Censored quantile instrumental variable estimation with Stata

Victor Chernozhukov
MIT
Cambridge, Massachusetts
vchern@mit.edu

Ivan Fernandez-Val
Boston University
Boston, Massachusetts
ivanf@bu.edu

Sukjin Han
UT Austin
Austin, Texas
sukjin.han@austin.utexas.edu

Amanda Kowalski
University of Michigan
Ann Arbor, Michigan
aekowals@umich.edu

Abstract. Many applications involve a censored dependent variable and/or an endogenous independent variable. Chernozhukov et al. (2015) introduced a censored quantile instrumental variable estimator (CQIV) for use in those applications, which has been applied by Kowalski (2016), among others. In this article, we introduce a Stata command, `cqiv`, that simplifes application of the CQIV estimator in Stata. We summarize the CQIV estimator and algorithm, we describe the use of the `cqiv` command, and we provide empirical examples.

Keywords: st0001, cqiv, quantile regression, censored data, endogeneity, instrumental variable, control function.

1 Introduction

Chernozhukov et al. (2015) introduced a censored quantile instrumental variable (CQIV) estimator. In this article, we introduce a Stata command, `cqiv`, that implements the CQIV estimator in Stata. Our goal is to facilitate the use of the `cqiv` command in a wide set of applications.

Many applications involve censoring as well as endogeneity. For example, suppose that we are interested in the price elasticity of medical expenditure, as in Kowalski (2016). Medical expenditure is censored from below at zero, and the price of medical care is endogenous to the level of medical expenditure through the structure of the insurance contract. Given an instrument for the price of medical care, the CQIV estimator facilitates estimation of the price elasticity of expenditure on medical care in a way that addressess censoring as well as endogeneity.

The CQIV estimator addresses censoring using the censored quantile regression (CQR) approach of Powell (1986), and it addresses endogeneity using a control function approach. For computation, the CQIV estimator adapts the Chernozhukov and Hong (2002) algorithm for CQR estimation. An important side feature of the `cqiv` stata command is that it can also be used in quantile regression applications that do not include censoring or endogeneity.

In section 2, we summarize the theoretical background on the CQIV command, following Chernozhukov et al. (2015). In section 3, we introduce the use of the CQIV command. We provide an empirical application with examples that involve estimation of Engel curves, as in Chernozhukov et al. (2015).

2 Censored quantile IV estimation

We first describe a model of triangular system for the CQIV regression. Suppose Y is an observed response variable obtained by censoring a continuous latent response Y^* from below at the level determined by the variable C. Let D be the continuous regressor of interest, possibly endogenous, and W be a vector of covariates, possibly containing C, and Z is a vector of (possibly discrete) IVs excluded from the equation for Y^*.[1] We observe $\{Y_i, D_i, W_i, Z_i, C_i\}_{i=1}^n$, a sample of size n of independent and identically distributed observations from the random vector (Y, D, W, Z, C), which obeys

$$\begin{aligned} Y &= \max(Y^*, C), & (1) \\ Y^* &= Q_{Y^*}(U \mid D, W, V) = X'\beta_0(U), & (2) \\ D &= Q_D(V \mid W, Z), & (3) \end{aligned}$$

where V is a latent unobserved variable that accounts for the possible endogeneity of D, $X = x(D, W, V)$ with $x(D, W, V)$ being a vector of transformations of (D, W, V), $Q_{Y^*}(u \mid D, W, V)$ is the u-quantile of Y^* conditional on (D, W, V), $Q_D(v \mid W, Z)$ is the v-quantile of D conditional on (W, Z), and

$$\begin{aligned} U &\sim U(0,1) \mid D, W, Z, V, C, \\ V &\sim U(0,1) \mid W, Z, C. \end{aligned}$$

This CQIV regression model nests the uncensored case of the quantile IV (QIV) regression by making C arbitrarily small. As an example for the CQIV model, in the Engel curve application of Chernozhukov et al. (2015), Y is the expenditure share in alcohol, bounded from below at $C = 0$, D is total expenditure on nondurables and services, W are household demographic characteristics, and Z is labor income measured by the earnings of the head of the household. Total expenditure is likely to be jointly determined with the budget composition in the household's allocation of income across consumption goods and leisure. Thus, households with a high preference to consume "non-essential" goods such as alcohol tend to expend a higher proportion of their incomes and therefore to have a higher expenditure. The control variable V in this case is the marginal propensity to consume, measured by the household ranking in the conditional distribution of expenditure given labor income and household characteristics. This propensity captures unobserved preference variables that affect both the level and composition of the budget. Under the conditions for a two stage budgeting decision process (Gorman (1959)), where the household first divides income between consumption and leisure/labor and then decide the consumption allocation, some sources of income can provide plausible exogenous variation with respect to the budget shares. For example, if preferences are weakly separable in consumption and leisure/labor, the consumption budget shares do not depend on labor income given the consumption expenditure (see, e.g., Deaton et al. (1980)). This justifies the use of labor income as an exclusion restriction.

