PROD. TYPE: FTP
PP: 1-14 (col.fig.: nil)
JSPI2988 MODC+
ED: Latha
PAGN: Vish – SCAN: ???

ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Two-stage Huber estimation

Tae-Hwan Kim[a,b], Christophe Muller[c,*]

[a] *Department of Economics, Yonsei University, Korea*
[b] *School of Economics, University of Nottingham, UK*
[c] *Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, Campus de San Vicente, 03080 Alicante, Spain*

Received 9 July 2003; accepted 23 January 2006

Abstract

In this paper we propose a new robust estimator in the context of two-stage estimation methods directed towards the correction of endogeneity problems in linear models. Our estimator is a combination of Huber estimators for each of the two stages, with scale corrections implemented using preliminary median absolute deviation estimators. In this way we obtain a two-stage estimation procedure that is an interesting compromise between concerns of simplicity of calculation, robustness and efficiency. This method compares well with other possible estimators such as two-stage least-squares (2SLS) and two-stage least-absolute-deviations (2SLAD), asymptotically and in finite samples. It is notably interesting to deal with contamination affecting more heavily the distribution tails than a few outliers and not losing as much efficiency as other popular estimators in that case, e.g. under normality. An additional originality resides in the fact that we deal with random regressors and asymmetric errors, which is not often the case in the literature on robust estimators.

MSC: 62F35; 62H12

Keywords: Two-stage estimation; Huber estimator; Robustness; Endogeneity

1. Introduction

The endogeneity problem in model estimation is usually dealt with by conducting the estimation in two stages. In the first stage, reduced form equations are estimated and the fitted values of endogenous variables are calculated. Then, these fitted values are used as regressors in the second stage and the covariance matrix of the estimated parameters is corrected for the replacement of the initial regressors by their fitted values. The use of two-stage least-squares to treat the endogeneity problem in a linear model is an example of this approach. Estimation methods in two stages have been studied for many M-estimators. For example, see Malinvaud (1970), Heckman (1978), Amemiya (1985), Krasker and Welsch (1985), Newey (1985, 1989, 1994), Krasker (1986), Pagan (1986), Duncan (1987).

The motivation of this research is to provide an estimation procedure that is at the same time (i) simple to implement, (ii) robust, and (iii) relatively efficient, notably under normality. Moreover, our estimation procedure based on the Huber estimator (Huber, 1964, 1981) deals with asymmetric errors and random regressors arising in two-stage setting of equations familiar to empirical researchers who focus their interest on the ‘structural equation’. This seems a reasonable

* Corresponding author. Tel.: +34 965 90 3400; fax: +34 965 90 3898.
E-mail address: cmuller@merlin.fac.ua.es (C. Muller).

0378-3758/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.jspi.2006.01.007

requirement for a method to be used in applied work in many situations. We use preliminary scale correction that is implemented at each estimation stage with median absolute deviation (MAD) estimator. The MAD is sometimes considered as the 'most robust estimator of scale.'

Our method is characterised by the implementation of the Huber estimator in the two stages. The Huber estimator is one of the main robust estimators and guarantees the robustness of the procedure with respect to error terms. We denote this procedure two-stage Huber (2SH) estimator. Typically, robust estimators attempt to respond to a variety of problems: outliers generating heavy tails errors in the dependent variable; true distributions of errors deviating from the assumed distribution (generally Gaussian distribution); other model misspecifications.[1] Using the same Huber estimator in the first stage not only makes the procedure simple to implement, but also eliminates the potential influence of non-normality or outliers in the first stage.

A few authors have tackled the problem of robust estimation for simultaneous equation systems; Krasker and Welsch (1985), Krasker (1986), Koenker and Portnoy (1990), Krishnakumar and Ronchetti (1997), Maronna and Yohai (1997), Flavin (1999). Krasker and Welsch (1985) propose weighted IV resistant estimators defined by implicit equations. These estimators have attractive robust properties but can only be calculated through an iterative procedure. Koenker and Portnoy (1990) study weighted LAD estimators that improve efficiency as compared to simple LAD estimators. Krishnakumar and Ronchetti (1997) propose B-robust (i.e. with bounded influence function) estimators as a good compromise between efficiency and robustness. Their estimators are based on the application of the Huber function to residuals that are affine combinations of the scores of the considered system when errors are normal. As for Krasker and Welsch estimator, the calculation of the estimator is not simple in that it requires an iterative procedure. In an empirical paper, Flavin (1999) robustifies the first-order conditions of an IV estimator, and uses the MAD as a scale estimator. Maronna and Yohai (1997) review some of these methods and identify three different estimating strategies; robustifying three-stage least squares, robustifying the full information maximum likelihood method and generalising multivariate τ-estimators.

Rather than directly following these strategies, we focus on a natural extension of the Huber estimator to the two-stage setting used in most applied empirical work. Indeed, the two-stage setting is the one likely to be used by applied researchers who are predominantly interested in the second stage estimates. Moreover, the successive Huber estimation is simple to implement, as opposed to many other robust estimators. Krasker (1986) proposes to robustify the two-stage least-square (2SLS) method by replacing each OLS stage by a bounded influence estimator and to apply it to simultaneous equations. We follow this approach by replacing each OLS stage with a modified Huber estimator for which we use the mean absolute deviation estimator as preliminary scaling estimator for the errors at each estimation stage. This feature modifying the classical Huber estimator has the advantage of being simple and quick to apply, and therefore to allow estimation procedures only based on usual Huber, LS and LAD estimators, all for which readily usable commands are available in many statistical packages (e.g. Stata). In a sense, our procedure is akin to that of Krishnakumar and Ronchetti, while it differs in that we apply the Huber function to each component of the modified system (the first- and second-stage equations in our case) and not to the score of the whole system together. Moreover, we replace the simultaneous estimation of scale and location by simple preliminary scale estimators. These modifications allow us to keep most of the robustness benefit of the Huber function in a context of two-stage estimation, while avoiding computational complications and making the resulting estimator regression-equivalent.

Our contribution in this paper is to develop the 2SH estimator, to derive its asymptotic and small sample properties. This work program is carried out in the case of random regressors and possibly asymmetric errors, weaker restrictions than what is often available in the literature for robust estimation. The simplicity of the estimation of parameter estimates is one of the main motivation of our work. Indeed, the requirement of simplicity is important for practitioners in the context of robust estimation, as stressed for example in Flavin (1999). Our estimator belongs to the small class of the available robust estimators for estimation problems with endogenous explanatory variables. The practical gain obtained by using a two-stage estimation approach and well-known procedures used by practitioners should help contributing to the dissemination of robust methods.

Another interest of our approach is that we keep a simple link with the most popular robust estimator, the Huber estimator, allowing for natural comparisons. Although different methods have been used to obtain robustness in system estimation, the Huber estimator remains a central reference because of its optimal minimax properties (established in

[1] See for example Huber (1964, 1981), Krasker and Welsch (1985).

Huber, 1964). It has been exploited in more complex frameworks in Krishnakumar and Ronchetti (1997) to derive B-robust estimators. However, the minimax properties of Huber and B-robust estimators have generally been established only in the case of symmetric errors and small contamination of normal errors. Although these cases are clearly central references, it is less clear if they correspond to the realistic situations of interest for the use of robust estimators. Moreover, when compared with the use of simple Huber estimators, B-robust estimators may be complicated to implement like other variants of the resistant estimators by Krasker and Welsch. We deal with this issue by sticking to simple estimation procedures without emphasising minimax properties that always depend on debatable distribution benchmarks. Other approaches could be to use least-squares or least absolute deviations regressions, which can both be readily implemented in several common statistical packages, as the basis of each estimation stage. Though, with 2SLS substantial robustness issues may arise, while with two-stage least-absolute-deviations (2SLAD) much efficiency may be lost under normality. Recently, more sophisticated two-stage estimators have been proposed in the framework of quantile regressions and LAD estimators in order to deal with endogeneity of treatment effects (Abadie et al., 2002; Chernozhukov and Hansen, 2001). We do not deal with these sophistications and remain in the usual setting of two-stage estimation methods, in part because we are concerned about robustness and not about treatment effects differing across quantiles. So, the estimator we propose corresponds to a convenient trade-off between simplicity of implementation, efficiency and robustness. However, as most of the other robust estimators in this context, our estimator only protects against local distribution alternatives, but not against large and pervasive departure from normality. The latter situation should be dealt differently, for example methods with high breakdown point.

