SET IDENTIFICATION WITH TOBIN REGRESSORS

VICTOR CHERNOZHUKOV, ROBERTO RIGOBON, AND THOMAS M. STOKER

Abstract. We give semiparametric identification and estimation results for econometric models with a regressor that is endogenous, bound censored and selected, called a Tobin regressor. We show how parameter sets are identified, and give generic estimation results as well as results on the construction of confidence sets for inference. The specific procedure uses quantile regression to address censoring, and a control function approach for estimation of the final model. Our procedure is applied to the estimation of the effects on household consumption of changes in housing wealth. Our estimates fall in plausible ranges, significantly above low OLS estimates and below high IV estimates that do not account for the Tobin regressor structure.

Date: October 2007.

This is a preliminary version for seminar presentations. The authors wish to thank the participants of the CEMMAP conference in June 2007 for helpful comments, in particular Andrew Chesher, Elie Tamer and Whitney Newey.

1. Introduction

In economic surveys, financial variables are often mismeasured in nonrandom ways. The largest values of household income and wealth are often eliminated by top-coding above prespecified threshold values. Income and wealth are also typically reported as nonnegative, which may neglect large transitory income losses large debts (negative components of wealth), or other aspects that could be modeled viewed as bottom-coding below a prespecified threshold value. In addition to mismeasurement problems related to upper and lower bounds, income and wealth are often missing due to nonresponse.[1]

These measurement problems are particularly onerous when they are systematic with respect to the economic process under study. For instance, suppose one is interested in the impact of liquidity constraints on consumption spending. The widespread practice of dropping all observations with top-coded income values seemingly eliminates households that are the least affected by liquidity constraints. Likewise, if one is studying the household demand for a luxury good, the most informative data is from rich households, who, for confidentiality reasons, often won't answer detailed questions about their income and wealth situations.

These problems can be compounded when the observed financial variable is itself an imperfect proxy of the economic concept of interest. For instance, suppose one is studying the impact of the availability of cash on a firm's investment decisions. Only imperfect proxies of "cash availability" are observed in balance sheet data, such as whether the firm has recently issued dividends. The mismeasurement of those

[1] For many surveys, extensive imputations are performed to attempt to "fill in" mismeasured or unrecorded data, often in ways that are difficult to understand. For instance, in the U.S. Consumer Expenditure (CEX) survey, every component of income is top-coded; namely wages, interest, gifts, stock dividends and gains, retirement income, transfers, bequests, etc., and there is no obvious relation between the top-coding on each component and the top-coding on total income. The CEX makes extensive use of ad hoc multiple imputation methods to fill in unrecorded income values.

proxies is not random; positive dividends indicate positive cash availability but zero dividends can indicate either mild cash availability or severe cash constraints. Thus, observed dividends represent a censored (bottom coded at zero) version of the cash availability status of a firm.

The study of mismeasurement due to censoring and selection was initiated by the landmark work of Tobin(1958). In the context of analyzing expenditures on durable goods, Tobin showed how censoring of a dependent variable induced biases, and how such bias could be corrected in a parametric framework This work has stimulated an enormous literature on parametric and semiparametric estimation with censored and selected dependent variables. The term "Tobit Model" is common parlance for a model with a censored or truncated dependent variable.

We study the situation where a regressor is censored or selected. This also causes bias to arise in estimation; bias whose sign and magnitude varies with the mismeasurement process as well as the estimation method used (Rigobon and Stoker (2006a)). With reference to the title, we use the term "Tobin regressor" to refer to a regressor that is bound censored, selected and (possibly) endogenous.

When the mismeasurement of the regressor is exogenous to the response under study – that is, both the correctly measured regressor and the censoring/selection process is exogenous – then consistent estimation is possible by using only the "complete cases," or dropping any observations with a mismeasured regressor. Even when valid, the complete cases are often a small fraction of the data, with the resulting estimates very imprecise. But the existence of consistent estimates allows for tests of whether bias is induced in estimates computed from the full data sample.[2]

[2] See Rigobon and Stoker (2006b) for regression tests and Nicholetti and Peracchi (2005) for tests in a GMM framework.

When endogenous regressors are censored or selected, the situation is considerably more complicated. Complete case analysis, or dropping observations with a mismeasured regressor, creates a selected sample for the response under study. Standard instrumental variables methods are biased when computed from the full data sample, and are also biased and inconsistent when computed using the complete cases only.

In this paper, we provide a full identification analysis and estimation solution for situations with Tobin regressors. We give results on the nonparametric identification and estimation of parameter sets. We apply recent work on confidence sets to provide a full theory of inference analogous to the case of point identification and inference. We carry out estimation and inference with quantile regression methods, although our generic theory is applicable with many types of flexible estimation methods. We see our methods as providing estimates and inference that make correct use of all observations with accurately measured data.

Our approach is related to several contributions in the literature. Without censoring or selection, our framework is in line with work on nonparametric estimation of triangular systems, as developed by Altonji and Matzkin (2005), Chesher (2003), Imbens and Newey (2005) and Chernozhukov and Hansen (2005), among others. Our accommodation of endogeneity uses the control function approach, as laid out by Blundell and Powell (2003). In terms of dealing with censoring, we follow Powell's (1984) lead in using monotonicity assumptions together with quantile regression methods (see Koenker's(2005) excellent review of quantile regression). Inference is possible in our framework following recent results on confidence intervals for sets of Chernozhukov, Hong and Tamer (2007).