A simple version of the model (1)–(3) is as follows:

$$Y^* = \beta_{00} + \beta_{01} D + \beta_{02} W + \Phi^{-1}(\epsilon), \quad \epsilon \sim U(0,1) \tag{4}$$

where Φ^{-1} denotes the quantile function of the standard normal distribution, and also assume that $(\Phi^{-1}(V),\ \Phi^{-1}(\epsilon))$ is jointly normal with correlation ρ_0. From the properties of the multivariate normal distribution, $\Phi^{-1}(\epsilon) = \rho_0 \Phi^{-1}(V) + (1 - \rho_0^2)^{1/2} \Phi^{-1}(U)$, where $U \sim U(0,1)$. This result yields a specific expression for the conditional quantile function of Y^*:

$$Q_{Y^*}(U \mid D, W, V) = X'\beta_0(U) = \beta_{00} + \beta_{01} D + \beta_{02} W + \rho_0 \Phi^{-1}(V) + (1 - \rho_0^2)^{1/2} \Phi^{-1}(U), \tag{5}$$

[1] We consider a single endogenous regressor D in the model and in the cqiv procedure.

where V enters the equation through $\Phi^{-1}(V)$.

Given this model, Chernozhukov et al. (2015) introduce the estimator for the parameter $\beta_0(u)$ as

$$\widehat{\beta}(u) = \arg \min_{\beta \in \mathbb{R}^{\dim(X)}} \frac{1}{n} \sum_{i=1}^{n} 1(\widehat{S}_i' \widehat{\gamma} > \varsigma) \rho_u (Y_i - \widehat{X}_i' \beta), \tag{6}$$

where $\rho_u(z) = (u - 1(z < 0))z$ is the asymmetric absolute loss function of Koenker and Bassett (1978), $\widehat{X}_i = x(D_i, W_i, \widehat{V}_i)$, $\widehat{S}_i = s(\widehat{X}_i, C_i)$, $s(X, C)$ is a vector of transformations of (X, C), ς is a positive cut-off, and $\widehat{V}_i$ is an estimator of V_i which is described below.

The estimator in (6) adapts the algorithm of Chernozhukov and Hong (2002) developed for the censored quantile regression (CQR) estimator to a setting where there is possible endogeneity. As described in Chernozhukov et al. (2015), this algorithm is based on the following implication of the model:

$$P(Y \leq X'\beta_0(u) \mid X, C, X'\beta_0(u) > C) = P(Y^* \leq X'\beta_0(u) \mid X, C, X'\beta_0(u) > C) = u,$$

provided that $P(X'\beta_0(u) > C) > 0$. In other words, $X'\beta_0(u)$ is the conditional u-quantile of the observed outcome for the observations for which $X'\beta_0(u) > C$, i.e., the conditional u-quantile of the latent outcome is above the censoring point. These observations change with the quantile index and may include censored observations. Chernozhukov et al. (2015) refer to them as the "quantile-uncensored" observations. The multiplier $1(\widehat{S}_i'\widehat{\gamma} > \varsigma)$ is a selector that predicts if observation i is quantile-uncensored. For the conditions on this selector, consult Assumptions 4(a) and 5 in Chernozhukov et al. (2015).

`cqiv` implements the censored quantile instrumental variable (CQIV) estimator which is computed using an iterative procedure where each step takes the form specified in equation (6) with a particular choice of $1(\widehat{S}_i'\widehat{\gamma} > \varsigma)$. We briefly describe this procedure here and then provide a practical algorithm in the next section. The procedure first selects the set of quantile-uncensored observations by estimating the conditional probabilities of censoring using a flexible binary choice model. Since $\{X'\beta_0(u) > C\} \equiv \{P(Y^* \leq C \mid X, C) < u\}$, quantile-uncensored observations have conditional probability of censoring lower than the quantile index u. The linear part of the conditional quantile function, $X_i'\beta_0(u)$, is estimated by standard quantile regression using the sample of quantile-uncensored observations. Then, the procedure updates the set of quantile-uncensored observations by selecting those observations with conditional quantile estimates that are above their censoring points, $X_i'\widehat{\beta}(u) > C_i$, and iterate.

`cqiv` provides different ways of estimating the control variable V, which can be chosen with option `firststage(`*string*`)`. Note that if $Q_D(v \mid W, Z)$ is invertible in v, the control variable has several equivalent representations:

$$V = \vartheta_0(D, W, Z) \equiv F_D(D \mid W, Z) \equiv Q_D^{-1}(D \mid W, Z) \equiv \int_0^1 1\{Q_D(v \mid W, Z) \leq D\} dv, \tag{7}$$

where $F_D(D \mid W, Z)$ is the distribution of D conditional on (W, Z). Different estimators of V can be constructed based on parametric or semiparametric models for $F_D(D \mid W, Z)$ and $Q_D(V \mid W, Z)$. Let $R = r(W, Z)$ with $r(W, Z)$ being a vector of collecting transformations of (W, Z) specified by the researcher. When *string* is `quantile`, a quantile regression model is assumed, where $Q_D(v \mid W, Z) = R'\pi_0(v)$ and

