U(\beta^0, \delta)$) can be written as

$$\begin{aligned} 0 = \frac{\partial \rho(\beta)}{\partial \beta} \quad = \quad & \sum_{i=1}^{n} \Big[s'_\beta(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \le s_{[h]}(x_i, y_i; \beta)\big) \\ & + s_i(x_i, y_i; \beta) \cdot \frac{\partial}{\partial \beta} I\big(s_i(x_i, y_i; \beta) \le s_{[h]}(x_i, y_i; \beta)\big) \Big]. \end{aligned}$$

Now, let us show that the continuity of $s_i(x_i, y_i; \beta)$ and order statistics $s_{[h]}(x_i, y_i; \beta)$ in β (Assumption H) guarantees that the second term is almost everywhere zero. Consider $j = 1, \ldots, p$ and an arbitrary, but fixed event $\omega \in \Omega^n$:

$$\begin{aligned} & \frac{\partial}{\partial \beta_j} I\big(s_i(x_i(\omega_i), y_i(\omega_i); \beta) \le s_{[h]}(x_i(\omega), y_i(\omega); \beta)\big) \\ = \quad & \lim_{\Delta \to 0} \frac{1}{\Delta} \Big[I\big(s_i(x_i(\omega_i), y_i(\omega_i); \beta^{(\Delta)}) \le s_{[h]}(x_i(\omega), y_i(\omega); \beta^{(\Delta)})\big) \\ & \qquad - I\big(s_i(x_i(\omega_i), y_i(\omega_i); \beta) \le s_{[h]}(x_i(\omega_i), y_i(\omega_i); \beta)\big) \Big], \end{aligned}$$

where $\beta^{(\Delta)} = (\beta_1, \ldots, \beta_{j-1}, \beta_j + \Delta, \beta_{j+1}, \ldots, \beta_p)$. As the ordering of terms $s_i(x_i(\omega_i), y_i(\omega_i); \beta)$ is constant in a neighborhood of β for all $\omega \in \Omega_1 \subseteq \Omega^n$, where $P(\Omega_1) = 1$ (see Lemma 1 below), the limit is equal to zero jointly for all $i = 1, \ldots, n$ and $j = 1, \ldots, n$ with probability

1. Consequently, it is enough to study the behavior of

$$\frac{\partial \rho(\beta)}{\partial \beta} = \sum_{i=1}^{n} s'_{\beta}(x_i, y_i; \beta) \cdot I(s_i(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)) \quad \text{a.s.}, \tag{13}$$

as the GTE estimator is a solution of $\frac{\partial \rho(\beta)}{\partial \beta} = 0$.

Lemma 1 *Let $n \in \mathbb{N}$ and $k_h(\beta) : \mathbb{R}^p \to \{1, \ldots, n\}$ be a function that represents an index of an observation such that $s_{k_h(\beta)}(x_i, y_i; \beta) = s_{[h]}(x_i, y_i; \beta)$, $h \in \{1, \ldots, n\}$. Under Assumptions D and H, there exists a set Ω_1, $P(\Omega_1) = 1$, such that for every $\omega \in \Omega_1 \subseteq \Omega^n$ there is some neighborhood $U(\beta^0, \varepsilon(\omega))$ of β^0 such that the function $k_h(\beta)$ is constant on $U(\beta^0, \varepsilon(\omega))$ for all $h \in \{1, \ldots, n\}$.*

Proof: See Appendix A. □

3.4 Consistency of general trimmed estimator

In Sections 2.1, I provided an intuitive argument why the consistency and asymptotic normality of (nonlinear) LTS and (nonlinear) least squares are equivalent. In this section, I will properly prove the consistency of GTE, and hence of nonlinear LTS as well.

To provide as complete a picture as possible about the consistency of GTE, I specify two sets of assumptions. The first group, Assumption NC, is as general as possible and is sufficient just for proving the consistency of GTE; the second group, Assumption NN, allows us to derive the $\sqrt{n}$-consistency of the estimator. In the presented form, Assumption NC corresponds mostly to the assumptions required for the uniform law of large numbers in nonlinear models, which is presented in a very general form in Andrews (1987).

Assumption NC. Let the following assumptions be satisfied for the function $q(x_i, y_i; \beta) = s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_{\beta}^{-1}(\lambda)\big)$.[8]

NC1 Let the parameter space B be a compact metric space (or a compact subset of $\mathbb{R}^p$).

NC2 Let $q(x_i, y_i; \beta)$, $q^*(x_i, y_i; \beta, \rho) = \sup\{q(x_i, y_i; \beta) : \beta' \in U(\beta, \rho)\}$, and $q_*(x_i, y_i; \beta, \rho) = \inf\{q(x_i, y_i; \beta) : \beta' \in U(\beta, \rho)\}$ be measurable random variables for all $\beta \in B, i \in \mathbb{N}$, and for all $\rho > 0$ sufficiently small.

[8] For the case of non-trimmed estimators, e.g., the nonlinear least squares or maximum likelihood, $\lambda = 1$ and $G_{\beta}^{-1}(\lambda) = \infty$. Therefore, this case corresponds to $q(x_i, \varepsilon_i; \beta) = s_i(x_i, y_i; \beta)$.

NC3 Let $\mathsf{E}\left\{\sup_{\beta\in B}|q(x_i,y_i;\beta)|\right\}^{1+\delta}<\infty$ for some $\delta>0$.

***Remark* 7** *Assumptions NC1–NC3 are necessary (together with the assumption concerning the differentiability of the function $s(x,y;\beta)$ with respect to β) for the uniform law of large numbers. Assumption NC3 is actually a standard condition used in this context to ensure that functions $q^*(x_i,y_i;\beta,\rho)$ and $q_*(x_i,y_i;\beta,\rho)$ satisfy pointwise the strong law of large numbers for any $\beta\in B$ and all ρ sufficiently small; see Andrews (1987), for instance. Moreover, note that the existence of an upper bound for $s_i(x_i,y_i;\beta)$ over β usually follows from the boundedness of the parametric space B and Assumption NC3 just requires additionally the existence of a certain expectation of this upper bound.*

NC4 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 that

$$\begin{aligned}\min_{\beta\in B-U(\beta^0,\varepsilon)}\mathsf{E}\,q(x_i,y_i;\beta)-\mathsf{E}\,q(x_i,y_i;\beta^0) \quad &= \\ \min_{\beta\in B-U(\beta^0,\varepsilon)}\mathsf{E}\left[s_i(x_i,y_i;\beta)\cdot I\big(s_i(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\big)\right]- & \\ -\mathsf{E}\left[s_i(x_i,y_i;\beta^0)\cdot I\Big(s_i(x_i,y_i;\beta^0)\le G_{\beta^0}^{-1}(\lambda)\Big)\right] \quad &> \quad \alpha(\varepsilon).\end{aligned}$$

***Remark* 8** *This is nothing but an analogy of the identification condition for the nonlinear least squares, see for example White (1980). However, it is one of the most important assumptions here and it is going to be even more important once we start to discuss GTE for limited-dependent-variable models.*

Now, the following theorem confirms that Assumption NC is sufficient for proving the consistency of GTE.

Theorem 1 *Let Assumptions D, H, and NC hold. Then, the general trimmed estimator defined for model (1) by*

$$\hat\beta_n^{(GTE,h_n)} = \arg\min_{\beta\in B}\sum_{i=1}^n s_i(x_i,y_i;\beta)\cdot I\big(s_i(x_i,y_i;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big) \tag{14}$$

$$= \arg\min_{\beta\in B}\frac{1}{n}\sum_{i=1}^n s_i(x_i,y_i;\beta)\cdot I\big(s_i(x_i,y_i;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big) \tag{15}$$

is consistent, i.e., $\hat\beta_n^{(GTE,h_n)}\to\beta^0$ in probability as $n\to+\infty$.

To prove this theorem, we need one additional lemma showing that we can use the uniform law of large numbers for sum (15) and that weak dependence among indicators

$$I\big(s_i(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)\big)$$

for $i = 1, \ldots, n$, does not spoil the result.

Lemma 6 *Let Assumptions D and H hold and assume that $t(x, y; \beta)$ is a real-valued function continuous in β uniformly in x and y over any compact subset of the support of (x, y). Moreover, assume that Assumptions NC1–NC3 hold for $t(x, y; \beta)$. Furthermore, let G_β denote the distribution function of $s_i(x_i, y_i; \beta)$ (for any $\beta \in B$). Finally, let $h_n/n \to \lambda \in \big(\frac{1}{2}, 1\big)$. Then*

$$\sup_{\beta \in B} \Bigg| \frac{1}{n} \sum_{i=1}^{n} \big[t(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big)\big] - \mathsf{E}\left[t(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right] \Bigg| \quad \to \quad 0$$

as $n \to +\infty$ in probability.

Proof: See Appendix A. □

Let us continue with the proof of Theorem 1 now.

Proof: The principle of the proof is actually very similar to the proof of the SLS consistency done by Ichimura (1993), and employs the theorem about uniform consistency in nonlinear models that is due to Andrews (1987) by means of Lemma 6. Let us denote ($s_i(x_i, y_i; \beta)$ are independent identically distributed random variables)

$$J(\beta) = \mathsf{E}\left\{s_1(x_i, y_i; \beta) \cdot I\big(s_1(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right\}, \quad \text{and}$$

$$J_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big).$$

By the definition of GTE, $P\Big(J_n\left(\hat{\beta}_n^{(GTE,h_n)}\right) < J_n\left(\beta^0\right)\Big) = 1$. For any $\delta > 0$ and an open neighborhood $U(\beta^0, \delta)$ of β^0, this probability can be decomposed as

$$\begin{aligned}
1 = P\Big(J_n\left(\hat{\beta}_n^{(GTE,h_n)}\right) < J_n\left(\beta^0\right)\Big) = \\
&= P\Big(J_n\left(\hat{\beta}_n^{(GTE,h_n)}\right) < J_n\left(\beta^0\right) \quad \text{and} \quad \hat{\beta}_n^{(GTE,h_n)} \in U(\beta^0,\delta)\Big) \\
&+ P\Big(J_n\left(\hat{\beta}_n^{(GTE,h_n)}\right) < J_n\left(\beta^0\right) \quad \text{and} \quad \hat{\beta}_n^{(GTE,h_n)} \in B - U(\beta^0,\delta)\Big) \\
&\leq P\Big(\hat{\beta}_n^{(GTE,h_n)} \in U(\beta^0,\delta)\Big) + P\Big(\inf_{\beta\in B-U(\beta^0,\delta)} J_n\left(\beta\right) < J_n\left(\beta^0\right)\Big).
\end{aligned}$$

Therefore, $P\big(\inf_{\beta\in B-U(\beta^0,\delta)} J_n\left(\beta\right) < J_n\left(\beta^0\right)\big) \to 0$ implies $P\Big(\hat{\beta}_n^{(GTE,h_n)} \in U(\beta^0,\delta)\Big) \to 1$ as $n \to +\infty$, that is, the consistency of $\hat{\beta}_n^{(GTE,h_n)}$ (δ was an arbitrary positive number). To verify $P\big(\inf_{\beta\in B-U(\beta^0,\delta)} J_n\left(\beta\right) < J_n\left(\beta^0\right)\big) \to 0$ note that

$$\begin{aligned}
&P\Big(\inf_{\beta\in B-U(\beta^0,\delta)} J_n\left(\beta\right) < J_n\left(\beta^0\right)\Big) \\
= \; &P\Big(\inf_{\beta\in B-U(\beta^0,\delta)} \left[J_n\left(\beta\right) - J(\beta) + J(\beta)\right] < J_n\left(\beta^0\right)\Big) \\
= \; &P\Big(\inf_{\beta\in B-U(\beta^0,\delta)} \left[J_n\left(\beta\right) - J(\beta)\right] < J_n(\beta^0) - \inf_{\beta\notin U(\beta^0,\delta)} J(\beta)\Big) \\
\leq \; &P\Big(\sup_{\beta\in B} \left|J_n\left(\beta\right) - J(\beta)\right| > \inf_{\beta\in B-U(\beta^0,\delta)} J(\beta) - J_n(\beta^0)\Big). \qquad (16)
\end{aligned}$$

Since $J_n(\beta^0) \to J(\beta^0)$ almost surely for $n \to \infty$ (see Assumption NC4 and remark 7) and the identification condition NC4 implies

$$(\forall\delta > 0)\,(\exists\alpha > 0)\left(\inf_{\beta\in B-U(\beta^0,\delta)} J(\beta) - J(\beta^0) > \alpha\right),$$

it immediately follows that

$$(\forall\delta > 0)\,(\exists\alpha > 0)\left(\lim_{n\to\infty}\left[\inf_{\beta\in B-U(\beta^0,\delta)} J(\beta) - J_n(\beta^0)\right] > \alpha\right)$$

almost surely for $n \to \infty$. Thus, to prove that (16) converges to zero as $n \to +\infty$ it is enough to show that for any $\alpha > 0$,

$$P\Big(\sup_{\beta\in B} \left|J_n\left(\beta\right) - J(\beta)\right| > \alpha\Big) \to 0 \text{ as } n \to +\infty.$$

This is indeed the result stated in Lemma 6 for the function $t(x_i, y_i; \beta) = s_i(x_i, y_i; \beta)$ ($s_i(x_i, y_i; \beta)$ is uniformly continuous in β on any compact subset of $\mathrm{supp}(x, \varepsilon)$ because of Assumption H1, and moreover, it satisfies Assumptions NC1–NC3). □

Next, let us recall that Assumption NC is sufficient for the consistency of GTE. However, if we enrich Assumption NC to obtain the below-stated Assumption NN, we are able to prove even the $\sqrt{n}$-consistence of a general trimmed estimator. Also, Assumption NN corresponds mostly to the assumptions required for the uniform law of large numbers in nonlinear models due to Andrews (1987), but they are applied additionally to the derivatives of the objective function.

Assumption NN.

NN1 Let Assumption NC hold, and additionally, Assumptions NC2–NC3 are satisfied for the function $q(x_i, \varepsilon_i; \beta) = s''_{\beta_j \beta_k}(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\big)$, where $j, k = 1, \ldots, p$.

NN2 Let $\mathsf{E}\left[s'_\beta(x_i, y_i; \beta^0) \cdot I(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda))\right] = 0$.

***Remark* 9** *This is actually a moment condition similar to those used in the (non)linear regression model. For example, for the nonlinear LTS estimator,* $s(x_i, y_i; \beta^0) = (y_i - h(x_i, \beta^0))^2$ *and* $s'_\beta(x_i, y_i; \beta^0) = (y_i - h(x_i, \beta^0)) \cdot h'_\beta(x_i, \beta^0) = \varepsilon_i \cdot h'_\beta(x_i, \beta^0)$. *Therefore,*

$$\begin{aligned} \mathsf{E}\left[s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right] &= \mathsf{E}\left[\varepsilon_i \cdot h'_\beta(x_i, \beta^0) \cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big)\right] \\ &= \mathsf{E}_x\left[h'_\beta(x_i, \beta^0) \cdot \mathsf{E}\left(\varepsilon_i \cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big) \,\middle|\, x_i\right)\right], \end{aligned}$$

where E_x *denotes the expectation taken over random variable* x_i. *Indeed,*

$$\mathsf{E}\left(\varepsilon_i \cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big) \,\middle|\, x_i\right) = 0 \tag{17}$$

is nothing but a typical moment condition used for the trimmed estimators (see, for example, Víšek (1996b) and Čížek (2001a)) and it is analogous to the usual orthogonality condition $\mathsf{E}(\varepsilon_i | x_i) = 0$. *All these conditions are clearly satisfied, for instance, if* ε_i *and* x_i *are independent random variables and* ε_i *has a symmetrical distribution. On the other hand, independence and symmetricity are not necessary conditions (*λ *is a fixed number).*

In the case of the MTLE estimator, the reasoning can be very similar (at least under the symmetry of ε_i*'s distribution) once we realize that* $s'_\beta(x_i, y_i; \beta^0) = l'_\beta(x_i, \varepsilon_i; \beta^0)/l(x_i, \varepsilon_i; \beta^0)$,

where $l(x_i, \varepsilon_i; \beta)$ is a likelihood function, and that the symmetricity of a likelihood function implies that the derivative of the likelihood function is symmetric with respect to the origin.

NN3 Assume that $\mathsf{E}\left(s'_\beta(x_i, y_i; \beta^0) \middle| s(x_i, \varepsilon_i; \beta^0) \in C\right)$ is uniformly bounded for all intervals C such that $G^{-1}(\lambda) \in C$.

***Remark* 10** *This assumption about conditional expectation does not limit us at all and represents just another regularity condition that cannot be expressed in another way in such a general setting. For example, in the case of LTS in classical nonlinear regression models, $s(x, y; \beta^0) = (y - h(x, \beta^0))^2$ and $s'_\beta(x, y; \beta^0) = (y - h(x, \beta^0)) \cdot h'_\beta(x, \beta^0) = \varepsilon_i \cdot h'_\beta(x, \beta^0)$; therefore*

$$\begin{aligned} \mathsf{E}\left(s'_\beta\big(x_i, y_i; \beta^0\big) \middle| s\big(x_i, \varepsilon_i; \beta^0\big) \in C\right) &= \mathsf{E}\left(\varepsilon_i \cdot h'_\beta(x, \beta^0) \middle| \varepsilon_i^2 \in C\right) \\ &\leq \mathsf{E}\left(\varepsilon_i \middle| \varepsilon_i^2 \in C\right) \cdot \mathsf{E}\left(h'_\beta(x, \beta^0)\right), \end{aligned}$$

where $\mathsf{E}\left(h'_\beta(x, \beta^0)\right) = const$ (provided that ε_i and x_i are independent). Therefore, we require $\mathsf{E}\left(\varepsilon_i \middle| \varepsilon_i^2 \in C\right)$ to be uniformly bounded over all intervals C containing a fixed point $G^{-1}(\lambda)$, which is guaranteed by Assumption D1 (ε_i has finite second moments).

Finally, combining all the conditions stated so far, namely, D, H, and NN, we can prove the $\sqrt{n}$-consistence of GTE.

Theorem 2 *Let Assumptions D, H, and NN hold. Then $\hat{\beta}_n^{(GTE,h_n)}$ is $\sqrt{n}$-consistent, i.e.,*

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

as $n \to +\infty$.

To prove this theorem, we need another assertion.

Lemma 4 *Under Assumptions D and H, for any fixed $i \in \mathbb{N}$ and $n \geq i$,*

$$P\Big(\exists \beta \in U(\beta^0, n^{-\frac{1}{2}}M) : I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big) \neq I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\Big) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$.

Proof: See Appendix A. □

Let us prove Theorem 2 now.

Proof: We already know that $\hat{\beta}_n^{(GTE,h_n)}$ is consistent (Theorem 1) because Assumption NC is a part of Assumption NN. Now, we shall use the normal equations presented in Section 3.3 to derive $\sqrt{n}$-consistency.

