Econometric Theory, **24**, 2008, 1500–1529. Printed in the United States of America.
doi:10.1017/S0266466608080596

GENERAL TRIMMED ESTIMATION: ROBUST APPROACH TO NONLINEAR AND LIMITED DEPENDENT VARIABLE MODELS

PAVEL ČÍŽEK
Tilburg University

High-breakdown-point regression estimators protect against large errors and data contamination. We generalize the concept of trimming used by many of these robust estimators, such as the least trimmed squares and maximum trimmed likelihood, and propose a general trimmed estimator, which renders robust estimators applicable far beyond the standard (non)linear regression models. We derive here the consistency and asymptotic distribution of the proposed general trimmed estimator under mild β-mixing conditions and demonstrate its applicability in nonlinear regression and limited dependent variable models.

1. INTRODUCTION

In econometrics, more and more attention is paid to techniques that can deal with data contamination, which can arise from miscoding or heterogeneity not captured or presumed in a model. Evidence about contamination of a part of data and its adverse effects on estimators such as (quasi-) maximum likelihood is provided, for example, by Gerfin (1996) in labor market data, by Sakata and White (1998) in financial time series, and by Čížek (2004a) in the prices of financial derivates. An effect of data contamination and large errors on an estimator, that is, its robustness, is often characterized by the breakdown point:[1] it measures the smallest fraction of a sample that can arbitrarily change the estimator under contamination (for the standard definition, see Rousseeuw and Leroy, 2003; for a discussion of the breakdown point under dependency, see Genton and Lucas, 2003). The breakdown point of standard regression methods, such as ordinary least squares, typically approaches zero with an increasing sample size. One way to construct a positive-breakdown-point method is to employ a standard (nonrobust) estimator and to trim some "unlikely" observations from its objective function. For example in linear regression, this is the case of the least trimmed squares (LTS) by Rousseeuw (1985), the least trimmed

Address correspondence to Pavel Čížek, Tilburg University, Faculty of Economics and Business Administration, Department of Econometrics and Operation Research, Room K616, P.O. Box 90153, 5000 LE Tilburg, The Netherlands; e-mail: P.Cizek@uvt.nl.

 © 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

absolute deviations (LTA) by Bassett (1991), and the maximum trimmed likelihood estimator (MTLE) by Neykov and Neytchev (1990) and Hadi and Luceno (1997). Here we generalize the concept of trimming, prove its consistency and asymptotic normality, and demonstrate its applicability in many econometric models including nonlinear regression, time series, and limited dependent variable models.

First, let us briefly review existing results concerning the LTS estimator, the LTA estimator, and the MTLE. The LTS estimator belongs to the class of affine-equivariant estimators that can achieve asymptotically the highest breakdown point $\frac{1}{2}$, and it is generally preferred to the similar, but slowly converging, least median of squares (LMS; Rousseeuw, 1984).[2] Thus, LTS has been receiving a lot of attention from the theoretical, computational, and application points of view. There are extensions involving nonlinear regression (Stromberg, 1993), weighted LTS (Víšek, 2002), and an adaptive choice of trimming (Čížek, 2002; Gervini and Yohai, 2002), and in most of these cases, the asymptotic and breakdown behavior is known in the standard regression model with independent and identically distribution (i.i.d.) regressors and errors. Simultaneously, there has been a significant development in computational methods (Agulló, 2001; Gilloni and Padberg, 2002). Last, but not least, there are also first applications of LTS in economics (Temple, 1998; Zaman, Rousseeuw, and Orhan, 2001) and finance (Knez and Ready, 1997; Kelly, 1997).

Next, the LTA estimator has not attracted much attention yet despite its favorable computational and robustness properties (Hawkins and Olive, 1999). The asymptotic properties are known in the univariate location model (Tableman, 1994) and linear regression (Hössjer, 1994). Finally, the MTLE, which can produce the LMS, LTS, maximum likelihood, and some other estimators in special cases (Hadi and Luceno, 1997), has been studied from the robustness point of view and applied in the context of (generalized) linear models (Vandev and Neykov, 1998; Müller and Neykov, 2003). Despite the appealing concept of the trimmed likelihood, the asymptotic results are known only in the case of linear regression with Gaussian errors (Vandev and Neykov, 1993).

The aim of this work is to generalize the principle of LTS, LTA, and MTLE, that is, trimming of "unlikely" observations from the model point of view. The proposed general trimmed estimator (GTE) not only includes LTS, LTA, and MTLE as special cases but also allows us to combine the trimming principle with many other existing parametric and semiparametric estimators in a variety of econometric models to make these estimators robust. For GTE based on extremum estimators defined by a smooth objective function, we prove its consistency and derive its asymptotic distribution under rather general conditions, which permit applying trimmed estimators in a wide range of econometric applications including time series, panel data, and limited dependent variable models. Thus, the application area of robust trimmed estimators is extended substantially. Another important consequence of the derived results is the consistency of LTA and the consistency and asymptotic normality of MTLE in

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

general multivariate location and regression models, which was not available up to now. The main tools in achieving this are the (uniform) law of large numbers (Andrews, 1988, 1992) and the uniform central limit theorem (Arcones and Yu, 1994) for mixing processes. On the other hand, the computational issues and robustness properties of GTE, which are presumably analogous to those of LTS, LTA, and MTLE, are not discussed here to a larger extent because of many existing studies that address the computation and breakdown behavior of trimmed estimators.

In the rest of the paper, we first propose the GTE in Section 2, where we also extensively discuss assumptions needed for studying asymptotic properties of GTE. Asymptotic results are summarized in Section 3. A number of specific trimmed estimators in various econometric models are presented in Section 4. The proofs are provided in two Appendixes.

2. GENERAL TRIMMED ESTIMATOR

Let us now introduce the GTE (Sect. 2.1). Later, the assumptions used in the paper and an alternative definition of GTE are discussed (Sects. 2.2 and 2.3).

2.1. General Trimmed Estimator

Let us consider a sample $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^k$ represents a vector of explanatory variables and $y_i \in \mathbb{R}$ denotes the dependent variable. Such data can be described, for example, by a (non)linear regression model

$$y_i = m(x_i, \beta^0) + \varepsilon_i, \tag{1}$$

where $m(x_i, \beta)$ is a regression function of explanatory variables x_i and unknown parameters β and ε_i is a continuously distributed error term. To estimate, we assume that $s(x_i, y_i; \beta)$ represents a loss function identifying the true value β^0 of parameter vector $\beta \in B$ in a compact parameter space $B \subset \mathbb{R}^p$.[3] For example, $s(x_i, y_i; \beta) = \{y_i - m(x_i, \beta)\}^2$ in the case of the least squares estimation of model (1), and $s(x_i, y_i; \beta) = -\ln l(x_i, y_i; \beta)$ in the case of the log-likelihood criterion. Further, let small values of $s(x_i, y_i; \beta)$ represent likely observations under a given model ("good fit," small squared residuals, high likelihood) and large values of $s(x_i, y_i; \beta)$ correspond to unlikely values ("bad fit," large squared residuals, low likelihood).

To achieve a high breakdown point, many robust methods such as LTS and MTLE trim unlikely observations, that is, observations (x_i, y_i) with large values of the loss function $s(x_i, y_i; \beta)$. Abstracting this concept, the general trimmed estimator $\hat{\beta}_n^{(GTE,h)}$ can therefore be defined as

$$\hat{\beta}_n^{(GTE,h)} = \arg\min_{\beta \in B} \sum_{j=1}^{h} s_{[j]}(\beta), \tag{2}$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

where $s_{[j]}(\beta)$ represents the jth smallest order statistics of $s(x_i, y_i; \beta), i = 1, \ldots, n$. Thus, the GTE estimate minimizes the loss of h most likely observations under a given parametric model. Apparently, this definition includes the LTS estimator and the MTLE as special cases for $s(x_i, y_i; \beta)$ being equal to $\{y_i - m(x_i, \beta)\}^2$ and $-\ln l(x_i, y_i; \beta)$, respectively.

The robust properties of trimmed estimators, especially their breakdown point, are closely related to the trimming constant h, which must satisfy $n/2 < h \leq n$ for an affine-equivariant estimator. This follows from definition (2), which implies that $n - h$ observations with the largest losses do not directly affect the estimator. In other words, the $n - h$ observations that are most unlikely in a given parametric model are dropped from the objective function. For example, in the case of the least-squares loss and $m(x, \beta) = g(x^\top \beta)$, where $g(t)$ is unbounded for $t \to \pm\infty$, Stromberg and Ruppert (1992) showed that the breakdown point equals asymptotically $\frac{1}{2}$ for $h = [n/2] + 1$ (most robust choice) and 0 for $h = n$ (nonlinear least squares). For more details on the properties of LTS in linear and nonlinear regression, see Čížek and Víšek (2000), Víšek (2000), and Čížek (2006a), Stromberg (1993), respectively. The robustness properties of MTLE are similar to those of LTS, and they were studied in (generalized) linear regression models (Vandev and Neykov, 1998; Müller and Neykov, 2003).

Despite the flexibility of definition (2), an even more general form of trimming is necessary to make trimmed estimation operational in some models (e.g., binary-choice or panel data models). Let us introduce an auxiliary trimming function $r(x_i, y_i; \beta)$, which also indicates likely and unlikely observations in a given model by small and large values, respectively, and let $r_{[j]}(\beta)$ denote the jth smallest order statistics of $r(x_i, y_i; \beta)$, $i = 1, \ldots, n$. Further, let $s_{r:[j]}(\beta)$ be the value of $s(x_l, y_l; \beta)$ at observation (x_l, y_l) corresponding to the jth-order statistics, $r_{[j]}(\beta) = r(x_l, y_l; \beta)$. Then the GTE is defined by

$$\hat{\beta}_n^{(GTE,h)} = \arg\min_{\beta \in B} \sum_{j=1}^{h} s_{r:[j]}(\beta). \tag{3}$$

In other words, the ordering of observations and their inclusion in the objective function are not determined by ordering values $s(x_i, y_i; \beta)$ of the loss function $s(x, y; \beta)$ but by ordering values $r(x_i, y_i; \beta)$ of the auxiliary trimming function $r(x, y; \beta)$. Although the existing trimmed estimators are based on $r(x, y; \beta) = s(x, y; \beta)$, using GTE in binary-choice models, for instance, will require $r(x, y; \beta) = \max_y s(x, y; \beta)$ (see Sect. 4 for details).

To provide an overview of how GTE nests and extends existing estimators, a summary of known and proposed trimmed estimators is presented in Table 1. Clearly, there are two important contributions of this paper: first, the asymptotic normality of MTLE, which was studied only from the robust point of view up to now, and second, the generalization of the trimmed estimators so that they can be employed in the context of limited dependent variable models. On the other hand, the breakdown point and other robust properties of trimmed

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

TABLE 1. Overview of existing and proposed trimmed estimators that are nested by the GTE concept

Model	Method	Robust properties	Asymptotic properties
Linear regression	Least trimmed squares (Rouseeuw, 1985)	Rousseeuw and Leroy (2003)	Čížek (2006a)
	Least trimmed absolute deviations (Bassett, 1991)	Hawkins and Olive (1999)	Hössjer (1994)
	Maximum trimmed likelihood (Hadi and Luceno, 1997)	Vandev and Neykov (1998)	Section 3
Generalized linear models	Maximum trimmed likelihood (Hadi and Luceno, 1997)	Müller and Neykov (2003)	Section 3
Limited dependent variable models	General trimmed estimator: examples for truncated regression (Čížek, 2007)	—	Section 3
Binary-choice models	General trimmed estimator: Maximum symmetrically trimmed likelihood (Sect. 4.2)	Čížek (2006b)	Section 3

Note: The previously published robust and asymptotic properties of the trimmed estimators are indicated by the reference to the corresponding publication, whereas new results are marked by reference to the section where they can be found in this paper.

estimators are only briefly mentioned in the rest of the paper, and an interested reader is referred to the sources mentioned in Table 1.

Before discussing assumptions concerning GTE, let us briefly return to the trimming constant h. Naturally, the choice of the trimming constant h should vary with the sample size n, and therefore we have to work with a sequence of trimming constants h_n. As h_n/n determines the fraction of a sample included in the GTE objective function and, consequently, the robust properties of GTE, we want to asymptotically fix this fraction at λ, $0 < \lambda \leq 1$. The trimming constant for a given sample size n can be then defined by $h_n = [\lambda n]$, where $[x]$ represents the integer part of x (in general, one can also consider any sequence $\{h_n\}_{n\in\mathbb{N}}$ such that $h_n/n \to \lambda$). In what follows, we derive asymptotic properties of GTE for any $0 < \lambda \leq 1$.[4] However, to ensure the robustness of affine-equivariant trimmed estimators, $\lambda \geq \frac{1}{2}$ must hold so that the estimator can "distinguish" the majority of correct observations from the minority of contaminated data points (Rousseeuw, 1997). Even though the value $\lambda = \frac{1}{2}$ corresponds to the most robust choice in models with continuously distributed response variable (for the case of generalized linear models, see Müller and Neykov, 2003), the most robust choice of λ can differ for other models. For example, the most robust choice of λ satisfies $\lambda \geq \frac{2}{3}$ in dichotomous binary-choice models (Čížek, 2006b).