$$V = \int_0^1 1\{R'\pi_0(v) \leq D\} dv.$$

The estimator of V then takes the form

$$\widehat{V} = \tau + \int_{\tau}^{1-\tau} 1\{R'\widehat{\pi}(v) \leq D\} dv, \tag{8}$$

where $\widehat{\pi}(v)$ is the Koenker and Bassett (1978) quantile regression estimator which is calculated within `cqiv` using the built-in `qreg` command in Stata, and τ is a small positive trimming constant that avoids estimation of tail quantiles. The integral in (8) can be approximated numerically using a finite grid of quantiles.[2] Specifically, the fitted values for pre-specified quantile indices (whose number n_q is controlled by option `nquant(#)`) are calculated, which then yields

$$\widehat{V}_i = \frac{1}{n_q} \sum_{j=1}^{n_q} 1\{R_i'\widehat{\pi}(v_j) \leq D_i\}.$$

For other related quantile regression models that can alternatively be used, see Chernozhukov et al. (2015).

When *string* is `distribution`, ϑ_0 is estimated using distribution regression. In this case we consider a semiparametric model for the conditional distribution of D to construct a control variable

$$V = F_D(D \mid W, Z) = \Lambda(R'\pi_0(D)),$$

where Λ is a probit or logit link function; this can be chosen using option `ldv1(`*string*`)` where *string* is either `probit` or `logit`. The estimator takes the form

$$\widehat{V} = \Lambda(R'\widehat{\pi}(D)), \tag{9}$$

where $\widehat{\pi}(d)$ is the maximum likelihood estimator of $\pi_0(d)$ at each d (see, e.g., Foresi and Peracchi (1995), and Chernozhukov et al. (2013)).[3] The expression (9) can be approximated by considering a finite grid of evenly-spaced thresholds for the conditional distribution function of D, where the number of thresholds n_t is controlled by option `nthresh(#)`. Concretely, for threshold d_j with $j = 1, ..., n_t$,

$$\widehat{V}_i = \Lambda(R_i'\widehat{\pi}(d_j)), \quad \text{for } i\text{'s s.t. } d_{j-1} \leq D_i < d_j \text{ with } d_0 = -\infty \text{ and } d_{n_t} = \infty,$$

where $\widehat{\pi}(d_j)$ is probit or logit estimate with $\tilde{D}_i(d_j) = 1\{D_i \leq d_j\}$ as a dependent variable and R_i as regressors.

Lastly, when *string* is `ols`, a linear regression model $D = R'\pi_0 + V$ is assumed and $\widehat{V}$ is a transformation of the OLS residual:

$$\widehat{V}_i = \Phi((D_i - R_i'\widehat{\pi})/\widehat{\sigma}), \tag{10}$$

where Φ is the standard normal distribution, $\widehat{\pi}$ is the OLS estimator of π_0, and $\widehat{\sigma}$ is the estimator of the error standard deviation. In estimation of (6) using `cqiv`, we assume that the control function $\widehat{V}$ enters the equation through $\Phi^{-1}(\widehat{V})$. This is motivated by the example (4)–(5).

[2] The use of the integral to obtain a generalized inverse is convenient to avoid monotonicity problems in $v \mapsto R'\widehat{\pi}(v)$ due to misspecification or sampling error. Chernozhukov et al. (2010) developed asymptotic theory for this estimator.

[3] Chernozhukov et al. (2013) developed asymptotic theory for this estimator.

2.1 CQIV algorithm

The algorithm recommended in Chernozhukov et al. (2015) to obtain CQIV estimates is similar to Chernozhukov and Hong (2002), but it additionally has an initial step to estimate the control variable V. This step is numbered as 0 to facilitate comparison with the Chernozhukov and Hong (2002) 3-Step CQR algorithm.

For each desired quantile u, perform the following steps:

0. Obtain $\widehat{V}_i = \widehat{\vartheta}(D_i, W_i, Z_i)$ from (8), (9) or (10) and construct $\widehat{X}_i = x(D_i, W_i, \widehat{V}_i)$.

1. Select a set of quantile-uncensored observations $J_0 = \{i : \Lambda(\widehat{S}_i'\widehat{\delta}) > 1 - u + k_0\}$, where Λ is a known link function, $\widehat{S}_i = s(\widehat{X}_i, C_i)$, s is a vector of collecting transformations specified by the researcher, k_0 is a cut-off such that $0 < k_0 < u$, and $\widehat{\delta} = \arg\max_{\delta \in \mathbb{R}^{\dim(S)}} \sum_{i=1}^{n} \{1(Y_i > C_i) \log \Lambda(\widehat{S}_i'\delta) + 1(Y_i = C_i) \log[1 - \Lambda(\widehat{S}_i'\delta)]\}$.