First, take a look at the derivatives of the objective function that form the normal equations. Let us denote the objective function of GTE by $S_n(x_i, y_i; \beta) = \frac{1}{n}\sum_{i=1}^{n} s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)\big)$. We have shown in Section 3.3 (see equation (13)) that

$$\frac{\partial S_n(x_i, y_i; \beta)}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} s'_\beta(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)\big)$$

almost surely, and by the same argument (Lemma 1), it follows that

$$\frac{\partial^2 S_n(x_i, y_i; \beta)}{\partial \beta \partial \beta^T} = \frac{1}{n}\sum_{i=1}^{n} s''_{\beta\beta}(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h]}(x_i, y_i; \beta)\big)$$

almost surely. Using Assumption NN and Lemma 6, we obtain for $S_n(x_i, y_i; \beta)$ and its second derivative that

$$\sup_{\beta \in B} \left| S_n(x_i, y_i; \beta) - \mathsf{E}\left[s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right]\right| \;\to\; 0, \tag{18}$$

$$\sup_{\beta \in B} \left| \frac{\partial^2 S_n(x_i, y_i; \beta)}{\partial \beta \partial \beta^T} - \mathsf{E}\left[s''_{\beta\beta}(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right]\right| \;\to\; 0 \tag{19}$$

in probability for $n \to \infty$ (the assumptions of Lemma 6 are satisfied for the loss function because of Assumptions H1 and NC and for its second derivative because of Assumptions H2 and NN1).

Next, $\hat{\beta}_n^{(GTE,h_n)}$ is a solution of the normal equations $\frac{\partial S_n(x_i, y_i; \beta)}{\partial \beta} = 0$. Thus, using Taylor's expansion theorem it holds

$$0 = \frac{\partial S_n(x_i, y_i; \hat{\beta}_n^{(GTE,h_n)})}{\partial \beta} = \frac{\partial S_n(x_i, y_i; \beta^0)}{\partial \beta} + \frac{\partial^2 S_n(x_i, y_i; \xi_n)}{\partial \beta \partial \beta^T} \cdot (\beta - \beta^0), \tag{20}$$

where $\xi_n \in \left[\beta^0, \hat{\beta}_n^{(GTE,h_n)}\right]_\varkappa$. Since $\hat{\beta}_n^{(GTE,h_n)} \to \beta^0$, the same holds for the sequence ξ_n: $\xi_n \to \beta^0$ in probability. Moreover, $\frac{\partial^2 S_n(x_i, y_i; \beta)}{\partial \beta \partial \beta^T}$ converges uniformly to a nonstochastic

function in β (see (19))

$$\mathsf{E}\left[s''_{\beta\beta}(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\big)\right],$$

which is continuous in β (see the verification of Assumption A3 in Lemma 5, Appendix A). Therefore,

$$\frac{\partial^2 S_n(x_i, y_i; \xi_n)}{\partial\beta\partial\beta^T} \to \mathsf{E}\left[s''_{\beta\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right] = Q_h$$

in probability as $n \to \infty$, where Q_h is a non-singular positive definite matrix (Assumption H4). Now, after rewriting (20) as

$$\sqrt{n}(\hat\beta_n^{(GTE,h_n)} - \beta^0) = -\left[\frac{\partial^2 S_n(x_i, y_i; \xi_n)}{\partial\beta\partial\beta^T}\right]^{-1}\left[\sqrt{n}\frac{\partial S_n(x_i, y_i; \beta^0)}{\partial\beta}\right],$$

it is clearly sufficient to verify only that $\sqrt{n}\frac{\partial S_n(x_i, y_i; \beta^0)}{\partial\beta} = \mathcal{O}_p(1)$ in order to prove $\sqrt{n}(\hat\beta_n^{(GTE,h_n)} - \beta^0) = \mathcal{O}_p(1)$ as $n \to \infty$.

So, let us analyze

$$\sqrt{n} \cdot \frac{\partial S_n(x_i, y_i; \beta^0)}{\partial\beta} = \frac{1}{n}\sum_{i=1}^{n} s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le s_{[h]}(x_i, y_i; \beta^0)\big)$$

and show that it behaves as $\mathcal{O}_p(1)$ for $n \to +\infty$. Apparently,

$$\begin{aligned}
\sqrt{n} \cdot \frac{\partial S_n(x_i, y_i; \beta^0)}{\partial\beta} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s'_\beta\big(x_i, y_i; \beta^0\big) \cdot \big[I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) \quad (21) \\
&\qquad\qquad -I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\big] \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big). \qquad (22)
\end{aligned}$$

First, we can employ once again the Chebyshev inequality for nonnegative random variables and write for the first term (21) and $j = 1, \ldots, p$

$$
\begin{aligned}
& P\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} s'_{\beta_j}\big(x_i, y_i; \beta^0\big) \cdot \times \right.\right. \\
& \qquad \left.\left. \times \left[I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) - I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right]\right| > K\right) \\
\le\ & \frac{1}{\sqrt{n}K}\,\mathsf{E}\left|\sum_{i=1}^{n} s'_{\beta_j}\big(x_i, y_i; \beta^0\big) \cdot \left[I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) - I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right]\right| \\
\le\ & \frac{\sqrt{n}}{K}\,\mathsf{E}\left|s'_{\beta_j}\big(x_i, y_i; \beta^0\big) \cdot \left[I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) - I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right]\right| \\
=\ & \frac{\sqrt{n}}{K}\,\mathsf{E}\left(\left|s'_{\beta_j}\big(x_i, y_i; \beta^0\big)\right| \,\middle|\, I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) \ne I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right) \\
& \times P\big(I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) \ne I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\big)
\end{aligned}
$$

Using Lemma 4 and Assumption NN3, we can rewrite this expectation as

$$
\begin{aligned}
& \frac{\mathcal{O}(1)}{K}\,\mathsf{E}\left(\left|s'_{\beta_j}\big(x_i, y_i; \beta^0\big)\right| \,\middle|\, I\big(s_i(x_i, y_i; \beta^0) \le s_{[h_n]}(x_i, y_i; \beta^0)\big) \ne I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right) \\
=\ & \frac{\mathcal{O}(1)}{K}\,\mathsf{E}\left(\left|s'_{\beta_j}\big(x_i, y_i; \beta^0\big)\right| \,\middle|\, s_i(x_i, y_i; \beta^0) \in \left[s_{[h_n]}(x_i, y_i; \beta^0), G^{-1}(\lambda)\right]_{\varkappa}\right) \\
=\ & \mathcal{O}(1)
\end{aligned}
$$

as $n \to \infty$. Hence, we can conclude that (21) behaves as $\mathcal{O}_p(1)$ as $n \to +\infty$.

Second, term (22) is bounded in probability as well: Assumption NN2 and H4 allows us to use the Feller-Lindenberg central limit theorem for (22) since

$$
\mathsf{E}\left[s'_{\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right] = 0
$$

and

$$
\mathsf{var}\left[s'_{\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right] \le \mathsf{var}\left[s'_{\beta}\big(x_i, y_i; \beta^0\big)\right]
$$

is finite (see Assumption H4). This in turn implies that (22) converges in distribution to a normally distributed random variable, and is, therefore, bounded in probability.

Thus, we have proved that $\sqrt{n}\frac{\partial S_n(x_i, \varepsilon_i; \beta^0)}{\partial \beta} = \mathcal{O}_p(1)$, $\frac{\partial^2 S_n(x_i, \varepsilon_i; \xi_n)}{\partial \beta \partial \beta^T} \to \lambda Q_h$, and consequently, $\sqrt{n}(\hat{\beta}_n^{(GTE, h_n)} - \beta^0) = \mathcal{O}_p(1)$ as $n \to +\infty$. $\square$

3.5 Identification condition

Altogether, we derived the consistency under Assumption NC and $\sqrt{n}$-consistency under Assumption NN. However, we did not discuss how restrictive all these assumptions are. We argued that most of them are just regularity conditions that are satisfied in practice, but there is one very important exception we did not discuss yet: the identification condition NC4. Therefore, we shall discuss its validity in a typical nonlinear regression model, both theoretically as well as specifically for nonlinear LTS and MTLE.

First, the identification condition NC4 can also be formulated so that

$$IC(\beta) = \mathsf{E}\left[s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right] \tag{23}$$

as a function of β has a unique minimum at β^0. A real-valued twice differentiable function, such as the objective function of GTE, has a unique minimum at $a \in \mathbb{R}$ if its first derivative equals zero at a and its second derivative is positive definite at a and positive semidefinite on the whole domain of the function. Thus, in order to verify the identification condition, it is sufficient to show that for all $\beta \in B$

$$\frac{\partial IC(\beta^0)}{\partial \beta} = 0, \qquad \frac{\partial^2 IC(\beta^0)}{\partial \beta^2} > 0, \quad \text{and} \quad \frac{\partial^2 IC(\beta)}{\partial \beta^2} \geq 0. \tag{24}$$

The first condition $\frac{\partial IC(\beta^0)}{\partial \beta} = 0$ actually says that β^0 is asymptotically a solution of normal equations, while the second condition $\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} > 0$ indicates that the objective function has a local minimum at β^0. The third condition $\frac{\partial^2 IC(\beta)}{\partial \beta^2} \geq 0$ guarantees the uniqueness of this solution.

Second, assuming the existence of derivatives of the objective function almost everywhere and their expectations (this is a part of Assumption H), it is possible to interchange the expectation and derivative: $\frac{\partial\, \mathsf{E}\, s(x_i,y_i;\beta)}{\partial \beta} = \mathsf{E}\, \frac{\partial s(x_i,y_i;\beta)}{\partial \beta}$. Moreover, we have shown in Section 3.3 that

$$\frac{\partial}{\partial \beta} I\big(s_i(x_i, y_i; \beta, \omega) \leq s_{[h]}(x_i, y_i; \beta, \omega)\big) = 0$$

almost surely. Therefore, we can write

$$\begin{aligned}
\frac{\partial IC(\beta)}{\partial \beta} &= \mathsf{E}\left[s'_\beta(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right], \quad \text{and} \\
\frac{\partial^2 IC(\beta)}{\partial \beta^2} &= \mathsf{E}\left[s''_{\beta\beta}(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right].
\end{aligned}$$

Finally, I summarize these results in the following Proposition 2.

Proposition 2 *Let Assumption H hold. Then, the sufficient conditions for Assumption NC4 are*

$$\frac{\partial IC(\beta^0)}{\partial \beta} = \mathsf{E}\left[s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] = 0, \tag{25}$$

$$\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} = \mathsf{E}\left[s''_{\beta\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] > 0, \quad \textit{and} \tag{26}$$

$$\frac{\partial^2 IC(\beta)}{\partial \beta^2} = \mathsf{E}\left[s''_{\beta\beta}(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right] \geq 0. \tag{27}$$

Alternatively, it is possible to say that a real-valued twice-differentiable function has a unique minimum at $a \in \mathbb{R}$ if its first derivative equals zero only at a and its second derivative is positive definite at a. Thus, in order to verify the identification condition, it is sufficient to show that for all $\beta \neq \beta^0$

$$\frac{\partial IC(\beta^0)}{\partial \beta} = 0, \qquad \frac{\partial IC(\beta)}{\partial \beta} \neq 0, \quad \text{and} \quad \frac{\partial^2 IC(\beta^0)}{\partial^2 \beta} > 0. \tag{28}$$

Similar to the previous case, the first condition $\frac{\partial IC(\beta^0)}{\partial \beta} = 0$ says that β^0 is a solution of the normal equations, the second condition $\frac{\partial IC(\beta)}{\partial \beta} \neq 0$ guarantees the uniqueness of this solution, and the third condition $\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} > 0$ shows that there is a local minimum at β^0. Consequently, this results in the following Proposition 3.

Proposition 3 *Let Assumption H hold. Then, the sufficient conditions for Assumption NC4 are*

$$\frac{\partial IC(\beta^0)}{\partial \beta} = \mathsf{E}\left[s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] = 0, \tag{29}$$

$$\frac{\partial IC(\beta)}{\partial \beta} = \mathsf{E}\left[s'_\beta(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right] \neq 0, \quad \textit{and} \tag{30}$$

$$\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} = \mathsf{E}\left[s''_{\beta\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] > 0. \tag{31}$$

Propositions 2 and 3 formulate sufficient conditions for the identification condition NC4 using the first and second derivatives of the objective function. The use of derivatives does not restrict us very much because of two reasons: first, the derivatives have to exist almost everywhere, but not at every point; second, although the assumption about the existence

of the second derivative is not needed for consistency, it is one of necessary assumptions for $\sqrt{n}$-consistency here, and as such, it is going to be required almost always.

***Remark* 11** *In the rest of the paper, I mostly verify only conditions (25) and (26), or alternatively (29) and (31). Thus, I only "prove" that there is a solution of the normal equations converging to the true value β^0, but not the uniqueness of this solution. This is a standard approach in such a general setting because there are no restrictions on the functional form, and hence, the uniqueness of the solution has to be assumed; see, for example, Amemiya (1983).*

3.5.1 Nonlinear least trimmed squares

In this section, I verify the identification condition for nonlinear LTS in the nonlinear regression model (1) under the standard assumptions used for the nonlinear least squares estimator. Until now we did not require the independence of error term ε_i and explanatory variables x_i, although it is one of most important conditions in regression analysis—together with $\mathsf{E}\,\varepsilon_i = 0$, it usually guarantees that the least-squares estimator is unbiased. Alternatively, moment conditions can be used, such as $\mathsf{E}(\varepsilon_i|x_i) = 0$. We show now that these assumptions directly imply that the identification condition for nonlinear LTS is satisfied; notice that (23) can be written as

$$IC(\beta) = \mathsf{E}\left[(y_i - h(x_i,\beta))^2 \cdot I\big((y_i - h(x_i,\beta))^2 \le G_\beta^{-1}(\lambda)\big)\right].$$

First, let us verify condition (25). The first derivative

$$\begin{aligned}
\frac{\partial IC(\beta^0)}{\partial \beta} &= \mathsf{E}\left[s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big)\right] \\
&= \mathsf{E}\left[-2\left(y_i - h(x_i,\beta^0)\right) h'_\beta(x_i,\beta^0) \cdot I\Big(\big(y_i - h(x_i,\beta^0)\big)^2 \le G^{-1}(\lambda)\Big)\right] \\
&= \mathsf{E}_x\left\{-2h'_\beta(x_i,\beta^0)\,\mathsf{E}\left[\varepsilon_i \cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big)\,\middle|\, x_i\right]\right\}
\end{aligned}$$

(E_x denotes the expectation taken over explanatory variables x_i) equals zero at β^0 if $\mathsf{E}[\varepsilon_i \cdot I(\varepsilon_i^2 \le G^{-1}(\lambda))|x_i] = 0$ for any x_i, for instance. This is an analogy to the standard orthogonality condition $\mathsf{E}(\varepsilon_i|x_i) = 0$ as discussed in remark 9 to Assumption NN2, equation (21). This "trimmed" orthogonality condition, $\mathsf{E}\left[\varepsilon_i \cdot I(\varepsilon_i^2 \le G^{-1}(\lambda))\,|\,x_i\right] = 0$, is satisfied, for example, when ε_i and x_i are independent random variables and ε_i is symmetrically distributed around zero.

Second, let us verify the second condition (26). It is apparently equivalent to Assumption H4:

$$\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} = \mathsf{E}\left[s''_{\beta\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] = Q_h > 0.$$

However, stronger, but more usual assumptions leading to this result are the moment condition

$$\mathsf{E}\left[\varepsilon_i \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \,\middle|\, x_i\right] = 0$$

together with the spheriality condition

$$\mathsf{E}\left[h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big)\right] = Q_{hh} > 0$$

is a positive definite matrix. Let us rewrite $\frac{\partial^2 IC(\beta^0)}{\partial \beta^2}$ to verify it:

$$\begin{aligned}
\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} &= \mathsf{E}\left[s''_{\beta\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \leq G^{-1}(\lambda)\big)\right] \\
&= \mathsf{E}\left[\left(2h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T\right) \cdot I\left(\big(y_i - h(x_i, \beta^0)\big)^2 \leq G^{-1}(\lambda)\right)\right] \\
&\quad - \mathsf{E}\left[\left(2\left(y_i - h(x_i, \beta^0)\right) h''_\beta(x_i, \beta^0)\right) \cdot I\left(\big(y_i - h(x_i, \beta^0)\big)^2 \leq G^{-1}(\lambda)\right)\right] \\
&= \mathsf{E}\left[\left(2h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T\right) \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big)\right] \\
&\quad - \mathsf{E}_x\left\{-2h''_\beta(x_i, \beta^0)\, \mathsf{E}\left[\varepsilon_i \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \,\middle|\, x_i\right]\right\} \\
&= Q_{hh} > 0.
\end{aligned}$$

Additionally, both the moment and spheriality conditions are trivially satisfied when ε_i and x_i are independent random variables, ε_i is symmetrically distributed around zero, and $\mathsf{E}\left(h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T\right)$ is a positive definite matrix:

$$\begin{aligned}
\mathsf{E}\left(h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T \cdot I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big)\right) &= \mathsf{E}\left(h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T\right) \cdot \mathsf{E}\, I\big(\varepsilon_i^2 \leq G^{-1}(\lambda)\big) \\
&= \mathsf{E}\left(h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T\right) \cdot \lambda > 0.
\end{aligned}$$

Finally, let me show that the identification assumption NC4 (locally in the sense of remark 11) can be obtained even without derivatives under the same conditions. Let us assume that ε_i and x_i are independent random variables, whereby ε_i is symmetrically

distributed around zero. Then

$$\begin{aligned}
& \mathsf{E}\left[s_i(x_i,y_i;\beta)\cdot I\big(s_i(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\big)\right]-\mathsf{E}\left[s_i(x_i,y_i;\beta^0)\cdot I\Big(s_i(x_i,y_i;\beta^0)\le G_{\beta^0}^{-1}(\lambda)\Big)\right] \\
= \;& \mathsf{E}\left[(y_i-h(x_i,\beta))^2\cdot I\big((y_i-h(x_i,\beta))^2\le G_\beta^{-1}(\lambda)\big)\right]-\mathsf{E}\left[\varepsilon_i^2\cdot I\Big(\varepsilon_i^2\le G_{\beta^0}^{-1}(\lambda)\Big)\right] \\
= \;& \mathsf{E}\left[(\varepsilon_i+h(x_i,\beta^0)-h(x_i,\beta))^2\cdot I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right]-\mathsf{E}\left[\varepsilon_i^2\cdot I\Big(\varepsilon_i^2\le G_{\beta^0}^{-1}(\lambda)\Big)\right] \\
= \;& \mathsf{E}\left[\varepsilon_i^2\cdot I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right]-\mathsf{E}\left[\varepsilon_i^2\cdot I\Big(\varepsilon_i^2\le G_{\beta^0}^{-1}(\lambda)\Big)\right] \\
& +2\,\mathsf{E}\left[\varepsilon_i(h(x_i,\beta^0)-h(x_i,\beta))\cdot I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right] \\
& +\mathsf{E}\left[(h(x_i,\beta^0)-h(x_i,\beta))^2\cdot I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right].
\end{aligned}$$

Next, $\mathsf{E}\left[\varepsilon_i(h(x_i,\beta^0)-h(x_i,\beta))\cdot I(r_i^2(\beta^0)=\varepsilon_i^2\le G^{-1}(\lambda))\right]=0$ because of the independence of ε_i and x_i and the symmetry of ε_i's distribution. Using Lemma 4 and the fact that random variables ε_i^2, ε_i and $h(x_i,\beta^0)-h(x_i,\beta)$ have finite first moments, we can rewrite this expression as

$$\begin{aligned}
& \mathsf{E}\left[s_i(x_i,y_i;\beta)\cdot I\big(s_i(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\big)\right]-\mathsf{E}\left[s_i(x_i,y_i;\beta^0)\cdot I\Big(s_i(x_i,y_i;\beta^0)\le G_{\beta^0}^{-1}(\lambda)\Big)\right] \\
= \;& \mathsf{E}\left[(h(x_i,\beta^0)-h(x_i,\beta))^2\cdot I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right]+o_p\big(\|\beta-\beta^0\|\big) \qquad (32)
\end{aligned}$$

as $n\to+\infty$. Unless $h(x_i,\beta)$ is a constant function on the domain defined by

$$I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)=1,$$

the expression $\mathsf{E}\left[(h(x_i,\beta^0)-h(x_i,\beta))^2\cdot I\big(r_i^2(\beta)\le G_\beta^{-1}(\lambda)\big)\right]$ is positive, and hence, the identification condition is (locally) satisfied.

