연세대학교 상경대학

경제연구소

Economic Research Institute

Yonsei Universit

서울시 서대문구 연세로 50

50 Yonsei-ro, Seodaemun-gu, Seoul, Korea

TEL: (+82-2) 2123-4065 FAX: (+82-2) 364-9149

E-mail: yeri4065@yonsei.ac.kr

http://yeri.yonsei.ac.kr/new

A test for Endogeneity in Conditional Quantile Model

Tae-Hwan Kim, Christophe Muller

May 2012

2012 RWP-49

Economic Research Institute
Yonsei University

A test for endogeneity in conditional quantile models

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

[a]School of Economics, Yonsei University, Seoul 120-749, Korea
tae-hwan.kim@yonsei.ac.kr

[b]DEFI, University of Aix-Marseille, 14, avenue Jules Ferry, F-13621 Aix-en-Provence cedex, France.
christophe.muller@univmed.fr

May 2012

Abstract: In this paper we develop a test to detect the presence of endogeneity in conditional quantile models. The proposed test is a Hausman-type test in that it is based on the distance between two estimators in which one is consistent only under no endogeneity while the other estimator is consistent regardless of the presence of endogeneity in conditional quantile models. We derive the asymptotic distribution of the test statistic under the null hypothesis of no endogeneity. The finite sample properties of the test is investigated by Monte Carlo simulations and it is found that the test shows reasonably good size and power properties in finite samples. Finally, we apply our approach to test for endogeneity in a conditional quantile model for estimating Engel curves using the UK consumption and expenditure data. It is revealed that the pattern of endogeneity found in the Engel curve varies across different quantiles.

Key words: regression quantile, endogeneity, two-stage estimation, Hausman test, Engel curve.

JEL codes: C21.

*Corresponding author. Tel.: +82-2-2123-5461; fax: +82-2-2123-8638. Tae-Hwan Kim is grateful for the financial support from the National Research Foundation of Korea – a grant funded by the Korean Government (NRF-2009-327-B00088).

1 Introduction

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 the 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. To name just a few, see Amemiya (1982), Powell (1983), Chen and Portnoy (1996), Kemp (1999), Sakata (2001), Arias et. al. (2001), Garcia et. al. (2001), Kim and Muller (2004), Chernozhukov and Hansen (2005, 2006, 2008), Ma and Koenker (2006) and Lee (2007).

However, not much attention has been paid to the issue of testing for the presence of endogeneity in conditional quantile models. It has been implicitly assumed in the previous literature that the presence of endogeneity in either the conditional mean or a particular conditional quantile automatically implies that the entire conditional distribution (i.e., all other conditional quantiles) is contaminated by endogeneity. Such a restriction appears too strong and it is more general and realistic to allow for the possibility that endogeneity is present at some part of the conditional distribution.

Let us for example consider a typical wage equation where the logarithm of the wage rate of a worker is linearly explained by the education level and other explanatory factors. The latter factors are often considered as independent of the error term. By contrast, the independence of the education variable and the error is generally disputed. One reason for this is that some unobservable genetic ability may be simultaneously related to both wage and education.

In that case, the model may be subject to the endogeneity problem. In such a model, one can speculate that the poor may have little access to secondary school, perhaps because it is costly, while both the poor and the non-poor are covered by universal mandatory primary education Let us further simplify that there are only two education levels: one for primary education and one for secondary education. In that case, the endogenous genetic ability may affect only the non-poor if such ability is useful only for skills learned at secondary school. Given that the non-poor mostly correspond to high quantiles in the quantile model for the wage equation, endogeneity is likely to be present mostly in high quantiles in this hyperthetical environment.

In this paper, we propose a formal test for the presence of endogeneity in a given specific conditional quantile level. The proposed test is a Hausman-type test in that it is based on the distance between two distinct estimators of the same quantile parameters. One estimator is consistent only when there is no endogeneity, while the other one remains consistent regardless of the presence of endogeneity. In our case, the first estimator is the standard quantile regression estimator and the second estimator is the double-stage quantile estimator developed in Kim and Muller (2004).

We present the model and discuss how the model is estimated in Section 2. The test statistic is proposed and its asymptotic distribution is derived in Section 3. The finite sample properties of the proposed test are studied through Monte Carlo simulations in Section 4. In Section 5, we apply our approach to test for endogeneity in a conditional quantile model for estimating Engel curves using the UK consumption and expenditure data. Finally, Section 6 concludes.

2 The Model and Estimation

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

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

where $[y_t, Y_t']$ is a $(G+1)$ row vector of possible endogenous variables, x_{1t}' is a K_1 row vector of exogenous variables, $Z_t = [x_{1t}', Y_t']'$, $\alpha_0 = [\beta_0', \gamma_0']'$, and u_t is an error term. We denote by x_{2t}' the row vector of $K_2 (= K - K_1)$ exogenous variables absent from (1).

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

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

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-stage quantile estimator for α_0. The one-stage estimator $\tilde{\alpha}$ is consistent if the following zero conditional expectation condition holds:

$$E(\psi_\theta(u_t)|Z_t) = 0. \qquad (3)$$

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. However, the condition in (3) is generally violated if there is endogeneity in Y_t. In that case, $\tilde{\alpha}$ is inconsistent and a two-stage estimation method can be employed to obtain a consistent estimator. In practice, we are often unsure whether there is endogeneity in the regression equation. In this paper we develop a procedure to test for possible endogeneity in Y_t.

