A Test for Endogeneity in Quantiles

Tae-Hwan Kim[a] and Christophe Muller[b]
[b] Department of Economics, Yonsei University,
Seoul 120-749, Korea
Tae-Hwan.Kim@yonsei.ac.kr; thkim8316@gmail.com
[b]DEFI, University of Aix-Marseille,
14, avenue Jules Ferry, F - 13621 Aix-en-Provence cedex, France.
Email: Christophe.muller@u-cergy.fr; christophe.muller@univmed.fr

Keyword: Quantile regression, endogeneity test, Hausman test. **Mots clé de la conférence:** Econométrie, Modèles semi et non paramétriques

Abstract: In this paper we develop a test to detect the presence of endogeneity in different quantiles in the conditional distribution of a variable of interest. This Hausman test type is based on one estimator consistent only under no endogeneity at the examined quantile and another estimator consistent in both the null and the alternative hypotheses. We derive the asymptotic distribution of the test statistic. Moreover, we study the finite sample properties of this test with Monte Carlo simulations of which results exhibit substantial power in the studied cases. Finally, we apply our test to Engel curve estimation with UK data. We find that the pattern of the endogenenity of the total expenditure for various commodities (food, alcohol, fuel, transport, services) is complex when examining it across quantiles.

Résumé: Dans ce papier nous développons un test pour détecter la présence d'endogénéité en différents quantiles de la distribution conditionnelle d'une variables d'intérêt. Ce test a la d'Hausman est basé sur un estimateur convergent seulement sous absence d'endogénéité et un autre estimateur convergent a la fois dans les hypothèses nulles et alternatives. Nous déerivons la distribution asymptotique de la statistique de test. De plus nous étudions les propriétés à distance finie de ce test à l'aide de simulations de Monte Carlo dont les réesultats montrent une puissance considérable dans les cas étudiés. Finalement, nous appliquons notre test à l'estimation de courbes d'Engel avec des données britaniques. Nous trouvons que la structure de l'endogénéité de la dépense totale pour divers biens (alimentation, alcool, fuel, transport, services) est complexe lorsqu'elle est examinée à travers les quantiles.

In recent years quantile regression has become a popular estimation method among applied economists. The main reason of this popularity is the flexibility of this method which, unlike the traditional mean-like estimators such as OLS, MLE or GMM, allows researchers to investigate every single corner of the conditional distribution of a variable of interest.

The issue of endogeneity in the context of quantile regression has been well recognized and different methods to deal with such a issue have been proposed, including the friendly two-stage fitted-value procedure. There are two trends in the literature about quantile regressions with endogeneity problem. The first one corresponds to models specified in terms of the conditional quantile of the structural equation (the 'structural conditional quantile'). The second set of works is anchored on conditional quantile restrictions applied to the reduced-form equation (the 'fitted-value approach') and has been employed by some empirical researchers. This allows the replacement of the endogenous regressors in the second stage with their fitted values obtained in the first stage. We follow this approach in this paper.

We are interested in the parameter (α_0) in the following equation for T observations:

$$\begin{aligned} y_t &= x_{1t}'\beta_0 + Y_t'\gamma_0 + u_t \\ &= Z_t'\alpha_0 + u_t \end{aligned} \tag{1}$$

where $[y_t, Y_t']$ is a $(G+1)$ row vector of endogenous variables, x_{1t}' is a K_1 row vector of exogenous variables, $\mathrm{Z}_t = [x_{1t}', Y_t']'$, $\alpha_0 = [\beta_0', \gamma_0']'$ is assumed bounded in R^{K_1+G}, and u_t is an error term, $t = 1, \ldots, T$ vector. We denote by x_{2t}' the row vector of $K_2 (= K - K_1)$ exogenous variables absent from (1).

We shall discuss later what we mean by exogeneity for given estimation problems. Although weak exogeneity is an issue here, more specific exogeneity notions can be used for given estimators, e.g. implying that the regressors are orthogonal to the errors for LS methods.