Notice that the orthogonality condition

$$\mathsf{E}\left[\varepsilon_i\cdot I\big(\varepsilon_i^2\le G^{-1}(\lambda)\big)\,\big|\,x_i\right]=0 \qquad (33)$$

has to be satisfied in any case. Therefore, the distribution of ε_i conditional on x_i has to be symmetric around zero, even if we do not require the independence of ε_i and x_i.

3.5.2 Maximum trimmed likelihood

Now, I show that the assumption used in subsection 3.5.1, most importantly the independence of the error term ε_i and explanatory variables x_i together with the assumption about the symmetricity of a distribution function, are sufficient also for the identification

of MTLE. Provided that f is the probability density function of ε_i, where f is symmetric around zero, the probability density function of $y_i - h(x_i, \beta)$ conditional on x_i (in model (1)) can be written as $f(y_i - h(x_i, \beta))$. Therefore, the conditional likelihood function $l_i = l(x_i, y_i; \beta)$ for an observation (x_i, y_i) can be written as $l_i = f(y_i - h(x_i, \beta))$ and equation (23) as

$$IC(\beta) = \mathsf{E}_x \, \mathsf{E} \left[-\ln f(y_i - h(x_i, \beta)) \cdot I\big(-\ln f(y_i - h(x_i, \beta)) \le G_\beta^{-1}(\lambda)\big) \,\middle|\, x_i \right].$$

We verify first both conditions (25) and (26) conditionally on x_i, which then implies that they are also valid unconditionally (we check the equality to zero and the positiveness of conditional expectations).

First, let us verify condition (25). The first derivative equals

$$\begin{aligned}
\frac{\partial IC(\beta^0)}{\partial \beta} &= \mathsf{E}_x \, \mathsf{E} \left[s'_\beta\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \\
&= \mathsf{E}_x \, \mathsf{E} \left[\frac{f'(y_i - h(x_i, \beta^0))}{f(y_i - h(x_i, \beta^0))} h'_\beta(x_i, \beta^0) \cdot I\big(-\ln f(y_i - h(x_i, \beta^0)) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \\
&= \mathsf{E}_x \left\{ h'_\beta(x_i, \beta^0) \cdot \mathsf{E} \left[\frac{f'(\varepsilon_i)}{f(\varepsilon_i)} \cdot I\big(-\ln f(\varepsilon_i) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \right\}.
\end{aligned}$$

Since f is symmetric around zero, its first derivative f' is symmetric around the origin. The trimming $I(-\ln f(\varepsilon_i) \le G^{-1}(\lambda))$ is then symmetric as well, and hence,

$$\mathsf{E} \left[\frac{f'(\varepsilon_i)}{f(\varepsilon_i)} \cdot I\big(-\ln f(\varepsilon_i) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] = 0.$$

Second, condition (26) can be checked using the same kind of argument. It holds that

$$\begin{aligned}
\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} &= \mathsf{E}_x \, \mathsf{E} \left[s''_{\beta\beta}\big(x_i, y_i; \beta^0\big) \cdot I\big(s_i(x_i, y_i; \beta^0) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \\
&= \mathsf{E}_x \, \mathsf{E} \left[\frac{f'(\varepsilon_i)}{f(\varepsilon_i)} h''_\beta(x_i, \beta^0) \cdot I\big(-\ln f(\varepsilon_i) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \\
&+ \mathsf{E}_x \, \mathsf{E} \left[\left(\frac{-f'' \cdot f + f' \cdot f'}{f^2} \right)(\varepsilon_i) \cdot h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T \cdot I\big(-\ln f(\varepsilon_i) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \\
&= \mathsf{E}_x \left\{ h'_\beta(x_i, \beta^0) h'_\beta(x_i, \beta^0)^T \cdot \mathsf{E} \left[[-\ln f(\varepsilon_i)]'' \cdot I\big(-\ln f(\varepsilon_i) \le G^{-1}(\lambda)\big) \,\middle|\, x_i \right] \right\}.
\end{aligned}$$

Therefore, $\frac{\partial^2 IC(\beta^0)}{\partial \beta^2}$ is positive definite if $\mathsf{E} \left[\big(\ln f(\varepsilon_i)\big)'' \cdot I(-\ln f(\varepsilon_i) \le G^{-1}(\lambda)) \,\middle|\, x_i \right] < 0$,

that is, if $\ln f(\varepsilon_i)$ is "on average" concave on the domain defined by $-\ln f(\varepsilon_i) \leq G^{-1}(\lambda)$. This assumption is equivalent to $\mathsf{E}\left[\left(\ln f(\varepsilon_i)\right)'' \middle| x_i\right] < 0$ in the case of MLE and guarantees that the solution of the MTLE normal equations corresponds to a maximum of the trimmed likelihood function.

Thus, we have shown that the identification condition is satisfied under the standard assumptions used for non-trimmed estimators (NLS, MLE) in the nonlinear regression model (1). Notice that the symmetry of the distribution of ε_i (conditional on x_i) is crucial for the consistency of GTE in both cases.

4 Consistency of GTE in limited-dependent-variable models

Let us now turn our attention to the properties of the general trimmed estimator (GTE) in limited-dependent-variable models. In Section 1, we showed that structural model (2) can be transformed to reduced model (3)

$$y_i = h(x_i, \beta) + \nu_i, \tag{34}$$

where $y_i = \tau(\tilde{y}_i)$ ($\tilde{y}_i$ is the original unobservable dependent variable), $h(x_i, \beta) = \mathsf{E}(y_i|x_i) = \mathsf{E}(\tau(\tilde{y}_i)|x_i)$, and $\nu_i = y_i - \mathsf{E}(y_i|x_i) = y_i - h(x_i, \beta)$. The only important difference in comparison with the classical nonlinear regression model (1) is the error structure: ν_i and x_i are not independent random variables here. Nevertheless, Assumptions D, H, and NC used to prove consistency of GTE do not require ν_i and x_i to be independent—the only requirement regarding the variables entering the regression model is that random vectors (y_i, x_i) form a sequence of independent and identically distributed random variables. Hence, the theorems presented in Section 3.4 can be applied in the limited-dependent-variable framework as well. We just have to use the same set of assumptions as in Section 3.4, where the error term ε_i is replaced by ν_i.

Theorem 3 *Let Assumptions D, H, and NC hold for reduced model (3). Then the general trimmed estimator defined for model (3) by*

$$\hat{\beta}_n^{(GTE,h_n)} = \arg\min_{\beta \in B} \sum_{i=1}^{n} s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big)$$

is consistent, i.e., $\hat{\beta}_n^{(GTE,h_n)} \to \beta^0$ *in probability as* $n \to +\infty$.

The same argument applies also to $\sqrt{n}$-consistency.

Theorem 4 *Let Assumptions D, H, and NN hold for reduced model (3). Then* $\hat{\beta}_n^{(GTE,h_n)}$ *is* $\sqrt{n}$*-consistent, i.e.,*

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

as $n \to +\infty$.

Although we can use the same theoretical results in the limited-dependent-variable framework, the most important question is whether Assumptions D, H, and NC (or NN) can be satisfied. For example, Assumption D requires ν_i to be an absolutely continuous random variable. Thus, either dependent variable y_i or at least one of the explanatory variables x_i has to be continuously distributed.[9] Nevertheless, most of the conditions are just usual regularity assumptions, which do not restrict us, and the only important assumption that has to be verified is the identification condition NC4. It requires 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 that

$$\begin{aligned}\min_{\beta \in B - U(\beta^0,\varepsilon)} & \mathsf{E}\left[s_i(x_i, y_i; \beta) \cdot I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right] - \\ & - \mathsf{E}\left[s_i(x_i, y_i; \beta^0) \cdot I\Big(s_i(x_i, y_i; \beta^0) \leq G_{\beta^0}^{-1}(\lambda)\Big)\right] \quad > \quad \alpha(\varepsilon).\end{aligned}$$

Similarly to nonlinear regression, we can use here Propositions 2 and 3 to verify the identification condition. However, in contrast to the case of the nonlinear regression model (1) discussed in Section 3, we cannot assume that the explanatory variables x_i and the error term ν_i are independent. On the other hand, we have shown in Section 3.5 that a sufficient assumption for the identification of GTE can be the symmetry of the distribution of ν_i conditional on x_i; for example, it is sufficient to require in the case of the LTS estimator that

$$\mathsf{E}\left[\nu_i \cdot I\big(\nu_i^2 \leq G^{-1}(\lambda)\big) \,\middle|\, x_i\right] = 0 \tag{35}$$

(see equation (33)). Therefore, the identification condition NC4 can be treated in the same way as in Section 3.5, but contrary to the nonlinear regression model, it cannot be verified under some general conditions—we have to check it for every class of limited-dependent-variable models separately. Examples are provided in Section 5.

[9] This is a typical assumption used for the semiparametric estimation of limited-dependent-variable models, see Ichimura (1993), for instance.

5 Examples of trimmed estimators

In this section, I provide several typical examples of nonlinear regression, limited-dependent-variable, and binary-choice models. Additionally, if it is not possible to directly use nonlinear LTS or MTLE for some models, a general-trimmed-estimator concept is used to design a robust estimator for such models. In such cases, the identification assumption NC4 is verified.

5.1 Nonlinear regression models

I proved the consistency of nonlinear LTS and MTLE for a general nonlinear model $y_i = h(x_i, \beta) + \varepsilon_i$ including the verification of the identification condition in Section 3. Therefore, I mention here directly several examples of econometric applications, where an application of LTS or MTLE is meaningful.

First, it is sometimes not clear, for instance, which functional form best describes the dependence on an explanatory variable. To resolve this point, the Box-Cox transformation can be used (see Box and Cox (1964)), which is a transformation of a random variable Z parameterized by $\lambda \in \mathbb{R}$ having the following form:

$$Z^{(\lambda)} = \begin{cases} \dfrac{Z^\lambda - 1}{\lambda} & \text{for} \quad \lambda \neq 0, \\ \ln Z & \text{for} \quad \lambda = 0. \end{cases}$$

Its advantage is that $Z^{(\lambda)}$ represents various functions of Z for different values of λ: linear ($\lambda = 1$), square root ($\lambda = 1/2$), logarithmic ($\lambda = 0$), inversely proportional ($\lambda = -1$), and so on. Applying the transformation either to the dependent or to independent variables provides then a parameterized choice between different regression models (linear, log-linear, semi-logarithmic, reciprocal, etc.) by means of a nonlinear regression model such as $y_i = \beta_0 + x_i^{(\lambda)}\beta_1 + \varepsilon_i$.

Next, another example of an intrinsically nonlinear model can be a model with an exponential regression function but an additive error term. For example, the estimation of a CES production function leads to the regression function $y_i = \alpha \cdot \left(\sum_{i=1}^k \delta_i x_i^{-\gamma}\right)^{-\lambda/\gamma}$, which is usually rewritten and estimated as

$$\ln y_i = \ln \alpha - \frac{\lambda}{\gamma} \ln \left[\sum_{i=1}^k \delta_i x_i^{-\gamma}\right] + \varepsilon_i,$$

where $\gamma \geq -1, 0 < \delta < 1$, and $\alpha > 0$. This model is intrinsically nonlinear because we cannot rewrite the regression function as a linear function of the parameters $\alpha, \gamma, \delta, \lambda$.

Further, in the analysis of economic time series, models allowing for a state-dependent regression are very popular. An example of such models is the self-exciting threshold autoregressive specification (SETAR):

$$y_t = \begin{cases} \alpha_0 + \sum_{i=1}^{p} y_{t-i}\alpha_i + \varepsilon_t & \text{if} \quad y_{t-d} \in (-\infty, c), \\ \beta_0 + \sum_{i=1}^{p} y_{t-i}\beta_i + \varepsilon_t & \text{if} \quad y_{t-d} \in (c, \infty) \end{cases}$$

(see Tong (1990)), where c is the threshold and $d \in \{1, \ldots, p\}$ is the delay parameter. Its main feature is that switching depends on a past realization y_{t-d}. This specification can be generalized to the smooth threshold autoregressive model (STAR), which allows for a smooth transition between states by means of a general function $h(y_{t-d}; c, \delta) : \mathbb{R} \to \langle 0, 1 \rangle$:

$$y_t = \alpha_0 + \sum_{i=1}^{p} y_{t-i}\alpha_i + \left(\delta_0 + \sum_{i=1}^{p} y_{t-i}\delta_i \right) \cdot h(y_{t-d}; c, \delta) + \varepsilon_t.$$

An extensive review of existing variants of STAR models is given by Dijk, Terasvirta, and Franses (2000). Because of the nonlinear nature of these models, nonlinear least squares is typically used for their estimation, and thus, nonlinear LTS can be a more robust candidate for the estimation of STAR. STAR was used, for example, by Proietti (1998) to model business-cycle asymmetries or by Terasvirta and Anderson (1992) to characterize differences in the dynamics of industrial production indices during expansion and recession periods.

5.2 Limited-dependent-variable models

5.2.1 Truncated regression

Suppose there are a dependent variable $\tilde{y}_i$ and explanatory variables x_i, but we cannot see observations with values of $\tilde{y}_i$ below some $c \in \mathbb{R}$. Thus, the regression model is truncated and we observe only $(y_i = \tilde{y}_i, x_i)$ such that $\tilde{y}_i > c$. This can happen, for example, if we study an individual's utility (in monetary terms) from an object—we can observe the individual's utility and other characteristics only if it exceeds the price of the item and he buys the object. The described situation corresponds to a truncation from below. Although there are

other possibilities (truncation from above, double truncation), I only deal with truncation from below in this example. Truncated regression models are typically estimated by MLE (see Maddala (1983)) or by symmetrically trimmed least squares (STLS), see Powell (1986). Moreover, there are many semiparametric methods available. Hausman and Wise (1976) used a truncated regression to estimate earnings functions.

It is possible to modify both parametric approaches to create corresponding trimmed estimators. The most important part of this adaptation is to verify the identification condition. Because we cannot assume that the error term and explanatory variables are independent, we have to make sure that the error term conditional on the explanatory variables has a symmetric distribution around zero (see Section 4). I demonstrate this in the rest of this section.

First, consider the STLS estimator. Its main idea is to trim the dependent variable (truncated from below) additionally from above to make it symmetrically distributed. Consider a linear regression model, $\tilde{y}_i = x_i^T\beta + \varepsilon_i$, where the dependent variable $\tilde{y}_i$ is truncated at c, and let y_i be the observed response. Then the error term ε_i conditional on x_i is truncated at $c - x_i^T\beta$. Now, the principle of STLS is to truncate $\varepsilon_i|x_i$ symmetrically at $x_i^T\beta - c$. This corresponds to the truncating of $\tilde{y}_i$ at c and $2x_i^T\beta - c$, or equivalently, to the truncating of y_i at $2x_i^T\beta - c$. Powell (1986) showed that this can be achieved by minimizing

$$\sum_{i=1}^{n} \left(y_i - \max\left\{0.5y_i + 0.5c, x_i^T\beta\right\}\right)^2$$

with respect to β. Since the objective function is continuous and differentiable in β almost everywhere, it is possible to propose the corresponding trimmed STLS estimator minimizing

$$\sum_{i=1}^{h} r_{[i]}^2(\beta) = \sum_{i=1}^{n} r_i^2(\beta) \cdot I\left(r_i^2(\beta) \leq r_{[i]}^2(\beta)\right)$$

with respect to β, where $r_i^2(\beta) = \left(y_i - \max\left\{0.5y_i + 0.5c, x_i^T\beta\right\}\right)^2$.

The next step is to verify the identification condition, that is, to check conditions (25) and (26) for $IC(\beta) = r_i^2(\beta) \cdot I\left(r_i^2(\beta) \leq G_\beta^{-1}(\lambda)\right)$. To do so, I use the same assumptions about the distribution function f of ε_i as Powell (1986), namely, that f is symmetric and

unimodal[10] around zero. The first derivative

$$\begin{aligned}\frac{\partial IC(\beta^0)}{\partial \beta} &= \mathsf{E}_x \mathsf{E}\left[-2r_i(\beta^0)\cdot x_i I\big(y_i \le 2x_i^T\beta^0 - c\big)\cdot I\big(r_i^2(\beta^0) \le G^{-1}(\lambda)\big)\big|\, x_i\right] \\ &= \mathsf{E}_x\left\{x_i \mathsf{E}\left[-2\left(y_i - x_i^T\beta^0\right)\cdot I\big(y_i \le 2x_i^T\beta^0 - c\big)\cdot I\big(r_i^2(\beta^0) \le G^{-1}(\lambda)\big)\big|\, x_i\right]\right\} \\ &= \mathsf{E}_x\left\{x_i \mathsf{E}\left[-2\varepsilon_i \cdot I\big(\varepsilon_i \le x_i^T\beta^0 - c\big)\cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big)\big|\, x_i\right]\right\} = 0,\end{aligned}$$

because ε_i conditionally on x_i is truncated from below at $c - x_i^T\beta^0$, the truncation from above is done by $I\big(\varepsilon_i \le x_i^T\beta^0 - c\big)$, and the trimming $I(\varepsilon_i^2 \le G^{-1}(\lambda))$ is symmetric. Similarly, the second derivative

$$\begin{aligned}\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} &= \mathsf{E}_x \mathsf{E}\left[2x_i x_i^T \cdot I\big(y_i \le 2x_i^T\beta^0 - c\big)\cdot I\big(r_i^2(\beta^0) \le G^{-1}(\lambda)\big)\big|\, x_i\right] \\ &= \mathsf{E}_x\left\{2x_i x_i^T \mathsf{E}\left[I\big(\varepsilon_i \le x_i^T\beta^0 - c\big)\cdot I\big(\varepsilon_i^2 \le G^{-1}(\lambda)\big)\big|\, x_i\right]\right\} > 0\end{aligned}$$

as long as $\mathsf{E}\left(x_i x_i^T P\big(\varepsilon_i \le x_i^T\beta^0 - c\big|\, x_i\big)\right)$ is a positive definite matrix. This is an analogy to the typically used spheriality condition $\mathsf{E}\, x_i x_i^T > 0$ (see Assumption D2).