The aim of this paper is to propose and study the 2SH estimator. In Section 2, we define the model and the estimation method. We derive the asymptotic representation of the 2SH estimator in Section 3. In Section 4, we analyse the asymptotic normality of the 2SH estimator. Some Monte Carlo simulation experiments are presented in Section 5. Both asymptotic and Monte Carlo results exhibit situations where the 2SH estimator is superior to the 2SLS and to the 2SLAD developed by Amemiya (1982) and Powell (1983). Finally, Section 6 concludes. All the technical proofs are collected in the Appendix.

2. The model and the estimator

2.1. The model

We consider a structural equation given by

$$y_t = Y_t'\gamma_0 + x_{1t}'\beta_0 + u_t, \quad t = 1, 2, \ldots, T, \tag{1}$$

where y_t is the dependent variable, Y_t is a $G \times 1$ vector of endogenous variables, x_{1t} is a $K_1 \times 1$ vector of random exogenous variables and u_t is the error term. We are interested in estimating and making inference on the structural parameters $\alpha_0 = (\gamma_0', \beta_0')'$. We denote by x_{2t} the $K_2 \times 1$ vector of exogenous variables that are absent from the equation in (1). Let us assume that $Y_t' = (Y_{1t}, Y_{2t}, \ldots, Y_{Gt})$ admits a reduced-form representation for each Y_{jt}:

$$Y_{jt} = x_t'\Pi_{0j} + V_{jt}, \quad j = 1, 2, \ldots, G, \tag{2}$$

where $x_t' = (x_{1t}', x_{2t}')$, Π_{0j} is a $K \times 1$ vector of unknown parameters with $K = K_1 + K_2$, and V_{jt} is the error term. (2) can be written as $Y_t' = x_t'\Pi_0 + V_t'$ where $\Pi_0 = [\Pi_{01}, \Pi_{02}, \ldots, \Pi_{0G}]$ and $V_t' = (V_{1t}, V_{2t}, \ldots, V_{Gt})$. Then, y_t has the following reduced-form representation:

$$y_t = x_t'\pi_0 + v_t, \tag{3}$$

where

$$\pi_0 = \left[\Pi_0, \begin{pmatrix} I_{K_1} \\ 0 \end{pmatrix}\right]\alpha_0$$

and $v_t = u_t + V_t'\gamma_0$. Note that (3) can be rewritten as

$$y_t = z_t'\alpha_0 + v_t, \tag{4}$$

where $z_t' = x_t' H(\Pi_0)$ and

$$H(\Pi_0) = \left[\Pi_0, \begin{pmatrix} I_{K_1} \\ 0 \end{pmatrix}\right]$$

is a $K \times (G + K_1)$ matrix. Hence, if the true value of Π_0 were known, the structural parameter vector α_0 could be directly estimated using (4). The essence of the two-stage approach is to replace Π_0 with an estimator $\hat{\Pi}$ from the first stage. We now turn to the definition of the 2SH estimator.

2.2. *The estimator*

The reduced-form equation in (2) is used for a first-stage Huber estimation that yields estimator $\hat{\Pi}_j$ of Π_{0j}, and delivers exogenous predictions of Y_{jt}: $\hat{Y}_{jt} = x_t'\hat{\Pi}_j$. In this situation, the usual Huber estimator would be obtained as the solution to: $\min_{\Pi_j} \sum_{t=1}^{T} \rho(Y_{jt} - x_t'\Pi_j)$, where

$$\rho(z) = \tfrac{1}{2}z^2 1_{[|z|<k]} + (k|z| - \tfrac{1}{2}k^2)1_{[|z|\geqslant k]},$$

where $k > 0$ is a threshold value fixed in advance. This function, often called the Huber function, has the first derivative

$$\psi(z) = z1_{[|z|<k]} + 2k(\tfrac{1}{2} - 1_{[z\leqslant 0]})1_{[|z|\geqslant k]}.$$

However, using the Huber function $\rho(z)$ does not deliver a regression-equivalent estimator. In order to obtain such an estimator, the Huber function must be modified using some scale estimator of the error term. Indeed, the basic Huber estimator, as most M-estimators, is not scale invariant. Following Bickel (1975) and Flavin (1999), we use the standardised MAD to obtain the required scale estimators as follows:

$$\hat{\sigma}^2 = \text{median}\{|e_t - \text{median}\{e_t\}|\}/\Phi^{-1}(\tfrac{3}{4}),$$

$$\hat{\sigma}_j^2 = \text{median}\{|e_{jt} - \text{median}\{e_{jt}\}|\}/\Phi^{-1}(\tfrac{3}{4}),$$

$$j = 1, 2, \ldots, G,$$

where the e_t are the residuals obtained from the LS regression of y_t on x_t and the e_{jt} are the residuals from the LS regression of Y_{jt} on x_t. Using a preliminary scale estimate is possible in our problem because $\hat{\sigma}^2$ and $\hat{\sigma}_j^2$ are consistent. Plugging the consistent scale estimators into the optimisation program defining Huber estimator does not affect the consistency of the estimation procedure as shown in Andrews (1994, p. 226).

For symmetric distributions, the MAD, which is asymptotically equivalent to half the interquartile range, is minimax with respect to bias and has the best breakdown properties under ε-contamination in the class of M-estimators (Huber, 1981). Note that this choice of scale estimator is different from Huber proposal,[2] or from B-estimators that follow Huber approach (Krishnakumar and Ronchetti, 1997). The advantage of Huber and B-estimator approaches is to ensure minimax properties for both scale and location parameters, namely to minimise the variance of these estimators under a 'least-informative distribution' of the contaminated errors. The drawback of these approaches is to lead to unattractive calculation complication, perhaps one reason why these estimators are relatively little used despite their interesting properties. One of the calculation complication comes from the simultaneous estimation of the location and the scale. Moreover, the optimality of these location and scale estimators is only reached when the reference distribution is the 'least-informative one,' and not for other alternatives that may be more plausible. Meanwhile, the minimax properties refer to symmetric contamination and normal uncontaminated errors. Finally, minimax approaches may not protect as well against heavy-tails errors that may be more likely than 'least-informative distributions.' Since the breakdown properties of simultaneous estimators of scale and location, or of estimators with preliminary scaling, are mainly

[2] An M-estimator S of scale defined by $\int \chi(x/S(F))F(\mathrm{d}x) = 0$ with $\chi(x) = x^2 - \beta$ for $|x| \leqslant \beta$ and $\chi(x) = k^2 - \beta$ for $|x| > \beta$, with β such that $\int \chi(x)\Phi(\mathrm{d}x) = 0$, all with Huber's obvious original notation.

determined by the breakdown properties of the scale estimator (Huber, 1981), it seems a good idea to use the MAD as scale estimator. Indeed, the MAD has the largest breakdown point ($\frac{1}{2}$) of all M-estimators. So, we explore a new route in this paper, namely using the MAD as preliminary scale estimator of each stage of a two-stage estimation context to deal with endogeneity issues. Another reason to use the MAD is that it can be considered as the 'most robust estimator of scale' (Huber, 1981, p. 122) in the sense that it is the limit of the Huber estimator of scale when the contamination grows up to almost all the observations.

