

Two-Stage M-Estimation with Composite Dependent Variable

Tae-Hwan Kim and Christophe Muller

School of Economics, University of Nottingham
Nottingham, NG7 2RD, UK
tae-hwan.kim@nottingham.ac.uk
christophe.muller@nottingham.ac.uk

(August 2000)

Revised January 2002

Abstract: In order to reduce the variance of two-stage M-estimators, we study the case where the dependent variable in the second stage is of the composite form $qy_t + (1 - q)X_t'\hat{\beta}$ with $\hat{\beta}$ the first stage estimator and X_t the vector of regressors in the first stage. We establish the conditions for consistency and asymptotic normality and we derive the formula of the asymptotic covariance matrix and that of the optimal value q^* that minimises this covariance matrix. Finally, simulation results show that using a consistent estimator of q^* can often substantially improve the accuracy of the two-stage estimation.

Keywords: Two-Stage Estimation, M-Estimation, Endogeneity, Variance Reduction.

JEL Classification: C30

We are grateful to J. Powell, H. White and O. Linton for useful discussions about a preliminary version of this paper and to participants in presentations in Nottingham University, CREST-INSEE in Paris, London School of Economics, University of California at San Diego for their comment. Remaining errors are ours. We acknowledge a grant from the British Academy.

1. Introduction

Model estimation is often perturbed by the presence of nuisance parameters stemming from endogeneity and other econometric problems. The econometricians generally deal with these difficulties by conducting the estimation in two stages. In the first step, they generate predictions of the nuisance parameters by using ancillary models that are estimated from the data. Then, they conduct the estimation of interest by replacing the nuisance parameters with their predictions and correcting the covariance matrix of the estimated parameters. The use of 2SLS to treat the endogeneity of some independent variables in a linear model is the most common example of this approach.

Unfortunately, two stage estimation methods often yield inaccurate estimates of the parameters of interest because of the information loss caused by mediocre instrumental variables. This problem can be alleviated by increasing the number and the quality of the instrumental variables used in the predictive equations, although this is not always possible. In this paper, we explore an alternative way of increasing accuracy that is to adjust the definition of the dependent variable by choosing different values for a weighting coefficient q . This method will naturally be fruitful only in some cases. However, successful results in studied cases would encourage similar investigations for more general specifications.

In the case of the two-stage least-squares estimator $\hat{\beta}_{(Y)}$, for the linear model $Y = X\beta + u$ where u is a vector of error terms and β is the parameter to

estimate, it is easy to verify that using for the dependent variable y_t or $qy_t + (1 - q)X'_t\hat{\beta}_{(Y)}$, where $q \neq 0$ and X'_t is the row vector of independent variables, delivers identical estimators¹. Similarly, for a consistent (respectively unbiased, respectively asymptotically normal) estimator $\check{\beta}_{(Y)}$ we have $\hat{\beta}_{(qY+(1-q)X\check{\beta}_{(Y)})} = (X'P_ZX)^{-1}X'P_Z(qY + (1 - q)X\check{\beta}_{(Y)}) = q\hat{\beta}_{(Y)} + (1 - q)\check{\beta}_{(Y)}$, which is another consistent (respectively unbiased, respectively asymptotically normal) estimator of β .

Clearly, this favourable situation is caused by the linearity of the 2SLS estimator with respect to Y . For some nonlinear two-stage estimator $\tilde{\beta}$ with a first-stage estimator of the reduced form, such as the LAD estimator (Amemiya, 1982, Powell, 1983), or quantile regressions (Kim and Muller, 2001), the properties obtained by using $qY + (1 - q)X\tilde{\beta}$ as a dependent variable have been studied. In general, the choice of q matters for the estimation and we shall show that optimal choices of q can be exhibited.

The estimation in two stages, in particular when using instrumental variables, has been analysed for many M-estimators². Most of these methods can be interpreted as a two-stage implementation of M-estimators (Two-Stage M-Estimators), first for the ancillary model, then for the model of interest. The aim of this paper is to explore the reformulation of the dependent variable in the case

¹Indeed, if Z is the matrix of instrumental variables, Y is the vector of the dependent variable and X is the matrix of independent variables, we have $\hat{\beta}_{(Y)} = (X'P_ZX)^{-1}X'P_ZY$ where $P_Z = Z(Z'Z)^{-1}Z'$, and $\hat{\beta}_{(qY+(1-q)X\hat{\beta}_{(Y)})} = (X'P_ZX)^{-1}X'P_Z(qY + (1 - q)X\hat{\beta}_{(Y)}) = \hat{\beta}_{(Y)}$.

²For example, Malinvaud (1970), Heckman (1976), Amemiya(1985), Krasker and Welsch (1985), Newey (1985, 89, 94), Krasker (1986), Pagan (1986), Duncan (1987).

of Two-Stage M-Estimators ($2SM(q)$) as a device to increase the accuracy of the estimators. Our contribution is not only to derive the conditions for consistency and asymptotic normality of the two-stage estimator with composite dependent variable, but also to derive the optimal q^* that minimises the asymptotic variance of the estimator of the parameter of interest.

Because our intention in this paper is to explore a new direction of research, we shall focus on the sensitivity of the two-stage estimator accuracy with respect to q . We leave to future work the theoretical analysis of the consequences of the imprecise estimation of the optimal value for q . However, we provide simulation results showing that much accuracy can often be gained by estimating the optimal q^* and using this estimate in the two stage estimation.

To make our method easier to grasp, we develop the examples of the Double-Stage Huber estimator and of the Two-Stage Quantile Regression when the first stage is the LS estimator. The favourable results obtained in the simulations should contribute to convince some readers to use our method or to adapt it to other problems.

In Section 2, we define the model and the estimation method. We derive in Section 3 the asymptotic representation of the $2SM(q)$ estimator. In Section 4 we analyse the asymptotic normality and we discuss the estimation of the asymptotic covariance matrix. Some Monte Carlo simulation experiments are presented in Section 5. Finally, Section 6 concludes.

2. The Model

We start with a general setting for 2SM(q) estimators. Further on we shall develop a few examples. Let us suppose that we are interested in the structural parameter (α_0) of an equation given in the following matrix form for a sample of T observations:

$$\begin{aligned} y &= Y\gamma_0 + X_1\beta_0 + u \\ &\equiv Z\alpha_0 + u \end{aligned} \tag{2.1}$$

where $[y, Y]$ is a $T \times (G+1)$ matrix of endogenous variables, X_1 is a $T \times K_1$ matrix of exogenous variables, $Z \equiv [Y, X_1]$, $\alpha'_0 \equiv [\gamma'_0, \beta'_0]$, and u is a $T \times 1$ vector. We denote by X_2 the matrix of $K_2 (\equiv K - K_1)$ exogenous variables that are absent from the equation. Let us assume that Y admits a reduced-form representation:

$$Y = X\Pi_0 + V \tag{2.2}$$

where $X \equiv [X_1, X_2]$ is a $T \times K$ matrix,³ Π_0 is a $K \times G$ matrix of unknown parameters and V is a $T \times G$ matrix of unknown error terms. We now specify the data generating process.

Assumption 1. *The sequence $\{(u_t, V_t)\}$ is independent and identically dis-*

³In this paper to simplify the presentation, X is assumed to be fixed. It is possible to extend the results to the case of X random as in Kim and Muller (2001) for the Two-Stage Quantile Regression.

tributed where u_t and V_t are the t^{th} elements in u and V respectively.