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

$$Y_t' = x_t'\Pi_0 + V_t' \qquad (4)$$

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

$$y_t = x_t'\pi_0 + v_t \qquad (5)$$

where

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

Equations in (4) and (5) are the basis of the first-stage estimation that yields consistent estimators $\hat{\pi}$, $\hat{\Pi}$, respectively of π_0, Π_0. 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

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

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

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 estimators, the second stage estimator $\hat{\alpha}$ is obtained as follows:

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

The resulting estimator $\hat{\alpha}$ is called the double-stage quantile estimator for α_0. 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 1 *The sequence $\{(Y_t', x_t', u_t, v_t, V_t')\}$ is independent and identically distributed (iid).*

Assumption 1 is imposed to ease the exposition of our results. It can be relaxed to include additional serial correlation and heteroskedasticity.

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 $g_j(\cdot|x) = g_j(\cdot)$ where $h(\cdot)$, $f(\cdot)$ and $g_j(\cdot)$ are assumed to be continuous. Moreover, all densities are positive when evaluated at zero; i.e., $h(0) > 0$, $f(0) > 0$ and $g_j(0) > 0$.*
(iv) *All densities are bounded above; i.e., 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 $E\{\psi_\theta(V_{jt}) \mid x_t\} = 0$ $(j = 1, \ldots, G)$.*

Assumption 2(i), the moment condition on the exogenous variables, is necessary for the stochastic equicontinuity of our empirical process in the dependent case, which is used for the asymptotic representation. We also use it to bound the asymptotic covariance matrix of the parameter estimators. Assumption 2(ii) is analog to the usual identification condition for simultaneous equations models. Assumption 2(iii) is needed to simplify the asymptotic covariance matrix of the double-stage estimator. Otherwise, the covariance has a very complicated form as shown in Kim and Muller (2004). Assumption 2(iv) simplifies the demonstration of convergence of remainder terms to zero for the calculation of the asymptotic representation. Assumption 2(v) is the counterpart of the usual condition for OLS that the sample second moment matrix of the regressor vectors converges towards a finite positive definite matrix. Lastly, Assumption 2(vi) is the assumption that zero is the θ^{th}-quantile of the conditional distribution of v_t and of each V_{jt}.[1] It identifies the coefficients of the model.

The asymptotic representation of the one-step quantile estimator $\tilde{\alpha}$ is well known:

$$T^{1/2}(\tilde{\alpha} - \alpha_0) = Q_z^{-1} T^{-1/2} \sum_{t=1}^{T} Z_t \epsilon_{1t} + o_p(1), \tag{9}$$

where $\epsilon_{1t} = h(0)^{-1}\psi_\theta(u_t)$. From (9), it is easily seen that $\tilde{\alpha}$ is consistent if $T^{-1}\sum_{t=1}^{T} Z_t \epsilon_{1t}$ vanishes in probability. Since the probability limit $E(Z_t \epsilon_{1t})$ is zero in the absence of endogeneity, we have

[1] Note that in the iid case, the term $f(F^{-1}(\theta))^{-1}$ typically appears in the variance formula of a quantile estimator (Koenker and Bassett, 1978). However, due to Assumption 3(iv), $F^{-1}(\theta)$ is now zero so that we have $f(0)^{-1}$ instead, in this case.

in that case

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

where $\sigma_{11} = E(\epsilon_{1t}^2) = h(0)^{-2}\theta(1-\theta)$ 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 $\hat{\alpha}$ (see Kim and Muller, 2004 for more detail):

$$T^{1/2}(\hat{\alpha}-\alpha_0) = Q_{zz}^{-1}H(\Pi_0)'T^{-1/2}\sum_{t=1}^T x_t\epsilon_{2t} + o_p(1), \tag{10}$$

where $Q_{zz} = H(\Pi_0)'Q_x H(\Pi_0)$, $Q_x = E(x_t x_t')$ and $\epsilon_{2t} = f(0)^{-1}\psi_\theta(v_t) - \sum_{i=1}^G \gamma_{0i}g_i(0)^{-1}\psi_\theta(V_{it})$. Therefore, we have the following result;

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

where $\sigma_{22} = E(\epsilon_{2t}^2)$. 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).

3 Test for Endogeneity

The null hypothesis we wish to test is

$$H_0 : \text{There is no endogeneity in the } \theta^{th} - \text{quantile},$$

which is equivalent to

$$H_0 : E(\psi_\theta(u_t)|Z_t) = 0 \text{ for a given } \theta. \tag{11}$$

Examining more closely the shape of this relationship will allow us to exhibit how endogeneity at different quantiles can be understood in a fashion similar to what is done for the exogeneity notion typically used in LS estimation. Equation (11) for a given θ implies $E\left\{Y_t(\theta - I_{[u_t<0]})\right\} = 0$, if we believe that the only variable likely of endogeneity problem is Y_t. For non-centered Y_t, this is equivalent to $\theta = \frac{E(Y_t f_t)}{EY_t}$, where f_t denotes the index variable $I_{[u_t<0]}$. In a sense, f_t appears to be an index of a 'failure' to reach the benchmark zero for the error associated to θ. Let us call f_t the 'failure index.' Under exogeneity of Y_t at quantile of order θ, the quantile index θ can be seen as the weighed average of the failure index.