2. Obtain the 2-step CQIV coefficient estimates: $\widehat{\beta}^0(u) = \arg\min_{\beta \in \mathbb{R}^{\dim(X)}} \sum_{i \in J_0} \rho_u(Y_i - \widehat{X}_i'\beta)$, and update the set of quantile-uncensored observations, $J_1 = \{i : \widehat{X}_i'\widehat{\beta}^0(u) > C_i + \varsigma_1\}$.

3. Obtain the 3-step CQIV coefficient estimates $\widehat{\beta}^1(u)$, solving the same minimization program as in step 2 with J_0 replaced by J_1.[4]

Remark 1 (Step 1). To predict the quantile-uncensored observations, a probit, logit, or any other model that fits the data well can be used. `cqiv` provides option `ldv2(`*string*`)` where *string* can be either `probit` or `logit`. Note that the model does not need to be correctly specified; it suffices that it selects a nontrivial subset of observations with $X_i'\beta_0(u) > C_i$. To choose the value of k_0, it is advisable that a constant fraction of observations satisfying $\Lambda(\widehat{S}_i'\widehat{\delta}) > 1 - u$ are excluded from J_0 for each quantile. To do so, one needs to set k_0 as the q_0th quantile of $\Lambda(\widehat{S}_i'\widehat{\delta})$ conditional on $\Lambda(\widehat{S}_i'\widehat{\delta}) > 1 - u$, where q_0 is a percentage (10% worked well in our simulation with little sensitivity to values between 5 and 15%). The value for q_0 can be chosen with option `drop1(`*#*`)`.

Remark 2 (Step 2). To choose the cut-off ς_1, it is advisable that a constant fraction of observations satisfying $\widehat{X}_i'\widehat{\beta}^0(u) > C_i$ are excluded from J_1 for each quantile. To do so, one needs to set ς_1 to be the q_1th quantile of $\widehat{X}_i'\widehat{\beta}^0(u) - C_i$ conditional on $\widehat{X}_i'\widehat{\beta}^0(u) > C_i$, where q_1 is a percentage less than q_0 (3% worked well in our simulation with little sensitivity to values between 1 and 5%). The value for q_1 can be chosen with option `drop2(`*#*`)`.[5]

Remark 3 (Steps 1 and 2). In terms of the notation of (6), the selector of Step 1 can be expressed as $1(\widehat{S}_i'\widehat{\gamma} > \varsigma_0)$, where $\widehat{S}_i'\widehat{\gamma} = \widehat{S}_i'\widehat{\delta} - \Lambda^{-1}(1-u)$ and $\varsigma_0 = \Lambda^{-1}(1 - u + k_0) - \Lambda^{-1}(1-u)$. The selector of Step 2 can also be expressed as $1(\widehat{S}_i'\widehat{\gamma} > \varsigma_1)$, where $\widehat{S}_i = (\widehat{X}_i', C_i)'$ and $\widehat{\gamma} = (\widehat{\beta}^0(u)', -1)'$.

[4] As an optional fourth step, one can update the set of quantile-uncensored observations J_2 replacing $\widehat{\beta}^0(u)$ by $\widehat{\beta}^1(u)$ in the expression for J_1 in step 2, and iterate this and the previous step a bounded number of times. This optional step is not incorporated in `cqiv` command, as Chernozhukov et al. (2015) find little gain of iterating in terms of bias, root mean square error, and value of Powell objective function in their simulation exercise.

[5] In practice, it is desirable that $J_0 \subset J_1$. If this is not the case, Chernozhukov et al. (2015) recommend altering q_0, q_1, or the specification of the regression models. At each quantile, the percentage of observations from the full sample retained in J_0, the percentage of observations from the full sample retained in J_1, and the percentage of observations from J_0 not retained in J_1 can be computed as simple robustness diagnostic tests. The estimator $\widehat{\beta}^0(u)$ is consistent but will be inefficient relative to the estimator obtained in the subsequent step because it uses a smaller conservative subset of the quantile-uncensored observations if $q_0 > q_1$.

2.2 Weighted Bootstrap Algorithm

Chernozhukov et al. (2015) recommend obtaining confidence intervals through either weighted bootstrap or nonparametric bootstrap procedures. We focus on weighted bootstrap here. To speed up the computation, a procedure is proposed that uses a one-step CQIV estimator in each bootstrap repetition.

For $b = 1, \dots, B$, repeat the following steps:

1. Draw a set of weights $(e_{1b}, \dots, e_{nb})$ i.i.d from the standard exponential distribution.
2. Reestimate the control variable in the weighted sample, $\widehat{V}^e_{ib} = \widehat{\vartheta}^e_b(D_i, W_i, Z_i)$, and construct $\widehat{X}^e_{ib} = x(D_i, W_i, \widehat{V}^e_{ib})$.
3. Estimate the weighted quantile regression: $\widehat{\beta}^e_b(u) = \arg\min_{\beta \in \mathbb{R}^{\dim(X)}} \sum_{i \in J_{1b}} e_{ib} \rho_u (Y_i - \beta' \widehat{X}^e_{ib})$, where $J_{1b} = \{i : \widehat{\beta}(u)' \widehat{X}^e_{ib} > C_i + \varsigma_1\}$, and $\widehat{\beta}(u)$ is a consistent estimator of $\beta_0(u)$, e.g., the 3-stage CQIV estimator $\widehat{\beta}^1(u)$.