Then, from eqs. 2.1 and 2.2, y has also a reduced-form representation:

$$y = X\pi_0 + v \quad (2.3)$$

where $\pi_0 \equiv \left[\Pi_0, \begin{pmatrix} I_{K_1} \\ 0 \end{pmatrix} \right] \alpha_0 \equiv H(\Pi_0)\alpha_0$ and $v \equiv u + V\gamma_0$. Equations 2.2 and 2.3 are used for a first-stage estimation that yields some estimators $\hat{\pi}$, $\hat{\Pi}$ respectively of π_0 , Π_0 . We start with first-stage estimators converging to the value of interest, as shows the next assumption.

Assumption 2. $T^{1/2}(\hat{\pi} - \pi_0) = O_p(1)$ and $T^{1/2}(\hat{\Pi} - \Pi_0) = O_p(1)$.

The Two-Stage M-estimator (2SM(q)) $\hat{\alpha}$ of α_0 is the solution to the following minimisation programme:

$$\min_{\alpha} S_T(\alpha, \hat{\pi}, \hat{\Pi}, q) \equiv \sum_{t=1}^T \rho(qy_t + (1-q)\hat{y}_t - \hat{X}'_t\alpha)$$

where ρ is a real-valued non-constant function, y_t and X'_t are the t^{th} elements in y and X respectively, $\hat{y}_t = X'_t\hat{\pi}$ and $\hat{X}'_t = X'_tH(\hat{\Pi})$. The combination weight q is a non-zero constant that will be used to reduce the variance of the estimator of α .

The 2SM(q) estimator can be alternatively defined as a solution to the pseudo-

likelihood equations□:

$$\mathcal{L}(\alpha) \equiv T^{-1/2} \sum_{t=1}^T H(\hat{\Pi})' X_t \psi(qy_t + (1-q)\hat{y}_t - \hat{X}_t' \alpha) = 0 \quad (2.4)$$

where ψ is a real-valued non-constant function. If ρ is strictly quasi-convex and $\psi = \rho'$, these definitions are equivalent. In general, ψ does not need to be a derivative of some function. Because of its generality, we will use the second definition of the 2SM(q) estimator.

Let us now consider a few examples of M-estimators:

(1) The least square estimator is such that $\rho(z) = \frac{1}{2}z^2$ and $\psi(z) = z$. Here, when $q = 1$, the 2SM(q) is the 2SLS estimator for which the first stage is implemented by using ordinary least squares with a set of instruments.

(2) The least absolute deviation estimator corresponds to $\rho(z) = \frac{1}{2}|z|$ and $\psi(z) = (\frac{1}{2} - 1_{[z \leq 0]})$.

(3) The quantile estimator is associated with $\rho(z) = z(\theta - 1_{[z \leq 0]})$ and $\psi(z) = \theta - 1_{[z \leq 0]}$. In our simulation study, we shall use the case of Two-Stage Quantile estimator when the first stage is the LS estimator (LSQR(θ, q)).

(4) The Huber estimator⁴ is defined by $\rho(z) = \frac{1}{2}z^2 1_{[|z| < k]} + (k|z| - \frac{1}{2}k^2) 1_{[|z| \geq k]}$ and $\psi(z) = z 1_{[|z| < k]} + 2k(\frac{1}{2} - 1_{[z \leq 0]}) 1_{[|z| \geq k]}$. We shall emphasize the case of the Double Huber Estimator (DH(k, q)), where the first and second stages are composed of the same Huber estimator, thereby ensuring the robustness of the global

⁴See Huber (1964, 1981).

estimation.

In the next section, we discuss the asymptotic representation of the 2SM(g) estimator $\hat{\alpha}$. We shall show that the following conditions are sufficient for the asymptotic representation.

Assumption 3. (i) $T^{-1} \sum_{t=1}^T X_t X_t' \rightarrow Q$ where Q is finite and positive-definite.

(ii) $H(\Pi_0)$ is of full column rank.

(iii) v_t has a density f . The cumulative density is denoted as F .

(iv) $E(\psi(v_t)) = 0$.

Assumptions 3(i)-3(ii) are standard. For least-square estimators or quantile regressions, Assumption 3(iv) is satisfied as soon as there is an intercept term in the linear model. In these cases, this condition ensures that the intercept term can be easily calculated and is innocuous. Because this condition will be imposed later on V_{jt} for $j = 1, \dots, G$, we need to examine it closely. Under regularity conditions, the sample roots of eq. 2.4 converge to their population analogues and are therefore consistent in that sense. However, these population analogues may have different meanings for different influence functions. If the influence function admits 3 (iv)* below, then in the cases of interest the meanings of the coefficients in all the variants are the same, except for the intercept.

Assumption 3 (iv)*. $E(\psi(v_t - c)) = 0$ admits a solution in c .

Assumption 3 (iv)* can be satisfied if the three following conditions are met: (i) ψ is monotonic, which is the case for many M-estimators, (ii) $\psi(v_t - c)$ takes at least one negative and one positive value on the support of v_t and (iii) between these two points the density of v_t does not cancel. This result can be derived from the mean value theorem.

Unfortunately, in general cases nothing ensures that Assumption 3 (iv)* is satisfied. This may happen because of the non-monotonicity of ψ , for example with asymptotic least square estimators, or because of the discontinuity of the support of v_t , for example when two separate and concentrated populations are mixed, or finally from the use of an influence function of constant sign. Then, we need to impose it in Assumption 3(iv)* that restricts the scope of 2SM(q) estimators under consideration. In that case, incorporating an intercept in the model makes it satisfy Assumption 3(iv). To be able to derive our asymptotic results, we further restrict the set of influence functions by adding the two following assumptions.

Assumption 4. (i) *The function ψ is of bounded variations in every interval, i.e., it can be written as*

$$\psi = \psi^+ - \psi^-$$

where ψ^\pm is monotone and further

$$\int_{-\infty}^{\infty} (\psi^\pm(x+h) - \psi^\pm(x-h))^2 f(x) dx = O(1)$$

as $h \rightarrow 0$ and

$$\sup \frac{1}{|h|} \left\{ \int_{-\infty}^{\infty} (\psi^\pm(x+q+h) - \psi^\pm(x+q)) f(x) dx : |q| \leq \epsilon, |h| \leq \epsilon \right\} < \infty$$

for some $\epsilon > 0$.

(ii) The function ψ is such that $E(\psi(qv_t)) = 0$ for the considered q .

Assumption 4(i) has been used in Bickel (1975) and, in many cases, Assumption 4(ii) is implied by Assumption 3 (iv). For example, Assumption 4(ii) is fulfilled for LS estimators and quantile regression estimators. It ensures that the transformation of the dependent variable does not perturb the scale of the parameter to estimate. Moreover, Assumption 4(ii) is satisfied for any arbitrary q when the influence function ψ is separable, that is: $\forall \lambda \neq 0$ and x , $\psi(\lambda x) = K(\lambda)\psi(x)$ for some function K . LS estimators and quantile regression estimators satisfy the above separability condition. The influence function of the Huber estimator does not satisfy the separability condition, but satisfies it for any symmetric density.

