Symmetry on Reporting Noncompliance of Sole Proprietor Income: A Censored Quantile Regression Analysis of the TCMP Data

William W. S. Chen and Chih-Chin Ho, Internal Revenue Service

Current tax laws impose no limit on the amount of net losses from a non-farm business income for offsetting individual taxable income. We have developed a semi-parametric model to examine symmetry in reporting noncompliance of sole proprietor income using quantile regression models.

We begin by estimating a quantile regression model for underreporting business income using the statistical procedure developed by Koenker and Bassett (1978) and the linear programming algorithm developed by Koenker and d'Orey (1987). We estimate the model to ascertain the significance of the impact of selected taxpayer characteristics and reporting attributes on the conditional quantile of reporting noncompliance. We compare the OLS results with the median regression results to see any differences between "average-cheaters" and "moderate-cheaters." We examine the results of different percentiles to test for symmetry on noncompliance between "low-cheaters" and "high-cheaters."

We then estimate censored quantile regression models for understating business income using a statistical procedure developed by Powell (1986). Since censored quantile regression does not have an attractive representation of linear programming, we implement the Powell's estimator by a non-linear algorithm developed by Fitzenberger (1996). We compare the model estimates with those from a Tobit model where the dependent variable is the censored conditional mean of reporting noncompliance.

We estimate the model using 1988 Taxpayer Compliance Measurement Program post-audit data on individual Schedule C filers. We follow the bootstrap procedure outlined in Buchinksky (1994) to compute standard errors of both censored and uncensored models.

◆ Model Framework

Quantile Regression

Let us consider a regression model

$$[1] \quad Y_i = \beta + \eta_i \quad \text{for } i = 1,2,\ldots n$$

Just as the OLS regression line gives a model for the mean of the distribution of the dependent variable Y conditional on the independent variables X, quantile regressions give models for different percentiles of the conditional distribution of Y given X.

Now, let us consider $q \in (0,1)$ and Prob{Y<y*) >= q,

$Q_Y(q) = F^{-1}{}_Y(q)$ gives the q-th quantile of Y, $Q_Y(q \mid X) = F^{-1}{}_Y(q \mid X)$ gives the q-th conditional (on X) quantile of Y. A quantile regression model assumes that the q-th conditional quantile of Yi is given by

$$[2] \quad Q_{Yi}(q \mid Xi) = X'_i \beta_q \quad \text{for } i = 1,2,\ldots n$$

Equation [2] is just the natural analogue of the zero conditional mean assumption on the errors in ordinary regression. If X only consists of only a constant term, then $\beta_{0.5}$ is just the sample median. If X represents a vector of regressors, β_q is the vector of slopes of the quantile regression line and gives the effect of changes in X on the q-th conditional quantile of the dependent variable Y.

It then follows that by estimating β_q for different values of $q \in (0,1)$, one can obtain a family of quantile regression curves that characterize the impact of changes in the regressors on different percentiles of the empirical conditional distribution of the dependent variable.

$$[3] \quad Y_i = X'_i \beta + (X'_i \delta)\, \mu_i \quad \text{for } i = 1,2,\ldots n$$

where μ_i is identically and independently distributed (iid) and δ is a vector of coefficients. In this case, X affects not only the mean of the conditional distribution of Y but also its scale; for instance, it may affect the dispersion as well as the skewness of the distribution.

From [2], it then follows that the effect of a change in X_i on Yi is given by

$$[4] \quad \partial\{Q_{Yi}(q \mid Xi)\} / \partial\{X_i\} = \beta_q = \beta + \delta\, Q_{\mu i}(q \mid Xi) \quad \text{for } i = 1,2,\ldots n$$

How can the quantile regression coefficients be interpreted? Consider the partial derivatives of the conditional quantile of Y with respect to one of the regressors, say j,

$$[5] \quad \partial Q_{Yi}(q \mid Xi) / \partial X_{ij} = \beta_{qj} \quad j = 1,2,\ldots k$$

This derivative is to be interpreted as the marginal change in the q-th conditional quantile due to marginal change in the j-th element of X. If X contains k distinct variables, then this derivative is given simply by β_{qj}, the coefficient of the j-th variable.