The first-stage Huber estimator $\hat{\Pi}_j$ is defined as the solution to the following:

$$\min_{\Pi_j} \sum_{t=1}^{T} \rho_{\hat{\sigma}_j}(Y_{jt} - x_t'\Pi_j),$$

where $\rho_{\hat{\sigma}_j}(z)=\rho(z/\hat{\sigma}_j)$. The corresponding first-order condition is: $T^{-1/2}\sum_{t=1}^{T} x_t\psi_{\hat{\sigma}_j}(Y_{jt}-x_t'\hat{\Pi}_j)=0$ for $j=1,\ldots,G$, where $\psi_{\hat{\sigma}_j}(z)=\psi(z/\hat{\sigma}_j)$, dropping the factor $1/\hat{\sigma}_j$.

Let us now turn to the second stage of the estimation. The 2SH estimator $\hat{\alpha}$ of α_0 is the solution to the following minimisation programme:

$$\min_{\alpha} S_T(\alpha,\hat{\Pi}) = \sum_{t=1}^{T} \rho_{\hat{\sigma}}(y_t - \hat{z}_t'\alpha), \tag{5}$$

where $\hat{z}_t' = x_t'H(\hat{\Pi})$ with $\hat{\Pi}=[\hat{\Pi}_1,\hat{\Pi}_2,\ldots,\hat{\Pi}_G]$ incorporates the first-stage predictions of the endogenous variables. Alternatively, the 2SH estimator can be defined as a solution to the first-order condition: $\sum_{t=1}^{T} H(\hat{\Pi})'x_t\psi_{\hat{\sigma}}(y_t-\hat{z}_t'\alpha)=0$. The same Huber estimator (i.e. the same parameter k) is used for both stages consistently with our requirement of simplicity. Alternatively, different values for k could be used for the two stages, little changing the properties of the estimator. In the next section, we derive the asymptotic representation of the 2SH estimator. The following conditions are needed for this task.

Assumption 1. (i) The entire sequence $\{(v_t, V_t, x_t)\}$ is independent and identically distributed with $\sigma^2 = E(v_t^2) \in (0,\infty)$, $\sigma_j^2 = E(V_{jt}^2) \in (0,\infty)$ and $\sigma_u^2 = E(u_t^2) \in (0,\infty)$.
(ii) $E(\|x_t\|)^3 < \infty$ where $\|x_t\| = (x_t'x_t)^{1/2}$.

Assumption 2. (i) v_t has a conditional cdf $F(\cdot|x)$ that is Lipschitz continuous for all x.
(ii) $E[\psi_\sigma(v_t)|x_t]=0$.
(iii) $Q=E\{g(0|x_t)x_tx_t'\}$ is finite and positive-definite where $g(0|x_t)=G'(0|x_t)$ with $G(z|x_t)=E[\psi_\sigma(v_t+z)|x_t]$.
(iv) $H(\Pi_0)$ is of full column rank.

The iid condition in Assumption 1(i), which may be interpreted as a description of the sampling scheme, is imposed for presentation simplicity. It can be relaxed to include heteroskedasticity and serial correlation. In such a case, our covariance matrix in Proposition 3 further on should be modified by using the Newey–West type covariance matrix. Assumption 1(ii) is the moment condition on the vector of exogenous variables. It is useful to establish the asymptotic representation of all considered estimators by applying a theorem of stochastic equicontinuity for the relevant empirical process. We also use it to obtain the boundedness of the asymptotic covariance matrix of parameter estimates. This condition and the other conditions of the exogenous regressors are weaker than what is sometimes done in two-stage estimation papers, where the exogenous regressors are assumed fixed (e.g. Powell, 1983).

Assumption 2(i) simplifies the demonstration of convergence of certain remainder terms to zero. It is used to derive the asymptotic representation of our estimator. Assumption 2(ii) says that the trimmed conditional mean of v_t is zero. If ψ were instead the quantile function (as in Powell, 1983), it would mean that the conditional quantile of v_t is zero. In the case of OLS, ψ would be the identity and the restriction would correspond to the nullity of the conditional mean. Assumption 2(ii) is satisfied if v_t is symmetric and $E(v_t|x_t)=0$, which is commonly assumed for Huber estimators (e.g. Bickel, 1975; Carroll and Ruppert, 1982). Here, even if the distribution of v_t is not symmetric, when the conditional density of v_t is not equal to zero over a large support and when there is an intercept in the model, this condition can be considered as an innocuous normalisation of the intercept. Therefore, it should be satisfied in applications where there are intercepts in the model equations, the usual case. Indeed, in that case $E(\psi(v_t)|x_t) = 0$ and $E(\psi(v_t)|x_t) \neq 0$ correspond to isomorphic statistical structures that distinguish themselves only by the value of the intercept term. They are observationally equivalent structures. Therefore, it is possible to impose $E(\psi(v)|x_t) = 0$, and thus to fix the value of the intercept, without loss of generality. An additional requirement to make this assumption useful is naturally that there are enough observations corresponding to errors between $-k$ and k after rescaling by $\hat{\sigma}_j$ and $\hat{\sigma}$. This is not an issue for asymptotic results but may cause difficulties for very small samples.

Assumption 2(iii) is akin to the usual condition for OLS that the mean cross-product of the regressor vectors converges towards a finite positive definite matrix. Here, the cross-product matrix is weighted by a coefficient that characterises how fast the trimmed conditional mean of v_t changes around zero along the change in v_t. This conditions is necessary for consistency and for the inversion of the relevant empirical process in order to establish the asymptotic normality.

Assumption 2(iv) is an identification condition, which is standard for simultaneous equations models. The structural equation in (1) is identified if the number of zero restrictions (K_2) is not less than the number of endogenous variables (G). Noting that $H(\Pi_0)$ is a $(K_1 + K_2) \times (K_1 + G)$ matrix, Assumption A2(iv) implies $K_2 \geqslant G$, which includes the exact identified case ($K_2 = G$) and the over-identified case ($K_2 > G$). This assumption is needed when proving that the 2SH estimator $\hat{\alpha}$ is consistent for the true parameter α_0.

It is now time to return to the consequences of the possible endogeneity of Y_t. If the Y_t are exogenous in the general sense, we have $E[\psi_{\sigma_u}(u_t)|Y_t, x_t] = 0$ where $\psi_{\sigma_u}(z) = \psi(z/\sigma_u)$ and σ_u is the preliminary scale estimator. Then, the truncation of outliers, via the Huber function, in terms of the error terms makes intuitive sense since it eliminates observations corresponding to errors that are discrepant relative to a central tendency. However, if the Y_t are endogenous, we expect in general that $E[\psi_{\sigma_u}(u_t)|Y_t, x_t] \neq 0$ as a consequence of the non-separation of the marginal laws of the Y_t and of the u_t. It is unclear what the truncation of the error terms means in this situation, and therefore what is estimated. Our 2SH estimator ensures that at each stage of the estimation, the truncation eliminating outliers applies to well-defined error terms that can be interpreted as prediction errors from a set of exogenous variables.

3. The asymptotic representation

The first step of the analysis is the derivation of an asymptotic representation of the 2SH estimator. For this, we define an empirical process $M_T(\cdot)$ as follows:

$$M_T(\Delta) = T^{-1/2} \sum_{t=1}^{T} x_t \psi_\sigma(v_t - T^{-1/2} x_t' \Delta) = T^{-1/2} \sum_{t=1}^{T} m_\sigma(w_t, \Delta),$$

where Δ is a $K \times 1$ vector, $w_t = (v_t, x_t')'$ and $m_\sigma(w_t, \Delta) = x_t \psi_\sigma(v_t - T^{-1/2} x_t' \Delta)$. We can apply Theorems 1–3 in Andrews (1994) because of the bounded variations of function ψ_σ, of the iid condition in Assumption 1(i) and of the moment condition on x_t in Assumption 1(ii). As a result, we have the following preliminary lemma.