2.2. Assumptions

Let us now complement the GTE definition first by some notation and definitions and later by assumptions on the random variables and loss and trimming functions needed for further analysis.

First, we refer to the distribution functions of $s(x_i, y_i;\beta)$ and $r(x_i, y_i;\beta)$ as $F_\beta(z)$ and $G_\beta(z)$ and to the corresponding probability density functions, if they exist, as $f_\beta(z)$ and $g_\beta(z)$, respectively. At the true parameter value β^0, we also use a simpler notation $F \equiv F_{\beta^0}$ and $G \equiv G_{\beta^0}$, and similarly for density functions, $f \equiv f_{\beta^0}$ and $g \equiv g_{\beta^0}$. Whenever we need to refer to the quantile functions corresponding to F_β and G_β, notation F_β^{-1} and G_β^{-1} is used. Next, because the derivatives of functions $s(x, y;\beta)$ and $r(x, y;\beta)$ are taken only with respect to β here, we denote them simply by $s'(x, y;\beta)$, $r'(x, y;\beta)$, and so on. Two purely mathematical symbols we need are the indicator function $I(A)$, which equals 1 if A is true and 0 otherwise, and an open δ-neighborhood of a point x in a euclidean space $\mathbb{R}^l$: $U(x, \delta) = \{z \in \mathbb{R}^l | \|z - x\| < \delta\}$.

Second, let us introduce the concept of β-mixing, which is central to the distributional assumptions made in this paper. A sequence of random variables $\{X_i\}_{i\in\mathbb{N}}$ is said to be absolutely regular (or β-mixing) if

$$\beta_m = \sup_{t\in\mathbb{N}} \mathsf{E} \sup_{B\in\sigma_{t+m}^{l}} |P(B|\sigma_t^p) - P(B)| \to 0$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

as $m \to \infty$, where the σ-algebras $\sigma_t^p = \sigma(X_t, X_{t-1}, \ldots)$ and $\sigma_t^f = \sigma(X_t, X_{t+1}, \ldots)$; see Davidson (1994) and Arcones and Yu (1994) for details. Numbers $\beta_m, m \in \mathbb{N}$, are called mixing coefficients.

Now, I specify all the assumptions necessary to derive the consistency and asymptotic normality of GTE (a smaller subset of assumptions sufficient for the consistency of GTE is discussed at the end of the section). They form three groups: distributional Assumptions D for random variables (x_i, y_i), Assumptions F concerning properties of the loss function $s(x, y; \beta)$ and auxiliary trimming function $r(x, y; \beta)$, and finally, identification Assumptions I.

Assumptions D.

D1. Random variables $\{y_i, x_i\}_{i \in \mathbb{N}}$ form a strongly stationary absolutely regular sequence of random vectors with mixing coefficients satisfying $m^{r_\beta/(r_\beta-2)}(\log m)^{2(r_\beta-1)/(r_\beta-2)}\beta_m \to 0$ as $m \to +\infty$ for some $r_\beta > 2$.

D2. The distribution function G_β of $r(x_i, y_i; \beta)$ is absolutely continuous for any $\beta \in B$.

D3. Assume that for $m_G = \inf_{\beta \in B} G_\beta^{-1}(\lambda)$ and $M_G = \sup_{\beta \in B} G_\beta^{-1}(\lambda)$, it holds that

$$M_{gg} = \sup_{\beta \in B} \sup_{z \in (m_G - \delta_g, M_G + \delta_g)} g_\beta(z) < \infty$$

and

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

for some $\delta_g > 0$.

Having a general objective function $s(x, y; \beta)$, Assumption D1 is one of rather weak conditions for the uniform central limit theorem used by Andrews (1993) and Arcones and Yu (1994), for instance. Assumption D2 indicates that at least one random variable has to be continuously distributed. Note though that the absolute continuity of G_β is necessary only in a neighborhood of its λ-quantile $G_\beta^{-1}(\lambda)$, as used in Assumption D3. Assumption D3 formalizes two things: first, the density function g_β has to be bounded uniformly in $\beta \in B$, which prevents distribution G_β from becoming or being arbitrarily close to a discrete or singular one for some $\beta \in B$. Second, the density function has to be positive in a neighborhood of the λ-quantile of G_β, that is, around the chosen "trimming" point of the $r(x_i, y_i; \beta)$ distribution. In a less general setting when the structure of a model is known, Assumption D3 is usually implied by $G \equiv G_{\beta^0}$ being absolutely continuous with a density function $g \equiv g_{\beta^0}$ positive, bounded, and differentiable around $G^{-1}(\lambda)$; see Čížek (2006a) for the case of nonlinear regression. Differentiability of the density function g around the point corresponding to the λ-quantile of the $r(x_i, y_i; \beta^0)$ distribution is a

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

standard condition needed for the analysis of rank statistics (see, e.g., Hössjer, 1994; Zinde-Walsh, 2002).

Next, several conditions on the loss function $s(x, y;\beta)$ and auxiliary trimming function $r(x, y;\beta)$ have to be specified. The GTE concept aims to add robust qualities to extremum estimators that lack robustness but preferably possess other desirable properties such as asymptotic normality and some kind of optimality. Because the loss function defining an extremum estimator typically has to be smooth to guarantee such properties and because the trimming function is usually closely related to the loss function (see Sect. 2.1 and Čížek, 2007), we will assume that both functions $s(x, y;\beta)$ and $r(x, y;\beta)$ are differentiable, at least in a neighborhood $U(\beta^0, \delta)$ of β^0. Similarly to other extremum estimators, the asymptotic variance of GTE will then depend on the expectations of these derivatives (cf. Manski, 1988). Specifically, it depends on the variance of the normal equations (see Lemma 2.1 in Sect. 2.3),

$$V_s(\lambda) = \mathsf{E}\{s'(x_i, y_i;\beta^0)s'(x_i, y_i;\beta^0)^\top I(r(x_i, y_i;\beta^0) \leq G^{-1}(\lambda))\}, \tag{4}$$

and on the expected value of the derivative of the normal equations with respect to parameters β, which, by the product rule, consists of the trimmed second derivative of the loss function,

$$Q_s(\lambda) = \mathsf{E}\{s''(x_i, y_i;\beta^0)I(r(x_i, y_i;\beta^0) \leq G^{-1}(\lambda))\}, \tag{5}$$

and the derivative of the expectation of the trimming indicator in the normal equations,

$$J_s(\lambda) = \frac{\partial}{\partial \beta^\top} \mathsf{E}\{s'(x_i, y_i;\beta^0)I(r(x_i, y_i;\beta) \leq G_\beta^{-1}(\lambda))\}\bigg|_{\beta=\beta^0}. \tag{6}$$

Assumptions F. Let us assume that there are a positive constant $\delta > 0$, a neighborhood $U(\beta^0, \delta)$, and an integer $n_0 \in \mathbb{N}$ such that the following assumptions hold.

- F1. Let $s(x_i, y_i;\beta)$ and $r(x_i, y_i;\beta)$ be continuous (uniformly over any compact subset of the support of (x, y)) in $\beta \in B$, $r(x_i, y_i;\beta)$ be differentiable in β on $U(\beta^0, \delta)$ almost surely, and $s(x_i, y_i;\beta)$ be twice differentiable in β on $U(\beta^0, \delta)$ almost surely with the locally Lipschitz second derivative $s''(x_i, y_i;\beta)$.
- F2. Let $\{r(x_i, y_i;\beta)|\beta \in U(\beta^0, \delta)\}$ and $\{s'(x_i, y_i;\beta)|\beta \in U(\beta^0, \delta)\}$ form Vapnik-Červonenkis (VC) classes of functions. Moreover, let us assume that the trimmed envelope $E_s(x) = \sup_{\beta\in U(\beta^0,\delta)} \sup_{n\geq n_0}|s'(x_i, y_i;\beta) I(r(xi, yi;\beta) \leq r_{[h_n]}(\beta))|$ has finite r_βth moments.
- F3. Expectations $\mathsf{E}\sup_{\beta\in B}|r_{[h_n]}(\beta)|$, $\mathsf{E}\sup_{\beta\in B}\sup_{n\geq n_0}|s(x_i, y_i;\beta)I(r(x_i, y_i;\beta)\leq r_{[h_n]}(\beta))|$, $\mathsf{E}\sup_{\beta\in U(\beta^0,\delta)}\sup_{n\geq n_0}|\partial s(x_i, y_i;\beta)/\partial\beta_k \cdot \partial s(x_i, y_i;\beta)/\partial\beta_l \cdot I(r(x_i, y_i;\beta) \leq r_{[h_n]}(\beta))|$, and $\mathsf{E}\sup_{\beta\in U(\beta^0,\delta)}\sup_{n\geq n_0}|\partial^2 s(x_i, y_i;\beta)/\partial\beta_k\partial\beta_l \cdot I(r(x_i, y_i;\beta) \leq r_{[h_n]}(\beta))|$ exist and are finite for $k, l = 1, \ldots, p$.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Moreover, assume that $Q_s(\lambda) + J_s(\lambda)$ and $V_s(\lambda)$ are nonsingular positive definite matrices.

F4. Conditional expectation

$$\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\delta)}|s'(x_i,y_i;\beta)I(r(x_i,y_i;\beta)\in\mathcal{I}(\beta))|\right.$$
$$\left.\Big|\exists\beta\in U(\beta^0,\delta): r(x_i,y_i;\beta)\in\mathcal{I}(\beta)\right\}, \tag{7}$$

where $\mathcal{I}(\beta)=\{z:|z-G^{-1}(\lambda)|\le|z-r_{[h_n]}(\beta)|\}$ is uniformly bounded for $n\ge n_0$.

As already discussed, the differentiability of the loss function and differentiability of the trimming function are standard assumptions. On the other hand, Assumption F2, which allows us to derive the convergence rate of the order statistics in this general framework, deserves further comment, because it limits the class of functions $s'(x,y;\beta)$ and $r(x,y;\beta)$ to VC classes (for a definition, see Pollard, 1984; Van der Vaart and Wellner, 1996). Although limited, VC classes cover many common functions, including polynomial, logarithmic, and exponential functions, their sums, products, maxima and minima, monotonic transformations, and so on. For example, trimming functions having a single-index form $\tau(x_i^\top\beta)$ with a monotonic link function τ are covered by Assumption F2. Even though this assumption is not necessarily restrictive in many contexts and it is not needed for the proof of consistency, it can be omitted as long as we impose stronger distributional assumptions. An alternative set of assumptions can be used to prove the L^{r_β}-continuity of $I(r(x_i,y_i;\beta)\le G_\beta^{-1}(\lambda))$ in $U(\beta^0,\delta)$ and to limit the bracketing cover numbers following Andrews (1993). Consequently, the results of Doukhan, Massart, and Rio (1995) could be employed instead of those of Arcones and Yu (1994) that are used in the current paper.

Further, let us comment on Assumption F3 concerning the existence of various expectations. First, the expectations $V_s(\lambda)$, $Q_s(\lambda)$, and $J_s(\lambda)$ are trimmed forms of the standard expectations (variances) that appear in the asymptotic variances of extremum estimators (see, e.g., Pakes and Pollard, 1989). Next, we assume that the trimmed derivatives of the loss function $s(x,y;\beta)$ have an integrable majorant in some small neighborhood $U(\beta^0,\delta)$. This is not very restrictive given that those expectations have to exist at β^0, that is, for $\delta=0$, and the derivatives are continuous. Additionally, we have to assume the existence of integrable majorants of the trimming function and trimmed loss function on the whole parametric space B. The identification assumptions presented subsequently, however, require that the parametric space B is compact and thus bounded, which makes Assumption F3 much less strict (alternatively, one can assume, e.g., $\sup_{\beta\in B}\mathsf{E}|r(x_i,y_i;\beta)|^{1+\varepsilon}$ for some $\varepsilon>0$). The assumptions of the bounded parametric space and the existence of the integrable majorants

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

of $r(x, y;\beta)$ and trimmed $s(x, y;\beta)$ can be relaxed only in the case of linear regression (see Manski, 1988) or models and estimators that generate normal equations with a structure very similar to the linear-regression case (note that Assumption F2 can be relaxed in such cases also). This holds, for example, in the censored regression estimated by semiparametrically censored least squares (Powell, 1986) and in the logistic regression (see Gourieroux and Monfort, 1981).