There is a great deal of literature on mismeasured data, some focused on regressors. Foremost is Manski and Tamer (2002), who use monotonicity restrictions to propose consistent estimation with interval data. For other contributions in econometrics, see Ai (1997), Chen, Hong and Tamer (2005), Chen, Hong and Tarozzi (2004), Liang, Wang, Robins and Carroll (2004), Tripathi (2004), among many others, which are primarily concerned with estimation when data is missing at random. The large literature in statistics on missing data is well surveyed by Little and Rubin (2002), and work focused on mismeasured regressors is surveyed by Little (1992).

The exposition proceeds by introducing our approach in a simple framework, in Section 2. Section 3 gives our general framework and a series of generic results on identification and estimation. Section 4 contains an empirical application, where we show how accommodating censoring and selection gives rise to a much larger estimates of the impact of housing wealth on consumption.

2. A Basic Discussion of The Model and Identification Procedures

2.1. **Linear Modeling Setup.** We introduce the main ideas of our approach using the simplest possible framework. Without censoring or selection, we assume a linear model with (potentially) endogenous regressor:

$$Y = X^*\alpha + U^* \tag{2.1}$$

$$X^* = Z'\gamma + V^* \tag{2.2}$$

$$U^* = \beta V^* + \varepsilon \tag{2.3}$$

where

$$\varepsilon \text{ is mean (or median or quantile) independent of } (V^*, X^*), \tag{2.4}$$

$$V^* \text{ is median independent of } Z \tag{2.5}$$

Here, X^* is the uncensored regressor, which is endogenous when $\beta \neq 0$, and Z represents valid instruments (without censoring or selection). We make no further assumption on the distribution of ε or V^*.

The observed regressor X is given as

$$X = I\{R=1\}I\{X^*>0\}X^* \tag{2.6}$$

$$R = \begin{cases} 1 & \text{with prob} \quad 1-\pi \\ 0 & \text{with prob} \quad \pi \end{cases}, \text{ independent of } Z. \tag{2.7}$$

That is, there are two sources of censoring of X^* to 0. First is bound censoring, which occurs when $X^* \leq 0$ or $V^* \leq -Z'\gamma$. Second is independent selection censoring, which occurs if $R = 0$. We further assume that

$$P\left[Z'\gamma > 0\right] > 0 \tag{2.8}$$

which is convenient as well as empirically testable.

Censoring is modeled with the lower bound (bottom-coding) of 0, but top-coding or different bound values are straightforward to incorporate. Selection censoring occurs with constant probability π here, but π will be modeled as varying with controls in our general framework. The observed regressor X is censored, selected and endogenous, which we refer to as a *Tobin regressor*. While there exists an instrument Z for the uncensored regressor X^*, that instrument will typically be correlated with $X - X^*$. Therefore, Z will not be a valid instrument if X is used in place of X^* in the response equation (2.1).

It is also straightforward to include additional controls in the response equation. With that in mind, we develop some examples for concreteness.

Example 1. Income and Consumption: Suppose X is income and Y is household consumption expenditure. X is typically endogenous, top-coded and missing for various households.. Bound censoring arises for large income values, and selection refers to missing values, possibly due to households declining to report their income. For instance, if one is estimating a permanent income model of consumption, then X would be observed permanent income (or wealth). If one is investigating excess sensitivity (or liquidity constraints), then X would be observed current income (and the equation would include lagged consumption). Finally, the same issues can arise in an Engel curve analysis, where Y is the expenditure on some commodity and X is total expenditures on all commodities.

Example 2: Dividends and Firm Investment: Suppose X is declared dividends and Y is investment, for individual firms. Here X^* is the level of cash availability (or opposite of cash constraints). Positive dividends X indicate positive cash availability, but zero dividends arises with either mild or severe cash constraints (small or large negative X^*).

Example 3: Day Care Expenditures and Female Wages: Suppose you are studying the economic situation faced by single mothers, where Y is expenditure on day care and X is the observed wage rate. X is potentially endogenous (work more to pay for higher quality day care), and is selected due to the labor participation choice.

2.2. **Basic Identification and Estimation Ideas.** The strategy for identification of the model is to set the amount of selection first, which allows the rest of the model to be identified. So, suppose we set a value π^* for $Pr[R = 1]$. The following steps give identification:

1) Observe that the conditional median curve $Q_{X^*}(\frac{1}{2}|Z) = Z'\gamma$ is partially identified from the estimable curve

$$Q_X\left(\frac{1}{2}(1-\pi^*) + \pi^*|Z\right) = \max[Z'\gamma, 0], \tag{2.9}$$

provided $Z'\gamma > 0$ with positive probability.

2) We can then estimate the control function

$$V^* = X^* - Z'\gamma = X - Z'\gamma,$$

whenever $X > 0$.

3) Given the control function V^*, we can recover the regression function of interest (mean, median or quantile) for the sub-population where $X > 0$ and $Z'\gamma > 0$. Namely, if ε is mean independent of (V^*, X^*), we can estimate the mean regression

$$E[Y|X, V^*] = X'\alpha + \beta V^* \tag{2.10}$$

or if ε is quantile independent of (V^*, X^*), we can estimate the quantile regression

$$Q_Y(\tau|X, V^*) = X'\alpha + \beta V^*. \tag{2.11}$$

4) All of the above parameters depend on the guess π^*. We recognize this functional dependence by writing $\alpha(\pi^*)$, $\beta(\pi^*)$, $\gamma(\pi^*)$ for solutions of steps 1), 2), and 3). For concreteness, suppose the particular value $\alpha_0 = \alpha(\pi_0)$ is of interest. Given a set $\mathcal{P}_0$ that contains π_0, the set

$$\mathcal{A}_0 = \{\alpha(\pi), \pi \in \mathcal{P}_0\}$$

clearly contains α_0. Likewise, if we denote $\theta(\pi) = \{\alpha(\pi), \beta(\pi), \gamma(\pi)\}$, then $\theta_0 = \theta(\pi_0)$ is contained in the set

$$\Theta_0 = \{\theta(\pi), \pi \in \mathcal{P}_0\}.$$

5) It remains to find the set $\mathcal{P}_0$. In the absence of further information, this set is given by:

$$\mathcal{P}_0 = [0, \inf_{z \in \text{support}(Z)} \Pr[D > 0|Z = z]]. \tag{2.12}$$

where

$$D \equiv 1 - I\{R = 1\}I\{X^* > 0\}$$

is the index of observations that are censored.