Remark 4 (Step 2). The estimate of the control function $\widehat{\vartheta}^e_b$ can be obtained by weighted least squares, weighted quantile regression, or weighted distribution regression, depending upon which *string* is chosen among `ols`, `quantile`, or `distribution` in option `firststage(`*string*`)`.

Remark 5 (Step 3). A computationally less expensive alternative is to set $J_{1b} = J_1$ in all the repetitions, where J_1 is the subset of selected observations in Step 2 of the CQIV algorithm. This alternative is not considered in the `cqiv` routine, because while it is computationally faster, it sacrifices accuracy.

Remark 6 As discussed in Chernozhukov et al. (2015), we focus on weighted bootstrap, partly because it has practical advantages over nonparametric bootstrap to deal with discrete regressors with small cell sizes, since it avoids having singular designs under the bootstrap data generating process. The `cqiv` procedure allows both weighted and nonparametric bootstraps.

Remark 7 For a cluster bootstrap procedure with clustered data, the bootstrap weights are generated after treating the cluster unit as the unit at which observations are assumed to be independent. In this procedure, the same weight is drawn for all the observations within each cluster.

3 The cqiv command

3.1 Syntax

The syntax for `cqiv` is as follows:

```
cqiv depvar [ varlist ] (endogvar = instrument) [ if ] [ in ] [ weight ] [ , quantiles(numlist)
   censorpt(#) censorvar(varname) top uncensored exogenous firststage(string)
   firstvar(varlist) nquant(#) nthresh(#) ldv1(string) ldv2(string) corner drop1(#)
   drop2(#) viewlog confidence(string) cluster(string) bootreps(#) setseed(#)
   level(#) norobust ]
```

3.2 Description

`cqiv` conducts CQIV estimation. This command can implement both censored and uncensored quantile IV estimation either under exogeneity or endogeneity. The estimators proposed by Chernozhukov et al. (2015) are used if CQIV estimation or QIV without censoring estimation are implemented. The estimator proposed by Chernozhukov and Hong (2002) is used if CQR is estimated without endogeneity. Note that all the variables in the parentheses of the syntax are those involved in the first stage estimation of CQIV and QIV.

3.3 Option

Model

`quantiles(`*numlist*`)` specifies the quantiles at which the model is estimated and should contain percentage numbers between 0 and 100. Note that this is not the list of quantiles for the first stage estimation with the quantile regression specification.

`censorpt(`#`)` specifies the fixed censoring point of the dependent variable, where the default is 0; inappropriately specified censoring point will generate errors in estimation.

`censorvar(`*varname*`)` specifies the censoring variable (i.e., the random censoring point) of the dependent variable.

`top` sets right censoring of the dependent variable; otherwise, left censoring is assumed as default.

`uncensored` selects uncensored quantile IV (QIV) estimation.

`exogenous` selects censored quantile regression (CQR) with no endogeneity, which is proposed by Chernozhukov and Hong (2002).

`firststage(`*string*`)` determines the first stage estimation procedure, where *string* is either `quantile` for quantile regression (the default), `distribution` for distribution regression (either probit or logit), or `ols` for OLS estimation. Note that `firststage(distribution)` can take a considerable amount of time to execute.

`firstvar(`*varlist*`)` specifies the list of variables other than instruments that are included in the first stage estimation; default is all the variables that are included in the second stage estimation.

`nquant(`#`)` determines the number of quantiles used in the first stage estimation when the estimation procedure is `quantile`; default is 50, that is, total 50 evenly-spaced quantiles from 1/51 to 50/51 are chosen in the estimation; it is advisable to choose a value between 20 to 100.

`nthresh(`#`)` determines the number of thresholds used in the first stage estimation when the estimation procedure is `distribution`; default is 50, that is, total 50 evenly-spaced thresholds (i.e., the sample quantiles of *depvar*) are chosen in the estimation; it is advisable to choose a value between 20 and the value of the sample size.

`ldv1(`*string*`)` determines the LDV model used in the first stage estimation when the estimation procedure is `distribution`, where *string* is either `probit` for probit estimation (the default), or `logit` for logit estimation.

`ldv2(`*string*`)` determines the LDV model used in the first step of the second stage estimation, where *string* is either `probit` (the default), or `logit`.

CQIV estimation

`corner` calculates the (average) marginal quantile effects for censored dependent variable when the censoring is due to economic reasons such are corner solutions. Under this option, the reported coefficients are the average corner solution marginal effects if the underlying function is linear in the endogenous variable, i.e., the average of $1\{Q_{Y^*}(u \mid D, W, V) > C\}\partial_D Q_{Y^*}(u \mid D, W, V) = 1\{x(D, W, V)'\beta_0(u) > C\}\partial_D x(D, W, V)'\beta_0(u)$ over all observations. If the underlying function is nonlinear in the endogenous variable, average marginal effects must be calculated directly from the coefficients without `corner` option. For details of the related concepts, see Section 2.1 of Chernozhukov et al. (2015). The relevant example can be found in Section 3.5.