3. The Asymptotic Representation

The first step of the analysis is the derivation of an asymptotic representation of the $2SM(q)$ as an intermediate result from which we shall derive the asymptotic normality and the asymptotic variance of our estimator. We particularly need the formula of the asymptotic variance of the estimator of the parameter of interest to be able to reduce this variance by selecting a value for q .

The function $G_q(z) \equiv E[\psi(qv_t + z)]$ is crucial in the derivation of the asymptotic representation because the equation $G_q(0) = 0$ statistically defines the true structural parameter α_0 . We need that it satisfies the following regularity assumption for the next propositions.

Assumption 5. *For any $q \neq 0$, $G_q(z)$ is differentiable with respect to z in R : $g_q(z) \equiv G_q'(z)$. It is also assumed that $g_q(0) \neq 0$ and g_q is continuous at $z = 0$.*

Proposition 1. *Suppose that Assumptions 1-5 hold. Then, the $2SM(q)$ estimator $\hat{\alpha}$ has the asymptotic representation*

$$\begin{aligned} T^{1/2}(\hat{\alpha} - \alpha_0) &= Q_{zz}^{-1} H(\Pi_0)' \left\{ T^{-1/2} \sum_{t=1}^T X_t g_q(0)^{-1} \psi(v_t) \right. \\ &\quad \left. + (1-q)QT^{1/2}(\hat{\pi} - \pi_0) - QT^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0 \right\} + o_p(1). \end{aligned}$$

All the technical proofs are collected in Appendix B. The asymptotic repre-

sensation shows that the asymptotic distribution of the second stage estimator $T^{1/2}(\hat{\alpha} - \alpha_0)$ depends on the asymptotic distribution of the first stage estimators $T^{1/2}(\hat{\pi} - \pi_0)$ and $T^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0$. Naturally, for $q = 1$, the influence of $\hat{\pi}$ disappears. The asymptotic representation of the 2SM(q) estimator is composed of three additive terms. The first term does not perturb consistency under Assumption 3(iv) and represents the contribution of the second stage to the uncertainty of the estimator. The second and third terms represent the contributions of respectively $\hat{\pi}$ and $\hat{\Pi}$ to the uncertainty of the estimator.

The asymptotic representation can be extended to the more general case where the first step estimators $\hat{\pi}$ and $\hat{\Pi}$ converge towards inappropriate values. For this purpose we define the following assumption.

Assumption 2'. *There exist $|B_\pi| < \infty$ and $|B_\Pi| < \infty$ such that $T^{1/2}(\hat{\pi} - \pi_0 - B_\pi) = O_p(1)$ and $T^{1/2}(\hat{\Pi} - \Pi_0 - B_\Pi) = O_p(1)$.*

Using Assumption 2', we obtain the following asymptotic representation with a possible bias B_α .

Proposition 2. *Suppose that Assumptions 1, 2', 3-5 hold. Then, the 2SM(q) estimator $\hat{\alpha}$ has the asymptotic representation*

$$T^{1/2}(\hat{\alpha} - \alpha_0 - B_\alpha) = R\{T^{-1/2} \sum_{t=1}^T X_t g_q(0)^{-1} \psi(v_t) + (1-q)QT^{1/2}(\hat{\pi} - \pi_0 - B_\pi)$$

$$-QT^{1/2}(\hat{\Pi} - \Pi_0 - B_{\Pi}\gamma_0) + o_p(1)$$

where $B_{\alpha} \equiv RQ\{(1 - q)B_{\pi} - B_{\Pi}\gamma_0\}$, $R = Q_{zz}^{*-1}H(\Pi_0^*)'$,

$$Q_{zz}^* \equiv H(\Pi_0^*)'QH(\Pi_0^*) \text{ and } \Pi_0^* \equiv \Pi_0 + B_{\Pi}.$$

Example 1. DH(k, q): In that case, $g_q(0) = F(kq^{-1}) - F(-kq^{-1})$ where F is the cdf of v_t . Because the same estimators are used with the same k for the first and the second stages, and since the distribution of v_t is symmetric, we have here $B_{\pi} = 0$ and $B_{\Pi} = 0$.

Example 2. LSQR(θ, q) : Here, $g_q(0) = qf(0)^{-1}$ where f is the pdf of v_t . $B_{\Pi} \equiv [E(V_t)', 0', \dots, 0']'_{(K \times G)}$ and $B_{\pi} \equiv [E(v_t)', 0', \dots, 0']'_{(K \times 1)}$.

4. Asymptotic Normality and Covariance Matrix

4.1. With First Step Estimators Converging Towards Parameters of Interest

To obtain the asymptotic representations of the 2SM(q) estimator $\hat{\alpha}$ in Section 3, it was sufficient to impose Assumption 2 or 2' regarding the preliminary estimators $\hat{\pi}$ and $\hat{\Pi}$. We now derive the first stage estimation procedure so as to explicitly derive the asymptotic representations of $T^{1/2}(\hat{\pi} - \pi_0)$ and $T^{1/2}(\hat{\Pi} - \Pi_0)$, which are to be substituted into the asymptotic representation of $T^{1/2}(\hat{\alpha} - \alpha_0)$. From now, we restrict our attention to the class of M-estimators in the first stage. The

first-step estimators, $\hat{\pi}$ and $\hat{\Pi}$, are defined as follows:

$$T^{-1/2} \sum_{t=1}^T X_t \psi_{\pi}(y_t - X_t' \hat{\pi}) = 0$$

$$T^{-1/2} \sum_{t=1}^T X_t \psi_{\Pi_j}(Y_{jt} - X_t' \hat{\Pi}_j) = 0$$

for $j = 1, 2, \dots, G$ where $Y_j \equiv (Y_{j1}, \dots, Y_{jT})'$ is the j^{th} column in Y , $\hat{\Pi}_j$ is the j^{th} column in $\hat{\Pi}$ and ψ_{π}, ψ_{Π_j} are the influence functions of the first-step estimators.

Note that the influence functions ψ_{π}, ψ_{Π_j} in the first step are not necessarily the same as the one ψ used in the second step. However, they are characterised by the following assumptions, similar to the ones stated for ψ and that we use for deriving the asymptotic representations of the first stage estimators of Lemma 1.

Assumption 6. (i) V_{jt} has a differentiable density h_j for $j = 1, 2, \dots, G$. The cumulative density of V_{jt} is denoted as H_j .

(ii) $E(\psi_{\pi}(v_t)) = 0$ and $E(\psi_{\Pi_j}(V_{jt})) = 0$ for $j = 1, 2, \dots, G$.

(iii) The influence functions ψ_{π}, ψ_{Π_j} ($j = 1, 2, \dots, G$) satisfy the conditions in Assumption 4(i).

(iv) $G_{\pi}(z) = E[\psi_{\pi}(v_t + z)]$ is differentiable with respect to z in R : $g_{\pi}(z) \equiv G'_{\pi}(z)$ (ii) $G_{\Pi_j}(z) = E[\psi_{\Pi_j}(V_{jt} + z)]$ is differentiable with respect to z in R : $g_{\Pi_j}(z) \equiv G'_{\Pi_j}(z)$ for $j = 1, 2, \dots, G$. It is also assumed that (i) $g_{\pi}(0) \neq 0$ and $g_{\Pi_j}(0) \neq 0$ (ii) g_{π} and g_{Π_j} are continuous at $z = 0$ for $j = 1, 2, \dots, G$.