This is the basic identification strategy. It is clear that point identification is achieved if π is a known value. In particular, if there is only bound censoring, then

$\pi = 0$, and estimation (step 1) uses median regression to construct the (single) control function for estimation.

For estimation, the population curves above will be replaced by empirical curves. In Section 4 we will discuss several ways of flexibly estimating the model as well as its non-additive generalizations. Confidence regions can be constructed as follows. Suppose α_0 is of interest and the set $\mathcal{P}_0$ is known. A standard confidence region for $\alpha(\pi)$ is

$$CR_{1-\alpha}(\alpha(\pi)) = [\hat{\alpha}(\pi) \pm c_{1-\alpha} s.e.(\hat{\alpha}(\pi))].$$

This implies that an $1-\alpha$-confidence region for $\alpha_0 = \alpha(\pi_0)$ is merely

$$\cup_{\pi \in \mathcal{P}_0} CR_{1-\alpha}(\alpha(\pi)).$$

A simple way to report such a confidence region is to report its largest and smallest elements.

We will discuss further adjustments because of the estimation of $\mathcal{P}_0$, the range of selection probabilities. Because of independence, the selection probability is a lower bound on the probability of censoring for all observations. Therefore, we can estimate the upper bound for the selection probability with an estimate of the minimum probability of censoring across the data. With this logic, a confidence region for $\mathcal{P}_0$ can be developed, as well as adjustments for the level of significance of the parameter confidence regions given above.

2.3. **A Geometric View of Identification and Estimation.** We illustrate the basic idea of identification through a sequence of figures that illustrate a simple one-regressor version of our empirical example. In the first step, we fix a set of values of π in a set from 0 to .09 (the range for the true π_0) and fit a family of censored conditional quantile estimates. Thus, we obtain a family of "first stage" estimates, shown in Figure 1, indexed by the admissible values of π. We compute these estimates using the three-step estimation procedure described in Chernozhukov and Hong (2002), which is a computationally attractive approximation to the estimator of Powell (1984). In the second step, we form a control function V_π using the results of the first step, and then we run mean regressions of Y on X and V_π. The results are indexed by the values of $\pi \in [0, 0.09]$. Thus, we obtain a family of “second stage" estimates, shown in Figure 2, indexed by the admissible values of π. Finally, Figure 3 shows the construction of a conservative though consistent upper bound on π. The underlying specification here corresponds to the one used in the empirical section of the paper. The panels of Figure 3 show the fitted probabilities of missing data on X. The top panel shows a naive plug-in upper bound on π. The bottom panel shows the upper bound of π adjusted up by the two times standard error times a logarithmic factor in the sample size.

3. Generic Set Identification and Inference

3.1. Set Identification without Functional Form Assumptions.

The general stochastic model we consider is the following system of quantile equations:

$$Y = Q_Y(U|X^*, W, V) \tag{3.1}$$

$$X^* = Q_{X^*}(V|W, Z) \tag{3.2}$$

where Q_Y is the conditional quantile function of Y given X^*, W, V and Q_{X^*} is the conditional quantile function of X^* given Z. Here U is Skorohod disturbance such that $U \sim U(0,1)|X^*, W, V$, and V is Skorohod disturbance such that $V \sim U(0,1)|W, Z$. The latent true regressor is X^*, which is endogenous when V enters the first equation nontrivially. Z represents “instruments" and W represents covariates.

The observed regressor X, the *Tobin regressor*, is given by the equation

$$X = I\{R = 1\}I\{X^* > 0\}X^* \tag{3.3}$$

where

$$R = \begin{cases} 1 & \text{with probability} \quad 1-\pi(W) \\ 0 & \text{with probability} \quad \pi(W) \end{cases} \tag{3.4}$$

conditional on W, Z, V.

There are two sources of censoring of X^* to 0. First there is bound censoring, occurring when $X^* \leq 0$. Second is selection censoring, occurring when $R = 0$, independently of the first source of censoring.

The model (3.1), (3.2) is quite general, encompassing a wide range of nonlinear models with an endogenous regressor. The primary structural restriction is that the system is triangular; that is, V can enter (3.1) but that U does not enter (3.2). The Skohorod disturbances U and V index the conditional quantiles of Y and X^*. We have by definition that

$$\begin{aligned} U &= F_Y(Y|X^*, W, V) \\ V &= F_{X^*}(X^*|W, Z) \end{aligned}$$

where F_Y is the conditional distribution function of Y given X^*, W, V and F_{X^*} is the conditional distribution function of X^* given W, Z. The random variables U and V provide an equivalent parameterization to the stochastic model as additive disturbances or other (more familiar) ways capturing randomness. For example, the linear model (2.1)-(2.2) is written in the form of (3.1), (3.2) as

$$\begin{aligned} Y &= X^*\alpha + Q_{U^*}\left(U|V\right) \\ X^* &= Z'\gamma + Q_{V^*}\left(V|Z\right) \end{aligned}$$

where the additive disturbances U^* and V^* have been replaced by U and V through the (equivalent) quantile representations $U^* = Q_{U^*}\left(U|V\right)$ and $V^* = Q_{V^*}\left(V|Z\right)$.