Additionally, the proof of $\sqrt{n}$ consistency requires an unusual regularity assumption, Assumption F4, which is one of the (weak) links between the loss function $s(x, y;\beta)$ and auxiliary trimming function $r(x, y;\beta)$. Considering small intervals around $G^{-1}(\lambda)$, Assumption F4 just expresses the idea that the loss function should not behave "wildly" around the trimming point; that is, the trimmed derivative $s'(x_i, y_i;\beta)$ should be bounded on average for x_i, y_i, and β such that $r(x_i, y_i;\beta)$ is close to $G^{-1}(\lambda)$. To exemplify, let us use a linear regression model with $s(x, y;\beta) = r(x, y;\beta) = (y - x^\top\beta)^2$. Then $s'(x_i, y_i;\beta) = -2(y_i - x_i^\top\beta)x_i$, and the condition $r(x_i, y_i;\beta) \in \mathcal{I}(\beta)$ has the form $|(y_i - x_i^\top\beta)^2 - G^{-1}(\lambda)| \le |r_{[h_n]}(\beta) - G^{-1}(\lambda)|$. Under this condition, the derivative $s'(x_i, y_i;\beta)$ is bounded in absolute value by $\|x_i\|\sqrt{\max\{G^{-1}(\lambda), r_{[h_n]}(\beta)\}}$, which converges to $\|x_i\|\sqrt{\max\{G^{-1}(\lambda), G_\beta^{-1}(\lambda)\}}$ as $n \to +\infty$ uniformly in $\beta \in U(\beta^0, \delta)$ (see Lemma A.2 in Appendix A). Thus for a sufficiently large n, Assumption F4 practically means that expectation $\mathsf{E}\|x_i\|$ is finite.

Finally, we introduce standard identification conditions.

Assumptions I.

I1. B is a compact parametric space.

I2. For any $\varepsilon > 0$ and $U(\beta^0, \varepsilon)$ such that $B \backslash U(\beta^0, \varepsilon)$ is compact, there is $\alpha(\varepsilon) > 0$ such that

$$\min_{\beta \in B\backslash U(\beta^0,\varepsilon)} \mathsf{E}[s(x_i, y_i;\beta)\cdot I(r(x_i, y_i;\beta) \le G_\beta^{-1}(\lambda))]$$

$$- \mathsf{E}[s(x_i, y_i;\beta^0)\cdot I(r(x_i, y_i;\beta^0) \le G_{\beta^0}^{-1}(\lambda))] > \alpha(\varepsilon).$$

I3. For any $n \in \mathbb{N}$, it holds that $\mathsf{E}[s'(x_i, y_i;\beta^0) I(r(x_i, y_i;\beta^0) \le r_{[h_n]}(\beta^0))] = 0$.

The identification assumptions form the second link between the general objective function $s(x, y;\beta)$ and the trimming function $r(x, y;\beta)$. Whereas Assumption I2 formalizes the notion of the trimmed objective function having a global minimum at β^0, Assumption I3 primarily states that the employed trimming does not invalidate the normal equations (see also Lemma 2.1). This can usually be achieved by trimming symmetrically with respect to $s'(x_i, y_i;\beta^0)$; see examples in Section 4.

To close this section, let us note that Assumptions D, F, and I are sufficient to prove the asymptotic normality of GTE. If only consistency is required, one

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

can omit all assumptions concerning the derivatives of the functions $s(x_i, y_i; \beta)$ and $r(x_i, y_i; \beta)$ (Assumptions F), omit Assumption F2 on VC classes, and also weaken Assumption D1, because centered $s(x_i, y_i; \beta)$ can form an $L^{1+\delta}$-mixingale in the most general case (Andrews, 1988).

2.3. Alternative Definition

Before proving the main results of the paper, some basic properties of the GTE objective function $S_n(\beta) = \sum_{j=1}^{h_n} s_{r:[j]}(\beta)$ and its alternative formulation, which is more suitable for deriving asymptotic results, are introduced.

LEMMA 2.1. *Under Assumptions D2 and F1, $S_n(\beta)$ is continuous on B, twice differentiable at $\hat{\beta}_n^{(GTE, h_n)}$ as long as $\hat{\beta}_n^{(GTE, h_n)} \in U(\beta^0, \delta)$, and almost surely twice differentiable at any fixed point $\beta \in U(\beta^0, \delta)$. Furthermore,*

$$S_n^{(l)}(\beta) = \sum_{i=1}^{n} s^{(l)}(x_i, y_i; \beta) \cdot I(r(x_i, y_i; \beta) \leq r_{[h_n]}(\beta)), \tag{8}$$

almost surely for $l = 0$ at any $\beta \in B$ and for $l = 1, 2$ at any $\beta \in U(\beta^0, \delta)$, where $s^{(l)}(x_i, y_i; \beta)$ represents the lth derivative of $s(x_i, y_i; \beta)$ with respect to β.

Proof. See Appendix A.

In general, this definition is not equivalent to the one used in (3) unless all the residuals are different from each other. However, Assumption D2 guarantees this with probability one. Hence, we will use this notation and definition of $S_n(\beta)$ in the rest of the paper.

3. ASYMPTOTIC PROPERTIES

Let us now present the main asymptotic results concerning GTE: its consistency and asymptotic distribution. In all cases, we split the GTE objective function into two parts:

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^{n} s(x_i, y_i; \beta) \cdot I(r(x_i, y_i; \beta) \leq r_{[h_n]}(\beta)) \\ &= \sum_{i=1}^{n} s(x_i, y_i; \beta) \\ &\quad \cdot [I(r(x_i, y_i; \beta) \leq r_{[h_n]}(\beta)) - I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda))] \qquad (9) \\ &\quad + \sum_{i=1}^{n} s(x_i, y_i; \beta) \cdot I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)). \qquad (10) \end{aligned}$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Whereas the first term (9) on the right-hand side will be shown to be small because of the convergence of order statistics to quantiles, $r_{[h_n]}(\beta) \to G_\beta^{-1}(\lambda)$, the second term (10) on the right-hand side will be dealt with by standard asymptotic tools and shown to converge to

$$S(\beta) = \mathsf{E}\{s(x_i, y_i;\beta)\cdot I(r(x_i, y_i;\beta) \leq G_\beta^{-1}(\lambda))\}.$$

First, using the uniform law of large numbers, we prove the consistency of the GTE $\hat{\beta}_n^{(GTE,h_n)}$ minimizing $S_n(\beta)$ on the parametric space B.

THEOREM 3.1. *Let $s(x_i, y_i;\beta)$ and $r(x_i, y_i;\beta)$ be continuous functions on B as specified in Assumption F1 and let Assumptions D, F3, and I hold. Then the GTE $\hat{\beta}_n^{(GTE,h_n)}$ is weakly consistent, that is, $\hat{\beta}_n^{(GTE,h_n)} \to \beta^0$ in probability as $n \to +\infty$.*

Proof. See Appendix B.

Next, the asymptotic distribution of GTE is of interest. To derive it, one has to study the behavior of the normal equations in a neighborhood of β^0 and to prove their asymptotic linearity. More specifically, we analyze the asymptotic behavior of $S_n'(\beta^0 - n^{-1/2}t) - S_n'(\beta^0)$ as a function of $t, \|t\| \leq M = \text{const.}$ (Lemma A.7 in Appendix A). Once the $\sqrt{n}$ consistency of GTE is proved (Lemma B.1 in Appendix B), the asymptotic-linearity result can be applied to the GTE estimates because $\hat{\beta}_n^{(GTE,h_n)} = \beta^0 - n^{-1/2}t, \|t\| \leq M$, with probability arbitrarily close to one. The application of the central limit theorem results then in the asymptotic normality of GTE.

THEOREM 3.2. *Let Assumptions D, F, and I hold. Then the GTE $\hat{\beta}_n^{(GTE,h_n)}$ is asymptotically normal, that is, $\sqrt{n}(\hat{\beta}_n^{(GTE,h_n)} - \beta^0) \xrightarrow{F} N(0, V(\lambda))$ as $n \to +\infty$, where*

$$V(\lambda) = \{Q_s(\lambda) + J_s(\lambda)\}^{-1}\cdot V_s(\lambda)\cdot\{Q_s(\lambda) + J_s(\lambda)\}^{-1}.$$

Proof. See Appendix B.

Although we proved the asymptotic normality of GTE, let us note that the use of the derived formula for the asymptotic variance of $\hat{\beta}_n^{(GTE,h_n)}$ is relatively limited because of matrix $J_s(\lambda)$, which is difficult to estimate. If the dependent variable conditionally on the explanatory variables is continuously distributed, for instance, which is the case of (non)linear or truncated regression, it is possible to derive a more specific of form of $J_s(\lambda)$ (Čížek, 2006a). Specifically, the expected value of the indicator function in $J_s(\lambda)$ (see eqn. (6)), is asymptotically linear in β and proportional to $g\{G^{-1}(\lambda)\}(\beta - \beta^0)$ (Čížek, 2004b, Lem. A.8). In such cases, one can then estimate the asymptotic variance $V(\lambda)$ even though it relies on a nonparametric estimate of the prob-

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

ability density function of regression residuals. In other cases, such as binary-choice regression (Sect. 4.2), $J_s(\lambda)$ can be expressed as a nontrivial function of the joint probability distribution of (y_i, x_i), which does not facilitate a practical computation.

In general, the estimation of the GTE asymptotic variance $V(\lambda)$ therefore has to be done by bootstrap. Theoretically, bootstrap can be used for GTE in the same situations as for the original nontrimmed estimator. There are, however, two issues that have to be accounted for. First, to preserve the robust properties of GTE also in the case of variance estimation, a weighted bootstrap has to be used to prevent bootstrap samples containing a large share of contaminated observations (Salibian-Barrera and Zamar, 2002) unless a parametric bootstrap can be employed. Second, the brute force application of the bootstrap principle to a trimmed estimator would be highly computationally demanding, and algorithms for GTE and bootstrap have to be integrated to achieve fast computation (Willems and van Aelst, 2005).

4. EXAMPLES OF TRIMMED ESTIMATORS

In this section, we discuss some trimmed estimators and models where they can be applied. To verify their feasibility, we check the identification Assumptions I2 and I3, as discussed in Section 4.1. Later, we present examples of trimmed estimators based on the likelihood function in nonlinear and binary-response regression (Sect. 4.2).

4.1. Identification Condition

A crucial ingredient of the consistency and asymptotic normality of GTE is the identification Assumptions I2 and I3, which differ from the usual least squares or maximum likelihood identification conditions by inclusion of trimming. The identification Assumption I2 can be formulated so that

$$IC(\beta) = \mathsf{E}[s(x_i, y_i; \beta) \cdot I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda))] \tag{11}$$

as a function of β has a unique minimum at β^0. Assumption I3 just means that

$$FOC(\beta^0) = \mathsf{E}[s'(x_i, y_i; \beta^0) \cdot I(r(x_i, y_i; \beta^0) \leq r_{[h_n]}(\beta^0))] = 0. \tag{12}$$

Whereas Assumption I3 can be checked by a straightforward evaluation of (12), Assumption I2 is more difficult to verify because of its global character. Let us therefore note that, if $s(x, y; \beta) \equiv r(x, y; \beta)$, proving Assumption I2 amounts to proving that $r(x_i, y_i; \beta)$ stochastically dominates $r(x_i, y_i; \beta^0)$ for any $\beta \neq \beta^0$. In the case of first-order stochastic dominance, this means that

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

$G_\beta(z) \leq G_{\beta^0}(z)$ for all $z \in \mathbb{R}$, where $G_\beta(z)$ represents the distribution function of $r(x_i, y_i; \beta), \beta \in B$. This follows from

$$IC(\beta) = \int_{-\infty}^{G_\beta^{-1}(\lambda)} r dG_\beta(r) = \int_0^\lambda G_\beta^{-1}(t)\, dt$$

and the fact that the stochastic dominance implies $G_\beta^{-1}(t) \geq G_{\beta^0}^{-1}(t)$ for any $t \in \langle 0,1 \rangle$. To guarantee that the minimum of $IC(\beta)$ at β^0 is unique, the stochastic dominance has to be strict for any $\beta \neq \beta^0$ at one or more points $z < G_{\beta^0}^{-1}(\lambda)$ (or $t < \lambda$).

For example, consider the nonlinear regression model (1) with ε_i being symmetrically distributed around zero and independent of x_i and the LTS estimator defined by $s(x, y; \beta) = r(x, y; \beta) = \{y - m(x, \beta)\}^2$. One can then easily see that condition (12), $\mathsf{E}[\varepsilon_i x_i I(\varepsilon_i^2 \leq \varepsilon_{[h_n]}^2)] = 0$, is satisfied because of the symmetry of ε_i distribution and that (11) is minimized at β^0 because $s(x_i, y_i; \beta) = \{y_i - m(x_i, \beta)\}^2 = \{\varepsilon_i + [m(x_i, \beta^0) - m(x_i, \beta)]\}^2$ first-order stochastically dominates $s(x_i, y_i; \beta^0) = \varepsilon_i^2$. For details, see the paper by Čížek (2007), who not only verifies these identification conditions but also proposes least-squares-based GTE for truncated and censored regression models.