Censored Quantile Regression

Another important property of quantiles is their equivariance to monotonic transformations, that is, the quantiles of a monotonic transformation of Y are the same as the quantiles of the original Y. As Powell (1994) observed, this feature has made the use of quantile regression specially useful in censored regression models. Let us consider a simple censored Tobit model

$$[6] \quad Yi = \max \{c_i, Y^*_i = X'_i \beta + \mu i\}$$

where c_i is the known censoring value for the I-th observation, Y* is the unobserved latent variable. Only observations on Y, c, and X are available to estimate β. In the case where censoring is fixed at zero (ci =0) for all observations, the model reduces to the usual Tobit model used by Clotfelter (1983).

For this model of fixed known censoring, Powell (1986) observed that as long as a positive fraction of the observations is uncensored, it suffices to impose a zero conditional quantile restriction on the error term to be able to identify β in [6] consistently. That is, if we assume the q_c-th conditional censored quantile of μ_i equals zero, then the equivariance of quantiles to monotonic transformations implies that

$$[7] \quad Q_{Yi}(q\_c \mid Xi) = \max \{ci, X'_i \beta_{q,c}\}.$$

As a result, the conditional quantile of Yi also gives the conditional quantile of the latent variable Y^*_i since the censoring transformation, max {0,b}, is monotone nondecreasing in b. Therefore, application of quantile regression to equation [7] allows us to obtain consistent estimates of $\beta_{q,c}$ without invoking any strong assumption about the shape of the unknown distribution of μ_i.

From (7), it then follows that estimates for different quantiles can also be used to uncover patterns of unobserved heterogeneity in the errors. In other words, they allow us to trace how changes in X affect the entire uncensored portion of the conditional distribution of the latent variable Y*.

What is the intuitive rationale behind the estimation procedure of $\beta_{q,c}$ in the censored regression model? Recall that $X'_i \beta_{q,c}$ is the conditional quantile of Y* given X. Now, two cases of left censoring are possible: (1) $X'_i \beta_{q,c} > c$, and (2) $X'_i \beta_{q,c} <= c$. In case (1), since the probability is greater than q_c that $Y^* = X' \beta_{q,c} + \mu_{q,c} > c$, the conditional quantile of Y* given X can be exactly identified. On the other hand, in case (2), the probability is less than q_c that $Y^* = X' \beta_{q,c} + \mu_{q,c} <= c$. In other words, the conditional quantile is in the unobserved part of the distribution. Consequently, nothing can be done with that portion of the data; we know only that its conditional quantile is greater than the censoring value. The implication is that one has to drop that portion of the data that cannot be used. As a result, the estimated asymptotic convariance matrix has to be adjusted for the fact that the estimation is conditioned on the inclusion of only the observations for which $X'_i \beta_{q,c} > c$.

◆ Estimation Algorithm

Quantile Regression

Given a sample of n observations on Y and X, Koenker and Bassett (1978) show that estimates of β_q in equation [4] can be solved by

$$[8] \quad \underset{\beta_q}{min} \sum^n_{i=1} h_q \mid \varepsilon i \mid \text{ for } i = 1,2,\ldots n$$

$$\text{where } \varepsilon i = Y_i - X'_i \beta$$

$$h_q = q \quad \text{if } \varepsilon i > 0$$

$$= (1-q) \text{ if } \varepsilon i < 0$$

Quantile regression, therefore, weights the absolute value of the residuals, with the weight depending upon

the quantile to be estimated. For q=0.5, this problem reduces to

$$[8.m] \quad \underset{\beta_{0.5}}{min} \ \Sigma^{n}_{i=1} \mid \varepsilon i \mid \text{ for } i=1,2,\ldots n$$

Equation [8.M] is simply median regression. In the general case, the function h_q weights negative and positive residuals of all observations asymmetrically to obtain the quantile of interest. Koenker and Bassett (1978) show under a set of regularity conditions the quantile regression estimator gives consistent estimates of β_q in [4]. Since quantile regressions have the attractive presentation of linear programming, equation [8] can be solved as a linear programming problem. An algorithm developed by Koenker and D'orey (1997) is used to implement the Koenker and Bassett (1978) quantile regression estimator defined in (1). Standard errors are estimated using the bootstrap procedure outlined by Buchinsky (1994).