The primary restriction of the *Tobin regressor* is that selection censoring is independent of bound censoring, (conditional on W, Z and V). We have left the selection probability in the general form $\pi(W)$, which captures many explicit selection models. For instance, we could have selection based on threshold crossing: suppose η is a disturbance and the selection mechanism is $R \equiv 1\left[W'\delta + \eta \geq 0\right]$, then $\pi(W) = \Pr\left\{\eta < -W'\delta\right\}$.

We now turn to the formal identification results. We require the following assumption

Assumption 1: *We assume that the systems of equations (3.1)-(3.4) and independence assumptions hold as specified above, and that $v \mapsto Q_{X^*}(v|W, Z)$ is strictly increasing in $v \in (0,1)$ almost surely.*

Our main identification result is

Proposition 1. *The identification regions for $Q_{Y^*}(\cdot|X^*, V, W)$ and $F_Y(\cdot|X^*, V, W)$ on the subregion of the support of (X^*, V, W) implied by $X > 0$ are given by*

$$\mathcal{Q} = \{Q_Y(\cdot|X, V_\pi, W), \pi \in \mathcal{P}\}$$

and

$$\mathcal{F} = \{F_Y(\cdot|X, V_\pi, W), \pi \in \mathcal{P}\}$$

where when $X > 0$

$$V_\pi = \frac{F_X(X|Z, W) - \pi(W)}{1 - \pi(W)}, \tag{3.5}$$

or equivalently when $X > 0$

$$V_\pi = \int_0^1 1\left\{Q_X\left((\pi(W) + (1 - \pi(W))v|W, Z\right) \leq X\right\} dv. \tag{3.6}$$

Finally,

$$\mathcal{P} = \left\{\pi(\cdot) \text{ measurable} : 0 \leq \pi(W) \leq \min_{z \in supp(Z)|W} F_X(0|W, z) \ a.s\right\}. \tag{3.7}$$

Proposition 1 says that given the level of the selection probability $\pi(W)$, we can identify the quantile function of Y with respect to X^* by using the (identified) quantile function of Y with respect to the observed Tobin regressor X, where we shift the argument V to V_π of (3.5,3.6). The identification region is comprised of the quantile functions for all possible values of $\pi(W)$. The proof is constructive, including indicating how the quantiles with respect to X^* and to X are connected.

Proof of Identification. We follow the logic of the identification steps outlined in the previous section. Suppose we first set a value $\pi(W)$ for $Pr[R = 1|W]$. For $x > 0$, we have that

$$\begin{aligned} \Pr[X \leq x|W,Z] &= \Pr[R=0|W,Z] + \Pr[R=1 \text{ and } X^* \leq x|W,Z] \\ &= \Pr[R=0|W,Z] + \Pr[R=1|W,Z] \cdot \Pr[X^* \leq x|W,Z] \end{aligned}$$

That is,

$$F_X[x|W,Z] = \pi(W) + (1-\pi(W))F_{X^*}[x|W,Z]$$

In terms of distributions, whenever $X > 0$,

$$V_\pi = F_{X^*}[X|W,Z] = \frac{F_X(X|Z,W) - \pi(W)}{1-\pi(W)}$$

Thus V_π is identified from the knowledge of $F_X(X|Z,W)$ and $\pi(W)$ whenever $X > 0$. In addition

$$X^* = X.$$

when $X > 0$

In terms of quantiles,

$$\begin{aligned} Q_{X^*}(V_\pi|W,Z) &= Q_X\left(\left. \frac{F_X(X|Z,W)-\pi(W)}{1-\pi(W)} \right| W,Z \right) \\ &= Q_X\left((\pi(W) + (1-\pi(W))V_\pi|W,Z\right). \end{aligned} \tag{3.8}$$

This implies that for any $X > 0$

$$\begin{aligned} V_\pi &= \int_0^1 1\{Q_{X^*}(v|W,Z) \leq X^*\}dv \\ &= \int_0^1 1\{Q_X\left((\pi(W) + (1-\pi(W))v|W,Z\right) \leq X\}dv \end{aligned}$$

Thus V_π is identified from the knowledge of $Q_X(\cdot|Z,W)$ whenever $X > 0$.

Inserting

$$X, V_\pi \text{ for cases } X > 0$$

into the outcome equation we have a point identification of the quantile functional

$$Q_Y(\cdot|X, V_\pi, W)$$

over the region implied by the condition $X > 0$. This functional is identifiable from the quantile regression of Y on X, V_π, W.

Likewise, we have the point identification of the distributional functional

$$F_Y(\cdot|X, V_\pi, W)$$

over the region implied by the condition $X > 0$. This functional is identified either by inverting the quantile functional or by the distributional regression of Y on X, V_π, W.

Now, since the (point) identification of the functions depends on the value $\pi(W)$, by taking the union over all $\pi(\cdot)$ in the class $\mathcal{P}$ of admissible conditional probability functions of W, we have the following identified sets for both quantities:

$$\{Q_Y(\cdot|X, V_\pi, W), \pi(\cdot) \in \mathcal{P}\}$$

and

$$\{F_Y(\cdot|X, V_\pi, W), \pi(\cdot) \in \mathcal{P}\}.$$

The quantities above are sets of functions or correspondences.