Let us now normalize Y_t as $n_t = Y_t/(EY_t)$. Then, we have $\theta = E(n_t f_t)$, which implies that $cov(n_t, f_t) = \theta - E(f_t)$. We note that $E(f_t)$ is the population proportion of errors below zero, which is exactly equal to θ as implied by exogeneity. Therefore, under exogeneity we have $cov(n_t, f_t) = 0$, which implies that $cov(Y_t, f_t) = 0$. Hence, we have shown that the quantile exogeneity condition can be interpreted as a linear orthogonality condition of the possibly endogenous variables with the failure index f_t. This interpretation is helpful because intuitive reasonings referring to the usual zero covariance condition can be used to select instruments in quantile regression as well in a fashion similar to what is typically carried out for 2SLS.

The idea driving our test is that both $\tilde{\alpha}$ and $\hat{\alpha}$ are consistent to the true value α_0 and asymptotically normal under the null hypothesis of no endogeneity while only $\hat{\alpha}$ is consistent to α_0 under the alternative hypothesis. Then, the distance between $\tilde{\alpha}$ and $\hat{\alpha}$ can be used to test consistently the null hypothesis of no endogeneity. We therefore place ourselves in the usual Hausman test framework (Hausman, 1978).

Note that under exogeneity both conditions $E(\psi_\theta(u_t)|Z_t) = 0$ (defining exogeneity in the $\theta^{th}-$quantile) and $E(\psi_\theta(v_t)|X_t) = 0$ (defining the double-stage quantile estimator in Kim and Muller, 2004) are compatible since no conditions are imposed on V_t. Under these conditions, both the one-stage quantile estimator and the double-stage estimator are consistent towards the same quantity. However, under $E(\psi_\theta(u_t)|Z_t) \neq 0$ and $E(\psi_\theta(v_t)|X_t) = 0$, the two estimators still converge, albeit towards different quantities. These properties are exploited to propose our test.

A simple specification is that of a quadratic distance between $\tilde{\alpha}$ and $\hat{\alpha}$, weighed by the variance matrix of $\tilde{\alpha} - \hat{\alpha}$. It will be shown below that the variance-covariance matrix of $\tilde{\alpha} - \hat{\alpha}$ is given by $RC^{-1}R'$, where $R = \left[\begin{array}{cc} I_{K_1+G}: & -I_{K_1+G} \end{array}\right]$ and C is defined in Theorem 1 below. Hence, the proposed test statistic is given by

$$T(\tilde{\alpha} - \hat{\alpha})[RC^{-1}R']^{-1}(\tilde{\alpha} - \hat{\alpha}).$$

The following theorem shows the null distribution of our proposed statistic. The proof of the theorem is provided in the Appendix. We allow for non-efficient estimators by looking at the joint distribution without having to invoke orthogonality conditions between estimators.

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

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

where

$$\begin{aligned} C &= \left[\begin{array}{cc} \sigma_{11}Q_z^{-1} & \sigma_{12}Q_z^{-1}Q_{zx}H(\Pi_0)Q_{zz}^{-1} \\ \sigma_{12}Q_{zz}^{-1}H(\Pi_0)'Q_{zx}'Q_z^{-1} & \sigma_{22}Q_{zz}^{-1} \end{array}\right], \\ Q_{zx} &= E(Z_t x_t'), \\ \sigma_{12} &= E(\epsilon_{1t}\epsilon_{2t}). \end{aligned}$$

Of course, the proposed test in Theorem 1 cannot be implemented unless C is known. However, C can be replaced with a consistent estimator $\hat{C}_T$ without affecting the limiting distribution. We use the plug-in principle to propose the following consistent estimator for C:

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

where

$$\begin{aligned} \hat{Q}_{zx} &= T^{-1}\sum_{t=1}^{T} Z_t x_t', \\ \hat{\sigma}_{12} &= T^{-1}\sum_{t=1}^{T} \hat{\epsilon}_{1t}\hat{\epsilon}_{2t}. \end{aligned}$$

We now establish the consistency of $\hat{C}_T$ in the following lemma.

Lemma 1. Suppose that kernel-type density estimators $\hat{h}(0)$, $\hat{f}(0)$ and $\hat{g}_i(0)$ are respectively consistent for $h(0)$, $f(0)$ and $g_i(0)$. Then, under Assumptions 1 and 2, we have

$$\hat{C}_T \xrightarrow{p} C.$$

Hence, an implementable test statistic (denoted by KM) is given by $KM = T(\tilde{\alpha}-\hat{\alpha})[R\hat{C}_T^{-1}R']^{-1}(\tilde{\alpha}-\hat{\alpha})$. It can be easily shown that the asymptotic distribution of KM under the null hypothesis is $\chi^2(K_1+G)$.