4.2. Maximum Trimmed Likelihood

Our main example concerns GTE based on the likelihood function, which in (non)linear regression coincides with MTLE. After mentioning briefly its identification in nonlinear regression, we focus on an example where standard MTLE does not apply but it is possible to construct a likelihood-based GTE: binary-choice regression. Other applications such as general trimmed estimation in truncated and censored regression are discussed in Čížek (2007). Note that we assume here for simplicity that data are i.i.d.

The MTLE in nonlinear regression model (1) is a special case of GTE for $r(x, y; \beta) = s(x, y; \beta) = -\ln \phi\{y - m(x, \beta)\}$, where ϕ denotes the density function of ε_i. Assumptions I2 and I3 can be verified in a similar way as for LTS in Section 4.1 because functions $r(x, y; \beta)$ and $s(x, y; \beta)$ are identical (this time using the fact that the likelihood function has a minimum at β^0). The most important additional assumption is again the (conditional) symmetry of the ε_i distribution, which implies that introducing "trimming" into the identification conditions does not invalidate them. For example, under conditional symmetry of $\phi(\varepsilon_i)$ given x_i, $\mathsf{E}[\phi'(\varepsilon_i)/\phi(\varepsilon_i)|x_i] = 0$ implies $\mathsf{E}[\phi'(\varepsilon_i)/\phi(\varepsilon_i) \cdot I(-\ln \phi(\varepsilon_i) \leq \{-\ln \phi(\varepsilon_i)\}_{[h_n]})|x_i] = 0$.

Applying the general trimmed estimation concept to the maximum likelihood estimation becomes less trivial once we consider discrete models, such as binary-choice models. In this case, the dependent variable takes on only two values, $y_i \in \{0, 1\}$, and its conditional expectation is described by $\mathsf{E}(y_i|x_i) = P(y_i = 1|x_i) = \Phi(x_i^\top \beta)$, where Φ is a symmetric absolutely continuous distri-

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

bution function with a differentiable density ϕ (e.g., Φ is the standard normal distribution function in the case of probit). The log-likelihood contribution is then described by

$$s(x_i, y_i; \beta) = -\ln l(x_i, y_i; \beta) = -y_i \ln \Phi(x_i^\top \beta) - (1 - y_i) \ln\{1 - \Phi(x_i^\top \beta)\}.$$

The MTLE, which sets $r(x, y; \beta) = s(x, y; \beta)$, cannot be applied here because by trimming unlikely observations, such as (y_i, x_i) with $y_i = 1$ and $\Phi(x_i^\top \beta)$ close to zero, MTLE induces separation of nontrimmed data with $y_i = 1$ and with $y_i = 0$ in the space of explanatory variables. Consequently, the maximum trimmed likelihood estimate is not identified (Albert and Anderson, 1984). Moreover, the identification condition (12) is not satisfied for $\lambda < 1$:

$$\begin{aligned} FOC(\beta^0) &= \mathsf{E}[-\{\ln l(x_i, y_i; \beta^0)\}' I(r(x_i, y_i; \beta^0) \le r_{[h_n]}(\beta^0))] \\ &= \mathsf{E}\left[\left\{ -\frac{y_i \phi(x_i^\top \beta^0)}{\Phi(x_i^\top \beta^0)} x_i + \frac{(1 - y_i)\phi(x_i^\top \beta^0)}{1 - \Phi(x_i^\top \beta^0)} x_i \right\} \right. \\ &\qquad \left. \cdot I(r(x_i, y_i; \beta^0) \le r_{[h_n]}(\beta^0)) \right] \qquad (13) \\ &= \mathsf{E}\left[-P(y_i = 1 | x_i) \frac{\phi(x_i^\top \beta^0)}{\Phi(x_i^\top \beta^0)} x_i I(r(x_i, 1; \beta^0) \le r_{[h_n]}(\beta^0)) \right] \qquad (14) \\ &\quad + \mathsf{E}\left[+P(y_i = 0 | x_i) \frac{\phi(x_i^\top \beta^0)}{1 - \Phi(x_i^\top \beta^0)} x_i I(r(x_i, 0; \beta^0) \le r_{[h_n]}(\beta^0)) \right] \qquad (15) \\ &= \mathsf{E}\{\phi(x_i^\top \beta^0) x_i [I(r(x_i, 0; \beta^0) \le r_{[h_n]}(\beta^0)) - I(r(x_i, 1; \beta^0) \le r_{[h_n]}(\beta^0))]\} \\ &\qquad (16) \end{aligned}$$

equals in general zero only if for all possible values of the random vector x_i

$$I(-\ln \Phi(x_i^\top \beta^0) \le r_{[h_n]}(\beta^0)) = I(-\ln\{1 - \Phi(x_i^\top \beta^0)\} \le r_{[h_n]}(\beta^0)) \qquad (17)$$

(after we substituted for $r(x_i, y_i; \beta^0)$); that is, if no trimming occurs, $h_n = n$ and $\lambda = 1$.[5]

On the other hand, this derivation hints that the identification condition would hold if the trimming function $r(x, y; \beta)$ satisfies $r(x, 0; \beta) = r(x, 1; \beta)$; see (14)–(16). Therefore, we propose to set $r(x_i, y_i; \beta) = r(x_i; \beta) = \max\{-\ln \Phi(x^\top \beta), -\ln[1 - \Phi(x^\top \beta)]\}$, for instance, and use GTE minimizing

$$\sum_{j=1}^{h} - \ln l(x_i, y_i; \beta) \cdot I\left(r(x_i; \beta) = -\max_{y \in \{0, 1\}} \ln l(x_i, y; \beta) \le r_{[h_n]}(\beta) \right). \qquad (18)$$

This GTE, which can be also called the maximum symmetrically trimmed likelihood estimator, provides an example of a model and robust method where the breakdown point is not maximized for $\lambda = \frac{1}{2}$. Čížek (2006b) showed that the

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

trimming constant λ maximizing the breakdown point of GTE has to satisfy $\lambda \geq \frac{2}{3}$ in this case.

For $r(x_i;\beta)$ used in (18), the first-order condition (12) is obviously satisfied (see (14)–(16)), and it remains to verify the global identification Assumption I2. Because at least one explanatory variable, let us say x_1, has to be continuously distributed by Assumption D2, we first verify that

$$\begin{aligned} IC(\beta) &= \mathsf{E}\{[-y_i \ln \Phi(x_i^\top\beta) - (1-y_i)\ln\{1-\Phi(x_i^\top\beta)\}] \\ &\qquad \cdot I(r(x_i,\beta) \leq G_\beta^{-1}(\lambda))\} \\ &= \mathsf{E}\{[-P(y_i=1|x_i)\ln \Phi(x_i^\top\beta) - P(y_i=0|x_i)\ln\{1-\Phi(x_i^\top\beta)\}] \\ &\qquad \cdot I(r(x_i,\beta) \leq G_\beta^{-1}(\lambda))\} \\ &= \mathsf{E}\{[-\Phi(x_i^\top\beta^0)\ln \Phi(x_i^\top\beta) - \{1-\Phi(x_i^\top\beta^0)\}\ln\{1-\Phi(x_i^\top\beta)\}] \\ &\qquad \cdot I(r(x_i,\beta) \leq G_\beta^{-1}(\lambda))\} \end{aligned} \tag{19}$$

has a minimum at β^0 conditionally on all other explanatory variables x_{i2}, $\ldots, x_{ip-1}$ (x_0 being the intercept). This will then imply the unconditional inequality in Assumption I2.

To prove it, note that function $-\Phi(t_0)\ln \Phi(t) - \{1-\Phi(t_0)\}\ln\{1-\Phi(t)\}$, where $t_0 \in \mathbb{R}$ is a constant and $t \in \mathbb{R}$, has a unique minimum at $t = t_0$ (property P1). This follows from the corresponding first-order condition

$$-\Phi(t_0)\phi(t)/\Phi(t) + \{1-\Phi(t_0)\}\phi(t)/\{1-\Phi(t)\} = 0,$$

providing that $\Phi(t)$ is strictly increasing and $-\ln \Phi(t)$ and $-\ln\{1-\Phi(t)\}$ are convex (these are sufficient conditions for the existence and uniqueness of a maximum likelihood estimate, see Silvapulle, 1981).

Now, keeping conditioning on $x_{i2}, \ldots, x_{ip-1}$ implicit, the index $t_i(\beta) = x_i^\top\beta$ consists only of the linear term $x_{i1}\beta_1$ and the constant term $x_i^\top\beta - x_{i1}\beta_1$. Denoting $t_i^0(\beta_1) = x_i^\top(\beta_0^0, \beta_1, \beta_2^0, \ldots, \beta_{p-1}^0)^\top = t_i\{(\beta_0^0, \beta_1, \beta_2^0, \ldots, \beta_{p-1}^0)^\top\}$ and the difference of the two indices by $\delta_i(\beta) = t_i(\beta) - t_i^0(\beta_1)$, property P1 implies that

$$\begin{aligned} &-\Phi\{t_i^0(\beta_1)\}\ln \Phi\{t_i^0(\beta_1) + \delta_i(\beta)\} - [1-\Phi\{t_i^0(\beta_1)\}] \\ &\qquad \cdot \ln[1-\Phi\{t_i^0(\beta_1) + \delta_i(\beta)\}] \end{aligned}$$

is uniquely minimized at $\delta_i(\beta) = 0$ irrespective of the value of index $t_i^0(\beta_1)$. Hence, the expectation $IC(\beta)$ can be at its minimum if and only if $\delta_i(\beta) = 0$. Moreover for any $\beta \in B$ minimizing $IC(\beta)$, the trimming set defined by $T_1(\beta_1) = \{x_1 : r(x_1,\beta_1) = r\{t_i^0(\beta_1)\} \leq G_\beta^{-1}(\lambda)\}$ does not depend on β_1 (property P2) because $P\{T_1(\beta_1)\} = \lambda$ by definition and $t_i^0(\beta_1)$ consists only of a constant independent of β_1 and the linear term $x_{i1}\beta_1$.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Moreover, property P1 further indicates that

$$-\Phi\{t_i^0(\beta_1^0)\}\ln\Phi\{t_i^0(\beta_1)\} - [1-\Phi\{t_i^0(\beta_1^0)\}]\ln[1-\Phi\{t_i^0(\beta_1)\}]$$

achieves its unique minimum at $\beta_1 = \beta_1^0$ for any value of x_1 (identity $t_i^0(\beta_1) = t_i^0(\beta_1^0)$ implies $\beta_1 = \beta_1^0$ because $t_i^0(\beta_1)$ is a linear function of β_1 and x_1). Together with property P2, it follows from (19) that (still implicitly conditioning on $x_{i2}, \ldots, x_{ip-1}$)

$$\begin{aligned} IC(\beta) &\geq \mathsf{E}\{[-\Phi(x_i^\top\beta^0)\ln\Phi(x_i^\top\beta^0) - \{1-\Phi(x_i^\top\beta^0)\}\ln\{1-\Phi(x_i^\top\beta^0)\}] \\ &\qquad \cdot I(r(x_i,\beta) \leq G_\beta^{-1}(\lambda))\} \\ &= \mathsf{E}\{[-\Phi(x_i^\top\beta^0)\ln\Phi(x_i^\top\beta^0) - \{1-\Phi(x_i^\top\beta^0)\}\ln\{1-\Phi(x_i^\top\beta^0)\}] \\ &\qquad \cdot I(r(x_i,\beta^0) \leq G^{-1}(\lambda))\} \\ &= IC(\beta^0). \end{aligned}$$

Hence, to verify the global identification Assumption I2 unconditionally, one only has to show that $\delta_i(\beta) = \delta_i(\beta^0) = 0$ almost surely if and only if $\beta_0 = \beta_0^0, \beta_2 = \beta_2^0, \ldots, \beta_{p-1} = \beta_{p-1}^0$. This can be however guaranteed by the usual full-rank condition $\mathsf{E}[x_i x_i^\top] > 0$, for instance.

5. CONCLUSION

Motivated by LTS, LTA, and MTLE, we proposed the GTE, which extends the applicability of high-breakdown-point methods to a wide range of econometric models, including nonlinear regression, time series, and limited dependent variable models. GTE can be combined with many parametric estimation methods, and it adds to them a protection against data contamination. The following conclusions concern robust properties of GTE, its extensions, and its use in applications.

Although we proved the consistency and derived the asymptotic distribution under rather general conditions, the choice of trimming and robust properties of GTE are rather specific to particular models. Such questions are currently addressed only in (generalized) linear models and binary-choice regression, and there are many areas for future research. In particular, they include GTE based on estimation methods modifying the error distribution, such as symmetrically trimmed least squares (Powell, 1986), and GTE in models including some form of time dependence, such as panel data and time series (see Genton and Lucas, 2003).