Second, MLE can also be generalized to MTLE in truncated regression models, but the generalization is not so straightforward as in the case of STLS. The identification condition NC4 requires

$$\mathsf{E}\left(\varepsilon_i I\big(s(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\big)\big|\, x_i\right) = 0,$$

so the distribution function F of ε_i and the trimming $I\big(s(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\big)$ conditional on x_i have to be symmetric around zero. Therefore, the likelihood has to be truncated in the same way as the error term in STLS—instead of the usual $l(x_i, y_i; \beta) = f(y_i - x_i^T\beta)/(1 - F(c - x_i^T\beta))\cdot I(y_i \ge c)$, the following truncated likelihood has to be used

$$l(x_i, y_i; \beta) = \frac{f(y_i - x_i^T\beta)}{(F(x_i^T\beta - c) - F(c - x_i^T\beta))}\cdot I\big(c \le y_i \le 2x_i^T\beta - c\big) \tag{36}$$

(i.e., the trimming at $y_i = 2x_i^T\beta - c$ is added). Then the corresponding MTLE estimator for a regression model with truncation from below at c is defined by

$$\hat\beta_n^{(MTLE,h)} = \arg\min_{\beta\in B} \sum_{i=1}^{h} s_{[i]}(x_i, y_i; \beta),$$

[10] Function f with mode at x_0 is unimodal if for any two values $x_1 < x_2$ and $f(x_1) < f(x_2)$ it follows that $x_1 < x_0$. Similarly, if $x_1 < x_2$ and $f(x_1) > f(x_2)$, then $x_0 < x_2$.

where $s_i(x_i, y_i; \beta) = -\ln l_i(x_i, y_i; \beta)$, as defined in (36). The identification condition can be verified in the same way as for the trimmed variant of STLS.

5.2.2 Censored regression

Similarly to truncated regression, consider a dependent variable $\tilde{y}_i$ and explanatory variables x_i such that we observe c instead of values $\tilde{y}_i$ smaller than c; let y_i denote the observed value of $\tilde{y}_i$. This is a special case of censored (Tobit) regression with censoring from below, which I discuss here. Other types of censoring, such as censoring from above, exist and can typically arise when we observe, for example, the duration of an event: if it lasts too long, we do not see the end of the event and its duration is censored (i.e., replaced by the time corresponding to the maximum duration we observed).

There is an extensive literature concerning the Tobit model. A summary of classical methods is given by Maddala (1983), for instance. On the other hand, it is beneficial in this context to use as a starting point an estimator motivated by STLS—symmetrically censored least squares (SCLS), see Powell (1986). It is widely used for censored regression, for example, by Lee (1995) to study female labor supply data. SCLS is a suitable candidate for generalizing to a trimmed estimator because it is also based on the symmetrization of a censored error-term distribution.

Similarly to STLS, the SCLS estimator is based on additional censoring of the dependent variable (censored from below) from above. Assuming a linear regression model $\tilde{y}_i = x_i^T\beta + \varepsilon_i$, and censoring y_i of the dependent variable $\tilde{y}_i$ at c, the error term ε_i conditional on x_i is censored at $c - x_i^T\beta$. Thus, SCLS censors $\varepsilon_i|x_i$ also at $x_i^T\beta - c$. This corresponds to the censoring of y_i at $2x_i^T\beta - c$. Powell (1986a) showed that this can be achieved by minimizing

$$\sum_{i=1}^{n} \Big\{ \big(y_i - \max\{0.5y_i + 0.5c, x_i^T\beta\}\big)^2 + I\big(y_i > 2x_i^T\beta - c\big) \cdot \big[(0.5y_i)^2 - \max\{c, x_i^T\beta\}^2\big] \Big\}$$

with respect to β. Since this objective function is continuous and differentiable almost everywhere with respect to β, it is possible to propose the corresponding trimmed SCLS estimator minimizing

$$\sum_{i=1}^{h} r_{[i]}^2(\beta) = \sum_{i=1}^{n} r_i^2(\beta) \cdot I\big(r_i^2(\beta) \le r_{[i]}^2(\beta)\big),$$

where

$$r_i^2(\beta) = \big(y_i - \max\{0.5y_i + 0.5c, x_i^T\beta\}\big)^2 + I\big(y_i > 2x_i^T\beta - c\big)\cdot\big[(0.5y_i + 0.5c)^2 - \max\{0.5c, x_i^T\beta\}^2\big].$$

The identification condition for this trimmed SCLS estimator can be verified in the same way as for the trimmed STLS.

5.3 Binary-choice models

Binary-choice models usually arise when we model a decision or a response of an individual; for example, Horowitz (1993) studies a binary trip-mode choice problem. In these models, the dependent variable y_i has just two possible values—zero and one—and its expectation is described by a function of an index $x_i^T\beta$. Moreover, we assume that there is a structural model (2) describing the decision, for example, a linear regression model characterizing the (unobservable) utility from decision $\tilde{y}_i$. If the error term in the corresponding structural model follows a symmetric distribution function F, then $P(y_i = 1|x_i) = F(x_i^T\beta)$ and the model (e.g., probit if F is the standard normal distribution) is estimated by MLE with a likelihood equal to

$$L(\beta) = \sum_{i=1}^{n} l_i(x_i, y_i; \beta), \quad \text{where} \quad l_i(x_i, y_i; \beta) = y_i \ln F(x_i^T\beta) + (1 - y_i)\ln(1 - F(x_i^T\beta)).$$

In this section, I design a trimmed MLE estimator for binary-choice models.

Now, consider a MTLE estimator based on the likelihood function $L(\beta)$, that is, the estimator minimizing $\sum_{i=1}^{h} -l_{[i]}(x_i, y_i; \beta)$. We can try to verify condition (25) first, where

$$IC(\beta) = -l_i(x_i, y_i; \beta)\cdot I\big(-l_i(x_i, y_i; \beta) \le G_\beta^{-1}(\lambda)\big).$$

Then

$$\frac{\partial IC(\beta^0)}{\partial \beta} = -\mathsf{E}_x \mathsf{E}\left[\left(\frac{y_i f(x_i^T\beta^0)}{F(x_i^T\beta^0)}x_i - \frac{(1-y_i)f(x_i^T\beta^0)}{1-F(x_i^T\beta^0)}x_i\right)\cdot I\big(-l_i(x_i,y_i;\beta^0)\le G^{-1}(\lambda)\big)\right] \quad (37)$$

$$= -\mathsf{E}_x\left(P(y_i=1|x_i)\frac{f(x_i^T\beta^0)}{F(x_i^T\beta^0)}x_i\cdot I\big(-l_i(x_i,1;\beta^0)\le G^{-1}(\lambda)\big)\right) \quad (38)$$

$$+\mathsf{E}_x\left(P(y_i=0|x_i)\frac{f(x_i^T\beta^0)}{1-F(x_i^T\beta^0)}x_i\cdot I\big(-l_i(x_i,0;\beta^0)\le G^{-1}(\lambda)\big)\right) \quad (39)$$

$$= \mathsf{E}_x\{f(x_i^T\beta^0)x_i\times \quad (40)$$

$$\times\left[I\big(-\ln(1-F(x_i^T\beta^0))\le G^{-1}(\lambda)\big) - I\big(-\ln F(x_i^T\beta^0)\le G^{-1}(\lambda)\big)\right]\}. \quad (41)$$

Apparently, this expectation equals zero if for all possible values of random variable x

$$I\big(\ln F(x^T\beta^0)\le G^{-1}(\lambda)\big) = I\big(\ln(1-F(x^T\beta^0))\le G^{-1}(\lambda)\big),$$

or equivalently, $\{x|x^T\beta^0\le C\} = \{x|-x^T\beta^0\le C\}$ for $C = F^{-1}\left(e^{G^{-1}(\lambda)}\right)$, where F^{-1} represents the inverse of the distribution function F. Such a condition cannot be satisfied unless $I\big(-\ln F(x^T\beta^0)\le G^{-1}(\lambda)\big)=1$ or $I\big(-\ln F(x^T\beta^0)\le G^{-1}(\lambda)\big)=0$ for all possible x. In other words, condition (25) is satisfied only if there is no trimming (this corresponds to MLE) or complete trimming (the objective function equals zero everywhere). Hence, a trimmed estimator cannot be designed this way.

On the other hand, equations (37)–(40) indicate that condition (25) would be satisfied if the trimming is independent of y_i. It is possible to achieve this by the symmetrization of the trimming part $I(l_i(x_i,y_i;\beta^0)\le G^{-1}(\lambda))$, for example, by replacing it with

$$I\big(-l_i(x_i,0;\beta^0)\le G^{-1}(\lambda)\big)\cdot I\big(-l_i(x_i,1;\beta^0)\le G^{-1}(\lambda)\big) = I\left(-\max_{y\in\{0,1\}} l_i(x_i,y;\beta^0)\le G^{-1}(\lambda)\right)$$

(the distribution function G should now describe random variable $\max_{y_i\in\{0,1\}} l_i(x_i,y_i;\beta^0)$). This way, I trim the observations for which y_i equals 0 or 1 with a probability very close to 1. Therefore, I propose the following MTLE estimator for binary-choice models:

$$\hat\beta_n^{(MTLE,h)} = \arg\min_{\beta\in B}\sum_{i=1}^n -l_i(x_i,y_i;\beta)\cdot I\left(s_i(x_i;\beta)=\max_{y\in\{0,1\}} -l_i(x_i,y;\beta)\le s_{[i]}(x_i;\beta)\right),$$

where $s_i(x_i;\beta)=\max_{y\in\{0,1\}} -l_i(x_i,y;\beta)$ and $l_i(x_i,y_i;\beta) = y_i\ln F(x_i^T\beta)+(1-y_i)\ln(1-F(x_i^T\beta))$. The corresponding function for the verification of the identification conditions

(25) and (26) is

$$IC(\beta) = -l_i(x_i, y_i; \beta) \cdot I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta) \leq G_\beta^{-1}(\lambda) \right).$$

So, let me verify conditions (25) and (26) for the proposed MTLE estimator. First, (25) can be expressed as

$$\begin{aligned}
\frac{\partial IC(\beta^0)}{\partial \beta} &= \mathsf{E}_x \, \mathsf{E} \left[- \left(\frac{y_i f(x_i^T \beta^0)}{F(x_i^T \beta^0)} x_i - \frac{(1-y_i) f(x_i^T \beta^0)}{1 - F(x_i^T \beta^0)} x_i \right) \cdot I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda) \right) \right] \\
&= - \mathsf{E}_x \left(P(y_i = 1 | x_i) \frac{f(x_i^T \beta^0)}{F(x_i^T \beta^0)} x_i \cdot I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda) \right) \right) \\
&\quad + \mathsf{E}_x \left(P(y_i = 0 | x_i) \frac{f(x_i^T \beta^0)}{1 - F(x_i^T \beta^0)} x_i \cdot I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda) \right) \right) \\
&= 0.
\end{aligned}$$

Similarly, the second derivative in (26) can be written as

$$\begin{aligned}
\frac{\partial^2 IC(\beta^0)}{\partial \beta^2} &= - \mathsf{E}_x \left\{ \mathsf{E} \left(y_i \left(\frac{f'F - f^2}{F^2} \right) (x_i^T \beta^0) x_i - (1-y_i) \left(\frac{f'(1-F) + f^2}{(1-F)^2} \right) (x_i^T \beta^0) x_i \right) \times \right. \\
&\qquad \left. \times I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda) \right) \right\} \\
&= - \mathsf{E}_x \left\{ \left(\left(\frac{f'F - f^2}{F} \right) (x_i^T \beta^0) x_i x_i^T - \left(\frac{f'(1-F) + f^2}{(1-F)} \right) (x_i^T \beta^0) x_i x_i^T \right) \times \right. \\
&\qquad \left. \times I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda) \right) \right\} \\
&= \mathsf{E}_x \left\{ \left(\left(\frac{f^2}{F} \right) (x_i^T \beta^0) x_i x_i^T + \left(\frac{f^2}{(1-F)} \right) (x_i^T \beta^0) x_i x_i^T \right) \times \right. \\
&\qquad \left. \times I\left(- \max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda) \right) \right\}.
\end{aligned}$$

Thus, the only difference between the standard maximum likelihood condition and condition (26) is the indicator $I\big(-\max_{y \in \{0,1\}} l_i(x_i, y; \beta^0) \leq G^{-1}(\lambda)\big)$, and in most cases, $\frac{\partial^2 IC(\beta^0)}{\partial \beta^2}$ will be a positive definite matrix as long as MLE is identified.

6 Computation of trimmed estimators

In the main part of this paper, I presented the concept of the general trimmed estimator and its asymptotical properties. Such a nontrivial estimator can be of any use in real applications only if it can be computed or approximated easily. Therefore, I describe in this section computational procedures that can be used for GTE and their advantages as well as potential weaknesses. There are two main estimation strategies: one follows the procedures used to estimate LTS and the other is based on the global optimization method called differential evolution.

6.1 Subsample selection and estimation

The traditional strategy how the least trimmed squares estimate can be determined relies on the search through subsamples of size h and the consecutive (nonlinear) LS estimation as described in Section 2. If we are able to examine the total of $\binom{n}{h}$ subsamples, we can obtain the precise solution in this way (neglecting ubiquitarian numerical errors). Unfortunately, this is hardly possible unless a very small sample is analyzed. Therefore, only an approximation can be computed in most cases. One kind of approximation can be obtained in the following way (see Čížek and Víšek (2000) for the case of LTS): let us choose randomly an h-tuple of observations, apply the nonlinear LS method on it, and evaluate 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 nonlinear 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 given 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) in the case of linear regression and a more efficient search method was described by Chen, Stromberg, and Zhou (1997) for the nonlinear regression. This approach can be naturally used also for other kinds of trimmed estimators.

The described estimation procedure has proven its qualities in the case of the linear regression model and it can certainly be applied in the case of the nonlinear regression as well. The main problem lies in the fact that the algorithm requires consecutive solving a large number of optimization problems (for many selected subsets of data). Although this

does not matter too much in the case of linear regression (the minimum of the objective function is unique in most cases and can be found easily, e.g., via least squares minimization), the situation in the case of nonlinear regression functions is just the opposite. For example, the minimization of the sum of squared residuals is time consuming and the speed of convergence of different estimation methods might differ significantly with respect to the structure of data. Therefore, the described approximation algorithm, which is based on the same idea as the mentioned algorithm for the LTS computation, can be used in the case of GTE, but it is going to be relatively slow and has most probably a lower accuracy compared to LTS in linear regression models.

6.2 Differential evolution

Facing the mentioned problems with the estimation of GTE, I think that it is beneficial to employ one of the global optimization methods—*differential evolution*—developed by Storn and Price (1995). The 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 the present 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 well the "combinatorial" nature of the GTE objective function.

The most important benefit of the differential-evolution algorithm over the algorithm described in subsection 6.1 is that it requires evaluating only the objective function instead of a complicated optimization problem. Therefore, it can be faster, especially as the size of data grows, because the complexity of the evaluation of the objective function is the same for both linear and nonlinear regression. To check whether this method is really suitable for the computation of GTE, I compared its performance in the case of the linear regression model with the existing algorithms for LTS.[11] For this purpose, I used simulated data as well

[11] The implementation of the variants of the differential-evolution algorithm is based on the source code

as the real data sets discussed in Víšek (1996b): in all cases, the estimates obtained by the differential-evolution algorithm (schemes DE/rand/1 and DE/best/1, see Storn (1996)) are identical to those obtained by the subsample-selection method described in subsection 6.1. Moreover, the speed of convergence seems to be quite high. Additionally, the comparison of the classical MLE algorithm and the differential evolution for the maximizing of a likelihood function leads to the same result. Hence, I conclude that this global optimization strategy suits well the type of minimization problems I deal with, namely, the minimization of the GTE objective function (h smallest residuals).

7 Conclusion

In this paper, I have introduced the nonlinear least trimmed squares estimator (LTS), the maximum trimmed likelihood estimator (MTLE), and the concept of general trimmed estimators incorporating both nonlinear LTS and MTLE. I also derived the $\sqrt{n}$-consistency of these estimators in nonlinear regression models and in limited-dependent-variable models.

Clearly, the formulation of the theorems is very general and encompasses many different models. On the one hand, I verified all the assumptions in the case of nonlinear regression. On the other hand, the assumptions needed for the main asymptotic results in the case of limited-dependent-variable models have to be checked on a case-by-case basis. Therefore, I demonstrated in Section 5 how the general concept of trimmed estimators (GTE) allows us to define and derive robust estimators analogous to LTS in various nonlinear and limited-dependent-variable models, and at the same time, I verified the identification condition for the proposed estimators in these models.

Finally, these robust procedures promise to provide more robust estimates in nonlinear models without the necessity to throw away classical parametric specification. Nevertheless, the presented research also poses many additional questions for further development. First, before a practical application, it is wise to study the behavior of the proposed estimators on both simulated data and real data that were already examined in the literature in order to find out more about the small-sample behavior of GTE. Second, there is certainly a possibility to extend the presented asymptotical results, for example, to study asymptotic distribution of estimators in a similar way as in Čížek (2001a) for nonlinear LTS. Third, although I demonstrated the concept of GTE and its use on several examples, it is probably necessary to derive specific trimmed estimators for most limited-dependent-

written by the authors of the method—Storn and Price (1995).

variable models and to verify their consistency using theorems presented in this paper. Moreover, for real econometric applications it is necessary to smooth GTE as proposed in Čížek (2001b) so that it can be applied to models with discrete explanatory variables. Finally, the implementation of computational procedures for GTE has to be created so that the method becomes widely available and applicable.

A Proofs of lemmas and other auxiliary propositions

In this appendix, I present the proofs of all lemmas used in this paper including all auxiliary propositions needed for the proofs.