We note that the intercept estimator in $\hat{\alpha}$ may not be 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 of θ. Then, they will not be satisfied for other subsequently chosen values of θ. 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 a test statistic (denoted by KM_s). 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. Then, we have the following corollary

Corollary 1. Suppose that kernel-type density estimators $\hat{h}(0)$, $\hat{f}(0)$ and $\hat{g}_i(0)$ are respectively consistent for $h(0)$, $f(0)$ and $g_i(0)$. Then, under Assumptions 1 and 2, we have

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

The result of Corollary 1 easily follows from Theorem 1 and Lemma 1. We now examine the finite-sample performance of the proposed test by Monte Carlo simulations in the next section.

4 Monte Carlo Simulations

The results obtained in the previous section hold in large samples. In this section, the finite sample properties in terms of size and power of the proposed test is studied through Monte Carlo simulations. We use a simultaneous equation system based on two equations. The first equation, which is the equation of interest, contains two endogenous variables and two exogenous variables including a constant. In total, four exogenous variables are present in the whole system. The structural simultaneous equation system can be written

$$B \begin{bmatrix} y_t \\ Y_t \end{bmatrix} + \Gamma x_t = U_t, \tag{12}$$

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. The error term $U_t = \begin{bmatrix} u_t \\ w_t \end{bmatrix}$ is a 2×1 vector of error terms. We specify the structural parameters as follows: $B = \begin{bmatrix} 1 & -0.3 \\ \delta & 1 \end{bmatrix}$ 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 $\Gamma_{13} = \Gamma_{14} = \Gamma_{22} = 0$.

We generate the error terms U_t using $\mathrm{N}(0_{2\times1}, I_{2\times2})$ and we draw the second to fourth elements x_t from the normal distribution with mean $(0.5, 1, -0.1)'$, variances equal to 1 for normalization,

$cov(x_{2t}, x_{3t}) = 0.3, cov(x_{2t}, x_{4t}) = 0.1$ and $cov(x_{3t}, x_{4t}) = 0.2$, where x_{2t}, x_{3t} and x_{4t} are the non-constant elements of x_t. Once x_t and U_t are generated, the endogenous variables y_t and Y_t are generated through (12). The first structural equation is

$$y_t = 0.3Y_t + 1 + 0.2x_{2t} + u_t, \tag{13}$$

where the presence of endogeneity will depend on the δ parameter in the second equation. Note that if $\delta = 0$, there is no endogeneity in (13). On the other hand, we have endogeneity if $\delta \neq 0$. The magnitude of δ determines the strength of endogeneity, which allows us to measure the empirical power of the proposed test. We select three values for δ; 0.00, 0.60, and 1.20. For each of the values, we compute the rejection probabilities by the test for the null hypothesis of no endogeneity at the 5% significance level based on 1,000 replications.

The results are displayed in Table 1 for $T = 100$. First, it is seen that when $\delta = 0$, the rejection probability in each case is close to the nominal 5% level. As the value of δ increases, the rejection probability also increases from 5% for all values of θ considered (0.25, 0.50 and 0.75); the empirical power is around 20-26% when δ is around 1.2. As we increase the sample size from 100 to 200 as shown in Table 2, the power of the test is further increased so that the rejection probability is in the range of 40-50% when δ is around 1.2.

5 An Application to Engel Curve Estimation

In this section we apply the proposed test for endogeneity in a quantile model for estimating Engel curves in the UK. The dataset is drawn from the UK Family Expenditure Survey (UKFES) conducted in 1995, which has been used in previous studies such as Blundell et. al. (2007) and Chen and Rouzo (2009).[2] As an illustration for our test, we consider the following quantile Engel curve equation:

$$y_i = \beta_{0\theta} + \beta_{1\theta}x_{1i} + \gamma_\theta Y_i + u_i, \tag{14}$$

where y_i is the budget share of household i on food in 1995, x_{1i} is a dummy variable for children (i.e., $x_{1i} = 0$ if without children and 1 if with children), Y_i is the log of total expenditure on both non-durables and services of household i in 1995. The total number of households surveyed is 1665. As mentioned in Blundell et. al. (2007), Y_i can be either exogenous or endogenous so that it might be empirically tested and determined. Following Blundell et. al. (2007), the earnings variable measuring the amount that the male of the household i earned during 1995 is used as an instrument; specifically its log is employed. Table 3 shows some summary statistics of the three main variables and their graphs are displayed in Figures 1-3.

Although our main focus is the quantile model in (14), we first estimate the corresponding conditional mean model and present the results in Table 4 where the conditional mean model is estimated twice by OLS and 2SLS, allowing for two possibilities that Y_i is either exogenous or endogenous. The coefficient estimates for the total expenditure have a negative sign as expected, indicating that the food share decreases as the total expenditure increases. The OLS estimate is much larger than the 2SLS estimate in absolute value, which leads to conjecture that the total expenditure variable may be endogenous. The p-value of the standard Hausman test is 0.005, empirically verifying the conjecture. Tables 5 and 6 shows the results based on one-stage quantile regression (assuming Y_i is exogenous) and double-stage quantile regression (assuming Y_i is endogenous), respectively. In line with the conditional mean regression results, all estimates for the total expenditure variable based on one-stage quantile regression are much larger than those based on double-stage quantile

[2] We are very grateful to Xiaohong Chen for kindly providing us with the dataset for this research.

regression. It is also noted that the conditional median regression results (i.e. $\theta = 0.5$) are quite comparable to the previous conditional mean regression results. Lastly, Table 7 presents the test results obtained from the application of the proposed test to the quantile model in (14) for each quantile index considered in the previous tables. As shown in Table 7, there seems no endogeneity in low quantiles whereas there exists evidence for the presence of endogeneity in the middle and high quantiles at the 10% significance level. In Table 7, only 9 quantile indexes are considered, but more quantiles over a finer grid (e.g. from 0.01 to 0.99 with increment of 0.01) can be investigated although its presentation in a table can be cumbersome. Hence, the test results over such a finer quantile grid are graphically displayed in Figure 4 from which we can obtain a more informative picture on which specific area over the entire conditional distribution is affected by endogeneity induced by the total expenditure variable.