Furthermore, we discussed only the most basic form of trimmed estimation, where observations are either included in or excluded from the GTE objective function. Nevertheless, various weighted trimmed estimators and data-adaptive choice of trimming, only recently introduced for LTS and MTLE, are straightforward to apply. In both cases, further research on the choice of the trimming

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

constant or weight function is necessary for practically any model outside of the generalized linear model class.

Finally, we argued that computational and finite-sample properties of GTE could be in many cases analogous to existing results concerning LTS, LTA, and MTLE. On the other hand, most existing robust estimators are studied and applied in the context of location or linear regression models, whereas possible applications of GTE also involve rather complex nonlinear models. Hence, simulation studies have to be employed to learn more about finite-sample behavior of GTE under different circumstances. Last, but not least, existing algorithms for evaluating trimmed estimators have to be adapted to many different models and integrated with a bootstrap procedure for variance estimation.

NOTES

1. There are also other concepts and measures of robustness; see Hampel, Ronchetti, Rousseeuw, and Stahel (1986) for an overview.

2. A significant improvement of LMS by smoothing its objective function was recently proposed by Zinde-Walsh (2002).

3. The assumption $x_i \in \mathbb{R}^k$ and $y_i \in \mathbb{R}$ introduced here corresponds to the most traditional use in regression models, but the presented results are valid also for $y_i \in \mathbb{R}^l$ and general multivariate models.

4. Theoretically, it is possible to also consider the case $h_n/n \to \lambda = 0$. Although it can also lead to consistent estimators, the convergence rate would be below the $\sqrt{n}$ rate characterizing the cases with $\lambda > 0$ and the presented proofs would not apply.

5. We neglect the other "solution," $\lambda = 0$, which results in objective function constantly equal to zero.

REFERENCES

Agulló, J. (2001) New algorithms for computing the least trimmed squares regression estimator. *Computational Statistics & Data Analysis* 36, 425–439.

Albert, A. & J.A. Anderson (1984) On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 71, 1–10.

Andrews, D.W.K. (1988) Laws of large numbers for dependent non-identically distributed random variables. *Econometric Theory* 4, 458–467.

Andrews, D.W.K. (1992) Generic uniform convergence. *Econometric Theory* 8, 241–257.

Andrews, D.W.K. (1993) An introduction to econometric applications of empirical process theory for dependent random variables. *Econometric Reviews* 12, 183–216.

Arcones, M.A. & B. Yu (1994) Central limit theorems for empirical and *U*-processes of stationary mixing sequences. *Journal of Theoretical Probability* 7, 47–71.

Bassett, G.W. (1991) Equivariant, monotonic, 50% breakdown estimators. *American Statistician* 45, 135–137.

Čížek, P. (2002) Robust estimation with discrete explanatory variables. In W. Härdle & B. Rönz (eds.), *COMPSTAT 2002—Proceedings in Computational Statistics*, pp. 509–514. Springer-Verlag.

Čížek, P. (2004a) Smoothed local L-estimation with an application. In M. Hubert, G. Pison, A. Struyf, & S. Aelst (eds.), *Theory and Applications of Recent Robust Methods*, pp. 59–70. Birhaüser.

Čížek, P. (2004b) Asymptotics of Least Trimmed Squares Regression. CentER Discussion paper 72/2004, Tilburg University.

Čížek, P. (2006a) Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference* 136, 3967–3988.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Čížek, P. (2006b) Trimmed likelihood-based estimation in binary regression models. *Austrian Journal of Statistics* 35, 223–232.
Čížek, P. (2007) General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models. CentER Discussion paper 65/2007, Tilburg University.
Čížek, P. & J.Á. Víšek (2000) Least trimmed squares. In W. Härdle, Z. Hlávka, & S. Klinke (eds.), *XploRe Application Guide*, pp. 49–64. Springer-Verlag.
Davidson, J. (1994) *Stochastic Limit Theory*. Oxford University Press.
Doukhan, P., P. Massart, & E. Rio (1995) Invariance principles for absolutely regular empirical processes. *Annales de l'Institut Henri Poincaré* 31, 393–427.
Genton, M.G. & A. Lucas (2003) Comprehensive definitions of breakdown points for independent and dependent observations. *Journal of the Royal Statistical Society, Series B* 65, 81–94.
Gerfin, M. (1996) Parametric and semi-parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometrics* 11, 321–339.
Gervini, D. & V.J. Yohai (2002) A class of robust and fully efficient regression estimators. *Annals of Statistics* 30, 583–616.
Gilloni, A. & M. Padberg (2002) Least trimmed squares regression, least median squares regression, and mathematical programming. *Mathematical and Computer Modelling* 35, 1043–1060.
Gourieroux, C. & A. Monfort (1981) Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics* 17, 83–97.
Hadi, A. & A. Luceno (1997) Maximum trimmed likelihood estimators: A unified approach, examples and algorithms. *Computational Statistics & Data Analysis* 25, 251–272.
Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, & W.A. Stahel (1986) *Robust Statistics, the Approach Based on Influence Function*. Wiley.
Hawkins, D.M. & D. Olive (1999) Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis* 32, 119–134.
Hössjer, O. (1994) Rank-based estimates in the linear model with high breakdown point. *Journal of the American Statistical Association* 89, 149–158.
Kelly, M. (1997) Do noise traders influence stock prices? *Journal of Money, Credit and Banking* 29, 351–363.
Knez, P.J. & M.J. Ready (1997) On the robustness of size and book-to-market in cross-sectional regressions. *Journal of Finance* 52, 1355–1382.
Manski, C.F. (1988) *Analog Estimation Methods in Econometrics*. Chapman and Hall.
Müller, C.H. & N.K. Neykov (2003) Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *Journal of Statistical Planning and Inference* 116, 503–519.
Neykov, N.M. & P.N. Neytchev (1990) A robust alternative of the maximum likelihood estimator. *Short Communications of COMPSTAT, Dubrovnik 1990*, 99–100.
Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.
Pollard, D. (1984) *Convergence of Stochastic Processes*. Springer-Verlag.
Powell, J.L. (1986) Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* 54, 1435–1460.
Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association* 79, 871–880.
Rousseeuw, P.J. (1985) Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze, & W. Wertz (eds.), *Mathematical Statistics and Applications, vol. B*, pp. 283–297. Reidel.
Rousseeuw, P.J. (1997) Introduction to positive-breakdown methods. In G.S. Maddala & C.R. Rao (eds.), *Handbook of Statistics, vol. 15: Robust Inference*. pp. 101–121. Elsevier.
Rousseeuw, P.J. & A.M. Leroy (2003) *Robust Regression and Outlier Detection*. Wiley.
Sakata, S. & H. White (1998) High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. *Econometrica* 66, 529–567.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Salibian-Barrera, M. & R.H. Zamar (2002) Bootstrapping robust estimates of regression. *Annals of Statistics* 30, 556–582.
Silvapulle, M.J. (1981) On the existence of maximum likelihood estimates for the binomial response models. *Journal of the Royal Statistical Society Series B* 43, 310–313.
Stromberg, A.J. (1993) High breakdown estimation of nonlinear regression parameters. *Journal of the American Statistical Association* 88, 237–244.
Stromberg, A.J. & D. Ruppert (1992) Breakdown in nonlinear regression. *Journal of the American Statistical Association* 87, 991–997.
Tableman, M. (1994) The asymptotics of the least trimmed absolute deviations (LTAD) estimator. *Statistics and Probability Letters* 19, 329–337.
Temple, J.R.W. (1998) Robustness tests of the augmented Solow model. *Journal of Applied Econometrics* 13, 361–375.
Van der Vaart, A.W. & J.A. Wellner (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag.
Vandev, D.L. & N.M. Neykov (1993) Robust maximum likelihood in the Gaussian case. In S. Morgenthaler, E. Ronchetti, & W.A. Stahel (eds.), *New Directions in Data Analysis and Robustness*, pp. 259–264. Birkhäuser.
Vandev, D.L. & N.M. Neykov (1998) About regression estimators with high breakdown point. *Statistics* 32, 111–129.
Víšek, J.Á. (2000) On the diversity of estimates. *Computational Statistics & Data Analysis* 34, 67–89.
Víšek, J.Á. (2002) The least weighted squares II. Consistency and asymptotic normality. *Bulletin of the Czech Econometric Society* 16, 1–28.
Willems, G. & S. van Aelst (2005) Fast and robust bootstrap for LTS. *Computational Statistics & Data Analysis* 48, 703–715.
Zaman, A., P.J. Rousseeuw, & M. Orhan (2001) Econometric applications of high-breakdown robust regression techniques. *Economics Letters* 71, 1–8.
Zinde-Walsh, V. (2002) Asymptotic theory for some high breakdown point estimators. *Econometric Theory* 18, 1172–1196.

APPENDIXES

The Appendixes present the proofs of lemmas and theorems on the order statistics of the trimming function and on the GTE objective function (Appendix A) and on the consistency and asymptotic normality of GTE (Appendix B). Note that the alternative definition (8) of GTE is employed in all proofs. Additionally, the following notation is used: the probability space, on which $\{x_i, y_i\}$ is defined, is denoted Ω; the loss and trimming functions are written as $s_i(\beta) = s(x_i, y_i; \beta)$ and $r_i(\beta) = r(x_i, y_i; \beta)$, respectively; and the asymptotic counterpart of the objective function $S_{nn}(\beta) = S_n(\beta)/n$ is referred to as $S(\beta) = \mathsf{E}\{s(x_i, y_i; \beta) \cdot I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda))\}$, and the same applies to the respective derivatives. Finally, because the behavior of the indicators $I(r(x_i, y_i; \beta) \leq r_{[h_n]}(\beta))$, $I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda))$, and their differences will be extensively studied, we define

$$\iota_{in}(\beta, K) = I(r(x_i, y_i; \beta) \leq r_{[h_n]}(\beta) + K),$$

$$\nu_{1i}(\beta, K) = I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda) + K), \qquad \nu_{2i}(\beta, K) = I(r(x_i, y_i; \beta) \geq G_\beta^{-1}(\lambda) - K),$$

$$\nu_{1i}(\beta) \equiv \nu_{1i}(\beta, 0), \qquad \nu_{2i}(\beta) \equiv \nu_{2i}(\beta, 0), \quad \text{and} \quad \iota_{in}(\beta) \equiv \iota_{in}(\beta, 0), \quad \text{and}$$

$$\nu_{in}(\beta) = I(r(x_i, y_i; \beta) \leq r_{[h_n]}(\beta)) - I(r(x_i, y_i; \beta) \leq G_\beta^{-1}(\lambda)).$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

APPENDIX A: Lemmas on Order Statistics and GTE Objective Function

Proof of Lemma 2.1. For a given sample size n, let us consider a fixed realization $\omega \in \Omega^n$. The objective function $S_n(\beta)$ at a particular point $\beta \in B$ equals to one of functions $T_1(\beta), \ldots, T_l(\beta)$, where $T_j(\beta) = \sum_{i=1}^{h_n} s_{k_{ji}}(\beta)$, $j = 1, \ldots, l = \binom{n}{h_n}$, and $\{k_{j1}, \ldots, k_{jh_n}\} \in \{1, \ldots, n\}^{h_n}$ are sets of h_n indices selecting observations from the sample. Each function $T_j(\beta)$ is uniformly continuous on B and twice differentiable in a neighborhood $U(\beta^0, \delta)$. There are two cases to discuss:

1. If one can find an index j and a neighborhood $U(\beta, \varepsilon)$ such that $S_n(\beta) = T_j(\beta)$ for all $\beta \in U(\beta, \varepsilon)$, $S_n(\beta)$ is continuous at β. Additionally, if $\beta \in U(\beta^0, \delta)$ there is a neighborhood $U(\beta, \varepsilon') \subset U(\beta^0, \delta)$ for some $\varepsilon' < \varepsilon$, and $S_n(\beta) = T_j(\beta)$ is even twice differentiable at β (almost surely).
2. In all other cases, β lies on a boundary in the sense that there are some $j_1, \ldots, j_m$ such that $S_n(\beta) = T_{j_1}(\beta) = \cdots = T_{j_m}(\beta)$. Because $S_n(\beta) = T_{j_1}(\beta) = \cdots = T_{j_m}(\beta)$ and all functions $T_{j_i}, i = 1, \ldots, m$, are continuous at β, $S_n(\beta)$ is continuous at β also. Furthermore, $S_n(\beta)$ is also differentiable provided that $T'_{j_1}(\beta) = \cdots = T'_{j_m}(\beta)$ and $\beta \in U(\beta^0, \delta)$. This condition is always satisfied at $\hat{\beta}_n^{(GTE, h_n)} \in U(\beta^0, \delta)$ as $T'_{j_1}(\hat{\beta}_n^{(GTE, h_n)}) = \cdots = T'_{j_m}(\hat{\beta}_n^{(GTE, h_n)}) = 0$; otherwise, $\hat{\beta}_n^{(GTE, h_n)}$ would not minimize $S_n(\beta)$.