Lemma 1. *Suppose that Assumptions* 1 *and* 2(i) *hold. Then, for some finite* $L > 0$, *we have the following*:

$$\sup_{\|\Delta\| \leqslant L} \|M_T(\Delta) - M_T(0) + Q\Delta\| = o_p(1).$$

Lemma 1 together with the remaining conditions in Assumption 2 is used to obtain the following asymptotic representation for the 2SH estimator.

Proposition 1. *Suppose that Assumptions* 1 *and* 2 *hold and that* $T^{1/2}(\hat{\Pi} - \Pi_0) = O_p(1)$. *Then, we have the following*:

$$T^{1/2}(\hat{\alpha} - \alpha_0) = Q_{zz}^{-1} H(\Pi_0)' \left\{ T^{-1/2} \sum_{t=1}^{T} x_t \psi_\sigma(v_t) - Q T^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0 \right\} + o_p(1),$$

where $Q_{zz} = H(\Pi_0)' Q H(\Pi_0)$.

In the above proposition, a high level condition on $\hat{\Pi}$ has been imposed, namely $T^{1/2}(\hat{\Pi} - \Pi_0) = O_p(1)$. This condition can be seen as a consequence of the following assumption.

Assumption 3. (i) V_{jt} has a conditional cdf $H_j(\cdot|x)$ for $j=1,\ldots,G$ that is Lipschitz continuous for all x.

(ii) $E[\psi_{\sigma_j}(V_{jt})|x_t]=0$ for $j=1,\ldots,G$, where $\psi_{\sigma_j}(z)=\psi(z/\sigma_j)$.

(iii) $Q_j=E\{g_j(0|x_t)x_tx_t'\}$ is finite and positive-definite, $j=1,\ldots,G$, where $g_j(0|x_t)=G_j'(0|x_t)$ with $G_j(z|x_t)=E[\psi_{\sigma_j}(V_{jt}+z)|x_t]$.

Assumptions 3(i)–(iii) are similar to Assumptions 2(i)–(iii). Hence,, the same discussion applies. We start the analysis of the asymptotic properties of the 2SH estimator with the derivation of its asymptotic representation from which we shall deduce the asymptotic normality and the asymptotic covariance matrix. It can easily be shown that $g(0|x_t)=F(k\,|x_t)-F(-k|x_t)$ and $g_j(0|x_t)=H_j(k|x_t)-H_j(-k|x_t)$ for $j=1,\ldots,G$. To simplify we eliminate from now the conditioning from the notations for functions G, G_j, g, g_j.

Proposition 2. *Suppose that Assumptions* 1–3 *hold. Then, the* 2SH *estimator* $\hat{\alpha}$ *has the asymptotic representation*:

$$T^{1/2}(\hat{\alpha}-\alpha_0)=DT^{-1/2}\sum_{t=1}^{T}Z_t+o_p(1),$$

where $D=Q_{zz}^{-1}H(\Pi_0)'[I,-QQ_1^{-1}\gamma_{01},\ldots,-QQ_G^{-1}\gamma_{0G}]$, $Z_t=W_t\otimes x_t$, *and* $W_t=[\psi_\sigma(v_t),\psi_{\sigma_1}(V_{1t}),\ldots,\psi_{\sigma_G}(V_{Gt})]'$.

Let us now compare this result with the following asymptotic representations of 2SLS, denoted $\hat{\alpha}_{2SLS}$, and 2SLAD, denoted $\hat{\alpha}_{2SLAD}$. It is easy to show under some regularity conditions[3] comparable to Assumptions 1–3 that

$$T^{1/2}(\hat{\alpha}_{2SLS}-\alpha_0)=D^{LS}T^{-1/2}\sum_{t=1}^{T}Z_t^{LS}+o_p(1),$$

where $D^{LS}=Q_{zz}^{LS-1}H(\Pi_0)'[I,-\gamma_{01}I,\ldots,-\gamma_{0G}I]$, $Q_{zz}^{LS}=H(\Pi_0)'Q^{LS}H(\Pi_0)$, $Q^{LS}=E\{x_tx_t'\}$ and $Z_t^{LS}=W_t^{LS}\otimes x_t$, and $W_t^{LS}=[v_t,V_{1t},\ldots,V_{Gt}]'$. Moreover,

$$T^{1/2}(\hat{\alpha}_{2LAD}-\alpha_0)=D^{LAD}T^{-1/2}\sum_{t=1}^{T}Z_t^{LAD}+o_p(1),$$

where $D^{LAD}=Q_{zz}^{LAD-1}H(\Pi_0)'[I,-Q^{LAD}(Q_1^{LAD})^{-1}\gamma_{01},\ldots,-Q^{LAD}(Q_G^{LAD})^{-1}\gamma_{0G}]$, $Q_{zz}^{LAD}=H(\Pi_0)'Q^{LAD}H(\Pi_0)$, $Q^{LAD}=E\{f(0|x_t)x_tx_t'\}$, $Q_j^{LAD}=E\{h_j(0|x_t)x_tx_t'\}$, for $j=1,\ldots,G$, $Z_t^{LAD}=W_t^{LAD}\otimes x_t$, $W_t^{LAD}=[\psi_{LAD}(v_t),\psi_{LAD}(V_{1t}),\ldots,\psi_{LAD}(V_{Gt})]'$, $\psi_{LAD}(z)=0.5-1[z\leqslant 0]$, and $f(\cdot|x)$ and $h_j(\cdot|x)$ are the conditional pdf's of v_t and V_{jt}, respectively.

4. Asymptotic normality and covariance matrix

We now show the asymptotic normality of the 2SH estimator by applying the Lindeberg–Levy CLT.

Proposition 3. *Suppose that Assumptions* 1–3 *hold. Then*,

$$T^{1/2}(\hat{\alpha}-\alpha_0)\xrightarrow{d}N(0,D\Omega D'),$$

where $\Omega=E(W_tW_t'\otimes x_tx_t')$.

[3] The complete list of conditions can be found in Kim and Muller (2004). Among the conditions, the crucial ones that ensure the consistency of the 2SLS and 2SLAD estimators are $E[v_t|x_t]=0$, $E[V_{jt}|x_t]=0$, $E[\psi_{LAD}(v_t)|x_t]=0$, and $E[\psi_{LAD}(V_{jt})|x_t]=0$. They are essentially two differences in the assumptions for the consistency and asymptotic normality of 2SLS, 2SLAD and 2SH: first, the semiparametric restriction, second the existence and positive definiteness of matrices incorporating cross-products of regressions. The only differences for these matrices lies in the presence of a different weighing factors. It does not seem to us that this fundamentally changes the scope of cases to consider. As for the semiparametric restriction, as long as the support of the errors is broad enough, and that there is an intercept term in the model, they can all be considered as mere normalisation of the intercept coefficient, and the slope estimates can be validly compared.

These asymptotic results show the differences in the asymptotic properties of 2SH, 2SLS and 2SLAD. The proof of the asymptotic representation for the 2SLAD with random regressors can be found for example in Kim and Muller (2004). The differences in assumptions in part boil down to different normalisation rules (of the type $E[\psi(v_t)|x_t]=0$) for the intercept, which we assume from now. So, the comparison of asymptotic properties makes sense for the slope coefficients at least, which are generally the coefficients of interest for applied researchers since they convey the effects of the explanatory variables. First, the influence function of 2SLS, easy to read from the empirical processes of the asymptotic representation, is not bounded, as opposed to that of 2SH and of 2SLAD that are robust methods. Then, the 2SLS is not appropriate for example when the errors are seriously contaminated. Second, low values of $f(0|x_t)$ and $h_j(0|x_t)$ for the 2SLAD can degrade the asymptotic efficiency of these estimators.