Censored Quantile Regression

Given a sample of n observations on Y, X, and C (censoring value), Powell (1996) shows that estimates of $\beta_{q\,c}$ can be solved by

$$[9] \quad \underset{\beta_{q\,c}}{min} \ \Sigma^{n}_{i=1} p_{q\_c} \mid \tau i \mid \qquad \text{for } i=1,2,\ldots n$$

$$\text{where } \tau i = Yi - \max\{ ci , X'_i \beta_{q\_c} \}$$

$$p_{q\_c} = q\_c \quad \text{if } \tau i > 0$$

$$= (1-q\_c) \text{ if } \tau i < 0$$

For q_c=0.5, this problem reduces to

$$[9.m] \quad \underset{\beta_{0.5\_c}}{min} \ \Sigma^{n} i=1 \mid \tau i \mid \qquad \text{for } i=1,2,\ldots n$$

Equation [9.M] is simply a censored median regression. Powell (1986) shows that the parameters solving [9] are consistent estimates of the true vector of quantile regression coefficients in equation [7]. An algorithm developed by Fitzenberger (1996) is used to implement the Powell (1986) censored quantile regression estimator defined in [9]. Standard errors are estimated using the bootstrap procedure outlined by Buchinsky (1994).

◆ Data Structure

We develop a simple model of tax evasion based on Clotfelter (1983) and Feinstein (1991) for analyzing the determinants of reporting noncompliance of sole proprietor income in Schedule C. We select taxpayer characteristics such as adjusted gross income, the marginal tax rate, net balance due, and marital status; and reporting attributes such as whether to offer a benefit plan for the employees, whether to incur substantial losses in capital income, or whether to itemize deductions. The definitions of the variables are listed in Figure 1.

Figure 1
Definition of Variables
Used in Model Estimation

Dependent Variable	
UNDER_INCOME	Examined Business Income minus Reported Business Income
Independent Variable	
DEPENDENT	Number of Dependents
NETDUE	$1,000 Net Tax Due
AGI	$1,000 Adjusted Gross Income
MTR	% Marginal Tax Rate (Combined Federal and State Rate)
SOUTH	Dummy Variable for Taxpayer Located in the South
AGE65	Dummy Variable for Taxpayer of Age 65 or Older
ITEMIZER	Dummy Variable for Filing Itemized Deductions (Schedule A)
MARRIED	Dummy Variable for Filing Married Jointly
MARRIED*AGI	Interactions between MARRIED and AGI
MARRIED*MTR	Interaction between MARRIED and MTR
BIG LOSS	Dummy Variable for Greater than 10,000 Combined Loss in Schedules C,D,E, and F
CAPITAL LOSS	Dummy Variable for Having Capital Losses on Schedule D
RENTAL LOSS	Dummy Variable for Having Rental Losses on Schedule E
BENEFIT	Dummy Variable for Having Employee Benefit Expense Deductions on Schedule C
LEGAL FEE	Dummy Variable for Having Legal Fee Deductions on Schedule C
INSURANCE	Dummy Variable for Having Insurance Premium Deductions on Schedule C
ADVERTISING	Dummy Variable for Having Advertising Expenditure Deductions on Schedule C
BAD DEBT	Dummy Variable for Having Bad Debt Deductions on Schedule C

◆ Estimation Results

Uncensored Quantile Regression

We first apply our model using both OLS and median regression to ascertain the determinants of underreporting total business income. Table 1 shows the results. Note that for most of independent variables, the signs are similar for the OLS and median regression. For example, both marginal tax rate (MTR) and net tax due (NETDUE) show positive impact on noncompliance, while both taxpayer age 65 or older (AGE65) and filing a Schedule A for itemized deductions (ITEMIZER) show negative impact instead.

However, for certain independent variables, the OLS and median regressions yield conflicting signs, meaning particular variables have opposite impact between conditional mean and conditional median of the dependent variable. For example, incurring a rental loss in Schedule E (RENTAL LOSS) reduces income underreporting for an average evader (negative impact on the conditional mean shown by the OLS regression), but increases income underreporting for a moderate evader (positive impact on the conditional median shown by the median regression). On the other hand, having bad debts would have positive impact on average underreporting, but negative impact on median underreporting.

To examine the symmetry in underreporting income, we estimate both lower quantile (25 percent) and upper quantile (75 percent) models. These two quantile regressions can represent the behavior of low evader and high evader, respectively. Table 2 presents the regression estimates for both quantiles.

All explanatory variables, with only one exception of the dummy for the married filing jointly status, exhibit the same signs for both quantiles. This finding indicates that the pattern of reporting compliance is quite similar for the low evader and the high evader.