Now, consider a fixed $\beta \in U(\beta^0, \delta)$ (n is still fixed). Assumption D2 implies that

$$P(\Omega_0 = \{\omega \in \Omega^n | \exists i, \quad j \in \{1, \ldots, n\}, \quad i \neq j, \quad \text{such that } r(\beta, \omega) = r_j(\beta, \omega)\}) = 0.$$

Moreover, there is a $\delta' > 0$ such that $r(x_i, y_i; \beta)$ is continuous on $\bar{U}(\beta, \delta')$, and therefore, it is also uniformly continuous on $\bar{U}(\beta, \delta')$, $i = 1, \ldots, n$. Therefore, for any given $\omega \notin \Omega_0$ and $\kappa(\omega) = \frac{1}{2}\min_{i,j=1,\ldots,n; i \neq j} |r_i(\beta, \omega) - r_j(\beta, \omega)| > 0$ we can find an $\varepsilon(\omega) > 0$ such that it holds that $\sup_{\beta' \in U(\beta, \varepsilon(\omega))} |r_i(\beta') - r_i(\beta)| < \kappa(\omega)$ for all $i = 1, \ldots, n$. Consequently, the ordering of $r_1(\beta), \ldots, r_n(\beta)$ is constant for all $\beta' \in U(\beta, \varepsilon(\omega))$, and there exist j such that $S_n(\beta) = T_j(\beta)$ almost surely as stated in point 1 ($P(\Omega \backslash \Omega_0) = 1$). Thus, $S_n(\beta)$ is twice differentiable at β almost surely.

Finally, the lemma directly follows from the two derived results: there are almost surely no i and j such that $r_i(\beta) = r_j(\beta)$ at any $\beta \in B$ and any fixed $n \in \mathbb{N}$, and $S_n(\beta)$ is almost surely twice differentiable at any $\beta \in U(\beta^0, \delta)$. ■

The next lemma just verifies that the uniform law of large numbers is applicable for trimmed sums.

LEMMA A.1. *Let Assumptions D, F1, and I1 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, for some $R \subseteq \mathbb{R}$, assume that $t_i(\beta) \equiv t(x_i, y_i; \beta)$ has an integrable majorant after trimming: (i)* $\mathsf{E} \sup_{\beta \in B, K \in R} |t_i(\beta) \nu_{1i}(\beta, K)| < \infty$, *or alternatively, (ii)* $\mathsf{E} \sup_{\beta \in B, K \in R} \sup_{n \geq n_0} |t_i(\beta) \iota_{in}(\beta, K)| < \infty$ *for some $n_0 \in \mathbb{N}$. Then*

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

$$\sup_{\beta\in B,\, K\in R}\left|\frac{1}{n}\sum_{i=1}^{n}[t_i(\beta)\nu_{1i}(\beta,K)]-\mathsf{E}[t_i(\beta)\nu_{1i}(\beta,K)]\right|\to 0$$

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

Proof. See Čížek (2007, Lem. A.1).

The following assertions present some fundamental properties of order statistics of regression residuals.

LEMMA A.2. *Let* $\lambda\in(0,1\rangle$ *and put* $h_n=[\lambda n]$ *for* $n\in\mathbb{N}$. *Under Assumptions D, F1, F3, and I1, it holds as* $n\to+\infty$ *that* $\sup_{\beta\in B}|r_{[h_n]}(\beta)-G_\beta^{-1}(\lambda)|\to 0$ *in probability and that*

$$E_{Gn}=\mathsf{E}\sup_{\beta\in B}|r_{[h_n]}(\beta)-G_\beta^{-1}(\lambda)|\to 0. \tag{A.1}$$

Proof. Let us recall that $r_i(\beta)\sim G_\beta$. Further, let us take an arbitrary $K_1>0$, set $K_\varepsilon=K_1 m_{gg}$ (see Assumption D3 for the definition of m_{gg}), and consider some $\varepsilon\in(0,1)$. For any $\varepsilon>0$, we will now find $n_0\in\mathbb{N}$ such that $P(\sup_{\beta\in B}|r_{[h_n]}(\beta)-G_\beta^{-1}(\lambda)|>K_1)<\varepsilon$ for all $n>n_0$. Without loss of generality, we can assume that $K_1<\delta_g$, where δ_g comes from Assumption D3.

First, note that for any $\beta\in B$ and and $K_1>0$

$$\mathsf{E}\nu_{1i}(\beta,K_1)=P(\nu_{1i}(\beta,K_1)=1)=P(r_i(\beta)\le G_\beta^{-1}(\lambda)+K_1)>\lambda.$$

Further, Lemma A.1 for $t(x,y;\beta)=1$ guarantees that we can use the weak law of large numbers for $\nu_{1i}(\beta,K_1)$ uniformly on $B\times\mathbb{R}_+$. Hence,

$$\sup_{\beta\in B,\,K_1\in\mathbb{R}_+}\left|\frac{1}{n}\sum_{i=1}^{n}\{\nu_{1i}(\beta,K_1)-\mathsf{E}\nu_{1i}(\beta,K_1)\}\right|\to 0$$

in probability. Consequently, we can find some n_0 such that it holds for all $n>n_0$ that

$$P\left(\sup_{\beta\in B,\,K_1\in\mathbb{R}_+}\left|\frac{1}{n}\sum_{i=1}^{n}\{\nu_{1i}(\beta,K_1)-\mathsf{E}\nu_{1i}(\beta,K_1)\}\right|\le\frac{1}{2}K_\varepsilon\right)>1-\frac{\varepsilon}{2}.$$

Thus, it holds uniformly in β and K_1 with probability greater or equal to $1-\varepsilon/2$ that

$$-\frac{1}{2}K_\varepsilon+\sum_{i=1}^{n}\mathsf{E}\nu_{1i}(\beta,K_1)\le\sum_{i=1}^{n}\nu_{1i}(\beta,K_1). \tag{A.2}$$

Second, because $K_1<\delta_g$, Assumption D3 implies $\mathsf{E}\nu_{1i}(\beta,K_1)>\lambda+K_1 m_{gg}=\lambda+K_\varepsilon$ for all $\beta\in B$ and $K_1<\delta_g$. This result together with equation (A.2) implies that for all $\beta\in B$

$$n\lambda+\left(n-\frac{1}{2}\right)K_\varepsilon=-\frac{1}{2}K_\varepsilon+n(\lambda+K_\varepsilon)<-\frac{1}{2}K_\varepsilon+\sum_{i=1}^{n}\mathsf{E}\nu_{1i}(\beta,K_1)\le\sum_{i=1}^{n}\nu_{1i}(\beta,K_1),$$

which means that at least $n\lambda\ge h_n$ of values $r_i(\beta)$ are smaller than $G_\beta^{-1}(\lambda)+K_1$. In other words, $r_{[h_n]}(\beta)\le G_\beta^{-1}(\lambda)+K_1$ with probability at least $1-\varepsilon/2$.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

The corresponding lower inequality, holding also with probability at least $1 - \varepsilon/2$, can be found by repeating these steps for $\nu_{2i}(\beta, K_1)$. Combining these two inequalities results in the first claim of the lemma. Further, because $r_{[h_n]}(\beta)$ is uniformly integrable as a result of Assumption F3 and Davidson (1994, Thm. 12.10), the second claim follows directly from the first one by Davidson (1994, Thm. 18.14): convergence in probability of uniformly integrable random variables implies the convergence in L^p-norm. ■

LEMMA A.3. *Let $\lambda \in (0,1\rangle$ and put $h_n = [\lambda n]$ for $n \in \mathbb{N}$. Under Assumptions D, F, and I1, there is $\varepsilon > 0$ such that, for $n \to +\infty$, $\sqrt{n}\sup_{\beta \in U(\beta^0,\varepsilon)} |r_{[h_n]}(\beta) - G_\beta^{-1}(\lambda)| = \mathcal{O}_p(1)$ and*

$$E_{Ln} = \mathsf{E}\left\{\sqrt{n}\sup_{\beta \in U(\beta^0,\varepsilon)} |r_{[h_n]}(\beta) - G_\beta^{-1}(\lambda)|\right\} = \mathcal{O}(1). \tag{A.3}$$

Proof. The proof has a structure rather similar to the proof of Lemma A.2. First, let us take a fixed $\varepsilon \in (0,1)$ and an arbitrary $K_1 > 0$ and set $K_\varepsilon = K_1 m_g$. Then $\mathsf{E}\nu_{1i}(\beta, K_1) > \lambda$.

Now, Assumption F2 and van der Vaart and Wellner (1996, Lems. 2.6.15 and 2.6.18) imply that $\{\nu_{1i}(\beta, K_1); \beta \in U(\beta^0, \delta), K_1 \in \mathbb{R}\}$ form a VC class that is uniformly bounded by 1. Because of Assumption D1 on the mixing coefficients, we can apply the uniform central limit theorem of Arcones and Yu (1994) to see that

$$\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\nu_{1i}(\beta, K_1) - \mathsf{E}\nu_{1i}(\beta, K_1)\} : \beta \in U(\beta^0, \delta), K_1 > 0\right\} \tag{A.4}$$

converges in distribution to a Gaussian process with bounded and uniformly continuous paths. To bound the expectation of the supremum of process (A.4), we now employ the maximal inequality of Doukhan et al. (1995). This is possible because (a) functions $\nu_{1i}(\beta, K_1)$ are uniformly bounded by 1, (b) the metric entropy with bracketing $H_2(u) = \mathcal{O}(|\log u|)$ by Assumption F2 and van der Vaart and Wellner (1996, Thm. 2.6.7), and (c) the mixing coefficients satisfy $\beta_m = o(m^{-r_\beta/(r_\beta - 2)})$ with $r_\beta/(r_\beta - 2)\cdot(1 - 1/r_\beta) = (r_\beta - 1)/(r_\beta - 2) > 1$. The properties (b) and (c) guarantee that condition (2.17) of Doukhan et al. (1995) is satisfied, which in turn implies the integrability condition (2.10) in the same paper. Therefore, we can use the maximum inequality by Doukhan et al. (1995, Thm. 2) and state for any $1 \le \pi < 2$ that

$$\sup_{n \in \mathbb{N}} \mathsf{E}\left[\sup_{\beta \in U(\beta^0,\varepsilon), K_1 > 0}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\nu_{1i}(\beta, K_1) - \mathsf{E}\nu_{1i}(\beta, K_1)\}\right|\right]^\pi < \infty.$$

This result is given in Doukhan et al. (1995) for $\pi = 1$, but all the proofs also hold for any $1 \le \pi < 2$. Consequently, the Markov-type inequality for nonnegative random variables, $P(X \ge K_\varepsilon) \le \mathsf{E}X^\pi/K_\varepsilon^\pi$, implies that there is some constant $U_\pi > 0$ such that for any $K_\varepsilon > 0$

$$P\left(\sup_{\beta \in U(\beta^0,\varepsilon), K_1 > 0}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\nu_{1i}(\beta, K_1) - \mathsf{E}\nu_{1i}(\beta, K_1)\}\right| > \frac{1}{2}K_\varepsilon\right) < \frac{U_\pi}{K_\varepsilon^\pi}$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

(note that this result is in the i.i.d. case directly implied by van der Vaart and Wellner, Thm. 2.14.9). Thus, it holds uniformly in $\beta \in U(\beta^0, \varepsilon)$ with probability greater than $1 - U_\pi / K_\varepsilon^\pi$

$$-\frac{1}{2}\sqrt{n} \cdot K_\varepsilon + \sum_{i=1}^{n} \mathsf{E}\, \nu_{1i}(\beta, K_1) \leq \sum_{i=1}^{n} \nu_{1i}(\beta, K_1). \tag{A.5}$$

Further, we can find n_0 such that $n^{-1/2} K_1 < \delta_g$ for all $n > n_0$ (δ_g comes from Assumption D3), and thus $\mathsf{E}\, \nu_{1i}(\beta, K_1) > \lambda + n^{-1/2} K_1 m_g = \lambda + n^{-1/2} K_\varepsilon$ for all $\beta \in U(\beta^0, \varepsilon)$ and $n > n_0$. This result together with equation (A.5) implies that for all $\beta \in U(\beta^0, \varepsilon)$

$$n\lambda + \frac{1}{2}\sqrt{n} K_\varepsilon = -\frac{1}{2}\sqrt{n} K_\varepsilon + n\lambda + \sqrt{n} K_\varepsilon < -\frac{1}{2}\sqrt{n} K_\varepsilon + \sum_{i=1}^{n} \mathsf{E}\, \nu_{1i}(\beta, K_1)$$

$$\leq \sum_{i=1}^{n} \nu_{1i}(\beta, K_1),$$

which means that at least $n\lambda \geq h_n$ of values $r_i(\beta)$ are smaller than $G_\beta^{-1}(\lambda) + n^{-1/2} K_\varepsilon$. In other words, $r_{[h_n]}(\beta) \leq G_\beta^{-1}(\lambda) + n^{-1/2} K_\varepsilon$ on $U(\beta^0, \varepsilon)$ with probability at least $1 - U_\pi / K_\varepsilon^\pi$. The corresponding lower inequality can be found by repeating these steps for $\nu_{2i}(\beta, K_1)$.