Lemma 1. *The asymptotic representation of the first-step estimators*

Suppose that Assumptions 1-3 and 6 hold. Then,

$$T^{1/2}(\hat{\pi} - \pi_0) = Q^{-1}T^{-1/2} \sum_{t=1}^T X_t g_{\pi}(0)^{-1} \psi_{\pi}(v_t) + o_p(1). \quad (4.1)$$

$$T^{1/2}(\hat{\Pi}_j - \Pi_{0j}) = Q^{-1}T^{-1/2} \sum_{t=1}^T X_t g_{\Pi_j}(0)^{-1} \psi_{\Pi_j}(V_{jt}) + o_p(1).$$

Lemma 1 is a straightforward consequence of Proposition 1. We now show the asymptotic normality of the 2SM(q). For this purpose, we apply the Liapounov CLT, for which we need the following additional assumption.

Assumption 7. (i) *There exists a positive constant Δ such that $\|X_t\| \leq \Delta < \infty$ for all t .*

(ii) *There exist positive constants δ and Δ such that $0 < E|\eta_t|^{2+\delta} < \Delta < \infty$*

where $\eta_t \equiv g_q(0)^{-1} \psi(v_t) + (1-q)g_{\pi}(0)^{-1} \psi_{\pi}(v_t) - \xi_t$ and

$$\xi_t \equiv \left[g_{\Pi_1}(0)^{-1} \psi_{\Pi_1}(V_{1t}), \dots, g_{\Pi_G}(0)^{-1} \psi_{\Pi_G}(V_{Gt}) \right] \gamma_0.$$

Proposition 3. *Suppose that Assumptions 1-7 hold. Then,*

$$T^{1/2}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, \sigma_0^2 Q_{zz}^{-1})$$

where $\sigma_0^2 \equiv E(\eta_t^2)$.

Note that the term σ_0^2 in the asymptotic covariance of the 2SM(q) estimator is a function of q . Thus, a researcher may choose some value for q based on his

own subject prior or experience. She can also minimise σ_0^2 with respect to q and use the optimal value in the estimation.

Example 1. DH(k, q): Here, $\psi_\pi = \psi_{\Pi_j} = \psi$. Then, $g_\pi(0) = F(k) - F(-k)$ and $g_{\Pi_j}(0) = H_j(k) - H_j(-k)$, where H_j is the cdf of V_j .

4.2. With First Step Estimators Converging towards Inappropriate Values

Assumption 6 (ii) has been used to obtain the consistency of the 2SM(q) estimator towards the values of interest. Depending on the type of data to be studied and the type of estimation method to be used, Assumption 6 (ii) might be too restrictive. Clearly, the consistency will be obtained as long as the first stage predictors are consistent for $E(Y)$ and $E(y)$. In this subsection we investigate the possibility of relaxing Assumption 6 (ii) by allowing first-stage estimators converging towards inappropriate values of the parameters. This type of situation may occur when the choices of the first stage and second stage estimators are done separately.

To simplify the analysis, we rewrite eq. 2.1 so as to put the constant term in first position. This yields

$$\begin{aligned} y &= X_1\beta_0 + Y\gamma_0 + u \\ &\equiv Z\alpha_0 + u \end{aligned}$$

with a new ordering for the components of Z and α_0 . Then, the matrix $H(\Pi_0)$

is now equal to

$$H(\Pi_0) \equiv \begin{bmatrix} I_{K_1} & \\ & \Pi_0 \\ 0 & \end{bmatrix}.$$

We need the following assumption similar to the one discussed in Section 2.

Assumption 6. (ii)' *The equations $E(\psi_\pi(v_t - c)) = 0$ and $E(\psi_{\Pi_j}(V_{jt} - c_j)) = 0$ admit a solution for c and c_j for $j = 1, 2, \dots, G$.*

If we consider a situation where $E(\psi_\pi(v_t)) = \mu \neq 0$ and $E(\psi_{\Pi_j}(V_{jt})) = \mu_j \neq 0$, for $j = 1, 2, \dots, G$, then, a solution for c and c_j can be obtained as a function of μ, F, μ_j and H_j . For example, with the influence function for quantile regressions, it can be shown that $c = F^{-1}(\mu + F(0))$ and $c_j = H_j^{-1}(\mu_j + H_j(0))$. Let $c = c^*(\mu, F)$ and $c_j = c_j^*(\mu_j, H_j)$ be the solutions as in Assumption 6 (ii)'

First, we define $V_t^* \equiv V_t - C$ where $C = (c_1, \dots, c_G)$ and $v_t^* \equiv v_t - c$. Then, the reduced forms for Y_t and y_t in (2.2) and (2.3) can be expressed as

$$Y_t = X_t' \Pi_0^* + V_t^* \tag{4.2}$$

where $\Pi_0^* \equiv \Pi_0 + B_\Pi$ and $B_\Pi \equiv [C', 0, \dots, 0]'_{(K \times G)}$

$$y_t = X_t' \pi_0^* + v_t^* \tag{4.3}$$

where $\pi_0^* \equiv \pi_0 + B_\pi$ and $B_\pi \equiv [c, 0, \dots, 0]'_{(K \times 1)}$. By construction, $E(\psi(V_{jt}^*)) = E(\psi(v_t^*)) = 0$. Let $\tilde{\Pi}$ and $\tilde{\pi}$ be M-estimators based on eqs. 4.2 and 4.3 and let $\tilde{\alpha}$ be a 2SM(q) estimator based on the M-estimators $\tilde{\Pi}$ and $\tilde{\pi}$ in the first step. Then, the asymptotic representation for the 2SM(q) estimator based on $\tilde{\Pi}$ and $\tilde{\pi}$ is given by Proposition 2 where $(\hat{\Pi}, \hat{\pi})$ is substituted with $(\tilde{\Pi}, \tilde{\pi})$. The asymptotic normality of $\tilde{\alpha} - \alpha_0 - B_\alpha$ can be easily derived from this result, provided that the following assumptions are satisfied.

Assumption 7. (ii)' *There exist positive constants δ and Δ such that $0 < E|\eta_t^*|^{2+\delta} < \Delta < \infty$ where $\eta_t^* \equiv g_q(0)^{-1}\psi(v_t) + (1-q)g_\pi(0)^{-1}\psi_\pi(v_t^*) - \xi_t^*$ and $\xi_t^* \equiv [g_{\Pi_1}(0)^{-1}\psi_{\Pi_1}(V_{1t}^*), \dots, g_{\Pi_G}(0)^{-1}\psi_{\Pi_G}(V_{Gt}^*)] \gamma_0$.*

Then, we have the following asymptotic normality for the 2SM(q) estimator.

Proposition 4. *Suppose that Assumptions 1, 2', 3-5, 6(i), (ii)', (iii), (iv) and 7(i), (ii)' hold. Then,*

$$T^{1/2}(\tilde{\alpha} - \alpha_0 - B_\alpha) \xrightarrow{d} N(0, \sigma_0^2 Q_{zz}^{*-1})$$

where $\sigma_0^2 \equiv E(\eta_t^{*2})$, $B_\alpha \equiv Q_{zz}^{*-1}H(\Pi_0^*)'Q\{(1-q)B_\pi - B_{\Pi}\gamma_0\}$ which affects only the intercept coefficient, $Q_{zz}^* \equiv H(\Pi_0^*)'QH(\Pi_0^*)$ and $\Pi_0^* \equiv \Pi_0 + B_{\Pi}$.