Censored Quantile Regression

Table 3 presents both Tobit and censored median model results for the pattern of underreporting income. Tobit model estimates provide the impacts of the determinants on the censored conditional mean of the dependent variable, while the censored median model provides the estimated impacts on the censored conditional median. Note that, for most of independent variables, the signs are the same for both models. For example, both NETDUE and MTR show positive impacts, while both INSURANCE and ADVERTISING show negative impacts on the censored conditional mean and median of underreporting income. However, while ITEMIZER and BIG LOSS reduce the conditional mean and raise the conditional median, RENTAL LOSS and BAD DEBT show conflicting impacts in the opposite direction.

Table 4 presents the lower (25 percent) and upper (75 percent) quantile results for underreporting income. Except for BIG LOSS, which shows positive impact for the lower quantile and negative impact for the higher quantile, all other determinants show the same sign for both quantiles. While MTR, NETDUE, and ITEMIZER, on one hand, show positive impacts for the "low" evader, AGI, MARRIED, and their interaction term (MARRIED*AGI), on the other hand, show negative impacts for the "high" evader.

◆ References

Buchinsky, M. (1994), "Changes in the U.S. Wage Structure 1963-1987: An Application of Quantile Regression," *Econometrica*, March, pp. 405-58.

Clotfelter, C. (1983), "Tax Evasion and Tax Rates," *Review of Economics and Statistics*, August, pp. 363-73.

Fitzenberger, B. (1996), "A Guide to Censored Quantile Regressions," in *Handbook of Statistics*, edited by C. Rao and G. Maddala, North-Holland, NY.

Koenker, R. and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, pp. 33-50.

Koenker, R. and D'Orey, V. (1987), "Computing Regression Quantiles," *Applied Statistics*, 36, pp. 383-93.

Powell, J. (1986), "Censored Regression Quantiles," *Journal of Econometrics*, 32, pp. 143-55. ■

TABLE 1
Uncensored Regression Model
Estimation Results

	Conditional Mean (OLS)	50% Quantile (Median)
CONSTANT TERM	**-6358.4** (1976.7)	278.90 (156.59)
DEPENDENT	471.66 (295.01)	**328.59** (98.178)
NET TAX DUE	**205.14** (28.114)	**241.81** (73.644)
AGI	**19.208** (8.8307)	25.298 (19.920)
MTR	**372.95** (89.423)	-11.090 (14.666)
SOUTH	760.86 (690.19)	206.08 (186.82)
AGE65	-813.85 (1230.8)	**-473.48** (220.80)
ITEMIZER	**-2693.4** (755.86)	**-509.41** (262.15)
MARRIED	439.89 (2406.7)	**-1470.9** (771.21)
BIGLOSS	**-3006.1** (1757.1)	**-1070.4** (439.39)
MARRIED* AGI	-21.301 (12.998)	-31.453 (20.873)
MARRIED* MTR	-7.6572 (105.72)	72.083 (46.358)
CAPITAL LOSS	1467.1 (1310.6)	371.51 (512.99)
RENTAL LOSS	-330.04 (1032.9)	506.35 (392.08)
BENEFIT	-744.33 (1859.3)	761.35 (938.01)
LEGAL FEE	965.20 (711.80)	213.89 (170.40)
INSURANCE	**1265.3** (732.11)	**492.22** (191.33)
ADVERTISE MENT	**2172.4** (709.13)	**629.88** (236.42)
BAD DEBT	893.81 (1673.3)	-233.44 (663.48)

*The standard error is in parentheses following the parameter estimate, and bold face designates the statistical significance at 10-percent level.