These inequalities can be rewritten as $Z_n = \sup_{\beta \in U(\beta^0, \varepsilon)} \sqrt{n} |r_{[h_n]}(\beta) - G_\beta^{-1}(\lambda)| \leq K_\varepsilon$, which holds with probability $1 - U_\pi / K_\varepsilon^\pi$. Thus, for any $\varepsilon > 0$ and $U_\pi > 0$, we find $K_\varepsilon = 1 + (2U_\pi / \varepsilon)^{1/\pi}$ such that $P(Z_n(\beta) \leq K_\varepsilon) > 1 - \varepsilon$, and so $Z_n = \mathcal{O}_p(1)$. Furthermore, denoting the cumulative distribution function of Z_n by $F_{z,n}$ and choosing $\pi = \frac{3}{2}$, the expectation

$$\mathsf{E}\, Z_n = \int_0^\infty \{1 - F_{z,n}(x)\}\, dx \leq 1 + \int_1^\infty \frac{2U_{3/2}}{x^{3/2}}\, dx = 1 + 4U_{3/2}$$

is finite, which concludes the proof. ■

The following lemma and corollaries translate the results on the convergence of the order statistics of residuals to the convergence of indicators $I(r_i(\beta) \leq r_{[h_n]}(\beta))$ to $I(r_i(\beta) \leq G_\beta^{-1}(\lambda))$ in probability and in mean.

LEMMA A.4. *Under Assumptions D, F1, F3, and I1, it holds that* $P_G = \mathsf{E} \sup_{\beta \in B} |\nu_{in}(\beta)| = o(1)$ *for any* $i \leq n \in \mathbb{N}$. *Additionally, under Assumptions D, F, and I1, there exists* $\varepsilon > 0$ *such that* $P_L = \mathsf{E} \sup_{\beta \in U(\beta^0, \varepsilon)} |\nu_{in}(\beta)| = \mathcal{O}(n^{-1/2})$ *as* $n \to +\infty$.

Proof. First notice that $\mathsf{E} \sup_{\beta \in B} |\nu_{in}(\beta)| = P(\exists \beta \in B : |\nu_{in}(\beta)| \neq 0)$ because $|\nu_{in}(\beta)| \in \{0, 1\}$. Without loss of generality, we discuss only the case $\nu_{in}(\beta) = -1$, which corresponds to $r_{[h_n]}(\beta) < r_i(\beta) \leq G_\beta^{-1}(\lambda)$. The other case $\nu_{in}(\beta) = 1$ can be derived analogously. Additionally, we assume without loss of generality that $i = n$.

Let Ω denote the probability space and let us consider an outcome $\omega = (\omega_1, \ldots, \beta_n) \in \Omega^n$, which generates observations $y_i(\omega_i), x_i(\omega_i)$, the corresponding residuals $r_i(\beta, \omega_i)$, and their order statistics $r_{[i]}(\beta, \omega)$, $i = 1, \ldots, n$. Given the first $n - 1$ observations determined by $\omega' = (\omega_1, \ldots, \omega_{n-1}) \in \Omega^{n-1}$ and ordered statistics $r_{[h_n]}(\beta, \omega')$ of residuals

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

$\{r_1(\beta,\omega_1),\ldots,r_{n-1}(\beta,\omega_{n-1})\}$, we can express the ordered statistics $r_{[h_n]}(\beta,\omega)$ for the whole sample as

$$r_{[h_n]}(\beta,\omega)=\begin{cases} r_{[h_n-1]}(\beta,\omega') & \text{if } r_n(\beta,\omega_n)<r_{[h_n-1]}(\beta,\omega') \\ r_n(\beta,\omega_n) & \text{if } r_{[h_n-1]}(\beta,\omega')\le r_n(\beta,\omega_n)<r_{[h_n]}(\beta,\omega') \\ r_{[h_n]}(\beta,\omega') & \text{if } r_{[h_n]}(\beta,\omega')\le r_n(\beta,\omega_n). \end{cases} \tag{A.6}$$

Denoting Ω_1, Ω_2, and Ω_3 subsets of Ω^n corresponding to the three (disjoint) cases in (A.6), we can write

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

and analyze this sum one by one.

(a) $P_1=P(\{\omega\in\Omega_1|\exists\beta\in B:\nu_{nn}(\beta)=-1\})\le P(\exists\beta\in B:r_{[h_n]}(\beta,\omega)<r_n(\beta,\omega_n)<r_{[h_n]}(\beta,\omega))=0$.

(b) $P_2=P(\{\omega\in\Omega_2|\exists\beta\in B:\nu_{nn}(\beta)=-1\})=P(\exists\beta\in B:r_{[h_n-1]}(\beta,\omega')\le r_n(\beta,\omega_n)=r_{[h_n]}(\beta,\omega)<G_\beta^{-1}(\lambda))$ can be analyzed in exactly the same way as $P(\{\omega\in\Omega_3|\exists\beta\in B:\nu_{nn}(\beta)=-1\})$; see point (c).

(c) $P_3=P(\{\omega\in\Omega_3|\exists\beta\in B:\nu_{nn}(\beta)=-1\})=P(\exists\beta\in B:r_{[h_n]}(\beta,\omega')=r_{[h_n]}(\beta,\omega)\le r_n(\beta,\omega_n)\le G_\beta^{-1}(\lambda))$. We can structure this term in the following way (Assumption D3):

$$P(\exists\beta\in B:r_{[h_n]}(\beta,\omega')<r_n(\beta,\omega_n)\le G_\beta^{-1}(\lambda)) \tag{A.7}$$

$$\begin{aligned} &= \int_{\omega'\in\Omega^{n-1}}\int_{\omega_n\in\Omega}\sup_{\beta\in B} I(r_{[h_n]}(\beta,\omega')<r_n(\beta,\omega_n)\le G_\beta^{-1}(\lambda)) \\ &\quad\cdot dP(\omega_n|\omega')\,dP(\omega') \\ &= \int_{\omega'\in\Omega^{n-1}} M_{gg}\cdot\sup_{\beta\in B}|r_{[h_n]}(\beta,\omega')-G_\beta^{-1}(\lambda)|\,dP(\omega') \\ &= M_{gg}\cdot\mathsf{E}\left\{\sup_{\beta\in B}|r_{[h_n]}(\beta)-G_\beta^{-1}(\lambda)|\right\}. \end{aligned} \tag{A.8}$$

The first claim of the lemma, $P_G=o(1)$, is then a direct consequence of Lemma A.2. The second result, $P_L=\mathcal{O}(n^{-1/2})$, can be derived analogously if we consider only a neighborhood $U(\beta^0,\varepsilon)$ instead of B, write the last expectation as

$$n^{-1/2}M_{gg}\cdot\mathsf{E}\left\{\sqrt{n}\sup_{\beta\in U(\beta^0,\varepsilon)}|r_{[h_n]}(\beta,\omega')-G_\beta^{-1}(\lambda)|\right\},$$

and employ Lemma A.3. ■

COROLLARY A.5. *Let Assumptions D, F1, F3, and I1 hold and assume that $t(x,y;\beta)$ is a real-valued function continuous in β uniformly in x and y over any*

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

compact subset of the support of (x, y). *Moreover, assume that* $\mathsf{E}\sup_{\beta\in B}\sup_{n\geq n_0}|t_i(\beta)\iota_{1i}(\beta)| < \infty$, *where* $t_i(\beta) = t(x_i, y_i;\beta)$ *and* $n_0 \in \mathbb{N}$. *Then it holds that* $\mathsf{E}\sup_{\beta\in B}|t_i(\beta)\nu_{in}(\beta)| = o(1)$. *Additionally, if Assumptions D, F, and I1 hold and there is* $\varepsilon > 0$ *such that* $\mathsf{E}\{\sup_{\beta\in U(\beta^0,\varepsilon)}|t_i(\beta)\nu_{in}(\beta)|\,|\sup_{\beta\in U(\beta^0,\varepsilon)}\nu_{in}(\beta)\neq 0\} < M_t$ *is bounded,* $\mathsf{E}\sup_{\beta\in U(\beta^0,\varepsilon)}|t_i(\beta)\nu_{in}(\beta)| = \mathcal{O}(n^{-1/2})$ *as* $n\to+\infty$.

Proof. This can verified along the same lines as in the proof of Lemma A.4. Defining functions $\nu_{in}(\beta)$ and sets Ω_1, Ω_2, and Ω_3 exactly the same way as in Lemma A.4, we can express the expectation of any random variable $\mathsf{E}X$ as $\{\int_{\Omega_1}+\int_{\Omega_2}+\int_{\Omega_3}\}x dF(x)$. By the same argument as in Lemma A.4, we will treat only the part concerning $\int_{\Omega_3}$ and assume without loss of generality that $i = n$.

First, because the expectation

$$\mathsf{E}\left\{\sup_{\beta\in B}|t_n(\beta)\nu_{in}(\beta)|\right\} = \mathsf{E}\left\{\sup_{\beta\in B}|t_n(\beta)\nu_{in}(\beta)|\cdot\sup_{\beta\in B}|\nu_{in}(\beta)|\right\}$$

is finite by the assumption of the corollary and the Lebesgue theorem (see Lemma A.1 for details) and $P(\sup_{\beta\in B}|\nu_{in}(\beta)| = 1)$ converges to zero as $n\to+\infty$ (Lemma A.4), the whole expectation converges to zero also, which is the first claim of this corollary.

Second, similarly to (A.7)–(A.8), we can write because of Assumption F4

$$\begin{aligned}
&\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}|t_n(\beta)\nu_{nn}(\beta)|\right\}\\
&\quad\leq\int_{\Omega_3}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}|t_n(\beta)\nu_{nn}(\beta)|\right\}dP(\omega)\\
&\quad\leq\int_{\omega'\in\Omega^{n-1}}\int_{\omega_n\in\Omega}\mathsf{E}\left\{\sup_{\beta\in U(\beta^0,\varepsilon)}|t_n(\beta)\nu_{nn}(\beta)|\;\middle|\;\sup_{\beta\in U(\beta^0,\varepsilon)}\nu_{nn}(\beta)\neq 0\right\}\\
&\qquad\cdot\sup_{\beta\in U(\beta^0,\varepsilon)}|\nu_{nn}(\beta)|dP(\omega_n|\omega')\,dP(\omega')\\
&\quad\leq M_t\int_{\omega'\in\Omega^{n-1}}\int_{\omega_n\in\Omega}\sup_{\beta\in U(\beta^0,\varepsilon)}|\nu_{nn}(\beta)|dP(\omega_n|\omega')\,dP(\omega')\\
&\quad\leq n^{-1/2}M_tM_{gg}\int_{\omega'\in\Omega^{n-1}}\sqrt{n}\sup_{\beta\in U(\beta^0,\varepsilon)}|r_{[h_n]}(\beta,\omega')-G_\beta^{-1}(\lambda)|dP(\omega').
\end{aligned}$$

Thus, we obtain from Lemma A.3

$$\mathsf{E}\left\{\sup_{\beta\in B}|t_n(\beta)\nu_{nn}(\beta)|\right\}\leq n^{-1/2}M_tM_{gg}E_{Ln} = \mathcal{O}(n^{-1/2}).$$ ■

COROLLARY A.6. *Under assumptions of Corollary A.5, it holds that* $\sup_{\beta\in B}|(1/n)\sum_{i=1}^n t_i(\beta)\nu_{in}(\beta)| = o_p(1)$ *and there is* $\varepsilon > 0$ *such that* $\sup_{\beta\in U(\beta^0,\varepsilon)}|(1/\sqrt{n})\sum_{i=1}^n t_i(\beta)\nu_{in}(\beta)| = \mathcal{O}_p(1)$ *as* $n\to+\infty$.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Proof. This result follows directly from the Markov inequality, $P(X \geq K) \leq \mathsf{E}X/K$, because

$$\mathsf{E} \sup_{\beta \in B} \left| \frac{1}{n} \sum_{i=1}^{n} t_i(\beta) \nu_{in}(\beta) \right| \leq \mathsf{E} \sup_{\beta \in B} |t_i(\beta) \nu_{in}(\beta)| = o(1)$$

and

$$\mathsf{E} \sup_{\beta \in U(\beta^0, \varepsilon)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} t_i(\beta) \nu_{in}(\beta) \right| \leq n^{1/2} \mathsf{E} \sup_{\beta \in U(\beta^0, \varepsilon)} |t_i(\beta) \nu_{in}(\beta)| = \mathcal{O}(1)$$

as $n \to +\infty$ by Corollary A.5, and the expectations are thus uniformly bounded in $n \in \mathbb{N}$. ■

LEMMA A.7. *Let Assumptions D, F, and I hold. Given constants $\lambda \in \langle 0, 1 \rangle$ and $M > 0$, it holds that*

$$n^{-1/2} \sup_{t \in T_m} |S_n'(\beta^0 - n^{-1/2} t) - S_n'(\beta^0) + n^{1/2} \{Q_s(\lambda) + J_s(\lambda)\} t| = o_p(1)$$

as $n \to +\infty$, where $T_m = \{t \in \mathbb{R}^p \,|\, \|t\| \leq M\}$ and $J_s(\lambda) = (\partial / \partial \beta^\top) \mathsf{E}[s_i'(\beta^0) \nu_{1i}(\beta)]|_{\beta = \beta^0}$.