Estimating α_0 at the θ^{th}-conditional distribution quantile of the dependent variable y can be achieved through the following minimization program:

$$\min_{\alpha} \sum_{t=1}^{T} \rho_\theta(y_t - Z_t'\alpha) \tag{2}$$

and $\rho_\theta(z) = z\psi_\theta(z)$ where $\psi_\theta(z) = \theta - 1_{[z \leq 0]}$ and $1_{[.]}$ is the Kronecker index. The solution from (2), denoted by $\tilde{\alpha}$, will be called the one-step quantile estimator for α_0. The one-step estimator $\tilde{\alpha}$ is consistent if the following zero conditional expectation condition holds:

$$E(\psi_\theta(u_t)|Z_t) = 0 \tag{3}$$

This condition is the assumption that zero is the given θ^{th}-quantile of the conditional distribution of u_t. It identifies the coefficients of the model. Since this identifying condition depends on the chosen θ, the parameter α_0 in the model depends on this θ too and would vary when considering other quantiles.

In this paper we develop a procedure to test for possible endogeneity in Y_t. As opposed to traditional endogeneity Hausman tests based on LS estimators, our specification allows for non-constant effects across quantile. This implies that we test a more general hypothesis of endogeneity than usual.

We assume that Y_t can be linearly predicted from the exogenous variables:

$$Y_t' = x_t'\Pi_0 + V_t' \tag{4}$$

where $x_t' = [x_{1t}', x_{2t}']$ is a K-rows vector, Π_0 is a $K \times G$ matrix of unknown parameters and V_t' is a G-rows vector of unknown error terms. By assumption the first element of x_{1t} is 1. Using (1) and (4), y can also be expressed as:

$$y_t = x_t'\pi_0 + v_t \tag{5}$$

where

$$\begin{aligned} \pi_0 &= H(\Pi_0)\alpha_0 \text{ with } H(\Pi_0) = \left[\begin{pmatrix} I_{K_1} \\ 0 \end{pmatrix}, \Pi_0\right] \\ \text{and } v_t &= u_t + V_t'\gamma_0. \end{aligned} \tag{6}$$

So far, we did not mention any restriction on errors. The error restrictions will be introduced below in Assumptions 2(vi). We now specify the data generating process.

Assumption 1 *The sequence* $\{(Y_t', x_t', u_t, v_t, V_t')\}$ *is independently and identically distributed (iid). Random vectors* Y_t', x_t', u_t, v_t, and V_t' *are the* t^{th} *elements in* Y, x, u, v, and V *respectively.*

More specifically, $\hat{\pi}$ and $\hat{\Pi}_j$ (the j^{th} column of $\hat{\Pi}$; $j = 1, \ldots, G$) are first stage estimators obtained by:

$$\min_{\pi} \sum_{t=1}^{T} \rho_\theta(y_t - x_t'\pi) \tag{7}$$

$$\min_{\Pi_j} \sum_{t=1}^{T} \rho_\theta(Y_{jt} - x_t'\Pi_j) \tag{8}$$

where π and Π_j are $K \times 1$ vectors and Y_{jt} is the $(j,t)^{th}$ element of Y. Based on these first-stage estimator, the second-stage estimator $\hat{\alpha}$ is defined and obtained as follows:

$$\min_{\alpha} \sum_{t=1}^{T} \rho_\theta(y_t - x_t'H(\hat{\Pi})\alpha).$$

In order to obtain the asymptotic distributions of $\tilde{\alpha}$ and $\hat{\alpha}$, we impose the following regularity conditions. Let $h(\cdot|x)$, $f(\cdot|x)$ and $g_j(\cdot|x)$ be the conditional densities respectively for u_t, v_t and V_{jt}.

Assumption 2 (i) $E(||x_t||^3) < \infty$ *and* $E(||Y_t||^3) < \infty$ *where* $||a|| = (a'a)^{1/2}$.
(ii) $H(\Pi_0)$ *is of full column rank.*

(iii) *There is no hetero-altitudinality. That is:* $h(\cdot|x) = h(\cdot)$, $f(\cdot|x) = f(\cdot)$ *and* $g_j(\cdot|x) = g_j(\cdot)$. *Moreover,* $h(\cdot)$, $f(\cdot)$ *and* $g_j(\cdot)$ *are continuous.*
(iii) *When evaluated at zero, all densities are positive;* $h(0) > 0$, $f(0) > 0$ *and* $g_j(0) > 0$.
(iv) *All densities are bounded above; that is, there exist constants* λ_h, λ_f, *and* λ_j *such that* $h(\cdot) < \lambda_h$, $f(\cdot) < \lambda_f$ *and* $g_j(\cdot) < \lambda_j$.
(v) *The matrices* $Q_x = E(x_t x_t')$ *and* $Q_z = E(Z_t Z_t')$ *are finite and positive definite.*
(vi) $E\{\psi_\theta(v_t) \mid x_t\} = 0$ *and* $E\{\psi_\theta(V_{jt}) \mid x_t\} = 0$ $(j = 1, \ldots, G)$.