6 Conclusion

In this paper we have proposed a test to detect the presence of endogeneity in conditional quantile models. The test is a Hausman-type test in that it is based on the distance between two estimators in which one is consistent only under no endogeneity while the other estimator is consistent regardless of the presence of endogeneity in conditional quantile models. The derived asymptotic null distribution of the test statistic is the usual Chi-sqaure distribution and Monte Carlo simulations indicate that the test has reasonably good size and power properties. By applying the proposed test to the UK consumption and expenditure data, it has been revealed that only some parts (including the median) of the conditional distribution of the food share is affected endogeneity.

References

[1] Abadie, A., J. Angrist and G. Imbens (2002), "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91-117.

[2] Amemiya, T. (1982). "Two stage least absolute deviations estimators." *Econometrica*, 50, 689-711.

[3] Arias, O., K.F. Hallock and W. Sosa-Escudero (2001). "Individual heterogeneity in the returns to schooling: Instrumental variables quantile regression using twins data." *Empirical Economics*, 26, 7-40.

[4] Blundell, R., X. Chen and D. Kristensen (2007), "Semi-nonparametric IV estimation of shape invariant Engel curves," *Econometrica*, 75, 1613-1669.

[5] Chen, L.-A. and S. Portnoy (1996). "Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models." *Commun. Statist.-Theory Meth.* 25, 1005-32.

[6] Chen, X., O. Linton and I. Van Keilegem (2003), "Estimation of semi-parametric models when the criterion function is not smooth," *Econometrica*, 71, 1591-1608.

[7] Chen, X., D. Pouzo (2009), "Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals," *Journal of Econometrics*, 152, 46-60.

[8] Chernozhukov, V. and C. Hansen (2005), "An IV model of quantile treatment effects," *Econometrica*, 73, 245-261.

[9] Chernozhukov, V. and C. Hansen (2006), "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491-525.

[10] Chernozhukov, V. and C. Hansen (2008), "Instrumental variable quantile regression: a robust inference approach," *Journal of Econometrics*, 142, 379-398.

[11] Chesher, A. (2003), "Identification in nonseparable models," *Econometrica*, 71, 1405-1441.

[12] Garcia, J., P.J. Hernandez and A. Lopez-Nicolàs (2001). "How wide is the gap? An investigation of gender wage differences using quantile regression." *Empirical Economics*, 26, 149-67.

[13] Hausman, J.A. (1978), "Specification tests in econometrics," *Econometrica*, 1251-1271.

[14] Hong, H. and E. Tamer (2003), "Inference in censored models with endogenous regressors," *Econometrica*, 71, 905-932.

[15] Kemp, G. (1999). "Least absolute error difference estimation of a single equation from a simultaneous equations system." Mimeo University of Essex.

[16] Kim, T.-H. and White, H (2003), "Estimation, inference, and specification testing for possibly misspecified quantile regression, *Advances in Econometrics*, 17, 107-132.

[17] 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.

[18] Kim, T. and C. Muller (2005). "Two-stage quantile regression." Working Paper.

[19] Lee, S. (2007), "Endogeneity in quantile regression models: a control function approach," *Journal of Econometrics*, 141, 1131-1158.

[20] Ma, L. and R. Koenker (2006), "Quantile regression methods for recursive structural equation models," *Journal of Econometrics,* 134, 471-506.

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

[22] Sakata, S. (2001). "Instrumental variable estimation based on mean absolute deviation." Working Paper.

Appendix

Proof of Theorem 1: Let $\hat{\delta} = (\tilde{\alpha}', \hat{\alpha}')'$ and $\delta_0 = (\alpha_0', \alpha_0')'$. Using (9) and (10), we have

$$\begin{aligned} T^{1/2}(\hat{\delta} - \delta_0) &= \begin{bmatrix} Q_z^{-1} T^{-1/2} \sum_{t=1}^{T} Z_t \epsilon_{1t} + o_p(1) \\ Q_{zz}^{-1} H(\Pi_0)' T^{-1/2} \sum_{t=1}^{T} x_t \epsilon_{2t} + o_p(1) \end{bmatrix} \\ &= D T^{-1/2} \sum_{t=1}^{T} S_t + o_p(1) \qquad (15) \end{aligned}$$

where

$$\begin{aligned} D &= \begin{bmatrix} Q_z^{-1} & 0 \\ 0 & Q_{zz}^{-1} H(\Pi_0)' \end{bmatrix} \\ S_t &= \begin{bmatrix} Z_t \epsilon_{1t} \\ x_t \epsilon_{2t} \end{bmatrix}. \end{aligned}$$

Let us now consider (15). S_t is iid by Assumption 1 and $E(S_t) = 0$ under the null of no endogeneity and Assumption 2(vi). Hence, in order to apply the Lindeberg-Levy's CLT to $T^{-1/2} \sum_{t=1}^{T} S_t$, it is sufficient to show that $var(S_t)$ is bounded. The moment conditions on x_t and Y_t in Assumption 2(i) are sufficient for this purpose because $\psi_\theta(\cdot)^2$ is bounded from above and all the densities evaluated at zero are bounded.

