NBER WORKING PAPER SERIES

CENSORED QUANTILE INSTRUMENTAL VARIABLE ESTIMATION WITH STATA

Victor Chernozhukov
Iván Fernández-Val
Sukjin Han
Amanda Kowalski

Working Paper 24232
http://www.nber.org/papers/w24232

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2018

We would like to thank Blaise Melly for helpful comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Censored Quantile Instrumental Variable Estimation with Stata
Victor Chernozhukov, Iván Fernández-Val, Sukjin Han, and Amanda Kowalski
NBER Working Paper No. 24232
January 2018
JEL No. C26,C31,C34

ABSTRACT

Many applications involve a censored dependent variable and an endogenous independent variable. Chernozhukov, Fernandez-Val, and Kowalski (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.

Victor Chernozhukov
Department of Economics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Mass. 02139
vchern@mit.edu

Sukjin Han
Department of Economics
University of Texas
2225 Speedway, BRB 3.152
Austin, TX 78712
sukjin.han@austin.utexas.edu

Iván Fernández-Val
Department of Economics
Boston University
270 Bay State Rd
Boston, MA 02215
ivanf@bu.edu

Amanda Kowalski
Department of Economics
Yale University
37 Hillhouse Avenue
Box 208264
New Haven, CT 06520
and NBER
amanda.kowalski@yale.edu

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
Yale University
New Haven, Connecticut
amanda.kowalski@yale.edu

Abstract. Many applications involve a censored dependent variable and 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

Assumption 1. (Model) 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 the model

assumptions

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

where $X = x(D, W, V)$ 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

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

Assumption 1 considers a triangular system, where Y^* is a continuous latent response variable, Y is an observed response variable obtained by censoring Y^* from below at the level determined by the variable C, D is the continuous regressor of interest, W is a vector of covariates, possibly containing C, V is a latent unobserved variable that accounts for the possible endogeneity of D, and Z is a vector of "instrumental variables" excluded from the equation for Y^*.[1]

Under Assumption 1, Chernozhukov et al. (2015) introduce the estimator for the parameter $\beta_0(u)$ as

$$\widehat{\beta}(u) = \arg \min_{\beta \in \mathbb{R}^{\dim(X)}} \frac{1}{n} \sum_{i=1}^{n} 1(\widehat{S}_i'\widehat{\gamma} > \varsigma)\rho_u(Y_i - \widehat{X}_i'\beta), \quad (4)$$

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 (4) 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) estimation which is computed using an iterative procedure where each step takes the form specified in equation (4) 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

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

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 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:

$$V = \vartheta_0(D, W, Z) \equiv F_D(D \mid W, Z) \equiv Q_D^{-1}(D \mid W, Z) \equiv \int_0^1 1\{Q_D(v \mid W, Z) \leq D\}dv, \tag{5}$$

where $F_D(D \mid W, Z)$ is the distribution of D conditional on (W, Z). For any estimator of $F_D(D \mid W, Z)$ or $Q_D(V \mid W, Z)$, denoted by $\widehat{F}_D(D \mid W, Z)$ or $\widehat{Q}_D(V \mid W, Z)$, based on any parametric or semi-parametric functional form, the resulting estimator for the control variable is

$$\widehat{V} = \widehat{\vartheta}(D, W, Z) \equiv \widehat{F}_D(D \mid W, Z) \text{ or } \widehat{V} = \widehat{\vartheta}(D, W, Z) \equiv \int_0^1 1\{\widehat{Q}_D(v \mid W, Z) \leq D\}dv.$$

Let $R = r(W, Z)$ with $r(W, Z)$ being a vector of transformations of (W, Z). 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

$$\widehat{V} = \tau + \int_\tau^{1-\tau} 1\{R'\widehat{\pi}(v) \leq D\}dv, \tag{6}$$

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 and τ is a small positive trimming constant that avoids estimation of tail quantiles. The integral in (6) 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

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

[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.