Proposition 1 *Let $x_1, x_2, \ldots$ be a sequence of independent identically distributed random variables with a distribution function $F(x)$. Let $b(x)$ be a lower bound for $F(x)$ in a neighborhood U_1 of $+\infty$. If $b(x)$ can be chosen as $1 - \frac{1}{P_6(x)}$, where $P_6(x)$ is a polynomial of the fourth order, then it holds that $n^{-\frac{1}{6}} \max_{i=1,\ldots,n} x_i = \mathcal{O}_p(1)$ as $n \to +\infty$. Analogously, let $c(x)$ be an upper bound for $F(x)$ in a neighborhood U_2 of $-\infty$. If $c(x)$ can be chosen as $\frac{1}{P_6(x)}$, where $P_6(x)$ is a polynomial of the fourth order, then it holds that $n^{-\frac{1}{6}} \min_{i=1,\ldots,n} x_i = \mathcal{O}_p(1)$ as $n \to +\infty$.*

Proof: I prove the lemma just for the case of the lower bound, $b(x)$, the other case can be derived similarly. The cumulative distribution function of $x_{max} = \max_{i=1,\ldots,n} x_i$ is $F_n(x) = F^n(x)$. I want to show that for any $\varepsilon > 0$ there is $K > 0$ such that $P(x_{max} > K\sqrt[6]{n}) = 1 - F_n(K\sqrt[6]{n}) < \varepsilon$. This is equivalent to the assertion that $F_n(K\sqrt[6]{n}) \to 1$ as $K \to +\infty$ uniformly for $n > n_0$ and some n_0. Because $b(x) < F(x)$, it also holds that $b^n(x) < F^n(x) = F_n(x)$ and thus it is enough to verify that $b^n(K\sqrt[6]{n}) \to 1$ as $K \to +\infty$ uniformly for $n > n_0$. In general, $P_6(x) = a_1x^6 + a_2x^5 + a_3x^4 + a_4x^3 + a_5x^2 + a_6x + a_7$ and its leading coefficient a_1 has to be positive—otherwise, $b(x) > 1$ for large enough x and it could not be a lower bound to a distribution function, which is at most equal to one. So, let us assume without loss of generality that $P_6(x) = x^6$ and $b(x) = 1 - \frac{1}{x^6}$. Hence,

$$b^n\left(K\sqrt[6]{n}\right) = \left(1 - \frac{1}{Kn}\right)^n = \left[\left(1 - \frac{1}{Kn}\right)^{Kn}\right]^{\frac{1}{K}} \to \left(\frac{1}{e}\right)^{\frac{1}{K}} = \sqrt[K]{\frac{1}{e}},$$

$b^n\left(K\sqrt[6]{n}\right)$ converges monotonically to a positive number smaller than one for a fixed $K > 0$; moreover, this number $\frac{1}{\sqrt[K]{e}}$ as well as $b^n\left(K\sqrt[6]{n}\right)$ increase with K. Therefore, we can find $n_0 > 0$ such that $b^n\left(K\sqrt[6]{n}\right) > \sqrt[K]{\frac{1}{3}}$ for all $n > n_0$ and $K > 1$. Since $\sqrt[K]{\frac{1}{3}} \to 1$ for $K \to \infty$, also $b^n(K\sqrt[6]{n}) \to 1$ as $K \to +\infty$ uniformly for $n > n_0$. This closes the proof. □

Lemma 1 *Let $n \in \mathbb{N}$ and $k_h(\beta) : \mathbb{R}^p \to \{1, \ldots, n\}$ be a function that represents an index of an observation such that $s_{k_h(\beta)}(x_i, y_i; \beta) = s_{[h]}(x_i, y_i; \beta)$, $h \in \{1, \ldots, n\}$. Under Assumptions D and H, there exists a set Ω_1, $P(\Omega_1) = 1$, such that for every $\omega \in \Omega_1 \subseteq \Omega^n$ there is some neighborhood $U(\beta^0, \varepsilon(\omega))$ of β^0 such that the function $k_h(\beta)$ is constant on $U(\beta^0, \varepsilon(\omega))$ for all $h \in \{1, \ldots, n\}$.*

Proof: Given our distributional assumptions about $s_i(x_i, y_i; \beta^0)$ (Assumptions D1 and D3) and an arbitrary fixed n, the probability that any two of the residuals $s_i(x_i, y_i; \beta^0), i = 1, \ldots, n$, have the same value is equal to zero ($s(x_1, y_1; \beta^0), \ldots, s(x_n, y_n; \beta^0)$ are independent identically distributed random variables that are continuously distributed). In other words, the set of events $\omega \in \Omega^n$ for which some residuals are equal at β^0 has probability zero—$P(\Omega_0 = \{\omega \in \Omega^n : \exists i, j \in \{1, \ldots, n\}, i \neq j, s(x_i, y_i; \beta^0, \omega) = s(x_j, y_j; \beta^0, \omega)\}) = 0$ (the event ω in $s(x_i, y_i; \beta^0, \omega)$ determines the realization of sample (x_i, y_i) used). Moreover, there is a $\delta' > 0$ such that $s(x_i, y_i; \beta)$ is continuous on $\bar{U}(\beta^0, \delta')$, and therefore also uniformly continuous on $\bar{U}(\beta^0, \delta')$. Therefore, for any given $\omega \notin \Omega_0$ and $\kappa(\omega) = \frac{1}{2} \min_{i,j=1,\ldots,n} |s(x_j, y_j; \beta^0, \omega) - s(x_i, y_i; \beta^0, \omega)| > 0$ we can find an $\varepsilon(\omega) > 0$ such that, for any $\beta \in U(\beta^0, \varepsilon(\omega))$, it holds that $|s(x_i, y_i; \beta, \omega) - s(x_i, y_i; \beta^0, \omega)| < \kappa(\omega)$ for all $i = 1, \ldots, n$. Consequently, mapping $k_h(\beta)$ is constant on $U(\beta^0, \varepsilon(\omega))$ for any $\omega \notin \Omega_0$ because the ordering of $s_i(x_i, y_i; \beta)$ is independent of β on this set. Thus, the set $\Omega_1 = \Omega - \Omega_0$. □

Lemma 2 *$P\big(\{\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n : s_i(x_i, y_i; \beta, \omega_i) = s_{[h]}(x_i, y_i; \beta, \omega)\}\big) = \frac{1}{n}$ for any $n \in \mathbb{N}$, $i, h \in \{1, \ldots, n\}$, and $\beta \in B$.*

Proof: Again, we use the extended notation $s_i(x_i, y_i; \beta, \omega)$, where the event ω indicates which realization of an obsevation (x_i, y_i) or the whole sample $(x_i, y_i)_{i=1}^n$ is meant. As $s_i(x_i, y_i; \beta, \omega_i), i = 1, \ldots, n$, form a vector of independent identically distributed random variables for given n and β,

$$\begin{aligned} &P\big(\{\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n : s_i(x_i, y_i; \beta, \omega_i) = s_{[h]}(x_i, y_i; \beta, \omega)\}\big) = \\ &= \quad P\big(\{\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n : s_j(x_i, y_i; \beta, \omega_j) = s_{[h]}(x_i, y_i; \beta, \omega)\}\big) \end{aligned}$$

for any $i, j \in \{1, \ldots, n\}$ and a fixed $h \in \{1, \ldots, n\}$. Moreover,

$$\sum_{i=1}^{n} P\big(\{\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n : s_i(x_i, y_i; \beta, \omega_i) = s_{[h]}(x_i, y_i; \beta, \omega)\}\big) = 1$$

since $P(\{\omega \in \Omega^n : \exists i \neq j \in \{1, \ldots, n\}, s_i(x_i, y_i; \beta, \omega_i) = s_j(x_j, y_j; \beta, \omega_j)\}) = 0$. Putting the last two equations together, we immediately get for any $i = 1, \ldots, n$

$$P\big(\{\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n : s_i(x_i, y_i; \beta, \omega_i) = s_{[h]}(x_i, y_i; \beta, \omega)\}\big) = \frac{1}{n},$$

which closes the proof. □

Lemma 3 *Let $1/2 < \lambda < 1$ and $0 < c < G_\beta^{-1}(\lambda)$ be a real constant, where G_β represents the distribution function of $s_i(x_i, y_i; \beta)$, $\beta \in B$. Then, under Assumption D, $P\big(s_{[h_n]}(x_i, y_i; \beta) \leq c\big) = \mathcal{O}\big(n^{-k}\big)$ for any $k \in \mathbb{N}$ as $n \to +\infty$.*

Proof: The distribution function of $s_{[h_n]}(x_i, y_i; \beta)$ is given by ($\beta \in B$ is a fixed parameter vector)

$$G_{\beta,h_n}(x) = \sum_{i=h_n}^{n} P_i(x), \quad P_i(x) = \binom{n}{i} G_\beta(x)^i (1 - G_\beta(x))^{n-i}. \tag{42}$$

Let $\lambda' = G_\beta(c) < \lambda < 1$ and let M_{h_n} be an upper bound for $G_{\beta,h_n}(x)$ on the interval $\langle 0, c\rangle$. We show that $\lim_{n\to\infty} n^k M_{h_n} = 0$ for any $k \in \mathbb{N}$.

First, I will draw attention to one fundamental property of the nth root of $G_\beta(x)^i(1 - G_\beta(x))^{n-i}$, the main element of $P_i(x)$ in (42). For any $a \in (\lambda, 1)$ (a represents here $\frac{i}{n}$ for any $i \in \{h_n, \ldots, n\}$, in fact)

$$\begin{aligned}
\big(G_\beta(x)^a (1 - G_\beta(x))^{1-a}\big)' &= aG_\beta(x)^{a-1} g_\beta(x)(1 - G_\beta(x))^{1-a} \\
&\quad -(1-a)G_\beta(x)^a (1 - G_\beta(x))^{-a\lambda} g_\beta(x) \\
&= G_\beta(x)^{a-1} g_\beta(x)(1 - G_\beta(x))^{-a} \left(a(1 - G_\beta(x)) - (1-a)G_\beta(x)\right) \\
&\geq 0, \quad \forall 0 < x \leq c < G_\beta^{-1}(\lambda) \leq G_\beta^{-1}(a)
\end{aligned}$$

($G_\beta(x)$ is monotonic). Therefore, $\left(\frac{G_\beta(x)}{a}\right)^a \left(\frac{1-G_\beta(x)}{1-a}\right)^{1-a} < \left(\frac{\lambda'}{a}\right)^a \left(\frac{1-\lambda'}{1-a}\right)^{1-a} = C(a) < 1$ as $x \in \langle 0, c\rangle$. Since $\lambda' < a$, the derivative of $\left(\frac{\lambda'}{a}\right)^a \left(\frac{1-\lambda'}{1-a}\right)^{1-a}$ with respect to a is negative:

$$\begin{aligned}
\left[\left(\frac{\lambda'}{a}\right)^a \left(\frac{1-\lambda'}{1-a}\right)^{1-a}\right]' &= \left[e^{a \ln \frac{\lambda'}{a} + (1-a)\ln\frac{1-\lambda'}{1-a}}\right]' \\
&= \left[e^{a\ln\lambda' - a\ln a + (1-a)\ln(1-\lambda') - (1-a)\ln(1-\alpha)}\right]' \\
&= (\ln\lambda' - \ln a - 1 - \ln(1-\lambda') + \ln(1-a) + 1) \cdot e^{-a\ln a - (1-a)\ln(1-\alpha)} \\
&= \ln\left(\frac{\lambda'}{1-\lambda'}\frac{1-a}{a}\right) \cdot \left(\frac{\lambda'}{a}\right)^a \left(\frac{1-\lambda'}{1-a}\right)^{1-a} < 0,
\end{aligned}$$

because $a \in (\lambda, 1) \subseteq (\frac{1}{2}, 1)$, $0 < \frac{\lambda'}{a} < 1$, and $0 < \frac{1-a}{1-\lambda'} < 1$. Hence, $C(a) < C(\frac{1}{2}(\lambda + \lambda')) = C < 1$. We show the usefulness of this result in a moment.

Now, we analyze the function $G_{\beta,h_n}(x)$ itself. Taking into account $h_n/n \to \lambda$ (h_n is defined as $[\lambda n]$), it follows that we can write $h_n/n = \lambda + a_n$, where $|a_n| < \frac{1}{n}$. Moreover, notice that $\frac{i}{n} > \lambda'' = \frac{1}{2}(\lambda + \lambda')$ for any n sufficiently high and $i \geq h_n$. Let us take some

$0 \leq x \leq c$ (therefore, $G_\beta(x) \neq 1$). Using the Stirling formula

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot \left(1 + \mathcal{O}\left(\frac{1}{12n}\right)\right),$$

we get for $n^{k+1/2} \cdot G_{\beta,h_n}(x)$

$$\begin{aligned}
& \lim_{n\to\infty} n^{k+1/2} \sum_{i=h_n}^{n} P_i(x) = \lim_{n\to\infty} n^{k+1/2} \sum_{i=h_n}^{n} \binom{n}{i-1} G_\beta(x)^i (1-G_\beta(x))^{n-i} \\
= & \lim_{n\to\infty} n^{k+1/2} \sum_{i=h_n}^{n} \frac{n!}{i!(n-i)!} G_\beta(x)^i [1-G_\beta(x)]^{n-i} \\
= & \lim_{n\to\infty} n^{k+1/2} \sum_{i=h_n}^{n} \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot \left(1+\mathcal{O}\left(\frac{1}{12n}\right)\right) \times G_\beta(x)^i [1-G_\beta(x)]^{n-i}}{\sqrt{2\pi i}\left(\frac{i}{e}\right)^i \cdot \left(1+\mathcal{O}\left(\frac{1}{12i}\right)\right) \times \sqrt{2\pi(n-i)}\left(\frac{n-i}{e}\right)^{n-i} \cdot \left(1+\mathcal{O}\left(\frac{1}{12(n-i)}\right)\right)} \\
= & \lim_{n\to\infty} n^{k} \sum_{i=h_n}^{n} \frac{\sqrt{2\pi} \cdot \left(1+\mathcal{O}\left(\frac{1}{12n}\right)\right) \times G_\beta(x)^i [1-G_\beta(x)]^{n-i}}{\sqrt{2\pi\frac{i}{n}}\left(\frac{i}{n}\right)^i \cdot \left(1+\mathcal{O}\left(\frac{1}{12i}\right)\right) \times \sqrt{2\pi\frac{n-i}{n}}\left(\frac{n-i}{n}\right)^{n-i} \cdot \left(1+\mathcal{O}\left(\frac{1}{12(n-i)}\right)\right)} \\
= & \lim_{n\to\infty} n^{k} \sum_{i=h_n}^{n} \left(\frac{G_\beta(x)}{\frac{i}{n}}\right)^i \left(\frac{1-G_\beta(x)}{\frac{n-i}{n}}\right)^{n-i} \cdot \frac{\sqrt{2\pi} \cdot \left(1+\mathcal{O}\left(\frac{1}{12n}\right)\right)}{\sqrt{2\pi\frac{i}{n}} \cdot \left(1+\mathcal{O}\left(\frac{1}{12i}\right)\right) \times \sqrt{2\pi\frac{n-i}{n}} \cdot \left(1+\mathcal{O}\left(\frac{1}{12(n-i)}\right)\right)} \\
= & \lim_{n\to\infty} n^{k} \sum_{i=h_n}^{n} \left(\frac{G_\beta(x)}{\frac{i}{n}}\right)^i \left(\frac{1-G_\beta(x)}{\frac{n-i}{n}}\right)^{n-i} \cdot \mathcal{O}(1) \\
= & \lim_{n\to\infty} n^{k} \sum_{i=h_n}^{n} \left[\left(\frac{G_\beta(x)}{\frac{i}{n}}\right)^{\frac{i}{n}} \left(\frac{1-G_\beta(x)}{\frac{n-i}{n}}\right)^{\frac{n-i}{n}}\right]^n \cdot \mathcal{O}(1) \\
\leq & \lim_{n\to\infty} n^{k} \sum_{i=h_n}^{n} C^n \cdot \mathcal{O}(1) \\
= & \lim_{n\to\infty} n^{k+1} C^n \cdot \mathcal{O}(1) \\
= & 0.
\end{aligned}$$

Therefore, we have proved that $\lim_{n\to\infty} n^k \cdot \sup_{x\in\langle 0,c\rangle} G_{\beta,h_n}(x) = 0$, which closes the proof as $P(s_{[h_n]}(x_i, y_i; \beta) \leq c) = G_{\beta,h_n}(c)$. □

Corollary 1 *Analogously, it is possible under Assumption D to show that for real constants* $1/2 < \lambda < 1$ *and* $G_\beta^{-1}(\lambda) < c < \infty$ *it holds under that* $P\left(s_{[h_n]}(x_i, y_i; \beta) \geq c\right) = \mathcal{O}\left(n^{-k}\right)$ *for any* $k \in \mathbb{N}$ *as* $n \to +\infty$.

Corollary 2 *Let* $1/2 < \lambda < 1$ *and* $0 < c < G^{-1}(\lambda) < c' < \infty$ *be real constants. Under*

Assumptions D and H, it holds that

$$P\Big(\exists\beta \in U(\beta^0, n^{-\frac{1}{2}}M) : s_{[h_n]}(x_i, y_i; \beta) \notin (c, c')\Big) = \mathcal{O}\big(n^{-k}\big) \tag{43}$$

for any $k \in \mathbb{N}$ *as* $n \to +\infty$.

Proof: First, note that $s_i(x_i, y_i; \beta, \omega) \to s_i(x_i, y_i; \beta^0, \omega)$[12] for $\beta \to \beta^0$ and any $\omega \in \Omega$ (convergence almost surely). So, for $\beta \to \beta^0$ and $G_\beta(x)$, being the cumulative distribution function of $s_i(x_i, y_i; \beta)$, it holds that $G_\beta(x) \to G_{\beta^0}(x) \equiv G(x)$ for all $x \in \mathbb{R}$ (convergence in distribution) because $G(x)$ is an absolutely continuous distribution function. Now, we show that this convergence of distribution functions $G_\beta(x) \to G_{\beta^0}(x) \equiv G(x)$ is uniform over all sequences $\beta_n \to \beta^0$ such that $\beta_n \in U(\beta^0, n^{-\frac{1}{2}}M)$. The reason is that $s_i(x_i, y_i; \beta) = s_i(x_i, y_i; \beta^0) - s'_\beta(x_i, y_i; \xi) \cdot (\beta - \beta^0)$, where the second term $s'_\beta(x_i, y_i; \xi) \cdot (\beta - \beta^0)$ can be bounded by a random variable of order $\mathcal{O}_p\Big(n^{\frac{1}{3}}\Big) \cdot \mathcal{O}\Big(n^{-\frac{1}{2}}\Big) = \mathcal{O}_p\Big(n^{-\frac{1}{6}}\Big)$ independently of β (see Assumption H3). Thus, the convergence of $s_i(x_i, y_i; \beta)$ to $s_i(x_i, y_i; \beta^0)$ is uniform in probability.