Given that

$$var(S_t) = \Omega = \begin{bmatrix} \sigma_{11} Q_z & \sigma_{12} Q_{zx} \\ \sigma_{12} Q_{zx}' & \sigma_{22} Q_x \end{bmatrix},$$

where $\sigma_{ij} = Cov(\varepsilon_{it}, \varepsilon_{jt})$, we have $T^{-1/2} \sum_{t=1}^{T} S_t \xrightarrow{d} N(0, \Omega)$ which implies that

$$T^{1/2}(\hat{\delta} - \delta_0) \xrightarrow{d} N(0, C)$$

where $C = D \Omega D'$. Noting that $T^{1/2}(\tilde{\alpha} - \hat{\alpha}) = R T^{1/2}(\hat{\delta} - \delta_0)$, we have

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

which in turn implies that

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

which completes the proof. QED.

Proof of Lemma 1: Under Assumptions 1 and 2, the cross product estimators $\hat{Q}_z$, $\hat{Q}_{zx}$ and $\hat{Q}_x$ are consistent almost surely due to the Kolmogorov law of large numbers. It is also obvious to show the consistency of $\hat{Q}_{zz}$ and $H(\hat{\Pi})$ because $\hat{\Pi}$ is consistent. Hence, it remains to show the consistency of $\hat{\sigma}_{11} = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{1t} \hat{\epsilon}_{1t}, \hat{\sigma}_{22} = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{2t} \hat{\epsilon}_{2t}$ and $\hat{\sigma}_{12} = T^{-1} \sum_{t=1}^{T} \hat{\epsilon}_{1t} \hat{\epsilon}_{2t}$. We provide the detailed proof only for $\hat{\sigma}_{12}$ since the same kind of argument is used for $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$.

Recalling the definition of $\sigma_{12} = E(\epsilon_{1t} \epsilon_{2t})$, the following comes directly from the Kolmogorov law of large numbers:

$$\tilde{\sigma}_{12} - \sigma_{12} \xrightarrow{p} 0$$

where $\tilde{\sigma}_{12} = T^{-1}\sum_{t=1}^{T}\epsilon_{1t}\epsilon_{2t}$ since (i) $\epsilon_{1t}\epsilon_{2t}$ is iid due to Assumption 1 and (ii) $E(|\epsilon_{1t}\epsilon_{2t}|) < \infty$ because of the boundedness of the $\psi_\theta(\cdot)-$function. Hence, it will be sufficient to show $\hat{\sigma}_{12}-\tilde{\sigma}_{12} \xrightarrow{p} 0$ to prove that $\hat{\sigma}_{12} - \sigma_{12} \xrightarrow{p} 0$.

Note that

$$
\begin{aligned}
& |\hat{\sigma}_{12} - \tilde{\sigma}_{12}| \\
\leq \; & a_T + b_T + c_T
\end{aligned}
$$

where $a_T = \theta T^{-1}\sum_{t=1}^{T}|1_{[\hat{v}_t\leq 0]} - 1_{[v_t\leq 0]}|$, $b_T = \theta T^{-1}\sum_{t=1}^{T}|1_{[\hat{u}_t\leq 0]} - 1_{[u_t\leq 0]}|$ and $c_T = T^{-1}\sum_{t=1}^{T}|1_{[\hat{u}_t\leq 0]}1_{[\hat{v}_t\leq 0]} - 1_{[u_t\leq 0]}1_{[v_t\leq 0]}|$. The proof for both $a_T \xrightarrow{p} 0$ and $b_T \xrightarrow{p} 0$ can be found in the proof of Proposition 3 in Kim and Muller (2004). Hence we only need to show that $c_T \xrightarrow{p} 0$. We note that

$$
\begin{aligned}
c_T &\leq T^{-1}\sum_{t=1}^{T}|1_{[\hat{u}_t\leq 0]}| \times |1_{[\hat{v}_t\leq 0]} - 1_{[v_t\leq 0]}| + T^{-1}\sum_{t=1}^{T}|1_{[v_t\leq 0]}| \times |1_{[\hat{u}_t\leq 0]} - 1_{[u_t\leq 0]}| \\
&\leq T^{-1}\sum_{t=1}^{T}|1_{[\hat{v}_t\leq 0]} - 1_{[v_t\leq 0]}| + T^{-1}\sum_{t=1}^{T}|1_{[\hat{u}_t\leq 0]} - 1_{[u_t\leq 0]}| \\
&\leq \theta^{-1}(a_T + b_T) \xrightarrow{p} 0.
\end{aligned}
$$

Hence, the proof is completed. QED.