As far as small sample properties are concerned, for each estimation method it is possible to exhibit corresponding error distributions for which the method is efficient and dominates the other two estimation methods: Normal distributions for the 2SLS, Laplace distributions for the 2SQR and minimax Huber distributions for the 2SH estimator (Huber, 1964). Therefore, no method can be considered to be best in all cases in terms of efficiency and the 2SH inherits minimax properties from the Huber estimator. However, the 2SLAD is known to suffer from particularly low efficiency in situations close to the Gaussian case, often considered as a major benchmark.

On the whole, the 2SH appears as an interesting compromise between the properties of robustness of the 2SLAD and the efficiency of the 2SLS under normality. In the next section, we present a simulation study that enables us to further compare small sample properties of the 2SH with that of the 2SLS and the 2SLAD.

5. Monte Carlo simulation

Using Monte Carlo simulations, we examine the small sample behaviour of the Huber estimator for the structural parameters (γ_0, β_0), on the one hand when the endogeneity problem is ignored and on the other hand when the problem is treated by using the 2SH estimator. We also compare this behaviour with that of the 2SLS and the 2SLAD.

We base our simulations on the simplest possible model: a simultaneous equation system with two simple equations so as to be able to naturally inject endogeneity into the model. The first equation, which is the equation of interest, contains two endogenous variables (including the dependent variable) and two exogenous variables including a constant. Four exogenous variables are present in the whole system.

The structural simultaneous equation system can then be written

$$B\begin{bmatrix} y_t \\ Y_t \end{bmatrix} + \Gamma x_t = U_t,$$

where $\begin{bmatrix} y_t \\ Y_t \end{bmatrix}$ is a 2×1 vector of endogenous variables, x_t is a 4×1 vector of exogenous variables with the first element equal to one. Then, we have an intercept term in the model and the semiparametric restrictions of the three estimators correspond to three different normalisation of the intercept coefficient. Therefore, we are mostly interested in the comparison of the slope coefficients. U_t is a 2×1 vector of error terms. We specify the structural parameters as follows: $B=\begin{bmatrix} 1 & -0.5 \\ -0.7 & 1 \end{bmatrix}$ and $\Gamma=\begin{bmatrix} -1 & -0.2 & 0 & 0 \\ -1 & 0 & -0.4 & -0.5 \end{bmatrix}$. The system is over-identified by the zero restrictions $\Gamma_{13}=\Gamma_{14}=\Gamma_{22}=0$. Hence, the first equation is

$$y_t = Y_t\gamma + (1, x_{1t})\beta + u_t = 0.5Y_t + 1 + 0.2x_{1t} + u_t,$$

where $\gamma=-B_{12}=0.5$, $\beta'=(\beta_0,\beta_1)=(-\Gamma_{11},-\Gamma_{12})=(1,0.2)$, x_{1t} is the second element of x_t and u_t is the first element of U_t. The second equation is therefore

$$Y_t = 0.7y_t + 1 + 0.4x_{2t} + 0.5x_{3t} + w_t,$$

where x_{2t} and x_{3t} are, respectively, the third and fourth elements of x_t, and w_t is the second element of U_t.

The choice of the parameter values is led by the following considerations. Since we are not so much interested in the intercepts, we simply choose them to be equal to 1 for all equations and all normalisations. Only slightly

Table 1
Simulation means and standard deviations of the deviations from the true value with: N(0, 1)

		Huber	2SH	2SLS	2SLAD
$k=2$					
$\hat{\gamma}$	Mean	−0.43	0.00	0.00	0.00
	Std	0.13	0.30	0.30	0.41
$\hat{\beta}_0$	Mean	1.38	−0.01	−0.02	−0.01
	Std	0.45	0.99	0.98	1.30
$\hat{\beta}_1$	Mean	0.16	0.00	0.00	0.00
	Std	0.16	0.21	0.21	0.26
$k=4$					
$\hat{\gamma}$	Mean	−0.44	0.00	0.00	0.00
	Std	0.13	0.30	0.30	0.41
$\hat{\beta}_0$	Mean	1.38	−0.02	−0.02	−0.01
	Std	0.44	0.98	0.98	1.30
$\hat{\beta}_1$	Mean	0.16	0.00	0.00	0.00
	Std	0.16	0.21	0.21	0.26

attenuated effects are chosen for the cross effects of the two endogenous variables (coefficients 0.5 and 0.7) so that the endogeneity be interesting but not extreme. Identification restrictions and overidentification drive the occurrence of exogenous variables in the equations. Moderate but non-negligible and relatively comparable effects are allowed for these variables (coefficients 0.2, 0.4, 0.5). In that way, a balanced situation is obtained where endogenous and exogenous influences should all matter. Thus, we expect that all parameter values in B and Γ play important roles, except perhaps for the intercept parameter, and especially the interaction between the endogenous variables.

We can rewrite the system in a matrix representation $[y\ \ Y]B' = -X\Gamma' + U$, which yields the reduced-form equations $[y\ \ Y] = X[\pi_0\ \ \Pi_0] + [v\ \ V]$, where $[\pi_0\ \ \Pi_0] = -\Gamma'(B')^{-1}$ and $[v\ \ V] = U(B')^{-1}$. Using $[\pi_0\ \ \Pi_0] = -\Gamma'(B')^{-1}$, we obtain $\pi_0' = (2.3, 0.3, 0.3, -0.15)$ and $\Pi_0' = (2.6, 0.2, 0.6, -0.3)$.

The errors $[v\ \ V]$ are successively drawn from the standard normal N(0, 1), the Student-t with 4 degrees of freedom, $t(4)$, and the Lognormal errors with log-mean $\mu = 0$ and log-standard deviation $\sigma = 1$, denoted by LN(0, 1), in such a way that Assumptions 2(ii) and 3(ii) are satisfied. We draw the second to fourth columns in X from the normal distribution with mean $(0.5, 1, -0.1)'$, variances equal to 1, $\mathrm{cov}(x_2, x_3) = 0.3$, $\mathrm{cov}(x_2, x_4) = 0.1$ and $\mathrm{cov}(x_3, x_4) = 0.2$. The variances are chosen equal to 1 for normalisation purposes, while the correlations between the exogenous variables are neither extreme nor negligible. Given X, $[v\ \ V]$ and $[\pi_0\ \ \Pi_0]$, we generate the endogenous variables $[y\ \ Y]$ by using the reduced-form equations.

We use 1000 replications and the sample size is 50. For each replication, we estimate the parameter values γ and $\beta = (\beta_0, \beta_1)'$ and we calculate the deviation of the estimates from the true values. Then, we compute the sample mean and sample standard deviation of those deviations based on the 1000 replications.

The simulated means and simulated standard deviation for the estimated γ, β_0 and β_1 of the one-stage Huber estimator, the 2SH, the 2SLS and the 2SLAD are displayed in Table 1 for Gaussian errors, Table 2 for $t(4)$ errors and Table 3 for LN(0, 1) errors. For Huber estimator, we use two different values of parameter k ($k = 2$ or $k = 4$).

As mentioned above, we are mostly interested in the slope coefficients since strictly speaking the intercept coefficients are not comparable across methods. However, it remains interesting to confront the accuracy of the various intercept estimators even when corresponding to different normalisations.