Now, let $\varepsilon > 0$ be chosen so that $(\lambda - 2\varepsilon, \lambda + 2\varepsilon) \subset (G(c), G(c'))$ and $(G^{-1}(\lambda) - 2\varepsilon, G^{-1}(\lambda) + 2\varepsilon) \subset (c, c')$ (remember, $G(c) < \lambda < G(c')$). Moreover, because of the described uniform convergence of $s_i(x_i, y_i; \beta)$ to $s_i(x_i, y_i; \beta^0)$ and $G_\beta(x)$ to $G_{\beta^0}(x)$, there exists $n_0 \in \mathbb{N}$ such that $G^{-1}_{\beta_n}(\lambda) \in (G^{-1}(\lambda) - \varepsilon, G^{-1}(\lambda) + \varepsilon)$ for any $\beta_n \in U(\beta^0, n^{-\frac{1}{2}}M)$ and $n > n_0$. Hence, $(G^{-1}_{\beta_n}(\lambda) - \varepsilon, G^{-1}_{\beta_n}(\lambda) + \varepsilon) \subset (c, c')$, and $|G_{\beta_n}(c) - \lambda| > \varepsilon$ and $|G_{\beta_n}(c') - \lambda| > \varepsilon$ for all $\beta_n \in U(\beta^0, n^{-\frac{1}{2}}M)$ and $n > n_0$. Hence, the constant λ' in the proof of Lemma 3 can be chosen equal to $\lambda - \varepsilon$ independently of $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$ and we can follow the same steps as in the proof of Lemma 3 to derive (43). □

Corollary 3 *Let* $1/2 < \lambda < 1$ *and* $\varepsilon > 0$ *be sufficiently small real constants. Under Assumptions D and H, it holds that*

$$P\big(\exists\beta \in B : s_{[h_n]}(x_i, y_i; \beta) \notin \big(G^{-1}_\beta(\lambda) - \varepsilon, G^{-1}_\beta(\lambda) + \varepsilon\big)\big) = \mathcal{O}\big(n^{-k}\big) \tag{44}$$

for any $k \in \mathbb{N}$ *as* $n \to +\infty$.

Proof: Let $\lambda'_\beta = G_\beta\left(G^{-1}_\beta(\lambda) - \varepsilon\right) < \lambda$ and $a \in (\lambda, 1)$. We know that $\left(\frac{G_\beta(x)}{a}\right)^a \left(\frac{1-G_\beta(x)}{1-a}\right)^{1-a} < \left(\frac{\lambda'_\beta}{a}\right)^a \left(\frac{1-\lambda'_\beta}{1-a}\right)^{1-a} = C(a, \beta) < 1$ as $x \in \left\langle 0, G^{-1}_\beta(\lambda) - \varepsilon\right\rangle$ (see Lemma 3). Furthermore,

[12] The event ω in $s(x_i, y_i; \beta^0, \omega)$ determines the realization of sample (x_i, y_i) used.

$C(a,\beta) < C(\frac{1}{2}(\lambda + \lambda'_\beta)) = C < 1$. Therefore, to prove the result of this corollary, we can follow the same steps as in Lemma 3 as long as we show that $\sup_{\beta\in B} C(a,\beta) < 1$, which is equivalent to

$$\Lambda = \sup_{\beta\in B} \lambda'_\beta = \sup_{\beta\in B} G_\beta\left(G_\beta^{-1}(\lambda) - \varepsilon\right) < \lambda.$$

Then we can choose $C(a,\beta) < C(\frac{1}{2}(\lambda+\Lambda)) = C < 1$ and complete the proof in the same way as the proof of Lemma 3. Because

$$G_\beta\left(G_\beta^{-1}(\lambda) - \varepsilon\right) = \lambda - g_\beta(\xi)\varepsilon,$$

it is sufficient to know that there is some $\delta > 0$ such that

$$\inf_{\beta\in B}\inf_{z\in(-\delta,\delta)} g_\beta\left(G_\beta^{-1}(\lambda) + z\right) > 0.$$

However, this is in Assumption NC, which closes the proof of the corollary. □

Lemma 4 *Under Assumptions D and H, for any fixed $i \in \mathbb{N}$ and $n \geq i$*

$$P\Big(\exists\beta \in U(\beta^0, n^{-\frac{1}{2}}M) : I\big(s_i(x_i,y_i;\beta) \leq s_{[h_n]}(x_i,y_i;\beta)\big) \neq I\big(s_i(x_i,y_i;\beta) \leq G_\beta^{-1}(\lambda)\big)\Big) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$, or analogously

$$P\left(\sup_{\beta\in U(\beta^0,n^{-\frac{1}{2}}M)} \left|I\big(s_i(x_i,y_i;\beta) \leq s_{[h_n]}(x_i,y_i;\beta)\big) - I\big(s_i(x_i,y_i;\beta) \leq G_\beta^{-1}(\lambda)\big)\right| \neq 0\right) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$.

Proof: Let us introduce a bit of notation first: $\upsilon_{in}(\beta) = I\big(s_i(x_i,y_i;\beta) \leq s_{[h_n]}(x_i,y_i;\beta)\big) - I\big(s_i(x_i,y_i;\beta) \leq G_\beta^{-1}(\lambda)\big)$. We have to derive, in fact, an upper bound for

$$\mathsf{E}\sup_{\beta\in U(\beta^0,n^{-\frac{1}{2}}M)} |\upsilon_{in}(\beta)| = P\left(\sup_{\beta\in U(\beta^0,n^{-\frac{1}{2}}M)} |\upsilon_{in}(\beta)| = 1\right).$$

For the sake of simplicity, we will omit in what follows the specification of the set across which the supremum is considered and write simply $P(\exists\beta : |\upsilon_{in}(\beta)| = 1) = P\big(\sup_\beta |\upsilon_{in}(\beta)| = 1\big)$ keeping in mind that we mean $P\Big(\sup_{\beta\in U(\beta^0,n^{-\frac{1}{2}}M)} |\upsilon_{in}(\beta)| = 1\Big)$ and $\beta \in U(\beta^0, n^{-\frac{1}{2}}M)$.

Without loss of generality, we derive only $P(\exists\beta : v_{in}(\beta) = -1)$, i.e.,

$$P\Big(\exists\beta \in U(\beta^0, n^{-\frac{1}{2}}M) : I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big) = 0 \wedge I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big) = 1\Big);$$

the other case can be analyzed analogously.

Before we start with the derivation, notice that the distribution function of an order statistics $s_{[h]}(x_i, y_i; \beta)$ for a given $h, 1 \leq h \leq n$, is given by (presuming that G_β and g_β are the c.d.f. and p.d.f. of $s_i(x_i, y_i; \beta)$)

$$G_{\beta,h}(x) = \sum_{j=h}^{n} P_j(x, \beta), \quad P_j(x, \beta) = \binom{n}{j} G_\beta(x)^j (1 - G_\beta(x))^{n-j}$$

and the corresponding probability density function is given by (for $n \geq 2$)

$$g_{\beta,h}(x) = n \binom{n-1}{h-1} g_\beta(x) G_\beta(x)^{h-1} (1 - G_\beta(x))^{n-h}.$$

Throughout this proof, I use notation G_{β^0} and $G_{\beta^0,h}$ instead of G and G_h to make it consistent with frequent use of G_β and $G_{\beta,h}$. The same applies for g_{β^0}. Moreover, I also use the extended notation $s_i(x_i, y_i; \beta, \omega)$, where the event ω indicates which realization of an obsevation (x_i, y_i) or the whole sample $(x_i, y_i)_{i=1}^n$ is meant.

Now, let us consider $\omega = (\omega_1, \ldots, \omega_n) \in \Omega^n$ and assume without loss of generality that $i = 1$. Given $\omega' = (\omega_2, \ldots, \omega_n) \in \Omega^{n-1}$ and $(s_2(x_i, y_i; \beta, \omega_2), \ldots, s_n(x_i, y_i; \beta, \omega_n))$

$$s_{[h]}(x_i, y_i; \beta, \omega) = \begin{cases} s_{[h-1]}(x_i, y_i; \beta, \omega') & \text{if } s_1(x_1, y_1; \beta, \omega_1) < s_{[h-1]}(x_i, y_i; \beta, \omega') \\ s_1(x_1, y_1; \beta, \omega_1) & \text{if } s_{[h-1]}(x_i, y_i; \beta, \omega') \leq s_1(x_1, y_1; \beta, \omega_1) \leq s_{[h]}(x_i, y_i; \beta, \omega') \\ s_{[h]}(x_i, y_i; \beta, \omega') & \text{if } s_{[h]}(x_i, y_i; \beta, \omega') < s_1(x_1, y_1; \beta, \omega_1). \end{cases} \tag{45}$$

Denoting Ω_1, Ω_2, and Ω_3 subsets of Ω^n corresponding to the three (disjoint) cases in (45), we can write

$$\begin{aligned} P(\{\omega \in \Omega^n | \exists\beta : v_{1n}(\beta) = -1\}) &= P(\{\omega \in \Omega_1 | \exists\beta : v_{1n}(\beta) = -1\}) \\ &+ P(\{\omega \in \Omega_2 | \exists\beta : v_{1n}(\beta) = -1\}) \\ &+ P(\{\omega \in \Omega_3 | \exists\beta : v_{1n}(\beta) = -1\}), \end{aligned}$$

and analyze this sum one by one.

1. $P(\{\omega \in \Omega_1 | \exists \beta : \upsilon_{1n}(\beta) = -1\}) \leq$
   $\leq P(\exists \beta : s_{[h_n]}(x_i, y_i; \beta, \omega) < s_1(x_1, y_1; \beta, \omega_1) < s_{[h_n]}(x_i, y_i; \beta, \omega)) = 0.$

2. $P(\{\omega \in \Omega_2 | \exists \beta : \upsilon_{1n}(\beta) = -1\}) =$
   $= P\big(\exists \beta : s_{[h_n-1]}(x_i, y_i; \beta, \omega') \leq s_1(x_1, y_1; \beta, \omega_1) = s_{[h_n]}(x_i, y_i; \beta, \omega) \leq G_\beta^{-1}(\lambda)\big)$ and can be analyzed in exactly the same way as $P(\{\omega \in \Omega_3 | \exists \beta : \upsilon_{1n} = -1\})$, see point 3.

3. $P(\{\omega \in \Omega_3 | \exists \beta : \upsilon_{1n}(\beta) = -1\}) =$
   $= P\big(\exists \beta : s_{[h_n]}(x_i, y_i; \beta, \omega') = s_{[h_n]}(x_i, y_i; \beta, \omega) < s_1(x_i, y_i; \beta, \omega_1) \leq G_\beta^{-1}(\lambda)\big)$. We can structure this last term in the following way ($1 \gg \varepsilon > 0$ is an arbitrary, but fixed real number; the choice of ε will be discussed later):

$$
\begin{aligned}
& P\big(s_{[h_n]}(x_i, y_i; \beta, \omega') < s_1(x_1, y_1; \beta, \omega_1) \leq G_\beta^{-1}(\lambda)\big) && (46)\\
\leq\ & P\Big(s_{[h_n]}(x_i, y_i; \beta, \omega') \leq G_{\beta^0}^{-1}(\lambda) - \varepsilon/2\Big) && (47)\\
& + P\Big(G_{\beta^0}^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i, y_i; \beta, \omega') < s_1(x_1, y_1; \beta, \omega_1) \leq G_\beta^{-1}(\lambda)\Big).
\end{aligned}
$$

   Please note that $G_\beta^{-1}(\lambda) \in (G_{\beta^0}^{-1}(\lambda) - \varepsilon/4, G_{\beta^0}^{-1}(\lambda) + \varepsilon/4)$ for n larger than a certain n_0 because $G_\beta(x) \to G_{\beta^0}(x)$ for all $x \in \mathbb{R}$ (remember, $G \equiv G_{\beta^0}$). Since Corollary 2 implies $P\Big(s_{[h_n]}(x_i, y_i; \beta, \omega) \leq G_{\beta^0}^{-1}(\lambda) - \varepsilon/2\Big) = o(\frac{1}{n})$ as $n \to +\infty$, we have to analyze just the second term on the right hand side of the equality:

$$
\begin{aligned}
& P\Big(G_{\beta^0}^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i, y_i; \beta, \omega) < s_1(x_1, y_1; \beta, \omega_1) \leq G^{-1}(\lambda)\Big) = && (48)\\
& = \int_{\omega' \in \Omega^{n-1}} \int_{\omega_1 \in \Omega} I\Big(G_{\beta^0}^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i, y_i; \beta, \omega') < s_1(x_1, y_1; \beta, \omega_1) \leq G_\beta^{-1}(\lambda)\Big)\\
& \qquad \mathrm{d}P(\omega_1)\mathrm{d}P(\omega'). && (49)
\end{aligned}
$$

   Let $1 \gg \varepsilon > 0$ be a small enough fixed real number and $n > n_0$ large enough so that (see Assumption D3 and related notation in Section 3.2 for the definition of functions G_β and g_β)

   (a) Assumption D4 implies that there is $M_g > 0$ such that $g_\beta(x) < M_g$ for all $G_{\beta^0}^{-1}(\lambda) - \varepsilon \leq x$,

   (b) $G_{\beta^0}(G_{\beta^0}^{-1}(\lambda) - \varepsilon) = \gamma > 0$ and $G_\beta(G_{\beta^0}^{-1}(\lambda) - \varepsilon) > \gamma/2 > 0$,

   (c) $g_\beta(x) > m_g > 0$ almost surely for all $G_{\beta^0}^{-1}(\lambda) - 2\varepsilon \leq x \leq G_{\beta^0}^{-1}(\lambda) + 2\varepsilon$, where $m_g > 0$ is a real constant (this again follows from Assumption D4),

(d) $M_g\varepsilon/\lambda < 1$ and $M_g\varepsilon/(1-\lambda) < 1$,

(e) $m_g^2/2 > \left|\frac{\lambda-1}{\lambda}\right|(1-M_g\varepsilon/\lambda)^{\lambda-2}M_g^3\varepsilon + \left|\frac{\lambda}{1-\lambda}\right| M_g^3\varepsilon + \left|\frac{\lambda}{1-\lambda}\right|\left|\frac{\lambda-1}{\lambda}\right|(1-M_g\varepsilon/\lambda)^{\lambda-2}M_g^4\varepsilon^2$, and

(f) $G_\beta(x) = G_\beta(G_\beta^{-1}(\lambda)) + g_\beta(\xi)(x - G_\beta^{-1}(\lambda)) = \lambda + g_\beta(\xi)(x - G_\beta^{-1}(\lambda))$ for all $x \in (G_{\beta^0}^{-1}(\lambda) - 2\varepsilon, G_{\beta^0}^{-1}(\lambda) + 2\varepsilon)$, where $\xi \in (x, G_\beta^{-1}(\lambda))$.

Suppose further that $n > n_0$, where n_0 is defined above. Then we can write (see equation (48))

$$
\begin{aligned}
& P\Big(\exists\beta : G_{\beta^0}^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i, y_i; \beta, \omega) < s_1(x_1, y_1; \beta, \omega_1) \le G_{\beta^0}^{-1}(\lambda)\Big) \le \\
= & \int_{\omega' \in \Omega^{n-1}} \int_{\omega_1 \in \Omega} I\Big(G_{\beta^0}^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i, y_i; \beta, \omega') < s_1(x_1, y_1; \beta, \omega_1) \le G_\beta^{-1}(\lambda)\Big) \\
& \qquad \mathrm{d}P(\omega_1)\mathrm{d}P(\omega') \qquad (50) \\
\le & \int_{\omega' \in \Omega^{n-1}} M_g \cdot \\
& \cdot \sup_\beta \Big\{ \big|G_\beta^{-1}(\lambda) - s_{[h_n]}(\beta, \omega')\big| \cdot I\Big(G_{\beta^0}^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i, y_i; \beta, \omega') \le G_\beta^{-1}(\lambda)\Big)\Big\}\, \mathrm{d}P(\omega') \\
= & M_g \cdot \int_{G_{\beta^0}^{-1}(\lambda)-\varepsilon/2}^{G_\beta^{-1}(\lambda)} \sup_\beta \big\{ \big|G_\beta^{-1}(\lambda) - x\big| \cdot g_{\beta,h_n}(x)\big\}\, \mathrm{d}x \qquad (51) \\
= & M_g \cdot \int_0^{G_\beta^{-1}(\lambda) - G_{\beta^0}^{-1}(\lambda) + \varepsilon/2} \sup_\beta \big\{ |y| \cdot g_{\beta,h_n}(G_\beta^{-1}(\lambda) - y)\big\}\, \mathrm{d}y \\
\le & M_g \cdot \int_0^\varepsilon y \cdot \sup_\beta \big\{ g_{\beta,h_n}(G_\beta^{-1}(\lambda) - y)\big\}\, \mathrm{d}y.
\end{aligned}
$$

To see how this integral behaves, it is necessary to analyze the function $g_{\beta,h_n}(\cdot)$ in a neighborhood of $G_\beta^{-1}(\lambda)$ given by 2ε for $n \to +\infty$. Since $x \in (G^{-1}(\lambda) - \varepsilon, G^{-1}(\lambda))$,

$$
\begin{aligned}
\frac{g_{\beta,h_n}(x)}{\sqrt{n}} &= \sqrt{n}\binom{n-1}{h_n-1} g_\beta(x) G_\beta(x)^{h_n-1}(1-G_\beta(x))^{n-h_n} \\
&\le M_g\sqrt{n}\binom{n-1}{h_n-1} G_\beta(x)^{h_n-1}(1-G_\beta(x))^{n-h_n}.
\end{aligned}
$$