Table 1. Rejection probabilities by the KM test for the null of no endogeneity at θ^{th} – quantile with T = 100

	δ	θ		
		0.25	0.50	0.75
Size	0.00	0.05	0.04	0.06
Power	0.60	0.11	0.09	0.10
	1.20	0.26	0.20	0.23

Table 2. Rejection probabilities by the KM test for the null of no endogeneity at θ^{th} – quantile with T = 200

	δ	θ		
		0.25	0.50	0.75
Size	0.00	0.06	0.05	0.06
Power	0.60	0.31	0.29	0.31
	1.20	0.53	0.57	0.52

Table 3. Summary statistics

	mean	Std.
Food share	0.2074	0.0971
Log expenditure	5.4215	0.4494
Log earnings	5.8581	0.5381
Sample size	1665	

Table 4. Conditional mean regression

	Estimates (Standard Errors)		
	Intercept	Dummy for Kids	Expenditure
OLS	0.761 (0.024)	0.056 (0.004)	-0.109 (0.004)
2SLS	0.614 (0.047)	0.054 (0.004)	-0.081 (0.009)
Hausman Test	Statistic = 13.02 P-value = 0.005.		

Table 5. Conditional quantile regression based on one-stage estimator

	Estimates (Standard Errors)		
Quantile (θ)	Intercept	Dummy for Kids	Expenditure
0.1	0.414 (0.044)	0.042 (0.008)	-0.061 (0.008)
0.2	0.499 (0.036)	0.046 (0.006)	-0.072 (0.007)
0.3	0.588 (0.032)	0.053 (0.005)	-0.084 (0.006)
0.4	0.658 (0.028)	0.057 (0.005)	-0.094 (0.005)
0.5	0.758 (0.028)	0.057 (0.005)	-0.109 (0.005)
0.6	0.849 (0.030)	0.063 (0.005)	-0.122 (0.006)
0.7	0.929 (0.035)	0.063 (0.006)	-0.133 (0.007)
0.8	1.012 (0.049)	0.065 (0.008)	-0.144 (0.009)
0.9	1.107 (0.105)	0.067 (0.018)	-0.156 (0.019)

Table 6. Conditional quantile regression based on double-stage estimator

	Estimates (Standard Errors)		
Quantile (θ)	Intercept	Dummy for Kids	Expenditure
0.1	0.275 (0.063)	0.049 (0.007)	-0.043 (0.013)
0.2	0.434 (0.058)	0.054 (0.006)	-0.066 (0.011)
0.3	0.499 (0.061)	0.057 (0.006)	-0.073 (0.012)
0.4	0.477 (0.057)	0.051 (0.005)	-0.063 (0.011)
0.5	0.629 (0.060)	0.058 (0.006)	-0.087 (0.011)
0.6	0.720 (0.065)	0.062 (0.006)	-0.098 (0.012)
0.7	0.761 (0.073)	0.068 (0.007)	-0.099 (0.013)
0.8	0.817 (0.091)	0.062 (0.008)	-0.100 (0.016)
0.9	0.850 (0.121)	0.045 (0.009)	-0.093 (0.020)

Table 7. Test for endogeneity in various quantiles

Quantile (θ)	Test Statistic (KM_s)	P-value
0.1	4.028	0 .133
0.2	2.902	0.234
0.3	2.582	0.275
0.4	11.829	0.003
0.5	4.742	0.093
0.6	4.952	0.084
0.7	9.362	0.009
0.8	7.587	0.023
0.9	7.898	0.019