The proof is similar to the one for Proposition 3. Hence, it is omitted. Since the asymptotic bias affects the intercept coefficient only, it may be useful to

separately investigate the asymptotic properties of the slope coefficient. This is done in the appendix.

Example 1. LSQR(θ, q) : Here, $\psi_\pi(z) = \psi_{\Pi_j}(z) = z$.

Then, $g_\pi(0) = g_{\Pi_j}(0) = 1$, $\eta_t^* \equiv qf(0)^{-1}\psi(v_t) + u_t - q(v_t - E(v_t))$ and $\sigma_0^2 = E(\eta_t^*)^2$.

5. Monte Carlo Simulations

In order to illustrate the impact of the choice of q for the accuracy of the 2SM(q), we conduct a few simulation experiments. We first consider cases where the M-estimators for the first and second stages are the same: least squares (DLS(q)), least absolute deviations (DLAD(q)), quantile (DQ(θ, q)) and Huber (DH(k, q)) estimators. When the regression model is correctly specified, each estimator normalised by $T^{1/2}$ and centred to the true value of the parameters follows asymptotically a normal distribution with asymptotic covariance of the form $\sigma_0^2 Q_{zz}^{-1}$. The covariance matrices of the estimators differ only by the scalar term σ_0^2 . The efficiency of each estimator depends on the distributions of v_t, u_t and V_{jt} , and on the shapes of the influence functions ψ, ψ_π and ψ_{Π_j} . The optimal q is not determined for DLAD(q) or DQ(θ, q) since σ_0^2 does not depend on q for those estimators under perfect identification. For DLS(q), $q^* = 1 + E(u_t, v_t)/E(v_t^2)$ and for DH(k, q), $q^* = E(\tilde{u}_t \tilde{v}_t)/E(\tilde{v}_t^2)$ and $q^* \neq 0$ where $\tilde{v}_t = g_\pi(0)^{-1}\psi(v_t)$ and $\tilde{u}_t = \{g_q(0)^{-1} + g_\pi(0)^{-1}\}\psi(v_t) - \xi_t$.

We examine the performance of DLS(q) and DH(5, q) in terms of the value of $\sigma_0^2(q)$ obtained for a range of values of q . We choose the value of γ_0 equal to 1 and

we only introduce one endogenous independent variable. The error terms v_t and V_t are chosen to follow first a bivariate normal law for a first set of simulations, then a bivariate Student law $t(5)$. Each set of simulations is performed for the following values of variances and correlations of the two error terms: $\sigma_{v_t} = 1, 5$; $\sigma_{V_t} = 1, 5$; $\text{corr}(v_t, V_t) = -0.5, 0, 0.5$. Other values for these parameters have been tried and yield qualitatively similar results. The number of replication for each simulation is 5,000.

The curves of $\sigma_0^2(q)$ presented in Figure 1 summarise the results of the simulations. Clearly, the choice of parameter q is crucial and can considerably improve or degrade the accuracy of the estimation. Almost always, the usual value $q = 1$ appears a severely suboptimal choice. Interestingly, cases where a negative value of q should be chosen appears to be quite possible as shown in some of the $\sigma_0^2(q)$ for $\text{DLS}(q)$ ⁵. It is the case for $\text{DLS}(q)$ in all situations when the errors are negatively correlated. By contrast, with Gaussian errors, $\text{DH}(5, q)$ corresponds to values of q^* well above 1, more so when error terms v_t and V_t are negatively correlated, than for positive correlations. Although the slope of the curve $\sigma_0^2(q)$ is not very sensitive to the chosen distribution (Gaussian or Student- t), the level of $\sigma_0^2(q)$ is generally higher for Student errors. The correlation between the two errors strongly affects the value of q^* . The more positive the correlation, the smaller q^* for $\text{DH}(5, q)$, but the larger q^* for $\text{DLS}(q)$. Increasing σ_{V_t} (respectively σ_{v_t})

⁵Note that $\sigma_{v_t} = 5$, $\sigma_{V_t} = 1$, $\text{corr}(v_t, V_t) = 0.5$, the optimal choice q^* is actually close to 0.1 and quite apart from 0.

makes the role of the correlation less strong (respectively stronger) for $\text{DH}(5, q)$, while the converse phenomenon appears with $\text{DLS}(q)$. Naturally, large σ_{v_t} or large σ_{V_t} can damage the accuracy of the $2\text{SM}(q^*)$ estimation.

The graphs in Figure 1 show how much we can reduce the asymptotic variance of $2\text{SM}(q)$ estimators if we knew the optimal q . In practice, the optimal q is not known. Because q^* is always expressed in terms of population expectations, it can be consistently estimated by using empirical means at the place of their expectations. Thus, a consistent estimator for the optimal value of the $\text{DLS}(q)$ estimator is given by

$$\hat{q}_{LS} = 1 + \frac{\sum_{t=1}^T \hat{v}_t \hat{u}_t}{\sum_{t=1}^T \hat{v}_t^2}$$

where \hat{v}_t and \hat{u}_t are respectively the residuals from the structural and reduced-form regressions. Similarly, a consistent estimator for the optimal value for the $\text{DH}(k, q)$ estimator is obtained by

$$\hat{q}_H = \frac{\sum_{t=1}^T \tilde{v}_t \tilde{u}_t}{\sum_{t=1}^T \tilde{v}_t^2}$$

where $\tilde{v}_t \equiv \hat{g}_\pi(0)^{-1} \psi(\hat{v}_t)$, $\tilde{u}_t \equiv \{\hat{g}_q(0)^{-1} + \hat{g}_\pi(0)^{-1}\} \psi(\hat{v}_t) - \hat{\xi}_t$ and

$\hat{\xi}_t \equiv [\hat{g}_{\Pi_1}(0)^{-1} \psi(\hat{V}_{1t}), \dots, \hat{g}_{\Pi_G}(0)^{-1} \psi(\hat{V}_{Gt})] \gamma_0$. Here, $\hat{g}_q(0)$, $\hat{g}_\pi(0)$ and $\hat{g}_{\Pi_j}(0)$ can be based on a kernel density estimation method. Consistent estimators for q^* for any general $2\text{SM}(q)$ estimator can be obtained in the same way. If a researcher has no prior information about the optimal combination weight q^* , then these consistent data-dependent choices can be used. This procedure not

only removes the problem of an arbitrary choice of q but also provides a basis for obtaining a consistent covariance matrix with maximum efficiency in large samples.

However, it is unclear what is the impact on the global estimation of the inaccuracy in the estimation of q , especially for finite samples. We now examine this question by focusing on the LSQR(θ, q) estimator. Here, the values of the parameters are $\gamma_0 = 0.5, \beta_{00} = 1, \beta_{01} = 0.2$. We do not comment the estimator of the intercept coefficient that is asymptotically biased and not very interesting. The small sample bias and the accuracy of the estimators of the two other parameters elicit the same type of properties across the simulation trials, and we discuss them together. Our main interest is to compare the cases $q = 1$ and $q = \hat{q}$ (estimated from the data).