Thus, it suffices to analyze the latter term. Using the Stirling formula

$$
n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot \left(1 + \mathcal{O}\left(\frac{1}{12n}\right)\right)
$$

and $h_n/n = \lambda + a_n$, where $|a_n| < \frac{1}{n}$, we can arrange the expression in question in the following way (notice that $G_\beta(x) \neq 0$ and $G_\beta(x) \neq 1$ for $x \in (G_\beta^{-1}(\lambda) - \varepsilon, G_\beta^{-1}(\lambda))$):

$$
\begin{aligned}
& \sqrt{n}\binom{n-1}{h_n-1} G_\beta(x)^{h_n-1}(1-G_\beta(x))^{n-h_n} = \\
= \;& n^{1/2}\frac{(n-1)!}{(h_n-1)!(n-h_n)!} G_\beta(x)^{h_n-1}[1-G_\beta(x)]^{n-h_n} \\
= \;& n^{1/2}\frac{\sqrt{2\pi(n-1)}\left(\frac{n-1}{e}\right)^{n-1}\cdot\left(1+\mathcal{O}\left(\frac{1}{12(n-1)}\right)\right)}{\sqrt{2\pi(h_n-1)}\left(\frac{h_n-1}{e}\right)^{h_n-1}\cdot\left(1+\mathcal{O}\left(\frac{1}{12(h_n-1)}\right)\right)} \times \\
& \times\frac{G_\beta(x)^{h_n-1}[1-G_\beta(x)]^{n-h_n}}{\sqrt{2\pi(n-h_n)}\left(\frac{n-h_n}{e}\right)^{n-h_n}\cdot\left(1+\mathcal{O}\left(\frac{1}{12(n-h_n)}\right)\right)} \\
= \;& \frac{\sqrt{2\pi}\cdot G_\beta(x)^{h_n-1}[1-G_\beta(x)]^{n-h_n}\cdot(1+o(1))}{\sqrt{2\pi\frac{h_n-1}{n-1}}\left(\frac{h_n-1}{n-1}\right)^{h_n-1}\cdot\sqrt{2\pi\frac{n-h_n+1}{n}}\left(\frac{n-h_n}{n-1}\right)^{n-h_n}} \\
= \;& \frac{(1+o(1))}{\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left(\frac{G_\beta(x)}{\frac{h_n-1}{n-1}}\right)^{h_n-1}\left(\frac{1-G_\beta(x)}{\frac{n-h_n}{n-1}}\right)^{n-h_n} \\
= \;& \frac{(1+o(1))}{\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left(\frac{G_\beta(x)}{\lambda}\right)^{h_n-1}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{n-h_n}\cdot\left(\frac{\lambda}{\frac{h_n-1}{n-1}}\right)^{h_n-1}\left(\frac{1-\lambda}{1-\frac{h_n-1}{n-1}}\right)^{n-h_n} \\
= \;& \frac{(1+o(1))}{\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left(\frac{G_\beta(x)}{\lambda}\right)^{h_n-1}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{n-h_n}\cdot \\
& \cdot\left(1+\frac{\lambda+na_n-1}{\lambda(n-1)}\right)^{-h_n+1}\left(1+\frac{1-\lambda-na_n}{(1-\lambda)(n-1)}\right)^{-n+h_n} \\
\geq \;& \frac{(1+o(1))}{\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left(\frac{G_\beta(x)}{\lambda}\right)^{h_n-1}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{n-h_n} \\
& \cdot\left(1+\frac{1}{\lambda(n-1)}\right)^{-\lambda(n-1)}\left(1+\frac{2}{(1-\lambda)(n-1)}\right)^{-(1-\lambda)(n-1)} \\
= \;& \frac{(1+o(1))}{e^3\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left(\frac{G_\beta(x)}{\lambda}\right)^{h_n-1}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{n-h_n} \\
= \;& \frac{(1+o(1))}{e^3\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left[\left(\frac{G_\beta(x)}{\lambda}\right)^{\lambda}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{1-\lambda}\right]^n\cdot\left(\frac{\lambda}{G_\beta(x)}\right)^{a_n-\frac{1}{n}}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{-a_n} \\
= \;& \frac{(1+o(1))}{e^3\sqrt{2\pi\lambda(1-\lambda)}}\cdot\left[\left(\frac{G_\beta(x)}{\lambda}\right)^{\lambda}\left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{1-\lambda}\right]^n\cdot
\end{aligned}
$$

Similarly, an upper bound for $\frac{g_{\beta,h_n}(x)}{\sqrt{n}}$ can be derived:

$$\frac{g_{\beta,h_n}(x)}{\sqrt{n}} \leq \frac{M_g(1+o(1))}{\sqrt{2\pi\lambda(1-\lambda)}} \cdot \left[\left(\frac{G_\beta(x)}{\lambda}\right)^\lambda \left(\frac{1-G_\beta(x)}{1-\lambda}\right)^{1-\lambda} \right]^n .$$

In the next step, we employ Taylor's expansion of functions $G_\beta(x)$ and $(1+x)^\lambda$ to analyze the behavior of $g_{\beta,h_n}(x)$ in a neighborhood of $G_\beta^{-1}(\lambda)$. The Taylor expansions $G_\beta(G_\beta^{-1}(\lambda) - x) = \lambda - g_\beta(\xi)x$ and $(1+x)^\lambda = 1 + \lambda x + \frac{1}{2}\lambda(\lambda-1)\zeta$ ($\xi \in (G_\beta^{-1}(\lambda) - x, G_\beta^{-1}(\lambda))$ and $\zeta \in (0,x)$) allow us to modify the bounds for $\frac{g_{\beta,h_n}(x)}{\sqrt{n}}$ as follows ($x \in (G_\beta^{-1}(\lambda) - \varepsilon, G_\beta^{-1}(\lambda))$, $\frac{1}{2} < \lambda < 1$):

$$\begin{aligned}
& \left(\frac{G_\beta(G_\beta^{-1}(\lambda)-x)}{\lambda}\right)^\lambda \left(\frac{1-G_\beta(G_\beta^{-1}(\lambda)-x)}{1-\lambda}\right)^{1-\lambda} = \\
= \; & \left(1 - \frac{g_\beta(\xi)}{\lambda}x\right)^\lambda \left(1 + \frac{g_\beta(\xi)}{1-\lambda}x\right)^{1-\lambda} \\
= \; & \left(1 - g_\beta(\xi)x + \frac{1}{2}\frac{\lambda-1}{\lambda}(1-\zeta)^{\lambda-2}g_\beta^2(\xi)x^2\right)\left(1 + g_\beta(\xi)x + \frac{1}{2}\frac{\lambda}{1-\lambda}(1+\zeta')^{-\lambda-1}g_\beta^2(\xi)x^2\right) \\
= \; & 1 - g_\beta^2(\xi)x^2 + \frac{1}{2}g_\beta^2(\xi)x^2\left[\frac{\lambda-1}{\lambda}(1-\zeta)^{\lambda-2} + \frac{\lambda}{1-\lambda}(1+\zeta')^{-\lambda-1}\right] + \mathcal{O}\big(x^3\big) \\
\geq \; & 1 - g_\beta^2(\xi)x^2 + \mathcal{O}\big(x^3\big) \geq 1 - \frac{3}{2}g_\beta^2(\xi)x^2 \geq 1 - \frac{1}{2}M_g^2x^2
\end{aligned}$$

because of assumptions on ε, and similarly

$$\left(\frac{G_\beta(G_\beta^{-1}(\lambda)-x)}{\lambda}\right)^\lambda \left(\frac{1-G_\beta(G_\beta^{-1}(\lambda)-x)}{1-\lambda}\right)^{1-\lambda} \leq 1 - \frac{1}{2}g_\beta^2(\xi)x^2 \leq 1 - \frac{1}{2}m_g^2x^2,$$

where $\xi \in (G_\beta^{-1}(\lambda) - x, G_\beta^{-1}(\lambda))$ and $\lambda\zeta, (1-\lambda)\zeta' \in (0, g_\beta(\xi)x)$. Having these results in hand, we can estimate the last integral in (51) from above:

$$M_g \cdot \int_0^\varepsilon y \cdot \sup_\beta g_{\beta,h_n}(G_\beta^{-1}(\lambda) - y)\mathrm{d}y \leq \sqrt{n}M_g^2\frac{(1+o(1))}{\sqrt{2\pi\lambda(1-\lambda)}} \cdot \int_0^\varepsilon y \cdot (1 - \frac{1}{2}m_g^2y^2)^n\mathrm{d}y$$

and similarly from below. As

$$\begin{aligned}\int_0^{\varepsilon} y \cdot (1 - \frac{1}{2} m_g^2 y^2)^n \mathrm{d}y &= \frac{1}{m_g^2} \int_{1-\frac{1}{2}m_g^2\varepsilon^2}^{1} u^n \mathrm{d}u \\ &= \left[\frac{1}{m_g^2} \frac{u^{n+1}}{n+1}\right]_{1-\frac{1}{2}m_g^2\varepsilon^2}^{1} \\ &= \frac{1}{m_g^2(n+1)} \left[1 - \left(1 - \frac{1}{2} m_g^2 \varepsilon^2\right)^{n+1}\right],\end{aligned}$$

it follows that

$$\begin{aligned}& P(\{\omega \in \Omega_3 | \exists \beta : v_{in}(\beta) = -1\}) \\ = \; & P\big(\exists \beta : s_{[h_n]}(x_i, y_i; \beta, \omega') = s_{[h_n]}(x_i, y_i; \beta, \omega) < s_1(x_1, y_1; \beta, \omega_1) \leq G_\beta^{-1}(\lambda)\big) \\ = \; & \mathcal{O}\Big(n^{-\frac{1}{2}}\Big).\end{aligned}$$

Thus, we finally get the result

$$\begin{aligned}P(\{\omega \in \Omega^n | \exists \beta : v_{in}(\beta) = -1\}) &= P(\{\omega \in \Omega_1 | \exists \beta : v_{in}(\beta) = -1\}) \\ &\quad + P(\{\omega \in \Omega_2 | \exists \beta : v_{in}(\beta) = -1\}) \\ &\quad + P(\{\omega \in \Omega_3 | \exists \beta : v_{in}(\beta) = -1\}) \\ &= \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)\end{aligned}$$

as $n \to +\infty$. $\square$

Corollary 4 *Under Assumptions D and H for any fixed $i \in \mathbb{N}$ and $n \geq i$*

$$P\big(\exists \beta \in B : I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big) \neq I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\big) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$, or analogously

$$P\left(\sup_{\beta \in B} \left| I\big(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta)\big) - I\big(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)\big)\right| \neq 0\right) = \mathcal{O}\Big(n^{-\frac{1}{2}}\Big)$$

as $n \to +\infty$.

Proof: To prove this result, we can follow the proof of Lemma 4, but we have to make sure that all steps are uniformly valid for all $\beta \in B$.

First, equation (46) can be written as

$$\begin{aligned} & P\big(s_{[h_n]}(x_i,y_i;\beta,\omega') < s_1(x_1,y_1;\beta,\omega_1) \le G_\beta^{-1}(\lambda)\big) \\ \le\ & P\big(s_{[h_n]}(x_i,y_i;\beta,\omega') \le G_\beta^{-1}(\lambda) - \varepsilon/2\big) \\ +\ & P\big(G_\beta^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i,y_i;\beta,\omega') < s_1(x_1,y_1;\beta,\omega_1) \le G_\beta^{-1}(\lambda)\big). \end{aligned} \tag{52}$$

Then we have to find out more about the two probabilities on the right side of the inequality. Due to Corollary 3, the first probability

$$P\big(s_{[h_n]}(x_i,y_i;\beta,\omega') \le G_\beta^{-1}(\lambda) - \varepsilon/2\big) = \mathcal{O}\big(n^{-1}\big)$$

uniformly over $\beta \in B$ for $n \to \infty$. Next, the second probability can be expressed as

$$\begin{aligned} & P\big(\exists\beta : G_\beta^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i,y_i;\beta,\omega) < s_1(x_1,y_1;\beta,\omega_1) \le G_\beta^{-1}(\lambda)\big) \le \\ =\ & \int_{\omega'\in\Omega^{n-1}} \int_{\omega_1\in\Omega} I\big(G_\beta^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i,y_i;\beta,\omega') < s_1(x_1,y_1;\beta,\omega_1) \le G_\beta^{-1}(\lambda)\big) \\ & \qquad \mathrm{d}P(\omega_1)\mathrm{d}P(\omega') \\ \le\ & \int_{\omega'\in\Omega^{n-1}} M_g \cdot \\ & \cdot \sup_\beta \big\{ \big| G_\beta^{-1}(\lambda) - s_{[h_n]}(x_i,y_i;\beta,\omega') \big| \cdot I\big(G_\beta^{-1}(\lambda) - \varepsilon/2 < s_{[h_n]}(x_i,y_i;\beta,\omega') \le G_\beta^{-1}(\lambda)\big) \big\}\, \mathrm{d}P(\omega') \\ =\ & M_g \cdot \int_{G_\beta^{-1}(\lambda)-\varepsilon/2}^{G_\beta^{-1}(\lambda)} \sup_\beta \big\{ \big| G_\beta^{-1}(\lambda) - x \big| \cdot g_{\beta,h_n}(x) \big\}\, \mathrm{d}x \\ =\ & M_g \cdot \int_0^{\varepsilon/2} \sup_\beta \big\{ |y| \cdot g_{\beta,h_n}(G_\beta^{-1}(\lambda) - y) \big\}\, \mathrm{d}y \\ \le\ & M_g \cdot \int_0^{\varepsilon} y \cdot \sup_\beta \big\{ g_{\beta,h_n}(G_\beta^{-1}(\lambda) - y) \big\}\, \mathrm{d}y. \end{aligned}$$

This second term can be treated in the same way as in the proof of Lemma 4 as long as we are able to find $\varepsilon > 0$ and $n_0 \in \mathbb{N}$ such that the requirements (a)–(f) on page 54 in Lemma 4 are satisfied uniformly for all $\beta \in B$.

1. Requirement (a) follows from Assumption D4—$g_\beta(z)$ is bounded on $\mathbb{R}$ by M_{gg} uniformly in β over B.

2. Requirement (b) follows again from Assumption D4:

$$m_{gg} = \inf_{\beta \in B} \inf_{z \in (-\delta, \delta)} g_\beta \left(G_\beta^{-1}(\lambda) + z \right) > 0$$

because $G_\beta \left(G_\beta^{-1}(\lambda) - \varepsilon \right) = \lambda - g_\beta(\xi)\varepsilon$.

3. Requirement (c) is equivalent to Assumption D4:

$$m_{gg} = \inf_{\beta \in B} \inf_{z \in (-\delta, \delta)} g_\beta \left(G_\beta^{-1}(\lambda) + z \right) > 0.$$

4. Requirement (d) and (e) are independent of β, only M_g and m_g are replaced by M_{gg} and m_{gg}.

5. Requirement (f) just requires the existence of the probability density function for any β, so it is satisfied as well.

Hence, the proof can follow along the same lines for all $\beta \in B$ and because the bounds are chosen independently of β, the result holds uniformly in $\beta \in B$. □

Lemma 5 *Let Assumptions D and H hold and assume that $t(x, y; \beta)$ is a real-valued function continuous in β uniformly in x and y over any compact subset of the support of (x, y). Moreover, assume that Assumptions NC1–NC3 hold for $t(x, y; \beta)$. Finally, let G_β denote the distribution function of $s_i^2(x, y; \beta)$ (for any $\beta \in B$). Then*

$$\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right.$$
$$\left. - \mathsf{E}\left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right| \quad \to \quad 0$$

as $n \to +\infty$ almost surely.

Proof: This result is nothing but an application of the uniform law of large numbers for nonlinear models and I here use its variant due to Andrews (1987). Therefore, we just have to verify that the assumptions of the uniform law of large numbers are satisfied. We verify here assumptions A1, B1, B2, and A3, and employ them together with Andrews (1987, Corollary 1). To do so, let us follow the notation used in Andrews (1987) and denote

$$q(x, y; \beta) = t(x, y; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right)$$

and

$$\begin{aligned} q_*(x,y;\beta,\rho) &= \inf_{\beta' \in U(\beta,\rho)} q(x,y;\beta'), \\ q^*(x,y;\beta,\rho) &= \sup_{\beta' \in U(\beta,\rho)} q(x,y;\beta'), \\ q(x,y) &= \sup_{\beta \in B} |q(x,y;\beta)|. \end{aligned}$$

Assumption A1 B is compact metric space: this is satisfied because of Assumption NC1.

Assumption B1 (x_i, y_i) should be a sequence of strongly mixing random variables with mixing numbers $\alpha(s), s = 1, 2, \ldots$, that satisfy $\alpha(s) = o(s^{-\alpha/(\alpha-1)})$ as $s \to \infty$ for some $\alpha \geq 1$: this condition of asymptotic weak dependence is satisfied for $\alpha = 1$, because (x_i, y_i) are independent random vectors in our case (Assumption D1).

Assumption B2, part a $q^*(x_i, y_i; \beta, \rho), q_*(x_i, y_i; \beta, \rho)$, and $q(x_i, y_i)$ are random variables and $q^*(\cdot, \cdot; \beta, \rho), q_*(\cdot, \cdot; \beta, \rho)$ are measurable functions for all $i \in \mathbb{N}$, all $\beta \in B$, and all ρ sufficiently small: this follows from Assumption NC2 and the fact that $(x_i, y_i), i = 1, \ldots, n$, is a sequence of identically distributed random variables.

Assumption B2, part b $\mathsf{E}\, q(x_i, y_i)^{1+\delta} < \infty$ for some $\delta > 0$: this follows from Assumption NC3 and the fact that (x_i, y_i) is a sequence of identically distributed random variables.

Assumption A3 For all $\beta \in B$,

$$\lim_{\rho \to 0} |\mathsf{E}\, q^*(x_i, y_i; \beta, \rho) - \mathsf{E}\, q(x_i, y_i; \beta)| = 0 \quad \text{and} \quad \lim_{\rho \to 0} |\mathsf{E}\, q_*(x_i, y_i; \beta, \rho) - \mathsf{E}\, q(x_i, y_i; \beta)| = 0. \tag{53}$$

Without loss of generality, we will prove this result for $i = 1$ and only for supremum q^*(the other part can be proved analogously). By the definition of $q^*(x_1, y_1; \beta, \rho)$,

$$\begin{aligned} q^*(x_1, y_1; \beta, \rho) = \sup_{\beta' \in U(\beta,\rho)} & t(x_1, y_1; \beta') \cdot I\big(s_1(x_1, y_1; \beta') \leq G_{\beta'}^{-1}(\lambda)\big) && (54) \\ = \; & t(x_1, y_1; \beta) \cdot I\big(s_1(x_1, y_1; \beta) \leq G_{\beta}^{-1}(\lambda)\big) && (55) \\ + \; & \sup_{\beta' \in U(\beta,\rho)} t(x_1, y_1; \beta) \cdot \big[I\big(s_1(x_1, y_1; \beta') \leq G_{\beta'}^{-1}(\lambda)\big) - I\big(s_1(x_1, y_1; \beta) \leq G_{\beta}^{-1}(\lambda)\big)\big] && (56) \\ + \; & \sup_{\beta' \in U(\beta,\rho)} [t(x_1, y_1; \beta') - t(x_1, y_1; \beta)] \cdot I\big(s_1(x_1, y_1; \beta') \leq G_{\beta'}^{-1}(\lambda)\big). && (57) \end{aligned}$$

Hence, to verify (53) we just need to show that the expectations of (56) and (57)

converge to zero for $\rho \to 0$.