It remains to characterize the admissible set $\mathcal{P}$. From the relationship

$$F_X[0|W,Z] = \pi(W) + (1-\pi(W)) \cdot F_{X^*}[0|W,Z]$$

we have

$$0 \leq \pi(W) = \frac{F_X\left[0|W,Z\right] - F_{X^*}[0|W,Z]}{1 - F_{X^*}[0|W,Z]} \leq F_X(0|W,Z),$$

where the last observation is by the equalities

$$0 = \min_{0 \leq x \leq F} \left(\frac{F-x}{1-x}\right) \leq \max_{0 \leq x \leq F} \left(\frac{F-x}{1-x}\right) = F.$$

Thus, taking the best bound over z:

$$0 \leq \pi(W) \leq \min_{z \in \mathcal{Z}|W} F_X(0|W,z),$$

Hence

$$\mathcal{P} = \left\{ \pi(\cdot) \text{ measurable : } 0 \leq \pi(W) \leq \min_{z \in \mathcal{Z}|W} F_X(0|W,z) \text{ a.s} \right\}.$$

□

As noted before, the proof makes it clear that point identification of the functions is possible where $X > 0$ and $\pi\left(W\right)$ is known (or point identified), including the no selection case with $\pi(W) = 0$.

3.2. **Generic Limit Theory and Inference.** We are typically interested in

$$\theta(\pi) = \theta\Big(Q(\cdot;\pi)\Big)$$

a functional of π taking values in Θ, where the quantile Q can either be the conditional quantile Q_Y or Q_{X^*}, or equivalently

$$\theta(\pi) = \ \theta^*\Big(F(\cdot;\pi)\Big)$$

where the conditional distribution F is either F_Y or F_{X^*} . For identification, we have the immediate corollary:

Corollary 1. *The identification region for the functional $\theta(\pi_0)$ is*

$$\{\theta(\pi), \pi \in \mathcal{P}\}.$$

For some generic results on estimation, we use a plug-in estimator

$$\widehat{\theta}(\pi) = \theta\Big(\widehat{Q}(\cdot;\pi)\Big) \text{ or } \theta^*\Big(\widehat{F}(\cdot;\pi)\Big).$$

We assume that the model structure is sufficiently regular to support the following standard estimator properties:

Assumption. 2.1 For any $\pi \in P$, suppose an estimate of $\widehat{Q}$ or $\widehat{F}$ is available such that

$$\mathcal{Z}_n(\pi) := A_n(\pi)\Big(\widehat{\theta}(\pi) - \theta(\pi)\Big) \Rightarrow \mathcal{Z}_\infty(\pi); \text{ for each } \pi \in \mathcal{P}$$

where convergence occurs in some metric space $(B, \|\cdot\|_B)$, where $A_n(\pi)$ is a sequence of scalers, possibly data dependent.

Assumption 2.2 Let

$$c(1-\alpha,\pi) := \ \alpha\text{-quantile of } \ \|\mathcal{Z}_\infty(\pi)\|_B$$

and suppose that the distribution function of $\|\mathcal{Z}_\infty(\pi)\|_B$ is continuous at $c(1-\alpha, P)$. Estimates are available such that $\widehat{c}(1-\alpha,\pi) \to_p c(1-\alpha,\pi)$ for each π.

With these assumptions, we can show the following generic result:, covering the case of known $\mathcal{P}$.

Proposition 2. *Let*

$$C_{1-\alpha}(\pi) := \Big\{\theta \in \Theta : \Big\|A_n(\pi)\Big(\widehat{\theta}(\pi) - \theta\Big)\Big\|_B \leq \widehat{c}(1-\alpha,\pi)\Big\}.$$

Let

$$CR_{1-\alpha} := \bigcup\nolimits_{\pi\in\mathcal{P}} C_{1-\alpha}(\pi).$$

Then

$$\liminf_{n\to\infty} P\Big\{\theta(\pi_0) \in CR_{1-\alpha}\Big\} \geq 1-\alpha.$$

Estimation of $\mathcal{P}$ poses more challenges. From (3.7), estimation of $\mathcal{P}$ is equivalent to estimation of the boundary function:

$$\ell(W) = \min_{z \in \mathcal{Z}} F_X[0|W, z].$$

Let $\widehat{\ell}(W)$ be a suitable estimate of this function. One example is

$$\widehat{\ell}(W) := \min_{z \in \mathcal{Z}} \widehat{F}_X[0|W, z]. \tag{3.9}$$

We assume that the model structure is sufficiently regular to permit the following estimator properties:

Assumption. 2.3 Let $\widehat{\kappa}_n(1-\beta)$ and the known scaler $B_n(W)$ to be such that

$$\ell(W) - \widehat{\ell}(W) \leq B_n(W)\widehat{\kappa}_n(1-\beta)$$

for all W with probability at least $1-\beta$.

Conservative forms of confidence regions of this type are available from the literature on simultaneous confidence bands. For instance, for $\widehat{\ell}(W) = \min_{z \in \mathcal{Z}} \widehat{F}_X[0|W, z]$. above, if we set

$$\hat{z} = arg \min_{z \in \mathcal{Z}} \widehat{F}_X[0|W, z]$$

and

$$B_n(W) := \left[\text{s.e.}(\widehat{F}(W, z))\right]_{z=\hat{z}_0(W)} \quad \kappa_n(1) = 2\sqrt{\log n}$$

then Assumption 2.3 holds. Sharper confidence regions for minimized functions are likely available, but their construction is relatively unexplored. For some initial results of this type, see Chernozhukov, Lee and Rosen (2008).