The case $q = 1$ is the benchmark case since it corresponds to the usual estimation procedures. Here, $\hat{q} = \frac{\frac{1}{T} \sum_{t=1}^T \left[(\hat{v}_t - \frac{1}{T} \sum_{s=1}^T \hat{v}_s) \hat{u}_t \right] - \frac{1}{T} \sum_{t=1}^T [\hat{f}(0)^{-1} \psi_\theta(\hat{v}_t) \hat{u}_t]}{\frac{1}{T} \sum_{t=1}^T \left[\hat{f}(0)^{-1} \psi_\theta(\hat{v}_t) - (\hat{v}_t - \frac{1}{T} \sum_{s=1}^T \hat{v}_s) \right]^2}$ is estimated from the residuals obtained from a preliminary LSQR($\theta, 1$) estimation of the model and where $\hat{f}(0)$ is the kernel estimator based on the residuals \hat{v}_t . Three distributions have been used: $N(0, 1), t(3), LN(0, 1)$, with 1000 replications for each simulation set. The correlation of V_t and v_t is fixed at 0.2.

The chosen values of θ are: 0.05, 0.25, 0.5, 0.75 and 0.95, so as to distinguish extreme and central quantiles. Finally, five sample sizes have been tried: 50, 100, 300, 500 and 1000. In some cases, the results for 50 and 100 observations yield too inaccurate estimates to consider them well adapted to many empirical purposes.

In these cases, the comparison of $\text{LSQR}(\theta, 1)$ and $\text{LSQR}(\theta, \hat{q})$ makes little sense and we focus on the cases of 300 and 500 observations in our comment. Table 1 shows the results corresponding to 100, 300 and 500 observations.

With normal and Student errors, using 300 observations yields relatively accurate estimates for γ_0 and β_{01} , which are our parameters of interest. In general $\text{LSQR}(\theta, \hat{q})$ is more concentrated and less biased than $\text{LSQR}(\theta, 1)$: $\text{LSQR}(\theta, \hat{q})$ clearly dominates $\text{LSQR}(\theta, 1)$. Estimated q are small for normal errors (between -0.05 and 0.08), but not necessarily for Student errors (between 0.04 and 0.65). They are nonetheless always very different from 1. Using 500 and 1000 observations increases the advantage of using \hat{q} over using $q = 1$. In contrast, with lognormal errors even a sample size of 300 is not enough to obtain high quality estimates with any method. Moreover, in that case $\text{LSQR}(\theta, \hat{q})$ and $\text{LSQR}(\theta, 1)$ cannot be clearly ranked in terms of their accuracy and their small sample bias. This is partly due to the fact that for small or average quantiles ($\theta \leq 0.5$), \hat{q} is close to 1 and both estimators are close. With 500 and 1000 observations, this ambiguity disappears and $\text{LSQR}(\theta, \hat{q})$ again dominates $\text{LSQR}(\theta, \hat{q})$ for all quantiles, much more when \hat{q} is far from 1.

Figure 2 shows the empirical distribution of the \hat{q} for the 3 distributions and the 5 quantiles in the case of 300 observations. In all cases the density estimates are remarkably symmetric and well concentrated. They suggest that the estimates of \hat{q} are satisfactory for our purpose in the studied cases and clearly apart from 1. With normal errors, \hat{q} is close to zero, while the general estimation results

are not very sensitive to the value of q chosen close to zero. With Student and Lognormal errors \hat{q} is also close to zero for extreme quantiles.

On the whole, the studied simulation results are encouraging in that a preliminary estimation of q enables us to increase the accuracy of the estimation, sometimes substantially, as soon as the sample size is sufficient to produce estimates accurate enough to be useful. However, when the sample is too small to produce accurate estimates with $q = 1$, using $q = \hat{q}$ instead does not help to increase the accuracy of the estimation and may even reduce it.

6. Conclusion

We consider in this article Two-Stage M-estimators where the dependent variable in the second stage is of the composite form $qy_t + (1 - q)X_t'\hat{\beta}$ with $\hat{\beta}$ the first stage estimator, $q \neq 0$, and X_t the vector of independent variables in the first stage. The case $q = 1$ is the usual one for the two-stage estimators in the literature. This approach is useful because an appropriate choice of q may improve the accuracy of the estimation. To explore this question, we derive the formula of the asymptotic covariance matrix for the parameters of interest, which depends on q . Then, it is possible to select an optimal value q^* that minimises this covariance matrix.

We distinguish the first-step estimators that converge towards inappropriate values of the parameters. This distinction is necessary if one wish freely choose the estimation method used in the first step. In that case, we show that the intercept coefficient of the model of interest can be asymptotically biased and we

give the formula of this bias. However, we also show that the slope coefficients are unbiased. Again, the value of q matters for the accuracy.

Finally, Monte Carlo simulation results show that the selection of q has generally a dramatic impact on the accuracy of the $2SM(q)$ estimator even for finite samples. Moreover, in these results a preliminary estimation of q improves the accuracy of the estimation, provided that the sample size is sufficient to produce estimates accurate enough to be useful.

All these results suggest that the reformulation of the dependent variable by combining it with a preliminary prediction is a fertile approach to variance reduction. Then, more general functional forms and procedures than what has been explored in this paper could be investigated to generate other variance reduction methods.

1 Appendix A: Properties of the slope coefficient

We first decompose $\tilde{\Pi} = \begin{bmatrix} \tilde{\Pi}_{(1)} \\ \tilde{\Pi}_{(2)} \end{bmatrix}$ where $\tilde{\Pi}_{(1)}$ is the first $1 \times G$ row and $\tilde{\Pi}_{(2)}$ is the remaining $(K-1) \times G$ matrix and $\tilde{\pi} = \begin{bmatrix} \tilde{\pi}_{(1)} \\ \tilde{\pi}_{(2)} \end{bmatrix}$ where $\tilde{\pi}_{(1)}$ is the first element and $\tilde{\pi}_{(2)}$ is the remaining $(K-1) \times 1$ vector. Hence, $\tilde{\Pi}_{(2)}$ and $\tilde{\pi}_{(2)}$ contain only the slope coefficients. We also decompose $Q^{-1} = \begin{bmatrix} \bar{Q}_1 \\ \bar{Q}_2 \end{bmatrix}$ where \bar{Q}_1 is the first $1 \times K$ row and \bar{Q}_2 is the remaining $(K-1) \times K$ matrix. Then, it is straightforward to see that

$$T^{1/2}(\tilde{\Pi}_{j(2)} - \Pi_{j0(2)}) = \bar{Q}_2 T^{-1/2} \sum_{t=1}^T X_t g_{\Pi_j}(0)^{-1} \psi_{\Pi_j}(V_{jt}^*) + o_p(1) \quad (1)$$

$$T^{1/2}(\tilde{\pi}_{(2)} - \pi_{0(2)}) = \bar{Q}_2 T^{-1/2} \sum_{t=1}^T X_t g_{\pi}(0)^{-1} \psi_{\pi}(v_t^*) + o_p(1) \quad (2)$$

where $\Pi_{0(2)}$ and $\pi_{0(2)}$ are the corresponding slope components of Π_0 and π_0 respectively.