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

$$\widehat{V} = \Lambda(R'\widehat{\pi}(D)), \tag{7}$$

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 (7) 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 $j = 1, ..., n_t$,

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

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:

$$\widehat{V}_i = \Phi((D_i - R_i'\widehat{\pi})/\widehat{\sigma}), \tag{8}$$

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 (4) using `cqiv`, we assume that the control function $\widehat{V}$ enters the equation through $\Phi^{-1}(\widehat{V})$. To motivate this, consider a simple version of the model (2)–(3):

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

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^*:

$$Q_{Y^*}(U \mid D, W, V) = X'\beta_0(U) = \beta_{00} + \beta_{01}D + \beta_{02}W + \rho_0\Phi^{-1}(V) + (1-\rho_0^2)^{1/2}\Phi^{-1}(U), \tag{9}$$

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

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 (6), (7) or (8) 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 transformations, k_0 is a cut-off such that $0 < k_0 < u$, and $\widehat{\delta} = \arg\max_{\delta \in \mathbb{R}^{\dim(S)}} \sum_{i=1}^n \{1(Y_i > C_i)\log\Lambda(\widehat{S}_i'\delta) + 1(Y_i = C_i)\log[1-\Lambda(\widehat{S}_i'\delta)]\}$.

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

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 (4), 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 $\widehat{\gamma} = (\widehat{\beta}^0(u)', -1)'$.

2.2 Weighted Bootstrap Algorithm

Chernozhukov et al. (2015) recommend obtaining confidence intervals through a weighted bootstrap procedure, though analytical formulas can also be used. If the estimation runs quickly on the desired sample, it is straightforward to rerun the entire CQIV algorithm B times weighting all the steps by the bootstrap weights. To speed up the computation, a procedure is proposed that uses a one-step CQIV estimator in each bootstrap repetition.

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

1. Draw a set of weights $(e_{1b}, \ldots, e_{nb})$ i.i.d from the standard exponential distribution or another distribution that satisfies Assumption 6.

2. Reestimate the control variable in the weighted sample, $\widehat{V}_{ib}^e = \widehat{\vartheta}_b^e(D_i, W_i, Z_i)$, and construct $\widehat{X}_{ib}^e = x(D_i, W_i, \widehat{V}_{ib}^e)$.

[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 $q_0 > q_1$.

3. Estimate the weighted quantile regression: $\widehat{\beta}_b^e(u) = \arg\min_{\beta\in\mathbb{R}^{\dim(X)}} \sum_{i\in J_{1b}} e_{ib}\rho_u(Y_i - \beta'\widehat{X}_{ib}^e)$, where $J_{1b} = \{i : \widehat{\beta}(u)'\widehat{X}_{ib}^e > C_i + \varsigma_1\}$, 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 $\widehat{\vartheta}_b^e$ 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.

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(#) top uncensored exogenous firststage(`*string*`) exclude nquant(#)`
`nthresh(#) ldv1(`*string*`) ldv2(`*string*`) corner drop1(#) drop2(#) viewlog`
`confidence(`*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 quantile specification.

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

`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.

exclude excludes exogenous regressors other than instruments from the first 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.

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)`	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(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 B1 of Chernozhukov et al. (2015)

*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 Chernozhukov et al. (2015) for detailed description of the data. Using this dataset, we

are interested in learning how alcohol share of consumption (`alcohol`) is affected by (logarithm of) expenditure (`logexp`), controlling for the number of children (`nkids`). For the endogenous expenditure, we use disposable income, i.e., (logarithm of) wages (`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 as 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*)

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 more possible examples 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) exclude
. cqiv alcohol logexp2 nkids (logexp = logwages nkids), confidence(weightboot) bootreps(10)
. cqiv alcohol nkids (logexp = logwages nkids), corner
```

In the second line of the examples, the option `exclude` is used. 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 Blaise Melly for helpful comments.

5 References

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* .

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.

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.