Let $\pi(W)$ belong to the parameter set $\mathcal{P}$. Then the confidence region for $\pi(W)$ is given by

$$CR'_{1-\beta} = \{\pi \in \Pi : \pi(W) - \widehat{l}(W) \leq B_n(W)\widehat{\kappa}_n(1-\beta)\}$$

We combine this with the previous proposition to obtain

Proposition 3. *Let*

$$CR_{1-\alpha} := \bigcup\nolimits_{\pi \in CR'_{1-\beta}} C_{1-\alpha}(\pi).$$

Then

$$\liminf_{n\to\infty} P\Big\{\theta(\pi_0) \in CR_{1-\alpha}\Big\} \geq 1-\alpha-\beta.$$

3.3. **Available Estimation Strategies.** We can proceed by either quantile or distribution methods, or some combination of the two. The key functions, used as inputs in quantile-based calculations given in the previous section, are as follows:

1. Conditional quantile function of selected regressor X given Z and other covariates W:

$$Q_X[\cdot|Z,W]$$

2. Conditional quantile function of outcome Y regressor X, covariates W, and the control function V_π:

$$Q_Y[\cdot|X,W,V_\pi]$$

   and its functionals θ_π such as average derivatives

$$E\partial Q_Y[\cdot|X,W,V_\pi]/\partial X.$$

3. Conditional mean function of outcome Y regressor X, covariates W, and the control function V_π:

$$E[Y|X,W,V_\pi]$$

   its functionals θ_π such as average derivatives

$$E\partial E[Y|X,W,V_\pi]/\partial X.$$

The key functions, used as inputs in distribution-based calculations given in the previous section, are as follows:

4. Conditional distribution function of selected regressor X given Z and other covariates W:

$$F_X[\cdot|Z, W]$$

5. Conditional distribution function of outcome Y regressor X, covariates W, and the control function V_π:

$$F_Y[\cdot|X, W, V_\pi]$$

and its functionals θ_π such as average derivatives

$$E\partial F_Y[\cdot|X, W, V_\pi]/\partial X.$$

6. Conditional mean function of outcome Y given regressor X, covariates W, and the control function V_π:

$$E[Y|X, W, V_\pi]$$

its functionals θ_π such as average derivatives

$$E\partial E[Y|X, W, V_\pi]/\partial X.$$

There are many available methods for all of the above options. Semiparametric methods include for the items listed in this order:

1 & 2. Censored quantile regression (Powell, 1984, Chernozhukov and Hong, 2002), and other quantile methods based on parametric models such as the classical Tobit. Two-step quantile regression with estimated regressor, as in Koenker and Ma (2006) and Lee (2006).

4 & 5. Various semi-parametric models for conditional distribution function, see e.g. Han and Hausman (1997), and Chenozhukov, Fernandez-Val, Melly (2007). Estimated regressors are covered by the theory in Newey and McFadden (1994).

3 & 6. Usual least squares with estimated regressor, as in Newey and McFadden (1994).

Nonparametric methods include for the items listed in this order:

1 & 2. Nonparametric quantile regression and its versions with estimated regressors. The former is available in Chaudhuri, Chaudhuri, Doskum, and Samarov, Belloni and Chernozhukov (2007), and the latter in Lee (2006).

4 & 5. Various nonparametric models for conditional distribution function, see e.g. Hall, Wolff, Yao (1997) and Chernozhukov and Belloni (2007). Estimated regressors are covered by the theory in Imbens and Newey (2006) for the cases of locally polynomial estimates.

3 & 6. Least squares with estimated regressor, as in Newey, Powell, Vella (2004).

Moreover, in quantile strategies it is easy to deal with either additive or nonadditive specifications. The distribution approaches are mostly geared towards nonadditive specifications, though location models allow to treat additive ones as well. By additivity here we mean the additivity or non-additivity of disturbances entering the first stage.

In the next version of the paper, we will give two complete algorithms of estimation, with all the details.

4. The Marginal Propensity to Consume out of Housing Wealth

Recent experience in housing markets has changed the composition of household wealth. In many countries such as the United States, housing prices have increased over a long period, followed by substantial softening. The market for housing debt, especially the risky subprime mortgage market, has experienced liquidity shortages that have resulted in increased volatility in many financial markets. From a policy perspective, the crisis in the subprime mortgage market is relevant in so far as it has an effect on consumption and overall economic activity. For instance, if a drop in housing prices occurs because of the crisis, and that drop leads to a severe contraction of consumption at the household level, then central bank intervention is likely to occur — and indeed, it has occurred recently.

Currently, two topics dominate the monetary policy discussion. First, what impact will the subprime crisis have on aggregate demand? Second, what distribution effects will occur because of the crisis? The first question requires an assessment of the marginal propensity to consume out of housing wealth, and the second depends on the exposure that households and banks had prior to the subprime crisis.

Surprisingly, the literature does not agree on the "right" measure of the marginal propensity of consumption out of housing wealth. Some papers find marginal propensities of 15 to 20 percent (e.g. Benjamin, Chinloy and Jud (2004)) while others report relatively low estimates of 2 percent in the short run and 9 percent in the long run (e.g.Carroll, Otsuka and Slacalek (2006)). Research in this area is very active, and what policy makers have done, is to take a conservative approach.[3]

One of the problems of estimation is the fact that variables such as income and housing wealth are endogenous and, in most surveys, also censored. The literature typically drops the censored observations, and tries to estimate the relationship by incorporating some non-linearities. As we have discussed, this is likely to bias the results, and therefore could have a role in why there is no agreement on a standard set of estimates. We feel that the estimation of the marginal propensity of consumption out of housing wealth is a good situation for using the methodologies developed here to shed light an a reasonable range of parameter values applicable to the design of policy.

[3] This impact of housing wealth is of primary interest for the world economy, not just the US. See, for instance, Catte, Girouard, Price and Andre (2004) and Guiso, Paiella and Visco (2005) for European estimates in the range of 3.5 percent. Asian estimates are in a simlar range; see Cutler (2004) for estimates of 3.5 percent for Hong Kong.