For the one-step quantile estimator $\tilde{\alpha}$ we have

$$T^{1/2}(\tilde{\alpha} - \alpha_0) \xrightarrow{d} N(0, \sigma_{11} Q_z^{-1})$$

where $\epsilon_{1t} = h(0)^{-1}\psi_\theta(u_t), \sigma_{11} = E(\epsilon_{1t}^2) = h(0)^{-2}\theta(1-\theta)$ and $Q_z = E(Z_t Z_t')$. The covariance estimator $\sigma_{11} Q_z^{-1}$ can be consistently estimated by $\hat{\sigma}_{11}\hat{Q}_z^{-1}$ where $\hat{Q}_z = T^{-1}\sum_{t=1}^T Z_t Z_t'$ and $\hat{\sigma}_{11} = T^{-1/2}\sum_{t=1}^T \hat{\epsilon}_{1t}^2 = \hat{h}(0)^{-2}\theta(1-\theta)$ with $\hat{\epsilon}_{1t} = \hat{h}(0)^{-1}\psi_\theta(\hat{u}_t), \hat{u}_t = y_t - Z_t\hat{\alpha}$. Here $\hat{h}(0)$ can be any kernel-type non-parametric estimator of the density h at zero.

A similar result can be obtained for the second-stage estimator $\hat{\alpha}$ (Kim and Muller, 2004):

$$T^{1/2}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, \sigma_{22} Q_{zz}^{-1}),$$

where $Q_{zz} = H(\Pi_0)' Q_x H(\Pi_0), Q_x = E(x_t x_t')$ and $\epsilon_{2t} = f(0)^{-1}\psi_\theta(v_t) - \sum_{i=1}^G \gamma_{0i} g_i(0)^{-1}\psi_\theta(V_{it})$, $\sigma_{22} = E(\epsilon_{2t}^2)$. As before, σ_{22} and Q_{zz} can be consistently estimated: $\hat{Q}_{zz} = H(\hat{\Pi})'\hat{Q}_x H(\hat{\Pi})$ with $\hat{Q}_x = T^{-1}\sum_{t=1}^T x_t x_t'$ and $\hat{\sigma}_{22} = T^{-1/2}\sum_{t=1}^T \hat{\epsilon}_{2t}^2$ with $\hat{\epsilon}_{2t} = \hat{f}(0)^{-1}\psi_\theta(\hat{v}_t) - \sum_{i=1}^G \hat{\gamma}_{0i}\hat{g}_i(0)^{-1}\psi_\theta(\hat{V}_{it})$ where $\hat{f}(0)$ and $\hat{g}_i(0)$ are kernel-type estimators of $f(0)$ and $g_i(0)$ respectively and $\hat{v}_t$ and $\hat{V}_{it}$ are the residuals from the first-stage regressions in (7) and (8).

The null hypothesis we wish to test is

$$H_0 : E(\psi_\theta(u_t)|Z_t) = 0, \text{ for a given } \theta \tag{9}$$

Theorem 1 *Suppose that Assumptions 1 and 2 hold. Then, under the null of no endogeneity, we have*

$$T(\tilde{\alpha} - \hat{\alpha})[RC^{-1}R']^{-1}(\tilde{\alpha} - \hat{\alpha}) \xrightarrow{d} \chi^2(K_1 + G),$$

where $R = \left[\ I_{K_1+G} : \ -I_{K_1+G}\ \right]$,

$$\begin{aligned} C &= \begin{bmatrix} \sigma_{11} Q_z^{-1} & \sigma_{12} Q_z^{-1} Q_{zx} H(\Pi_0) Q_{zz}^{-1} \\ \sigma_{12} Q_{zz}^{-1} H(\Pi_0)' Q_{zx}' Q_z^{-1} & \sigma_{22} Q_{zz}^{-1} \end{bmatrix}, \\ Q_{zx} &= E(Z_t x_t') \text{ and } \sigma_{12} = E(\epsilon_{1t}\epsilon_{2t}). \end{aligned}$$