1. Let us start with (56). First, observe that

$$\begin{aligned} & \sup_{\beta' \in U(\beta,\rho)} \left\{ t(x_1, y_1; \beta) \cdot \left[I\left(s_1(x_1, y_1; \beta') \leq G_{\beta'}^{-1}(\lambda)\right) - I\left(s_1(x_1, y_1; \beta) \leq G_{\beta}^{-1}(\lambda)\right) \right] \right\} \leq \\ \leq & \sup_{\beta \in B} |t(x_1, y_1; \beta)| \cdot \sup_{\beta' \in U(\beta,\rho)} \left| I\left(s_1(x_1, y_1; \beta') \leq G_{\beta'}^{-1}(\lambda)\right) - I\left(s_1(x_1, y_1; \beta) \leq G_{\beta}^{-1}(\lambda)\right) \right|, \end{aligned}$$

where $\sup_{\beta \in B} |t(x_1, y_1; \beta)|$ is a function independent of β and with a finite expectation (Assumption NC3). Because the difference

$$\left| I\left(s_1(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda)\right) - I\left(s_1(x_i, y_i; \beta) \leq G_{\beta}^{-1}(\lambda)\right) \right|$$

is always lower than or equal to one, (56) has an integrable majorant independent of β. Therefore, if we show that the probability

$$\lim_{\rho \to 0} P\left(\sup_{\beta' \in U(\beta,\rho)} \left| I\left(s_1(x_i, y_i; \beta') \leq G_{\beta'}^{-1}(\lambda)\right) - I\left(s_1(x_i, y_i; \beta) \leq G_{\beta}^{-1}(\lambda)\right) \right| = 1 \right) = 0,$$

it implies, that the expectation of (56) converges to zero for $\rho \to 0$ as well.

Second, let us derive an intermediate result regarding the convergence of distribution function $G_{\beta'}$ to G_{β}. Note that $s_1(x_1, y_1; \beta', \omega) \to s_1(x_1, y_1; \beta, \omega)$ for $\beta' \to \beta$ and any $\omega \in \Omega$ (convergence almost surely). So, for $\beta' \to \beta$ and $G_{\beta}(x)$ being the cumulative distribution function of $s_1(x_1, y_1; \beta)$, it holds that $G_{\beta'}(x) \to G_{\beta}(x)$ for all $x \in \mathbb{R}$ (convergence in distribution) because $G_{\beta}(x)$ is an absolutely continuous distribution function. Now, we show that this convergence of distribution function $G_{\beta'}(x) \to G_{\beta}(x)$ is uniform over all sequences $\beta_n' \to \beta$ such that $\beta_n' \in U(\beta, \rho_n)$, where ρ_n is a sequence of positive numbers such that $\rho_n \to 0$. The reason is that $s_1(x_1, y_1; \beta') = s_1(x_1, y_1; \beta) - s_{\beta}'(x_1, y_1; \xi) \cdot (\beta' - \beta)$, where the second term $s_{\beta}'(x_1, y_1; \xi) \cdot (\beta' - \beta)$ can be bounded by a random variable of order $\mathcal{O}_p\left(n^{\frac{1}{3}}\right) \cdot \mathcal{O}\left(n^{-\frac{1}{2}}\right) = \mathcal{O}_p\left(n^{-\frac{1}{6}}\right)$ (see Assumption H3). Thus, the second term converges to zero in probability uniformly in $\beta \in B$. The same is true about the convergence of $G_{\beta'}^{-1}(\lambda)$ to $G_{\beta}^{-1}(\lambda)$ because G_{β} is absolutely continuous; the convergence is uniform over all sequences $\beta_n' \to \beta$ such that $\beta_n' \in U(\beta, \rho_n)$.

Third, let us choose now an arbitrary, but fixed $\varepsilon > 0$. Then we can find $n_1 \in \mathbb{N}$ such that $\left| G_{\beta'}^{-1}(\lambda) - G_{\beta}^{-1}(\lambda) \right| < \frac{\varepsilon}{8 M_{gg}}$ for any $\beta' \in U(\beta, \rho^1)$, where $\rho^1 = \rho_{n_1}$ and M_{gg} is the uniform upper bound for the probability density functions of $r_1^2(\beta)$ over all

$\beta \in B$ (see Assumption D4). Further, $s_1(x_1, y_1; \beta') = s_1(x_1, y_1; \beta) + s'_\beta(x_1, y_1; \xi) \cdot (\beta' - \beta)$, where $\xi \in [\beta, \beta']_\varkappa$. So, we can find $n_2 \in \mathbb{N}$ and $\rho^2 = \rho_{n_2}$ such that $\left|s'_\beta(x_1, y_1; \xi) \cdot (\beta' - \beta)\right| < \frac{\varepsilon}{8M_{gg}}$ with probability greater than $1 - \frac{\varepsilon}{2}$ ($\beta' \in U(\beta, \rho^2)$). Hence, setting $\rho^\varepsilon = \min\{\rho^1, \rho^2\}$,

$$
\begin{aligned}
& P\left(\sup_{\beta' \in U(\beta, \rho^\varepsilon)} \left| I\big(s_1(x_1, y_1; \beta') \le G^{-1}_{\beta'}(\lambda)\big) - I\big(s_1(x_1, y_1; \beta) \le G^{-1}_{\beta}(\lambda)\big) \right| = 1 \right) \\
\le \;& P\left(|s_1(x_1, y_1; \beta') - s_1(x_1, y_1; \beta)| \ge \frac{\varepsilon}{8M_{gg}} \right) \\
& + P\left(s_1(x_1, y_1; \beta) \in \left(G^{-1}_\beta(\lambda) - \frac{\varepsilon}{4M_{gg}}, G^{-1}(\lambda) + \frac{\varepsilon}{4M_{gg}} \right) \right) \\
\le \;& \frac{\varepsilon}{2} + \frac{2\varepsilon}{4M_{gg}} \cdot M_{gg} = \varepsilon,
\end{aligned}
$$

because M_{gg} is the uniform upper bound for the probability density functions of $s_1(x_i, y_i; \beta)$ over all $\beta \in B$. Thus, we have shown that for any $\varepsilon > 0$ we can find $\rho^\varepsilon > 0$ such that

$$
P\left(\sup_{\beta' \in U(\beta, \rho^\varepsilon)} \left| I\big(s_1(x_1, y_1; \beta') \le G^{-1}_{\beta'}(\lambda)\big) - I\big(s_1(x_1, y_1; \beta) \le G^{-1}_{\beta}(\lambda)\big) \right| = 1 \right) < \varepsilon
$$

and thus

$$
\lim_{\rho \to 0} P\left(\sup_{\beta' \in U(\beta, \rho)} \left| I\big(s_1(x_1, y_1; \beta') \le G^{-1}_{\beta'}(\lambda)\big) - I\big(s_1(x_1, y_1; \beta) \le G^{-1}_{\beta}(\lambda)\big) \right| = 1 \right) = 0.
$$

We have verified that the expectation of (56) converges to zero for $\rho \to 0$.

2. We should deal now with (57) and prove that

$$
\begin{aligned}
\left| \lim_{\rho \to 0} \mathsf{E} \left\{ \sup_{\beta' \in U(\beta, \rho)} [t(x_1, y_1; \beta') - t(x_1, y_1; \beta)] \cdot I\big(s_1(x_1, y_1; \beta') \le G^{-1}_{\beta'}(\lambda)\big) \right\} \right| \;&\le \\
\lim_{\rho \to 0} \mathsf{E} \left\{ \sup_{\beta' \in U(\beta, \rho)} |t(x_1, y_1; \beta') - t(x_1, y_1; \beta)| \right\} \;&= \; 0.
\end{aligned}
$$

First, note that the difference

$$
|t(x_1, y_1; \beta') - t(x_1, y_1; \beta)| \le |t(x_1, y_1; \beta')| + |t(x_1, y_1; \beta)| \le 2 \sup_{\beta \in B} |t(x_1, y_1; \beta)|
$$

can be bounded from above by a function independent of β and having a finite expectation (Assumption NC3). Let $2\,\mathsf{E}\sup_{\beta\in B}|t(x_1,y_1;\beta)| = U_E$.
Second, for an arbitrary fixed $\varepsilon > 0$, we can find a compact subset A_ε of the support of (x_1, y_1) (and its complement $\overline{A_\varepsilon}$) such that $P((x_1,y_1)\in A_\varepsilon) > 1-\frac{\varepsilon}{2U_E}$ (both x_1 and y_1 are random variables with finite second moments) and $2\int_{\overline{A_\varepsilon}}\sup_{\beta\in B}|t(x_1,y_1;\beta)| < \frac{\varepsilon}{2}$. Given this set A_ε and $\beta \in B$, we can employ the continuity of $t(x_1,y_1;\beta)$ in β (uniform over $(x_1,y_1)\in A_\varepsilon$) and find an $\rho^\varepsilon > 0$ such that

$$\sup_{(x_1,y_1)\in A_\varepsilon}\sup_{\beta'\in U(\beta,\rho^\varepsilon)}|t(x_1,y_1;\beta') - t(x_1,y_1;\beta)| < \frac{\varepsilon}{2}.$$

Hence,

$$\begin{aligned}\mathsf{E}\left\{\sup_{\beta'\in U(\beta,\rho^\varepsilon)}|t(x_1,y_1;\beta') - t(x_1,y_1;\beta)|\right\} &\leq \int_{\overline{A_\varepsilon}} 2\sup_{\beta\in B}|t(x_1,y_1;\beta)|\,\mathrm{d}F_x(x_1)\mathrm{d}F_\varepsilon(\varepsilon_1)\\ &\quad + \int_{A_\varepsilon}\frac{\varepsilon}{2}\mathrm{d}F_x(x_1)\mathrm{d}F_\varepsilon(\varepsilon_1)\\ &\leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon,\end{aligned}$$

and consequently,

$$\lim_{\rho\to 0}\mathsf{E}\left\{\sup_{\beta'\in U(\beta,\rho)}|t(x_1,y_1;\beta') - t(x_1,y_1;\beta)|\right\} = 0.$$

We have verified that the expectation of (57) converges to zero for $\rho \to 0$. Thus, assumption A3 of Andrews (1987) is satisfied as well.

Since we have verified all assumptions needed for the uniform law of large numbers, we can use it for $\frac{1}{n}\sum_{i=1}^n\left[t(x_i,y_i;\beta)\cdot I\big(s_i(x_i,y_i;\beta)\leq G_\beta^{-1}(\lambda)\big)\right]$ to get the result of the lemma. □

Lemma 6 *Let Assumptions D and H hold and assume that $t(x,y;\beta)$ is a real-valued function continuous in β uniformly in x and y over any compact subset of the support of (x,y). Moreover, assume that Assumptions NC1–NC3 hold for $t(x,y;\beta)$. Furthermore, let G_β denote the distribution function of $s_i(x_i,y_i;\beta)$ (for any $\beta\in B$). Finally, let*

$h_n/n \to \lambda \in \left(\frac{1}{2}, 1\right)$. *Then*

$$\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta) \right) \right] \right.$$
$$\left. - \mathsf{E}\left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right| \to 0$$

as $n \to +\infty$ *in probability.*

Proof: Using the result of Lemma 5 (the assumptions of this lemma and Lemma 5 are identical) and

$$\begin{aligned}
& \sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta) \right) \right] \right. \\
& \qquad \left. - \mathsf{E}\left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right| \\
\leq \; & \sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right. \\
& \qquad \left. - \mathsf{E}\left[t(x_i, y_i; \beta) \cdot I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right| \\
+ \; & \sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} t(x_i, y_i; \beta) \cdot \left[I\left(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta) \right) - I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right|,
\end{aligned}$$

we just have to prove that

$$\sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} t(x_i, y_i; \beta) \cdot \left[I\left(s_i(x_i, y_i; \beta) \leq s_{[h_n]}(x_i, y_i; \beta) \right) - I\left(s_i(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) \right) \right] \right| \to 0$$

in probability for $n \to \infty$. The Chebyshev inequality for non-negative random variables—$P(X \geq K) \leq \mathsf{E}\, X/K$—implies for a sequence of non-negative random variables X_n that if expectations $\mathsf{E}\, X_n$ converges to zero for $n \to \infty$, then the sequence X_n converges to 0 in

probability. So, we will derive now that

$$\begin{aligned}
& \mathsf{E}\left\{\sup_{\beta\in B}\left|\frac{1}{n}\sum_{i=1}^{n} t(x_i,y_i;\beta)\cdot\left[I\big(s_i(x_i,y_i;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big)-I\big(s_i(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\big)\right]\right|\right\} \\
= \;& \mathsf{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\sup_{\beta\in B}|t(x_i,y_i;\beta)|\times\right. \\
& \qquad \left.\times\sup_{\beta\in B}\left|I\big(s_i(x_i,y_i;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big)-I\big(s_i(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda)\big)\right|\right\} \\
= \;& \mathsf{E}\left\{\sup_{\beta\in B}|t(x_1,y_1;\beta)|\cdot\sup_{\beta\in B}\left|I\big(s_1(x_1,y_1;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big)-I\big(s_1(x_1,y_1;\beta)\le G_\beta^{-1}(\lambda)\big)\right|\right\}
\end{aligned}$$

converges to zero for $n\to\infty$. Since we assume that $\sup_{\beta\in B}|t(x_1,y_1;\beta)|$ has a finite first moment, all we have to actually prove is

$$\begin{aligned}
\mathsf{E}\left\{\sup_{\beta\in B}\left|I\big(s_1(x_1,y_1;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big)-I\big(s_1(x_1,y_1;\beta)\le G_\beta^{-1}(\lambda)\big)\right|\right\} \;&=\; \\
P\left(\sup_{\beta\in B}\left|I\big(s_1(x_1,y_1;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big)-I\big(s_1(x_1,y_1;\beta)\le G_\beta^{-1}(\lambda)\big)\right|=1\right) \;&\to\; 0
\end{aligned}$$

for $n\to\infty$. But this is the claim of Corollary 4:

$$P\left(\sup_{\beta\in B}\left|I\big(s_1(x_1,y_1;\beta)\le s_{[h_n]}(x_i,y_i;\beta)\big)-I\big(s_1(x_1,y_1;\beta)\le G_\beta^{-1}(\lambda)\big)\right|=1\right)=\mathcal{O}\left(n^{-\frac{1}{2}}\right)$$

as $n\to\infty$. □

References

[1] Amemiya, T. (1983): Non-linear regression models. In Griliches, Z., and Intriligator, M. D., eds., *Handbook of Econometrics Vol. 1*, North-Holland, Amsterdam, 333–389.

[2] Andrews, D. W. K. (1987): Consistency in nonlinear econometic models: A generic uniform law of large numbers. *Econometrica*, Vol. 55 / 6, 1465–1471.

[3] Arabmazar, A., and Schmidt, P. (1981): Further evidence on the robustness of the Tobit estimator to heteroscedasticity. *Journal of Econometrics*, Vol. 17, 253–258.

[4] Arabmazar, A., and Schmidt, P. (1982): An investigation of the robustness of the Tobit estimator to non-normality. *Econometrica*, Vol. 50 / 4.

[5] Box, G. E. P., and Cox, D. (1964): An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 211–264.

[6] Brillinger, D. R. (1983): A generalized linear model with 'Gausian' regressor variables. In P. J. Bickel, K. A. Doksum, and J. L. Hodges eds., *A festschrift for Erich L. Lehman*, Woodsworth International Group, Belmont.

[7] Chatterjee, S., and Hadi, A. S. (1988): *Sensitivity analysis in linear regression.* Wiley, New York.

[8] Chen, Y., Stromberg, A., and Zhou, M. (1997): The least trimmed squares estimate in nonlinear regression. *Technical report*, Department of statistics, University of Kentucky, 365/2000.

[9] Čížek, P. (2001a): Robust estimation in nonlinear regression models. *SFB 373 Discussion paper,* 25/2001.

[10] Čížek, P. (2001b): Robust estimation with discrete explanatory variables. *mimeo.*

[11] Dick, D., Terasvirta, T., and Franses, P. H. (2000): Smooth transition autoregressive models—a survey of recent developments. *SSE/EFI Working paper series in Economics and Finance,* 380/2000.

[12] Griffiths, W. E., Hill, R. C., and Judge, G. G. (1993): *Learning and practicing econometrics.* Wiley, New York.

[13] 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.

[14] Härdle, W., Linton, O. (1994): Applied nonparametric methods. In R. F. Engle and D. L. McFadden, eds., *Handbook of Econometrics IV*, Elsevier science.

[15] Hausman, J. A., and Wise, D. A. (1976): The evaluation of results from truncated samples: The New Jeyrsey negative income tax experiment. *Annals of Economic and Social Measurement*, Vol. 5, 421–445.

[16] Horowitz, J. L. (1993): Semiparametric estimation of a work-trip mode choice model. *Journal of Econometrics*, Vol. 58, 49–70.

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

[18] Hurd, M. (1979): Estimation in truncated samples when there is heteroskedasticity. *Journal of Econometrics*, Vol. 11, 247–258.

[19] Ichimura, H. (1993): Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, Vol. 58, 71–120.

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

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

[22] Klein, R. W., and Spady, R. H (1993): An efficient semi-parametric estimator for binary response models. *Econometrica*, Vol. 61, 387–421.

[23] Lee, M. (1995): A semiparametric estimation of simultaneous equations with limited dependent variables: A case study of female labor supply. *Journal of Applied Econometrics*, Vol. 10, 187–200.

[24] Lee, M. (1996): *Methods of moments and semiparametric econometrics for limited dependent variable models.* Springer-Verlag, New York.

[25] Maddala, G. S. (1983): *Limited-dependent and qualitative variables in econometrics.* Press syndicate of the University of Cambridge, Cambridge.

[26] Manski, C. F., and Thompson, T. S. (1986): Operational properties of the maximum score estimator. *Journal of Econometrics*, Vol. 32, 65–108.

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

[28] Powell, J. L. (1986): Symmetrically trimmed least squares estimation for Tobit models. *Econometrica*, Vol. 54, 1435–1460.

[29] Powell, J. L., Stock, J. H., and Stoker, T. M. (1989): Semiparametric estimation of index coefficients. *Econometrica*, Vol. 57, 1403–1430.

[30] Proietti, T. (1998): Characterizing asymmetries in business cycles using smooth-transition structural time-series models. *Studies in Nonlinear Dynamics and Econometrics*, Vol. 3 / 3, 141–156.

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

[32] 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.

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

[34] Seber, G. A. F., and Wild, C. J. (1989): *Nonlinear regression.* Wiley, New York.

[35] Stromberg, A. J. (1993): High breakdown estimation of nonlinear regression parameters. *Journal of American Statistical Association*, Vol. 88, 237–244.

[36] Terasvirta, T., and Anderson, H. M. (1992): Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, Vol. 7, 119–136.

[37] Tong, H. (1990): *Nonlinear time series: A dynamical systems approach.* Oxford University Press, Oxford.

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

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

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

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