We have data on U.S. household consumption and wealth from Parker (1999). These data are constructed by imputing consumption spending for observed households in the Panel Survey of Income Dynamics (PSID), using the Consumer Expenditure Survey (CEX). Income data is preprocessed — original observations on income are top-coded, but all households with a top-coded income value have been dropped in the construction of our data.

We estimate a 'permanent income' style of consumption model:

$$\ln C_{it} = \alpha + \beta_{PY} \ln PY_{it} + \beta_H \ln H_{it} + \beta_{OW} \ln OW_{it} + \beta_Y \ln Y_{it} + U_{it}^* \tag{4.1}$$

Here C_{it} is consumption spending, PY_{it} is a constructed permanent component of income (human capital), H_{it} is housing wealth, OW_{it}, is other wealth and Y_{it} is current income. Our focus is the elasticity β_H, the propensity to consume out of housing wealth.

Log housing wealth takes on many zero values, which we model as the result of bound censoring and selection. These features arise first by the treatment of mortgage debt (we do not observe negative housing wealth values) and by the choice of renting versus owning of a household's residence. We view the composition of wealth between housing and other financial assets as endogenous, being chosen as a function of household circumstances and likely jointly with consumption decisions. For instruments, we use lagged values of current income, permanent income and other wealth.. Thus, we model log housing is a *Tobin regressor*.

One implication of the *Tobin regressor* structure is that all standard OLS and IV estimates are biased; including estimates that take into account either censoring or endogeneity, but not both. In Table 1, we present OLS and IV estimates for various subsamples of the data. The OLS estimates are all low; 2.8% for all data, 3.4% for households with observed lag values, and 5.3% for the "complete cases, " or households with nonzero housing values. The IV estimate for the complete cases is at least a five-fold increase, namely 28.1%.

Table 1. Basic Estimates of Housing Effects

	All Households	Households with observed IV	Nonzero Housing Wealth (Complete Cases)
Sample Size	6,647	3,256	2,126
OLS	.028	0.034	0.053
	(004)	(0.005)	(0.010)
IV (TSLS)			0.281
			(0.030)

Our procedure involves three steps. First, we establish a range for the selection probability by studying the probability of censoring. Second, we compute quantile regressions of the Tobin regressor for different values of the selection probability as in (2.9) or (3.8), and then estimate the control function for each probability value. Third, we estimate the model (4.1), including the estimated control function, as in (2.10) or (2.11). Our set estimates coincide with the range of coefficients obtained for all the different selection probability values. Their confidence intervals are given by the range of upper and lower confidence limits for coefficient estimates.

The first step is to estimate the probability that $\ln H = 0$ given values of PY, Y and OW, and find its minimum over the range of our data, implementing (3.9). Specifically, we used a probability model with Cauchy tails, including polynomial terms in the regressors. We used the procedure of Koenker and Yoon (2007), which is implemented in R. The minimum values were small, and, as a result we chose a rather low yet conservative value of $\widehat{\ell}(W) = .04$. After adjusting by two times standard error times a log factor, the upper bound estimate became .09. (We have illustrated this calculation graphically in Figure 3). Thus, for the remaining steps, we do computations for a grid of values over the range $\pi \in [0, .09]$. The procedure and results for the mean log consumption regressions are summarized in Figures 4 and 7. In particular, we plot the estimates for each value of π, as well as the associated

Table 2. Confidence Sets for Housing Effects

	Set Estimate	Confidence Region
Housing coefficient β_H		
Mean Outcome	$[.141, .149]$	$[.116, .175]$
90% Quantile	$[.165, .167]$	$[.127, .206]$
10% Quantile	$[.141, .188]$	$[.081, .263]$
Selection Probability π	$[0, .04]$	$[0, .09]$

confidence interval. The set estimates are the projections of those curves onto the left axis. As the Figure shows, there is relatively little variation in the coefficient estimates, so the set estimates are fairly sharp. We also computed estimates for the upper and lower ranges of consumption values, namely the 90% quantile and the 10% quantile. For the low ranges, there was more variation in the coefficients. We present the results for the housing effects in Figures 5, 6 and 8 and 9.

All the results on housing effects are summarized in Table 2. We note that all results are substantially (and significantly) larger than the OLS estimates (2.8%-5.3%), which ignore endogeneity. All results are substantially smaller than the IV estimate of 28.1%, which ignores censoring. Relative to the policy debate on the impact of housing wealth, our results fall in a very plausible range.

5. Summary and Conclusion

We have presented a general set of identification and estimation results for models with a Tobin regressor, a regressor that is endogenous and mismeasured by bound censoring and (independent) selection. Tobin regressor structure arises very commonly with observations on financial variables, and our results are the first to deal with endogeneity and censoring together. As such, we hope our methods provide a good foundation for understanding of how top-coding, bottom-coding and selection distort the estimated impacts of changes in income, wealth, dividends and other financial variables.

Our results are restricted to particular forms of censoring. It is not clear how to get around this issue, because endogeneity requires undoing the censoring, and undoing the censoring (seemingly) requires understanding its structure. Here we separate selection and bound censoring with independence, use quantile regression to address bound censoring, and identify parameter sets for the range of possible selection probability values. Recent advances in the theory of set inference allow straightforward construction of confidence intervals for inference on parameter values.