`drop1(`#`)` sets the proportion of observations q_0 with probabilities of censoring above the quantile index that are dropped in the first step of the second stage (See Remark 1 above for details); default is 10.

`drop2(`#`)` sets the proportion of observations q_1 with estimate of the conditional quantile above (below for right censoring) that are dropped in the second step of the second stage (See Remark 2 above for details); default is 3.

`viewlog` shows the intermediate estimation results; the default is no log.

Inference

`confidence(`*string*`)` specifies the type of confidence intervals. With *string* being `no`, which is the default, no confidence intervals are calculated. With *string* being `boot` or `weightedboot`, either nonparametric bootstrap or weighted bootstrap (respectively) confidence intervals are calculated. The weights of the weighted bootstrap are generated from the standard exponential distribution. Note that `confidence(boot)` and `confidence(weightboot)` can take a considerable amount of time to execute.

`cluster(`*string*`)` implements a cluster bootstrap procedure for clustered data when `confidence(weightboot)` is selected, with *string* specifying the variable that defines the group or cluster.

`bootreps(`#`)` sets the number of repetitions of bootstrap or weighted bootstrap if the `confidence(boot)` or `confidence(weightboot)` is selected. The default number of repetitions is 100.

`setseed(`#`)` sets the initial seed number in repetition of bootstrap or weighted bootstrap; the default is 777.

`level(`#`)` sets confidence level, and default is 95.

Robust check

`norobust` suppresses the robustness diagnostic test results. No diagnostic test results to suppress when `uncensored` is employed.

3.4 Saved results

`cqiv` saves the following results in `e()`:

Scalars	
`e(obs)`	number of observations
`e(censorpt)`	fixed censoring point
`e(drop1)`	q_0
`e(drop2)`	q_1
`e(bootreps)`	number of bootstrap or weighted bootstrap repetitions
`e(level)`	significance level of confidence interval
Macros	
`e(command)`	name of the command: `cqiv`
`e(regression)`	name of the implemented regression: either `cqiv`, `qiv`, or `cqr`
`e(depvar)`	name of dependent variable
`e(endogvar)`	name of endogenous regressor
`e(instrument)`	name of instrumental variables
`e(censorvar)`	name of censoring variable
`e(firststage)`	type of first stage estimation
`e(confidence)`	type of confidence intervals
Matrices*	
`e(results)`	matrix containing the estimated coefficients, mean, and lower and upper bounds of confidence intervals
`e(quantiles)`	row vector containing the quantiles at which CQIV have been estimated
`e(robustcheck)`	matrix containing the results for the robustness diagnostic test results; see Table 1 below

*Note that the entry `complete` denotes whether all the steps are included in the procedure; 1 when they are, and 0 otherwise. For other entries consult the paper.

3.5 Examples

We illustrate how to use the command by using some examples. For the dataset, we use a household expenditure dataset for alcohol consumption drawn from the British Family Expenditure Survey (FES); see Blundell et al. (2007) and Chernozhukov et al. (2015) for detailed description of the data. Using this dataset, we are interested in learning how the share of total expenditure on alcohol (`alcohol`) is affected by (logarithm of) total expenditure (`logexp`), controlling for the number of children (`nkids`). For the endogenous expenditure, we use disposable income, i.e., (logarithm of) gross earnings of the head of the household (`logwages`), as an excluded instrument. The dataset (`alcoholengel.dta`) can be downloaded from SSC as follows:

```
. ssc describe cqiv
. net get cqiv
```

The first line will show the dataset is accessible via the second line of the command. The second line will then download `alcoholengel.dta` to the current working directory. Given this dataset, we can generate part of the empirical results of Chernozhukov et al. (2015):

```
. cqiv alcohol logexp2 nkids (logexp = logwages nkids), quantiles(25 50 75)
  (output omitted)
```

Here, `logexp2` is the squared (logarithm of) total expenditure. Using `cqiv` command, the QIV estimation can be implemented with `uncensored` option:

```
. cqiv alcohol logexp2 nkids (logexp = logwages nkids), uncensored
  (output omitted)
```

And the CQR estimation with `exogenous` option:

```
. cqiv alcohol logexp logexp2 nkids, exogenous
  (output omitted)
```

Here are other possible examples of the CQIV estimation with different specifications and options. Outputs are all omitted.

```
. cqiv alcohol logexp2 (logexp = logwages), quantiles(20 25 70(5)90) firststage(ols)
. cqiv alcohol logexp2 (logexp = logwages), firststage(distribution) ldv1(logit)
. cqiv alcohol logexp2 nkids (logexp = logwages), firstvar(nkids)
. cqiv alcohol logexp2 nkids (logexp = logwages nkids), confidence(weightboot) bootreps(10)
. cqiv alcohol nkids (logexp = logwages nkids), corner
```

In order of appearance, the commands conduct: the estimation using OLS in the first stage; the estimation using distribution regression with logistic distribution; the estimation where `nkids` is the only variable other than the instrument that is included in the first stage estimation; the estimation with two instruments and calculating the confidence interval using the weighted bootstrap; and the estimation calculating the marginal effects when censoring is due to corner solutions. In this particular example, `logexp2` cannot be included in the first-stage regression when distribution regression is implemented. This is because `logexp2` is a monotone transformation of `logexp`, and thus the distribution estimation yields a perfect fit.