Figure 1. Food share of 1665 households

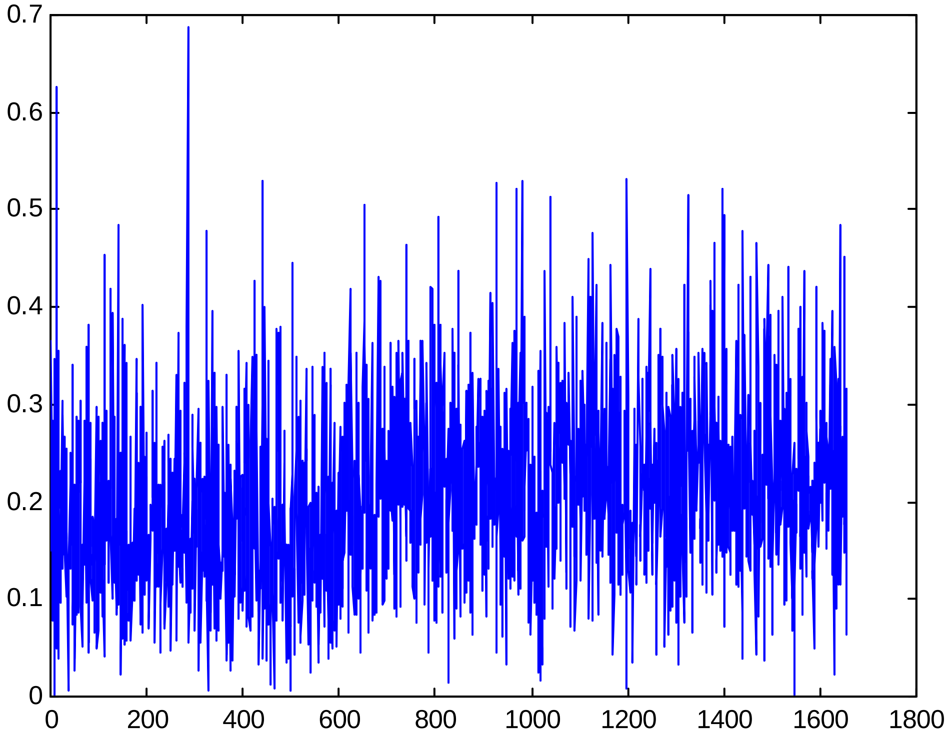

Figure 2. Total expenditure on durables and services of 1665 households

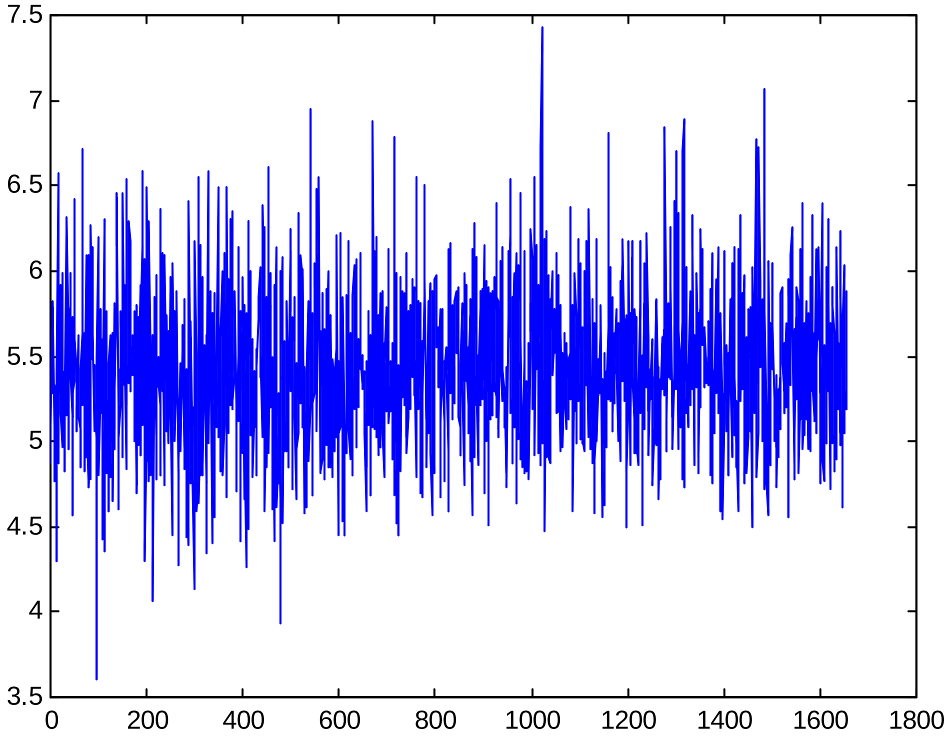

Figure 3. Earnings of 1665 households

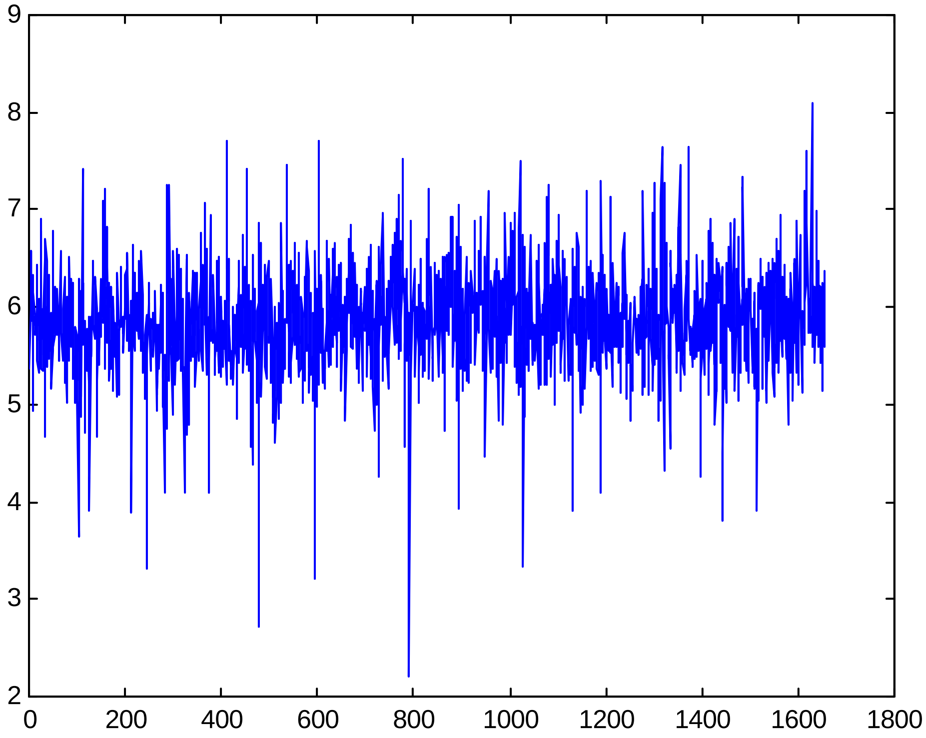

Figure 4. P-values of the KM_s test over a quantile grid of [0.01, 0.99]