One essential feature of our framework is that the censoring is not complete, namely that some true values of the censored variable are observed. Such "complete cases" provide the data for our estimation of the main equation of interest. However, not all forms of censoring involve observing complete cases. Suppose, for instance, that we were studying household data where all that we observe is whether the household is poor or not; or that their income falls below the poverty line threshold. In that case, using the "poor" indicator is a severely censored form of income, and no complete cases (income values) are observed. Our methods do not apply in this case, although it is of substantial practical interest.

References

[1] Ai, Chunrong (1997), "An Improved Estimator for Models with Randomly Missing Data," *Nonparametric Statistics*, 7, 331-347.

[2] Altonji, J. and R. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogneous Regressors," *Econometrica*, 73, 1053-1103.

[3] Belloni, A., and V. Chernozhukov (2007): "Conditional Quantile and Probability Processes under Increasing Dimension," preprint.

[4] Benjamin, J.D., P. Chinloy and G.D. Donald (2004) "Why Do Households Concentrate Their Wealth in Housing?" *Journal of Real Estate Research*, Oct-Dec.

[5] Blundell, R., and J.L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," chapter 8 in M. Dewatripont, L. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics*, Cambridge University Press, Cambridge, 312-357.

[6] Carroll, C., M. Otsuka and J. Slacalek (2006), "How Large Is the Housing Wealth Effect? A New Approach," NBER Working Paper.

[7] Catte, P., N. Girouard, R. Price and C. André (2004), "Housing Markets, Wealth and the Business Cycle," OECD Working Paper no. 394.

[8] Chaudhuri, P.; K. Doksum and A. Samarov (1997), "On Average Derivative Quantile Regression," *Annals of Statistics*,. 25 (2), , 715–744.

[9] Chaudhuri, S. (1991), "Nonparametric Estimates of Regression Quantiles and their Local Bahadur Representation," *Annals of Statistics,* 19(2), 760–777.

[10] Chen, X., H. Hong and E. Tamer (2005), "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343-366.

[11] Chen, X., H. Hong and A. Tarozzi (2004), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," Working Paper, November.

[12] Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245-261.

[13] Chernozhukov, V. and H. Hong (2002), "Three-step Censored Quantile Regression and Eextramarital Affairs,". *Journal of the American Statistical Association*, 97, 872–882.

[14] Chernozhukov, V., H. Hong and E. Tamer, (2007), "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243-1284.

[15] Chernozhukov, V., I. Fernadez-Val and B. Melly (2007). “Inference on Counterfactual Distributions,” preprint.

[16] Chernozhukov, V, S. Lee and A. Rosen (2008)

[17] Chesher, A. (2003), “Identification in Nonseparable Models,” *Econometrica* 71, 1405-1441.

[18] Cutler, J. (2004), “The Relationship between Consumption, Income and Wealth in Hong Kong,” HKIMR Working Paper No.1/2004.

[19] Guiso, L. M. Paiella and I. Visco (2005), “Do capital gains affect consumption? Estimates of wealth effects from Italian households’ behavior” Economic Working Paper no. 555, Bank of Italy.

[20] Hall, P., R. Wolff, and Q. Yao (1999), “Methods for estimating a conditional distribution function,” *Journal of the American Statistical Association* 94, pp. 154–163.

[21] He, X. and Q-M Shao (2000), “On Parameters of Increasing Dimensions,” *Journal of Multivariate Analysis* 73(1) , 120–135.

[22] Imbens, G.W. and W.K. Newey, (2005), “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” MIT Working Paper, revised September.

[23] Keonker, R. (2005), *Quantile Regression*, Cambridge University Press, Cambridge.

[24] Koenker R. and L. Ma (2006), "Quantile Regression Methods for Recursive Structural Equation Models", forthcoming *Journal of Econometrics.*

[25] Koenker R. and J. Yoon (2007), "Parametric Links for Binary Response Models", forthcoming, *Journal of Econometrics.*

[26] Lee, S. (2007), "Endogeneity in Quantile Regression Models: A Control Function Approach", *Journal of Econometrics*, forthcoming.

[27] Liang, H, S. Wang, J.M. Robins and R.J. Carroll (2004), "Estimation in Partially Linear Models with Missing Covariates," *Journal of the American Statistical Association*, 99, 357-367.

[28] Little, R. J. A. (1992), "Regression with Missing X’s: A Review," *Journal of the American Statistical Association*, 87, 1227-1237.

[29] Little, R. J. A. and D. B. Rubin (2002), *Statistical Analysis with Missing Data, 2nd edition*, John Wiley and Sons, Hoboken, New Jersey.

[30] Manski, C.F. (2003), *Partial Identification of Probability Distributions*, Springer, Springer Series in Statistics.

[31] Manski, C.F. and E. Tamer (2002) “Inference on Regressions with Interval Data on a Regressor or Outcome,” *Econometrica*, 70, 519-546.

[32] Newey, W. K, J. L. Powell and F.Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models,". *Econometrica* 67, 565–603.

[33] Newey, W. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, vol. IV*, North-Holland, Amsterdam, 2111–2245.

[34] Nicoletti, C. and F. Peracchi (2005), The Effects of Income Imputations on Micro Analyses: Evidence from the ECHP," Working Paper, University of Essex, August.

[35] Powell, J.L. (1984), "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303-325.

[36] Rigobon, R and T. M. Stoker (2006a) "Bias from Censored Regressors", MIT Working Paper, September.

[37] Rigobon, R and T. M. Stoker (2006b) "Testing for Bias from Censored Regressors", MIT Working Paper, revised February.

[38] Tripathi, G. (2004), "GMM and Empirical Likelihood with Incomplete Data," Working Paper.

[39] Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24-36.

Current address:, Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA, vchern@mit.edu;, Sloan School of Management, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA, rigobon@mit.edu;, Sloan School of Management, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA, tstoker@mit.edu.,