For all three parameters, the one-step Huber estimator (under the heading of 'Huber') is characterised by a systematic bias in finite samples, which does not vanish as the sample size increases from 50 (results are not shown, but available upon request). By contrast, the means of the 2SH, 2SLS and 2SLAD estimates (in deviation to the true value) are much closer to zero than the means of the one-stage Huber estimator for the three considered distributions. The endogeneity problem appears well corrected by the 2SH even with small samples in the studied cases.

As expected, for the Gaussian errors, the 2SLS performs better than the 2SLAD. Moreover, the 2SH is very close to the 2SLS. In the case of the Student errors, the 2SH is relatively more efficient than both the 2SLS and the 2SLAD and this relative efficiency is more pronounced with $k = 2$ than $k = 4$. In the case of the Lognormal distribution, the

Table 2
Simulation means and standard deviations of the deviations from the true value with: $t(4)$

		Huber	2SH	2SLS	2SLAD
$k=2$					
$\hat{\gamma}$	Mean	−0.48	0.00	−0.02	0.01
	Std	0.18	0.39	0.47	0.48
$\hat{\beta}_0$	Mean	1.53	−0.02	0.05	−0.05
	Std	0.59	1.30	1.59	1.53
$\hat{\beta}_1$	Mean	0.18	0.00	0.01	0.00
	Std	0.19	0.26	0.30	0.31
$k=4$					
$\hat{\gamma}$	Mean	−0.49	−0.01	−0.02	0.01
	Std	0.20	0.42	0.47	0.48
$\hat{\beta}_0$	Mean	1.55	0.02	0.05	−0.05
	Std	0.65	1.41	1.59	1.53
$\hat{\beta}_1$	Mean	0.18	0.01	0.01	0.00
	Std	0.21	0.29	0.30	0.31

Table 3
Simulation means and standard deviations of the deviations from the true value with: LN(0, 1)

		Huber	2SH	2SLS	2SLAD
$k=2$					
$\hat{\gamma}$	Mean	−0.52	0.00	−0.09	0.00
	Std	0.10	0.40	0.73	0.43
$\hat{\beta}_0$	Mean	1.69	0.02	0.28	0.01
	Std	0.42	1.34	2.37	1.42
$\hat{\beta}_1$	Mean	0.20	0.00	0.05	0.00
	Std	0.19	0.29	0.46	0.26
$k=4$					
$\hat{\gamma}$	Mean	−0.54	−0.04	−0.09	0.00
	Std	0.12	0.60	0.73	0.43
$\hat{\beta}_0$	Mean	1.75	0.15	0.28	0.01
	Std	0.48	2.00	2.37	1.42
$\hat{\beta}_1$	Mean	0.20	0.02	0.05	0.00
	Std	0.24	0.37	0.46	0.26

2SH estimator is more accurate than the 2SLS. When compared with the 2SLAD, the results are mixed; for some parameters (γ, β_0) with $k = 2$, the 2SH is more accurate while the reverse result is observed for other cases. Thus, adjusting parameter k may help to tune in the efficiency-robustness properties of 2SH as compared with 2SLS and 2SLAD. This is an additional advantage of 2SH, which could be systematically investigated to yield a rule for fixing k in different cases of interest.

Clearly, there exist cases easy to exhibit where 2SH dominates 2SLS or 2SLAD, even without introducing arbitrary outliers. This supports the contention that 2SH is a useful complement to the toolbox of two-stage estimation methods for systems of equations with endogeneity problems.

6. Conclusion

In this article we propose and study the two-stage Huber (2SH) estimator in the case of random regressors and possibly asymmetric errors. The error scale is corrected by preliminary uses of the MAD estimator at every stages. We derive the formula of the asymptotic covariance matrix for the parameters of interest. The comparison of the asymptotic properties and of Monte Carlo simulation results for the 2SH estimator, the 2SLS estimator and the 2SLAD estimator indicates that none of these estimators dominates the other ones in terms of both robustness and efficiency, even for a few simple distributions, whether asymptotically or for finite samples. In this situation, the 2SH estimator provides a convenient compromise between requirements of simplicity of implementation, robustness and efficiency. Asymptotic and finite sample results are important because they are first the base of inferences using our estimator, and second because they show that the 2SH estimator is a useful alternative to 2SLS and 2SLAD estimators.

In practice, when the data may be contaminated and that the 2SLAD may yield severely inefficient estimates (e.g. if the non-contaminated part of the data follows a distribution close to the normal law), then our estimator provides an interesting procedure for models where some explanatory variables are deemed to be endogenous and when ancillary equations based on exogenous regressors can be used to replace the endogenous regressors of the first stage with their fitted values.

A practical difficulty however, as for any other two-stage method, is that specific calculations must be carried out to estimate the asymptotic covariance matrix of the parameters. Arguably, the formula of the matrix is complex enough. However, convenient estimators of this matrix can be obtained by plugging residual and non-parametric estimators of expressions involving densities, similarly to what is done in Kim and Muller (2004) for two-stage quantile regressions. Alternatively, bootstrap estimators of the covariance matrix could be explored, perhaps on the lines proposed by Hahn (1995) for quantile regressions. Finally, an other useful extension would be the analysis of the small sample inaccuracy coming from the preliminary scale estimation for each stage.

Acknowledgements

This research has been supported under the British Academy Award No. 30,649. Tae-Hwan Kim is grateful to the College of Business and Economics of Yonsei University for financial support under the project code 2005-1-0244 and the second author is grateful for the financial support by Spanish Ministry of Sciences and Technology. Project No. BEC2002-03097 and by the Instituto Valenciano de Investigaciones Economicas. We thank participants in a seminar in Alicante and two referees for their useful comments. Usual disclaimers apply.

Appendix

Proof of Lemma 1. We first define $V_T(\Delta) = T^{-1/2}\sum_{t=1}^{T}[m_\sigma(w_t,\Delta) - E\{m_\sigma(w_t,\Delta)\}] = M_T(\Delta) - E(M_T(\Delta))$. In order to apply Theorem 1 in Andrews (1994) to the empirical process $V_T(\Delta)$, we need to check the following two conditions: (i) $m(w_t,\Delta)$ satisfies Pollard's entropy condition with some envelop $\bar{M}(w_t)$; (ii) For some $\delta > 2$, $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E[\{\bar{M}(w_t)\}^{\delta}] < \infty$.

We define $f_1(w_t,\Delta) = x_t$ and $f_2(w_t,\Delta) = \psi_\sigma(v_t - T^{-1/2}x'\Delta)$ so that $m_\sigma(w_t,\Delta) = f_1(w_t,\Delta) f_2(w_t,\Delta)$. We note that each element of $f_1(w_t,\Delta)$ is Type I class with envelope $\|x_t\|$ (see Andrews, 1994, for the definition of Type I class) and $f_2(w_t,\Delta)$ is also Type I class with envelope $C = 1 \vee k$ where $\vee$ is the maximum operator. Hence, the product $m_\sigma(w_t,\Delta)$ satisfies Pollard's entropy condition with envelope $\bar{M}(w_t) = C(\|x_t\| \vee 1)$ by Theorem 3 in Andrews (1994), which ensures the first condition. By Assumption 1(i) we have $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} E[\{\bar{M}(w_t)\}^{\delta}] = C^{\delta}E(\|x_t\| \vee 1)^{\delta}$, which is bounded by Assumption 1(ii). Hence, by applying Theorem 1 in Andrews (1994), we obtain the following:

$$\sup_{\|\Delta_1-\Delta_2\| \leqslant L} \|V_T(\Delta_1) - V_T(\Delta_2)\| = o_p(1), \tag{6}$$

for some finite $L > 0$. Setting $\Delta_1 = \Delta$ and $\Delta_2 = 0$ in (6) yields

$$\sup_{\|\Delta\| \leqslant L} \|M_T(\Delta) - M_T(0) - \{E(M_T(\Delta)) - E(M_T(0))\}\| = o_p(1). \tag{7}$$