C can be replaced with a consistent estimator $\hat{C}_T$ without affecting the limiting distribution. We use the plug-in principle to propose:

$$\hat{C}_T = \begin{bmatrix} \hat{\sigma}_{11}\hat{Q}_z^{-1} & \hat{\sigma}_{12}\hat{Q}_z^{-1}\hat{Q}_{zx} H(\hat{\Pi})\hat{Q}_{zz}^{-1} \\ \hat{\sigma}_{12}\hat{Q}_{zz}^{-1} H(\hat{\Pi})'\hat{Q}_{zx}'\hat{Q}_z^{-1} & \hat{\sigma}_{22}\hat{Q}_{zz}^{-1} \end{bmatrix},$$

$$\hat{Q}_{zx} = T^{-1}\sum_{t=1}^{T} Z_t x_t' \text{ and } \hat{\sigma}_{12} = T^{-1}\sum_{t=1}^{T} \hat{\epsilon}_{1t}\hat{\epsilon}_{2t}.$$

A first try of test statistics is: $\xi = T(\tilde{\alpha} - \hat{\alpha})[R\hat{C}_T^{-1}R']^{-1}(\tilde{\alpha} - \hat{\alpha})$,which follows asymptotically $\chi^2(K_1 + G)$ if $\tilde{\alpha}$ and $\hat{\alpha}$ are both consistent under the null.

However, the intercept estimator in $\hat{\alpha}$ is not consistent for all values of θ. It is because the semi-parametric restrictions $E\{\psi_\theta(v_t)\} = 0$ and $E\{\psi_\theta(V_{jt})\} = 0$ implied by Assumption 2(vi) are first imposed for a starting value θ_0. Then, they will not be satisfied for other values of θ. For another such θ, the quantile regression admits an asymptotic bias of $F^{-1}(\theta - \theta_0)$ on the intercept, while the Double-Stage quantile regression admits a bias $F^{-1}(\theta - \theta_0) + \gamma_0 G^{-1}(\theta - \theta_0)$, still on the intercept coefficient. **to check**

On the other hand, the slope estimator is consistent regardless of the value of θ. Hence, in order to propose a test for any value of θ, we use the slope estimators only to construct the test statistic (denoted by KM) and the null distribution is $\chi^2(K_1 + G - 1)$. Specifically, let $\alpha_{0(1)}$ and $\alpha_{0(2)}$ be the intercept and slope coefficients respectively and we also decompose the quantile estimators $\tilde{\alpha}$ and $\hat{\alpha}$ accordingly; that is, $\tilde{\alpha}' = (\tilde{\alpha}_{(1)}, \tilde{\alpha}_{(2)}')$ and $\tilde{\alpha}' = (\tilde{\alpha}_{(1)}, \tilde{\alpha}_{(2)}')$. Let $R_{(2)}$ be the matrix composed of the last $(K_1 + G - 1)$ rows in R; that is $R_{(2)} = [0 \;\; I]$ where the zero vector 0 and the identity matrix I are of size $K_1 + G - 1$. Then,

Theorem 2

$$KM = T(\tilde{\alpha}_{(2)} - \hat{\alpha}_{(2)})[R_{(2)}C^{-1}R_{(2)}']^{-1}(\tilde{\alpha}_{(2)} - \hat{\alpha}_{(2)}) \xrightarrow{d} \chi^2(K_1 + G - 1),$$

Bibliographie

References

[1] [1] Chernozhukov, V. and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, Vol. 73, No. 1, 245-261.

[2] Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica*, 1251-1271.

[3] Kim, T. and C. Muller (2004), "Two-Stage Quantile Regressions when the First Stage is Based on Quantile Regressions," *The Econometrics Journal*, 7, 218-231.

[4] Powell, J. (1983). The asymptotic normality of two-stage least absolute deviations estimators. *Econometrica* 51, 1569-75.