4 Acknowledgments

We would like to thank Simon Jäger and Blaise Melly for helpful comments.

5 References

Blundell, R., X. Chen, and D. Kristensen. 2007. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* 75(6): 1613–1669.

Chernozhukov, V., I. Fernández-Val, and A. Galichon. 2010. Quantile and probability curves without crossing. *Econometrica* 78(3): 1093–1125.

Chernozhukov, V., I. Fernández-Val, and A. E. Kowalski. 2015. Quantile regression with censoring and endogeneity. *Journal of Econometrics* 186(1): 201–221.

Chernozhukov, V., I. Fernández-Val, and B. Melly. 2013. Inference on counterfactual distributions. *Econometrica* 81(6): 2205–2268.

Chernozhukov, V., and H. Hong. 2002. Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association* .

Deaton, A., J. Muellbauer, et al. 1980. *Economics and consumer behavior.* Cambridge university press.

Foresi, S., and F. Peracchi. 1995. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association* 90(430): 451–466.

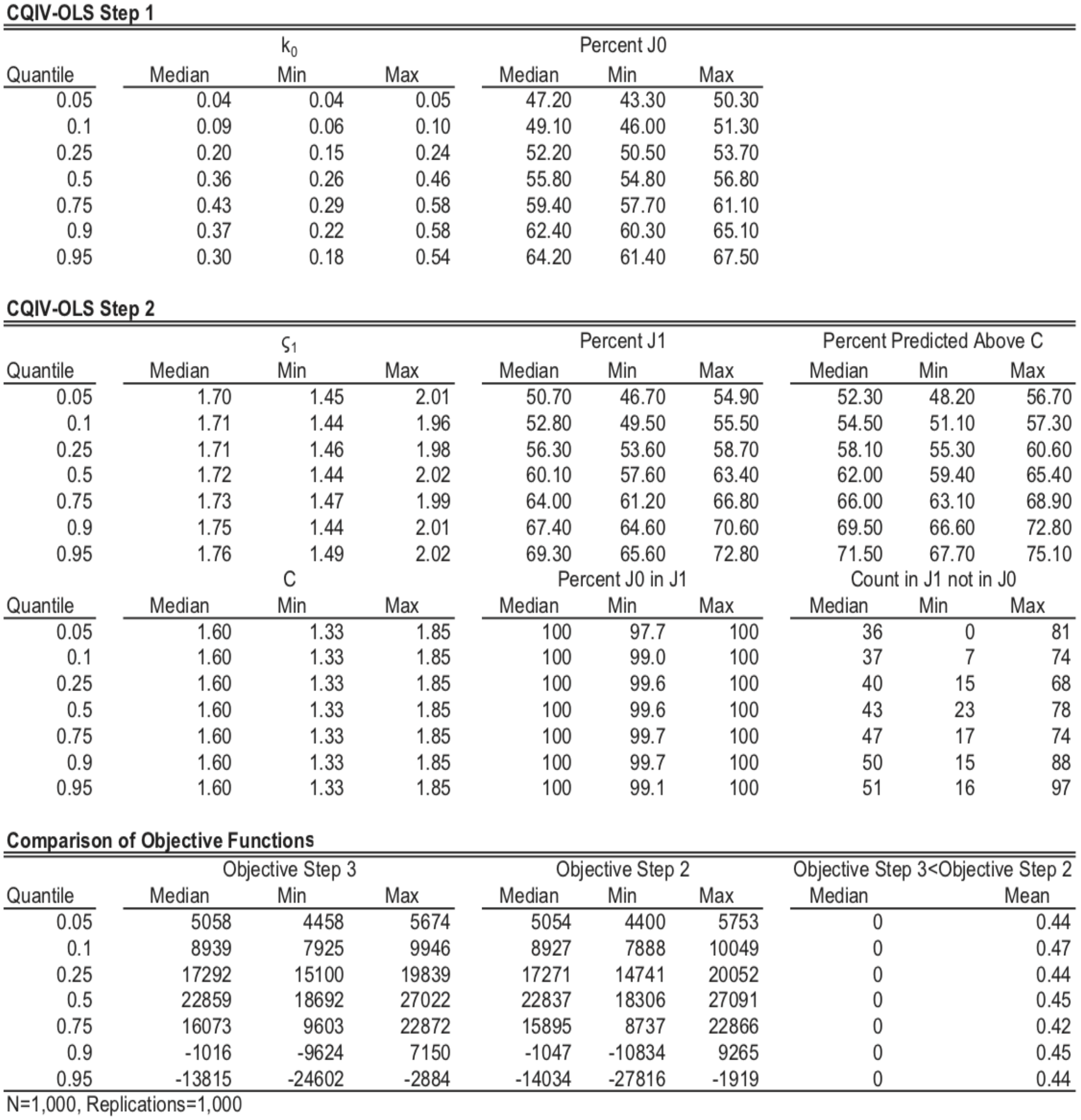
Gorman, W. M. 1959. Separable utility and aggregation. *Econometrica: Journal of the Econometric Society* 469–481.