We now show that $E(M_T(\Delta)) - E(M_T(0)) \to -Q\Delta$. Using the law of iterated expectation and the mean value theorem, we have that

$$\begin{aligned} E(M_T(\Delta)) - E(M_T(0)) &= E\left[T^{-1/2}\sum_{t=1}^{T} x_t\{G(-T^{-1/2}x_t'\Delta|x_t) - G(0|x_t)\}\right] \\ &= -E\left\{T^{-1}\sum_{t=1}^{T}\frac{G(-T^{-1/2}x_t'\Delta|x_t) - G(0|x_t)}{-T^{-1/2}x_t'\Delta}x_t x_t'\right\}\Delta \\ &= -E\left\{T^{-1}\sum_{t=1}^{T} g(\xi_T|x_t)x_t x_t'\right\}\Delta, \end{aligned} \tag{8}$$

where $\xi_{T,t}$ is between zero and $-T^{-1/2}x_t'\Delta$. Since one can show that $g(\lambda|x) = \sigma^{-1}\{F(-\lambda+\sigma k|x) - F(-\lambda-\sigma k|x)\}$, Assumption 2(i) implies that $g(\lambda|x)$ is Lipschitz continuous in λ for all x; that is, for some constant $L_0 \in (0,\infty)$ and for all x, $|g(\lambda_1|x) - g(\lambda_2|x)| \leqslant L_0|\lambda_1 - \lambda_2|$. Now we consider the (i,j)th element of $|E\{T^{-1}\sum_{t=1}^{T} g(\xi_T|x_t)x_t x_t'\} - E\{g(0|x_t)x_t x_t'\}|$, which is given by

$$\left|E\left\{T^{-1}\sum_{t=1}^{T}(g(\xi_T|x_t) - g(0|x_t))x_{ti}x_{tj}\right\}\right| \leqslant L_0 T^{-1}\sum_{t=1}^{T} E(|T^{-1/2}x_t'\Delta| \times |x_{ti}| \times |x_{tj}|), \tag{9}$$

where the inequality is obtained by the triangle inequality, the Jensen's inequality, the Lipschitz continuity and the fact that $0 \leqslant |\xi_{T,t}| \leqslant |T^{-1/2}x_t'\Delta|$. By the moment condition in Assumption 1(ii), the last expression in (9) converges to zero, which implies that $E\{T^{-1}\sum_{t=1}^{T} g(\xi_T|x_t)x_t x_t'\} \to E\{g(0|x_t)x_t x_t'\}$. Hence, we have $E(M_T(\Delta)) - E(M_T(0)) \to -Q\Delta$ as $T \to \infty$. Then, we substitute $E(M_T(\Delta)) - E(M_T(0))$ in (7) with its limit $-Q\Delta$ to obtain the conclusion of the lemma. □

Proof of Proposition 1. We first define $\hat{\Delta}_0 = \sqrt{T}(\hat{\Pi} - \Pi_0)\gamma_0$. Then, $\hat{\Delta}_0 = O_p(1)$ by assumption. Hence, by Lemma 1, we have $M_T(\hat{\Delta}_0) = M_T(0) - Q\hat{\Delta}_0 + o_p(1)$. The first term is given by

$$M_T(0) = T^{-1/2}\sum_{t=1}^{T} x_t\psi_\sigma(v_t),$$

which converges in distribution to a normal random variable by the Lindeberg–Levy CLT under Assumptions 1 and 2(ii). Since $M_T(0) = O_p(1)$ and $Q\hat{\Delta}_0 = O_p(1)$, using Lemma 1, we obtain

$$M_T(\hat{\Delta}_0) = O_p(1). \tag{10}$$

Next, we define

$$\hat{\Delta}_1(\delta) = H(\hat{\Pi})\delta + \hat{\Delta}_0 \tag{11}$$

where $\delta \in R^{G+K_1}$. It is easily shown that Lemma 1 implies

$$\sup_{\|\delta\| \leqslant L_1} \|M_T(\hat{\Delta}_1(\delta)) - M_T(0) + Q\hat{\Delta}_1(\delta)\| = o_p(1) \tag{12}$$

for some finite $L_1 > 0$. We further define $\tilde{M}_T(\delta) = H(\hat{\Pi})'M_T(\hat{\Delta}_1(\delta))$ and $\|H(\hat{\Pi})\|^2 = \mathrm{tr}(H(\hat{\Pi})H(\hat{\Pi})') = O_p(1)$. By using the argument displayed between (A.7) and (A.8) in Powell (1983), it is shown that (10) and (12) together imply that

$$\sup_{\|\delta\| \leqslant L_1} \|\tilde{M}_T(\delta) - H(\Pi_0)'M_T(\hat{\Delta}_0) + Q_{zz}\delta\| = o_p(1), \tag{13}$$

where $Q_{zz} = H(\Pi_0)'QH(\Pi_0)$. The next step of the proof is to show $\hat{\delta} = T^{1/2}(\hat{\alpha} - \alpha) = O_p(1)$ in order to plug into (13). For this, we use Lemma A.4. in Koenker and Zhao (1996), which can be applied under the following conditions:

(i) $-\delta'\tilde{M}_T(\lambda\delta) \geqslant -\delta'\tilde{M}_T(\delta)$ for $\lambda \geqslant 1$, (ii) $\|H(\Pi_0)'M_T(\hat{\Delta}_0)\| = \mathrm{O_p}(1)$, (iii) $\tilde{M}_T(\hat{\delta}) = \mathrm{o_p}(1)$, where $\hat{\delta} = T^{1/2}(\hat{\alpha} - \alpha_0)$, (iv) Q_{zz} is positive-definite. With these conditions, Lemma A.4 in Koenker and Zhao will deliver the desired results: $\hat{\delta} = T^{1/2}(\hat{\alpha} - \alpha) = \mathrm{O_p}(1)$.[4] First, we note that the following function $h(\lambda)$ is convex in λ: $h(\lambda) = \sum_{t=1}^{T} \rho_\sigma(v_t - T^{-1/2}x_t'H(\hat{\Pi})\delta\lambda - T^{-1/2}x_t'\hat{\Delta}_0)$. Since $-\delta'\tilde{M}_T(\lambda\delta)$ is the gradient of the above convex function, it is non-decreasing in λ. Hence, condition (i) is satisfied. The result in (10) implies condition (ii). To prove (iii), we note that

$$T^{1/2}\tilde{M}_T(\hat{\delta}) = \left[\left.\frac{\partial S_T}{\partial \alpha}\right|_{\alpha=\hat{\alpha}}\right]_- + \mathrm{o_p}(1), \tag{14}$$

where $[\partial S_T/\partial\alpha|_{\alpha=\hat{\alpha}}]_-$ is the vector of left-hand side partial derivatives of the objective function S_T in (5) evaluated at the solution $\hat{\alpha}$. Hence, $[\partial S_T/\partial\alpha|_{\alpha=\hat{\alpha}}]_- = \mathrm{o_p}(1)$. The difference of order $\mathrm{o_p}(1)$ in (14) comes from the fact that the scaled estimator $\hat{\sigma}$ is used on the right-hand side (i.e. in the objective function S_T) while the true value σ is used on the left-hand side (i.e. in the definition of $\tilde{M}_T(\hat{\delta})$). Since $\hat{\sigma} \xrightarrow{p} \sigma$, the difference converges to zero in probability as T goes to infinity. Hence, $T^{1/2}\tilde{M}_T(\hat{\delta}) = \mathrm{o_p}(1)$ and condition (iii) is satisfied. The final condition (iv) is ensured by the identification condition in Assumption 2(iv). Therefore, we have $T^{1/2}(\hat{\alpha} - \alpha_0) = \mathrm{O_p}(1)$. This result combined with (13) results in