TABLE 2
Uncensored Regression Model
Estimation Results

	50% Quantile (Low)	75% Quantile (High)
CONSTANT TERM	-3.5646 (33.434)	**581.19** (302.64)
DEPENDENT	33.706 (35.064)	**452.04** (266.97)
NET TAX DUE	**38.016** (20.115)	**862.64** (227.41)
AGI	1.2805 (3.2887)	**189.59** (75.924)
MTR	-0.71612 (2.9481)	-36.770 (48.878)
SOUTH	31.336 (46.621)	450.95 (376.23)
AGE65	-119.20 (57.424)	-611.42 (698.83)
ITEMIZER	-27.952 (74.718)	**-1064.1** (491.19)
MARRIED	-370.62 (279.67)	55.741 (234.85)
BIGLOSS	-25.515 (187.53)	736.33 (1587.9)
MARRIED* AGI	-6.1145 (4.9243)	**-144.04** (80.396)
MARRIED* MTR	21.881 (16.612)	48.878 (149.94)
CAPITAL LOSS	41.128 (134.48)	729.65 (834.09)
RENTAL LOSS	12.473 (116.91)	-684.87 (886.42)
BENEFIT	479.09 (371.36)	**2628.7** (1466.5)
LEGAL FEE	20.353 (58.437)	229.22 (527.89)
INSURANCE	**107.28** (61.927)	711.87 (569.56)
ADVERTISE MENT	68.288 (67.291)	**968.58** (571.45)
BAD DEBT	-83.752 (208.70)	791.39 (1320.8)

*The standard error is in parentheses following the parameter estimate, and bold face designates the statistical significance at 10-percent level.

TABLE 3
Censored Regression Model
Estimation Results

	Conditional Mean (TOBIT)	50% Quantile (Median)
CONSTANT TERM	**-12921** (2194.1)	-286.67 (1205.5)
DEPENDENT	**590.99** (321.17)	**718.60** (353.37)
NET TAX DUE	**542.80** (47.302)	**1922.1** (830.74)
AGI	**-42.217** (15.546)	-279.70 (206.47)
MTR	**524.38** (100.59)	31.043 (72.206)
SOUTH	1104.3 (733.22)	822.75 (535.88)
AGE65	**-3031.8** (1384.7)	-3375.3 (2678.4)
ITEMIZER	**-1912.6** (837.95)	683.92 (875.37)
MARRIED	-14.739 (2648.7)	**-12259** (5686.4)
BIGLOSS	**-2470.5** (1380.9)	-1197.5 (24897)
MARRIED* AGI	-20.293 (16.859)	-139.93 (93.671)
MARRIED* MTR	41.155 (117.55)	**723.68** (297.13)
CAPITAL LOSS	1289.7 (1435.6)	2055.2 (1728.9)
RENTAL LOSS	1704.7 (1094.1)	-1072.1 (1469.4)
BENEFIT	-611.15 (2012.2)	4452.3 (3233.3)
LEGAL FEE	478.88 (775.72)	920.05 (663.99)
INSURANCE	**1627.9** (798.38)	**1273.1** (682.53)
ADVERTISE MENT	**2557.0** (773.95)	272.64 (625.13)
BAD DEBT	522.99 (1820.9)	2230.7 (2240.3)

*The standard error is in parentheses following the parameter estimate, and bold face designates the statistical significance at 10-percent level.

TABLE 4
Censored Regression Model
Estimation Results

	25% Quantile (Low)	75% Quantile (High)
CONSTANT TERM	528.53 (360.19)	-21707 (9532.5)
DEPENDENT	**941.85** (363.56)	1310.1 (1340.5)
NET TAX DUE	**2456.5** (460.64)	98.138 (803.29)
AGI	-55.701 (138.13)	24.114 (260.81)
MTR	-41.062 (68.972)	304.42 (388.39)
SOUTH	**1153.7** (588.66)	1098.4 (1968.2)
AGE65	-1028.7 (1425.8)	-541.65 (6817.7)
ITEMIZER	370.93 (971.71)	-2251.6 (2168.3)
MARRIED	-8936.5 (6255.7)	3688.3 (11432)
BIGLOSS	-853.23 (7600.2)	-293.38 (11395)
MARRIED* AGI	**-275.69** (164.09)	-96.738 (267.37)
MARRIED* MTR	675.27 (440.23)	-403.72 (649.22)
CAPITAL LOSS	2261.5 (1635.6)	291.04 (3099.8)
RENTAL LOSS	-435.51 (1390.6)	-1814.9 (3077.2)
BENEFIT	**4721.2** (2528.8)	8350.1 (14982)
LEGAL FEE	70.871 (689.96)	2899.3 (2677.6)
INSURANCE	646.28 (781.88)	471.32 (2368.9)
ADVERTISE MENT	846.84 (658.04)	**3677.2** (1958.3)
BAD DEBT	91.697 (1786.1)	6317.3 (5258.2)

*The standard error is in parentheses following the parameter estimate, and bold face designates the statistical significance at 10-percent level.