We decompose $\tilde{\alpha} = \begin{bmatrix} \tilde{\alpha}_{(1)} \\ \tilde{\alpha}_{(2)} \end{bmatrix}$ where $\tilde{\alpha}_{(1)}$ is the first element and $\tilde{\alpha}_{(2)}$ is the remaining $(K_1 + G - 1) \times 1$ vector and $\alpha_0 = \begin{bmatrix} \alpha_{0(1)} \\ \alpha_{0(2)} \end{bmatrix}$ likewise. Using eqs. 1, 2 and the asymptotic representation in Proposition 2, the following lemma can be easily proved.

Lemma 2: *Suppose that the same assumptions as in Proposition 4 except γ (ii)' hold. Then, $T^{1/2}(\tilde{\alpha}_{(2)} - \alpha_{0(2)}) = R_2 T^{-1/2} \sum_{t=1}^T X_t \eta_t^* + o_p(1)$ where R_2 is the last $(K_1 + G - 1)$ rows in R , $\eta_t^* \equiv g_q(0)^{-1} \psi(v_t) + (1 - q)g_{\pi}(0)^{-1} \psi_{\pi}(v_t^*) - \xi_t^*$ and $\xi_t^* \equiv [g_{\Pi_1}(0)^{-1} \psi_{\Pi_1}(V_{1t}^*), \dots, g_{\Pi_G}(0)^{-1} \psi_{\Pi_G}(V_{Gt}^*)] \gamma_0$. From this lemma, the asymptotic normality of the slope coefficients is easily derived.*

Proposition 5: *Suppose that the same assumptions as in Proposition 4 hold. Then,*

$$T^{1/2}(\tilde{\alpha}_{(2)} - \alpha_{0(2)}) \xrightarrow{d} N(0, \sigma_0^2 R_2 Q R_2')$$

where $\sigma_0^2 \equiv E(\eta_t^{*2})$.

Proof of Lemma 2: First, note that the representation in (1) can be used to show that

$$T^{1/2}(\tilde{\Pi}_{(2)} - \Pi_{0(2)})\gamma_0 = \bar{Q}_2 T^{-1/2} \sum_{t=1}^T X_t \xi_t^* + o_p(1) \quad (3)$$

where ξ_t^* is given in Assumption 7 (ii)'.

Next, we decompose the matrix Q in the same way: $Q \equiv [Q_1 \quad Q_2]$ where Q_1 is a $K \times 1$ matrix and Q_2 is a $K \times (K - 1)$ matrix. Then, as discussed in Kim and Muller (2001), one can show that $RQ = \begin{bmatrix} 1 & R_1 Q_2 \\ 0 & R_2 Q_2 \end{bmatrix}$. Hence,

we have $B_\alpha = \begin{bmatrix} c_\alpha \\ 0 \end{bmatrix}$ where $c_\alpha = (1 - q)c - C\gamma_0$. Using these results, the representation from Proposition 2 can be decomposed as follows.

$$\begin{aligned} & \begin{bmatrix} T^{1/2}(\tilde{\alpha}_{(1)} - \alpha_{0(1)} - c_\alpha) \\ T^{1/2}(\tilde{\alpha}_{(2)} - \alpha_{0(2)}) \end{bmatrix} = \begin{bmatrix} R_1 T^{-1/2} \sum_{t=1}^T X_t g_q(0)^{-1} \psi(v_t) \\ R_2 T^{-1/2} \sum_{t=1}^T X_t g_q(0)^{-1} \psi(v_t) \end{bmatrix} \\ & + \begin{bmatrix} 1 & R_1 Q_2 \\ 0 & R_2 Q_2 \end{bmatrix} (1 - q) \begin{bmatrix} T^{1/2}(\tilde{\pi}_{(1)} - \pi_{0(1)} - c) \\ T^{1/2}(\tilde{\pi}_{(2)} - \pi_{0(2)}) \end{bmatrix} \\ & - \begin{bmatrix} 1 & R_1 Q_2 \\ 0 & R_2 Q_2 \end{bmatrix} \begin{bmatrix} T^{1/2}(\tilde{\Pi}_{(1)} - \Pi_{0(1)} - C)\gamma_0 \\ T^{1/2}(\tilde{\Pi}_{(2)} - \Pi_{0(2)})\gamma_0 \end{bmatrix} + o_p(1), \end{aligned}$$

from which we extract the only slope estimator $\tilde{\alpha}_{(2)}$. The asymptotic representation for the slope estimator together with (2) and (3) delivers the desired result. *QED*.

2 Appendix B: Proofs of the Propositions

Proof of Proposition 1:

We consider the following data generating process that is deduced from eq. 2.3:

$$\tilde{y}_t = \tilde{X}'_t \alpha_0 + \tilde{\epsilon}_t \quad (4)$$

where $\tilde{y}_t \equiv qy_t + (1 - q)\hat{y}_t$. A simple algebra shows that $\tilde{\epsilon}_t = qv_t$. Under Assumption 4(ii), we have $E(\psi(\tilde{\epsilon}_t)) = 0$.

The M-estimation of α_0 cannot be obtained from eq. 4 because both the dependent and independent variables include unknown auxiliary parameters (π_0, Π_0) . Nevertheless, we shall show that most asymptotic results for the M-estimator can be derived with slight modifications when replacing π_0, Π_0 with consistent and asymptotically normal estimators $\hat{\pi}, \hat{\Pi}$. Eq. 4 allows the direct application of Bickel's (1975) results from which we derive the asymptotic representation of the 2SM(q). Indeed, all the conditions for Lemma 4.1 in Bickel (1975) are satisfied: (i) $\tilde{\epsilon}_t$ is i.i.d. (by Assumption 1), (ii) $T^{-1} \sum_{t=1}^T \tilde{X}_t \tilde{X}'_t \rightarrow Q_{zz} \equiv H(\Pi_0)' Q H(\Pi_0)$ (by Assumption 3(ii)) and Q_{zz} is positive-definite (by Assumptions 3(i)-(ii)), (iii) $\max_{t,k} T^{-1/2} | \tilde{X}_{tk} | \rightarrow 0$ (as a consequence of Assumption 3(i)), (iv) ψ satisfies Conditions A and C1 in Bickel (1975) by Assumptions 3(iv) and 4(i).

In order to apply Bickel's lemma, we define

$$M_T(\Delta) \equiv T^{-1/2} \sum_{t=1}^T X_t \psi(\tilde{\epsilon}_t - T^{-1/2} X_t' \Delta)$$

where Δ is a $K \times 1$ vector. A direct application of Bickel's lemma yields the following lemma.

Lemma 3: *Suppose that Assumptions 1 and 3-5 hold. Then, for any $L > 0$,*

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

where $\|\Delta\| \equiv (\Delta' \Delta)^{1/2}$.