$$T^{1/2}(\hat{\alpha} - \alpha_0) = Q_{zz}^{-1}H(\Pi_0)'M_T(\hat{\Delta}_0) + \mathrm{o_p}(1),$$

which delivers the desired result in the proposition by using the above decomposition of $M_T(\hat{\Delta}_0)$ provided by Lemma 1:

$$T^{1/2}(\hat{\alpha} - \alpha_0) = Q_{zz}^{-1}H(\Pi_0)'\left\{T^{-1/2}\sum_{t=1}^{T} x_t\psi_\sigma(v_t) - QT^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0\right\} + \mathrm{o_p}(1), \tag{15}$$

which completes the proof.[5] □

Proof of Proposition 2. In a fashion similar to Proposition 1 the conditions in Assumption 3 together with Assumptions 1 and 2(iv) are sufficient to show that the first stage Huber estimator $\hat{\Pi}_j$ has the following representation: $T^{1/2}(\hat{\Pi}_j - \Pi_{0j}) = Q_j^{-1}T^{-1/2}\sum_{t=1}^{T} x_t\psi_{\sigma_j}(V_{jt}) + \mathrm{o_p}(1)$.

Hence, $T^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0 = \sum_{j=1}^{G} Q_j^{-1}T^{-1/2}\sum_{t=1}^{T} x_t\psi_{\sigma_j}(V_{jt})\gamma_{0j} + \mathrm{o_p}(1)$, which is in turn plugged into (15) to deliver the following result:

$$\begin{aligned} T^{1/2}(\hat{\alpha} - \alpha_0) &= T^{-1/2}\sum_{t=1}^{T} Q_{zz}^{-1}H(\Pi_0)'x_t\psi_\sigma(v_t) \\ &\quad - T^{-1/2}\sum_{t=1}^{T} Q_{zz}^{-1}H(\Pi_0)'QQ_1^{-1}x_t\psi_{\sigma_1}(V_{1t})\gamma_{01} \\ &\quad \ldots \\ &\quad - T^{-1/2}\sum_{t=1}^{T} Q_{zz}^{-1}H(\Pi_0)'QQ_G^{-1}x_t\psi_{\sigma_G}(V_{Gt})\gamma_{0G} + \mathrm{o_p}(1) \\ &= DT^{-1/2}\sum_{t=1}^{T} Z_t + \mathrm{o_p}(1), \end{aligned} \tag{16}$$

where $D = Q_{zz}^{-1}H(\Pi_0)'[I, -QQ_1^{-1}\gamma_{01}, \ldots, -QQ_G^{-1}\gamma_{0G}]$, $Z_t = W_t \otimes x_t$, and $W_t = [\psi_\sigma(v_t), \psi_{\sigma_1}(V_{1t}), \ldots, \psi_{\sigma_G}(V_{Gt})]'$. □

---

[4] Once we show $\hat{\delta} = \mathrm{O_p}(1)$, the consistency of $\hat{\alpha}$ for α_0 follows by a by-product.

[5] Since the objective function in the second stage is convex, an alternative proof to the presented one might have been obtained using the insightful approach of Hjort and Pollard (1993). We use Koenker and Zhao (1996) instead of Hjort and Pollard (1993) because we are more familiar with Koenker and Zhao (1996) on the one hand, and on the other hand because a parallel can thus be drawn with the proof in Kim and Muller (2004) for the two-stage quantile regression.

Proof of Proposition 3. We first derive the distributional limit of $T^{-1/2}\sum_{t=1}^{T} Z_t$. Assumption 1(i) implies that Z_t is independent and identically distributed. Using Assumptions 2(ii), 3(ii) and the law of iterated expectation, one can show that $E(Z_t)=0$. Finally, we note that $\mathrm{var}(Z_t)=E(W_t W_t' \otimes x_t x_t')$ and all the elements in $W_t W_t'$ are bounded by a constant. Hence, Assumption 1(ii) is sufficient to confirm that $\mathrm{var}(Z_t)$ is bounded. Therefore, we now can apply the Lindeberg–Levy's CLT to obtain $T^{-1/2}\sum_{t=1}^{T} Z_t \to \mathrm{N}(0, \Omega)$. The conclusion of the proposition follows. □

References

Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica 70, 91–117.
Amemiya, T., 1982. Two stage least absolute deviations estimators. Econometrica 50, 689–711.
Amemiya, T., 1985. The nonlinear two-stage least-squares estimation. J. Econometrics 2, 105–110.
Andrews, D.W.K., 1994. Empirical process methods in econometrics. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, New York, pp. 2247–2294.
Bickel, P.J., 1975. One-step Huber estimates in the linear model. J. Amer. Statist. Assoc. 70, 428–433.
Carroll, R.J., Ruppert, D., 1982. Robust estimation in heteroscedastic linear models. Ann. Statist. 10, 429–441.
Chernozhukov, V., Hansen, C., 2001. An IV model of quantile treatment effects. Working Paper 02-06, Massachussets Institute of Technology, Department of Economics, December.
Duncan, G.M., 1987. A simplified approach to M-estimation with application to two-stage estimators. J. Econometrics 34, 373–389.
Flavin, M., 1999. Robust estimation of the joint consumption/asset demand decision. NBER Working Paper No. 7011.
Hahn, J., 1995. Bootstrapping quantile regression estimators. Econometric Theory 11, 105–121.
Heckman, J.J., 1978. Dummy endogenous variables in a simultaneous equation system. Econometrica 46, 931–960.
Hjort, N.L., Pollard, D., 1993. Asymptotics for minimisers of convex processes. Department of Statistics, Yale University, Preprint.
Huber, P.J., 1964. Robust estimation of a location parameter. Ann. Math. Statist. 35, 73–101.
Huber, P.J., 1981. Robust Statistics. Wiley, New York.
Kim, T.-H., Muller, C., 2004. Two-stage quantile regression when the first-stage is a quantile regression. Econometrics J. 7, 218–231.
Koenker, R., Portnoy, S., 1990. M-estimation of multivariate regressions. J. Amer. Statist. Assoc. 85, 1060–1068.
Koenker, R., Zhao, Q., 1996. Conditional quantile estimation and inference for ARCH models. Econometric Theory 12, 793–813.
Krasker, W.S., 1986. Two-stage bounded-influence estimators for simultaneous-equations models. J. Business Economic Statist. 4, 437–444.
Krasker, W.S., Welsch, R.E., 1985. Resistant estimation for simultaneous-equations models using weighted instrumental variables. Econometrica 53, 1475–1488.
Krishnakumar, J., Ronchetti, E., 1997. Robust estimators for simultaneous equations models. J. Econometrics 78, 295–314.
Malinvaud, E., 1970. The consistency of nonlinear regressions. Ann. Math. Statist. 41, 956–969.
Maronna, R.A., Yohai, V.J., 1997. Robust estimation in simultaneous equations models. J. Statist. Plann. Inference 57, 233–244.
Newey, W.K., 1985. Semiparametric estimation of limited dependent variables models with endogenous explanatory variables. Annales de l'INSEE 59/60, 219–235.
Newey, W.K., 1989. A method of moments interpretation of sequential estimators. Economics Lett. 14, 201–206.
Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.
Pagan, A.R., 1986. Two stage and related estimators and their applications. Rev. Econom. Stud. 57, 517–538.
Powell, J., 1983. The asymptotic normality of two-stage least absolute deviations estimators. Econometrica 51, 1569–1575.