Proof. See Čížek (2007, Lem. A.7).

APPENDIX B: Proof of Consistency and Convergence Rate

Proof of Theorem 3.1. This is a standard proof of consistency based on the uniform law of large numbers and the convergence of the order statistics $r_{[h_n]}(\beta)$ to the corresponding quantile $G_\beta^{-1}(\lambda)$. By definition, $P(S_{nn}(\hat{\beta}_n^{(GTE, h_n)}) < S_{nn}(\beta^0)) = 1$. For any $\delta > 0$,

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

Hence, $P(\inf_{\beta \in B \backslash U(\beta^0, \delta)} S_{nn}(\beta) < S_{nn}(\beta^0)) \to 0$ as $n \to +\infty$ implies $P(\hat{\beta}_n^{(GTE, h_n)} \in U(\beta^0, \delta)) \to 1$ as $n \to +\infty$, that is, the consistency of $\hat{\beta}_n^{(GTE, h_n)}$ (δ was an arbitrary positive number). To verify $P(\inf_{\beta \in B \backslash U(\beta^0, \delta)} S_{nn}(\beta) < S_{nn}(\beta^0)) \to 0$ note that

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

$$
\begin{aligned}
&P\left(\inf_{\beta\in B\setminus U(\beta^0,\delta)} S_{nn}(\beta) < S_{nn}(\beta^0)\right)\\
&\quad = P\left(\inf_{\beta\in B\setminus U(\beta^0,\delta)} [S_{nn}(\beta) - S(\beta) + S(\beta)] < S_{nn}(\beta^0)\right)\\
&\quad \le P\left(\inf_{\beta\in B\setminus U(\beta^0,\delta)} [S_{nn}(\beta) - S(\beta)] < S_{nn}(\beta^0) - \inf_{\beta\in B\setminus U(\beta^0,\delta)} S(\beta)\right)\\
&\quad \le P\left(\sup_{\beta\in B}|S_{nn}(\beta) - S(\beta)| > \inf_{\beta\in B\setminus U(\beta^0,\delta)} S(\beta) - S_{nn}(\beta^0)\right)\\
&\quad \le P\left(2\sup_{\beta\in B}|S_{nn}(\beta) - S(\beta)| > \inf_{\beta\in B\setminus U(\beta^0,\delta)} S(\beta) - S(\beta^0)\right).
\end{aligned}
$$

Because Assumption I2 implies that there is $\alpha > 0$ such that $\inf_{\beta\in B\setminus U(\beta^0,\delta)} S(\beta) - S(\beta^0) > \alpha$, it is enough to show for all $\alpha > 0$ that $P(\sup_{\beta\in B}|S_n(\beta) - S(\beta)| > \alpha) \to 0$ as $n \to +\infty$. This is a direct consequence of Corollary A.6 and Lemma A.1 for function $t_i(\beta) = s_i(\beta)$ (see Assumptions D, F1, and F3), because

$$
\begin{aligned}
S_{nn}(\beta) - S(\beta) &= \frac{1}{n}\sum_{i=1}^{n} s_i(\beta)\nu_{in}(\beta)\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\{s_i(\beta)\nu_{1i}(\beta) - \mathsf{E}[s_i(\beta)\nu_{1i}(\beta)]\}.
\end{aligned}
$$

■

After proving the consistency of GTE, we aim to derive the asymptotic distribution of GTE using its asymptotic linearity (Lemma A.7). However, to use it, one has to show first that the general trimmed estimates converge at rate $n^{-1/2}$.

LEMMA B.1. *Let Assumptions D, F, and I hold. Then $\hat\beta_n^{(GTE,h_n)}$ is $\sqrt{n}$-consistent, that is, $\sqrt{n}(\hat\beta_n^{(GTE,h_n)} - \beta^0) = \mathcal{O}_p(1)$ as $n \to +\infty$.*

Proof. We already know that $\hat\beta_n^{(GTE,h_n)}$ is consistent. Hence $P(\|\hat\beta_n^{(GTE,h_n)} - \beta^0\| > \rho) \to 0$ as $n \to +\infty$ for any $\rho > 0$ (Theorem 3.1).

Further, we employ the almost sure second-order differentiability of $S_{nn}(\beta)$ at β^0 (see Lemma 2.1 and Assumption F1). Because $S_{nn}(\beta) = (1/n)\sum_{i=1}^n s_i(\beta)\nu_{in}(\beta) + (1/n)\sum_{i=1}^n s_i(\beta)\nu_{1i}(\beta)$, Assumptions F, Lemma A.1, and Corollary A.6 imply that $S_{nn}(\beta) \to S(\beta)$ as $n \to +\infty$ in probability. Using the same argument for the first two derivatives of $S_{nn}(\beta)$ (see Lemma 2.1), $S'_{nn}(\beta) \to S'(\beta)$ and $S''_{nn}(\beta) \to S''(\beta)$ as $n \to +\infty$ uniformly in $\beta \in U(\beta^0,\delta)$, whereby $S''(\beta^0) = \mathsf{E}\{s_i''(\beta^0)\nu_{1i}(\beta)\} = Q_s > 0$ by Assumptions D2 and F3. Because Q_s is a positive definite matrix by Assumption F3, there is a constant $\rho, \delta > \rho > 0$, such that $\|S'(\beta)\| \ge C\|\beta - \beta^0\|$ for all $\beta \in U(\beta^0,\rho)$ and some $C > 0$. Because of the consistency of $\hat\beta_n^{(GTE,h_n)}$, this implies that for any $\varepsilon > 0$ there is some $n_0 \in \mathbb{N}$ such that $\hat\beta_n^{(GTE,h_n)} \in U(\beta^0,\rho)$ and subsequently $\|S(\hat\beta_n^{(GTE,h_n)})\| \ge C\|\hat\beta_n^{(GTE,h_n)} - \beta^0\|$ for all $n > n_0$ with probability at least $1 - \varepsilon$. Therefore, it is sufficient to show that $\sqrt{n}\|S'(\hat\beta_n^{(GTE,h_n)})\| = \mathcal{O}_p(1)$ to prove the theorem.

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

Let us express $\sqrt{n}S'(\hat{\beta}_n^{(GTE,h_n)})$ for $n > n_0$ with probability greater than $1 - \varepsilon$ as

$$\sqrt{n}\mathsf{E}[s_i'(\hat{\beta}_n^{(GTE,h_n)})\nu_{1i}(\hat{\beta}_n^{(GTE,h_n)})]$$

$$\leq \sup_{\beta \in U(\beta^0,\rho)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{\mathsf{E}[s_i'(\beta)\nu_{1i}(\beta)] - s_i'(\beta)\nu_{1i}(\beta)\} \tag{B.1}$$

$$- \sup_{\beta \in U(\beta^0,\rho)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i'(\beta)\nu_{in}(\beta) \tag{B.2}$$

(recall that $S_{nn}'(\hat{\beta}_n^{(GTE,h_n)}) = 0$ by Lemma 2.1). We only have to show that both terms are bounded in probability. This result for the first term on the right-hand side of (B.2) is a consequence of Corollary A.6 together with Assumptions F1, F3, and F4. The other part (B.1) on the right-hand side can be bounded in probability by the following argument. Assumption F2 together with van der Vaart and Wellner (1996, Lem. 2.6.18) implies that $\mathcal{F}_{n,\delta} = \{s_i'(\beta)\nu_{1i}(\beta) : \beta \in U(\beta^0,\delta)\}$ form a VC class of functions. Therefore, Assumptions D1 and F2 permit the use of the uniform central limit theorem of Arcones and Yu (1994), which implies that $\mathcal{F}_{n,\delta}$ converges in distribution to a Gaussian process with uniformly bounded and continuous paths, which confirms that (B.1) is bounded in probability. ■

Proof of Theorem 3.2. The asymptotic normality of GTE is a direct consequence of the $\sqrt{n}$ consistency of the estimator (Lemma B.1) and its asymptotic linearity (Lemma A.7). Because $t_n = \sqrt{n}(\hat{\beta}_n^{(GTE,h_n)} - \beta^0) = \mathcal{O}_p(1)$ as $n \to +\infty$, we can write

$$S_n'(\beta^0 - n^{-1/2}t_n) - S_n'(\beta^0) + n^{1/2}\{Q_s(\lambda) + J_s(\lambda)\}t_n = o_p(n^{1/2})$$

with a probability arbitrarily close to one. Substituting for S_n' and t_n yields

$$\sum_{i=1}^{n} s_i'(\hat{\beta}_n^{(GTE,h_n)})\iota_{in}(\hat{\beta}_n^{(GTE,h_n)}) - \sum_{i=1}^{n} s_i'(\beta^0)\iota_{in}(\beta^0)$$

$$+ n\{Q_s(\lambda) + J_s(\lambda)\}(\hat{\beta}_n^{(GTE,h_n)} - \beta^0) = o_p(n^{1/2}).$$

Because the first sum is by the definition of GTE equal to zero (Lemma 2.1), it follows that

$$\sqrt{n}(\hat{\beta}_n^{(GTE,h_n)} - \beta^0) = n^{-1/2}\{Q_s(\lambda) + J_s(\lambda)\}^{-1} \sum_{i=1}^{n} s_i'(\beta^0)\iota_{in}(\beta^0)$$

$$= n^{-1/2}\{Q_s(\lambda) + J_s(\lambda)\}^{-1} \sum_{i=1}^{n} s_i'(\beta^0)\{\iota_{in}(\beta^0) - \nu_{1i}(\beta^0)\} \tag{B.3}$$

$$+ n^{-1/2}\{Q_s(\lambda) + J_s(\lambda)\}^{-1} \sum_{i=1}^{n} s_i'(\beta^0)\nu_{1i}(\beta^0) + o_p(1). \tag{B.4}$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press

First, we show that (B.3) is asymptotically negligible in probability. Expectations

$$\mathsf{E}|n^{1/4}s_i'(\beta^0)\{\iota_{in}(\beta^0) - \nu_{1i}(\beta^0)\}|^l$$
$$= \mathsf{E}\{n^{l/4}|s_i'(\beta^0)|^l|\iota_{in}(\beta^0) - \nu_{1i}(\beta^0)|\} = \mathcal{O}(1)$$

are bounded for $l = 1,2$ because of Corollary A.5. Assumption I3 further indicates that the summands in (B.3) multiplied by $n^{1/4}$ form a stationary sequence of random variables with zero means and finite variances. Thus, the law of large numbers for mixingales (e.g., Davidson, 1994, Cor. 20.16) leads to

$$n^{-3/4}\sum_{i=1}^{n} n^{1/4}s_i'(\beta^0)\{\iota_{in}(\beta^0) - \nu_{1i}(\beta^0)\} \to 0,$$

which implies that (B.3) is negligible in probability as $n \to \infty$. Hence,

$$\sqrt{n}(\hat{\beta}_n^{(GTE,h_n)} - \beta^0) = n^{-1/2}\{Q_s(\lambda) + J_s(\lambda)\}^{-1}\sum_{i=1}^{n} s_i'(\beta^0)\nu_{1i}(\beta^0) + o_p(1).$$

Second, the summands in (B.4), $s_i'(\beta^0)\nu_{1i}(\beta^0)$, form a sequence of identically distributed random variables with zero mean and finite second moments (Assumptions D1, F3, and I3). Because by the law of large numbers for L^1-mixingales (Andrews, 1988)

$$\frac{1}{n}\sum_{i=1}^{n} s_i'(\beta^0)s_i'(\beta^0)^\top \cdot \nu_{1i}(\beta^0) \to V_s(\lambda)$$

in probability as $n \to +\infty$, we can employ the central limit theorem for (B.4) (e.g., Arcones and Yu, 1994, by Assumptions D1 and F2). This results directly in the asymptotic normality of $\hat{\beta}_n^{(GTE,h_n)}$ with the asymptotic variance given by Davidson (1994, Thm. 22.8):

$$V(\lambda) = \{Q_s(\lambda) + J_s(\lambda)\}^{-1}\cdot V_s(\lambda)\cdot\{Q_s(\lambda) + J_s(\lambda)\}^{-1}. \quad \blacksquare$$

https://doi.org/10.1017/S0266466608080596 Published online by Cambridge University Press