Proof of Lemma 3: Since the proof is almost identical to the ones in Bickel (1975) and Ruppert and Carroll (1980), we just show the new derivation of the limit of $E[M_T(\Delta) - M_T(0)]$.

$$\begin{aligned} E[M_T(\Delta) - M_T(0)] &= T^{-1/2} \sum_{t=1}^T X_t \left\{ E[\psi(v_t - T^{-1/2} X_t' \Delta)] - E[\psi(v_t)] \right\} \\ &= T^{-1/2} \sum_{t=1}^T X_t \left\{ G_q(-T^{-1/2} X_t' \Delta) - G_q(0) \right\} \\ &= -T^{-1} \sum_{t=1}^T X_t X_t' g_q(\xi_T) \Delta \end{aligned}$$

where the last line is due to Assumption 5 and the mean value theorem. Since ξ_T is between 0 and $-T^{-1/2} X_t' \Delta$, ξ_T converges to zero in probability. Hence, we have that

$$T^{-1} \sum_{t=1}^T X_t X_t' g_q(\xi_T) \Delta \rightarrow -g_q(0)Q\Delta$$

since g_α is continuous at $z = 0$ by Assumption 5. *QED.*

Lemma 3 can be derived from Lemma 4.1 in Bickel (1975) by replacing $E[M_T(\Delta) - M_T(0)]$ with its limit $-g_q(0)Q\Delta$. The following lemma is an extension of Lemma 3 which will be used for the asymptotic representation of the 2SM(q) estimator.

Lemma 4: *Suppose that Assumptions 1 and 3-5 hold. Then, for any $L > 0$,*

$$\sup_{\|\delta\| \leq L} \|M_T(\hat{\Delta}(\delta)) - M_T(0) + g_q(0)Q\hat{\Delta}(\delta)\| = o_p(1)$$

where $\hat{\Delta}(\delta) \equiv \hat{\delta}_1 \delta + \hat{\delta}_2$ and $\hat{\delta}_1$ and $\hat{\delta}_2$ are $O_p(1)$ -variables.

Proof of Lemma 4: The detailed proof can be found in Kim and Muller (2001).

We combine Lemmas 3-4 and Assumption 2 to prove Proposition 1.

We define for $\|\delta\| \leq L_1$, $\hat{\Delta}_1(\delta) \equiv H(\hat{\Pi})\delta + \hat{\Delta}_0$, where $\delta \in R^{G+K_1}$ and $\hat{\Delta}_0 \equiv -(1-q)\sqrt{T}(\hat{\pi} - \pi_0) + T^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0$. Since $\hat{\Delta}_0 = O_p(1)$ by Assumption 2, Lemma 2 implies that

$$\sup_{\|\delta\| \leq L_1} \|M_T(\hat{\Delta}_1(\delta)) - M_T(0) + g_q(0)Q\hat{\Delta}_1(\delta)\| = o_p(1) \quad (5)$$

for any $L_1 > 0$. Next, we define $\hat{\Delta} \equiv T^{1/2}(\hat{\alpha} - \alpha_0)$. Then, one can show:

$$M_T(\hat{\Delta}_1(\hat{\Delta})) = o_p(1) \quad (6)$$

because $H(\hat{\Pi})M_T(\hat{\Delta}_1(\hat{\Delta})) = \mathcal{L}(\hat{\alpha})$. Indeed, $H(\hat{\Pi})$ is bounded in probability and $\mathcal{L}(\hat{\alpha})$ is $o_p(1)$ by our definition of the 2SM estimator in eq. 2.4 in the text. Hence, one can show using Lemma 5.2 in Jurečková (1977) that the results in (5) and (6) together imply that

$$\hat{\Delta} = O_p(1). \quad (7)$$

The final step in deriving the asymptotic distribution of the 2SM(q) estimator $\hat{\alpha}$ is to combine the results in eq. 2.4 in the text and (6) to obtain

$$g_q(0)Q\hat{\Delta}_1(\hat{\Delta}) = M_T(0) + o_p(1). \quad (8)$$

By rearranging terms in (8), we have the asymptotic representation for the 2SM(q) estimator:

$$\begin{aligned} T^{1/2}(\hat{\alpha} - \alpha_0) &= Q_{zz}^{-1}H(\Pi_0)' \left\{ T^{-1/2} \sum_{t=1}^T X_t g_q(0)^{-1} \psi(v_t) \right. \\ &\quad \left. + (1-q)QT^{1/2}(\hat{\pi} - \pi_0) \right. \\ &\quad \left. - QT^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0 \right\} + o_p(1). \quad QED. \end{aligned}$$

Proof of Proposition 2: The proof is similar to that of Proposition 1 with $\hat{\Delta}_1(\delta) \equiv H(\hat{\Pi})\delta - (1-q)T^{1/2}(\hat{\pi} - \pi_0 - B_\pi) + T^{1/2}(\hat{\Pi} - \Pi_0 - B_\Pi)\gamma_0$ and $\hat{\Delta} \equiv T^{1/2}(\hat{\alpha} - \alpha_0 - B_\alpha)$. *QED.*

Proof of Proposition 3: Lemma 1 shows the asymptotic representations for the first step estimators $T^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0$:

$$T^{1/2}(\hat{\Pi} - \Pi_0)\gamma_0 = Q^{-1}T^{-1/2} \sum_{t=1}^T X_t \xi_t + o_p(1) \quad (9)$$

where ξ_t is defined in Assumption 6(ii). Substituting eq. 4.1 in the text and (9) into the asymptotic representation for the 2SM(q) estimator in Proposition 1 and collecting terms gives $T^{1/2}(\hat{\alpha} - \alpha_0) = Q_{zz}^{-1}H(\Pi_0)'T^{-1/2} \sum_{t=1}^T X_t \eta_t +$

$o_p(1)$, where η_t is defined in Assumption 6(ii). By applying the Liapounov's central limit theorem to, we obtain $T^{-1/2} \sum_{t=1}^T X_t \eta_t \xrightarrow{d} N(0, \sigma_0^2 Q^{-1})$ which completes the proof. *QED.*

References

- Amemiya, T. (1982), "Two Stage Least Absolute Deviation Estimators," *Econometrica*, 50, 689-711.
- Amemiya, T. (1985), "The Nonlinear Two-Stage Least-Squares Estimation," *Journal of Econometrics*, 2, 105-110.
- Bickel, P.J. (1975), "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 70, 428-433.
- Duncan, G.M. (1987), "A Simplified Approach to M-Estimation with Application to Two-Stage Estimators," *Journal of Econometrics*, 34, 373-389.
- Heckman, J.J. (1976), "The Common Structure of Statistical Model of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for such Models," *Annals of Economic and Social Measurement*, 5, 475-492.
- Heckman, J.J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931-960.
- Huber, P.J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73-101.
- Huber, P.J. (1981), *Robust Statistics*, New York: Wiley.
- Kim, T and C. Muller (2001), "Two-Stage Quantile Regression," mimeo School of Economics, University of Nottingham.
- Krasker, W.S. and R.E. Welsch (1985), "Resistant Estimation for Simultaneous-Equations Models using Weighted Instrumental Variables," *Econometrica*, 53, 1475-1488.
- Krasker, W.S. (1986), "Two-Stage Bounded-Influence Estimators for Simultaneous-Equations Models," *Journal of Business & Economic Statistics*, 4, 437-444.
- Malinvaud, E. (1970), "The Consistency of Nonlinear Regressions," *Annals of Mathematical Statistics*, 41, 956-969.

- Newey, W.K. (1985), "Semiparametric Estimation of Limited Dependent Variables Models with Endogenous Explanatory Variables," *Annales de l'INSEE*, n. 59/60, 219-235.
- Newey, W.K. (1989), "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 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," *Review of Economic Studies*, 57, 517-538.
- Powell, J. (1983), "The Asymptotic Normality of Two-Stage Least Absolute Deviations Estimators," *Econometrica*, 51, 1569-1575.
- Ruppert, D. and Carroll. R.J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828-838.