Koenker, R., and G. Bassett, Jr. 1978. Regression quantiles. *Econometrica: journal of the Econometric Society* 33–50.

Kowalski, A. 2016. Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care. *Journal of Business & Economic Statistics* 34(1): 107–117. URL https://doi.org/10.1080/07350015.2015.1004072.

Powell, J. 1986. Censored regression quantiles. *Journal of Econometrics* 32(1): 143–155. Cited By 341. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-38249039685&doi=10.1016%2f0304-4076%2886%2990016-3&partnerID=40&md5=65bc49a1ededfc7bebae3335dc029e74.

Table 1: CQIV Robustness Diagnostic Test Results for CQIV with OLS Estimate of the Control Variable - Homoskedastic Design



CQIV-OLS Step 1

	k_0			Percent J0		
Quantile	Median	Min	Max	Median	Min	Max
0.05	0.04	0.04	0.05	47.20	43.30	50.30
0.1	0.09	0.06	0.10	49.10	46.00	51.30
0.25	0.20	0.15	0.24	52.20	50.50	53.70
0.5	0.36	0.26	0.46	55.80	54.80	56.80
0.75	0.43	0.29	0.58	59.40	57.70	61.10
0.9	0.37	0.22	0.58	62.40	60.30	65.10
0.95	0.30	0.18	0.54	64.20	61.40	67.50

CQIV-OLS Step 2

	ς_1			Percent J1			Percent Predicted Above C		
Quantile	Median	Min	Max	Median	Min	Max	Median	Min	Max
0.05	1.70	1.45	2.01	50.70	46.70	54.90	52.30	48.20	56.70
0.1	1.71	1.44	1.96	52.80	49.50	55.50	54.50	51.10	57.30
0.25	1.71	1.46	1.98	56.30	53.60	58.70	58.10	55.30	60.60
0.5	1.72	1.44	2.02	60.10	57.60	63.40	62.00	59.40	65.40
0.75	1.73	1.47	1.99	64.00	61.20	66.80	66.00	63.10	68.90
0.9	1.75	1.44	2.01	67.40	64.60	70.60	69.50	66.60	72.80
0.95	1.76	1.49	2.02	69.30	65.60	72.80	71.50	67.70	75.10

	C			Percent J0 in J1			Count in J1 not in J0		
Quantile	Median	Min	Max	Median	Min	Max	Median	Min	Max
0.05	1.60	1.33	1.85	100	97.7	100	36	0	81
0.1	1.60	1.33	1.85	100	99.0	100	37	7	74
0.25	1.60	1.33	1.85	100	99.6	100	40	15	68
0.5	1.60	1.33	1.85	100	99.6	100	43	23	78
0.75	1.60	1.33	1.85	100	99.7	100	47	17	74
0.9	1.60	1.33	1.85	100	99.7	100	50	15	88
0.95	1.60	1.33	1.85	100	99.1	100	51	16	97

Comparison of Objective Functions

	Objective Step 3			Objective Step 2			Objective Step 3<Objective Step 2	
Quantile	Median	Min	Max	Median	Min	Max	Median	Mean
0.05	5058	4458	5674	5054	4400	5753	0	0.44
0.1	8939	7925	9946	8927	7888	10049	0	0.47
0.25	17292	15100	19839	17271	14741	20052	0	0.44
0.5	22859	18692	27022	22837	18306	27091	0	0.45
0.75	16073	9603	22872	15895	8737	22866	0	0.42
0.9	-1016	-9624	7150	-1047	-10834	9265	0	0.45
0.95	-13815	-24602	-2884	-14034	-27816	-1919	0	0.44

N=1,000, Replications=1,000

In this table, we present the CQIV robustness diagnostic tests suggested in Chernozhukov et al. (2015) for the CQIV estimator with an OLS estimate of the control variable. See Section 2.1 of the present paper for the definitions of $k_0, \varsigma_1, J_0, J_1$. In our estimates, we used a probit model in the first step, and we set $q_0 = 10$ and $q_1 = 3$. In practice, we do not necessarily recommend reporting the diagnostics in Table 1, but we do recommend examining them.

In the top section of the table, we present diagnostics computed after CQIV Step 1. In the second section, we present robustness test diagnostics computed after CQIV Step 2. In the last section, we report the value of the Powell objective function obtained after CQIV Step 2 and CQIV Step 3. See Chernozhukov et al. (2015) for more discussion.