The Stata Journal (2019)
19, Number 4, pp. 768–781 DOI: 10.1177/1536867X19893615

Censored quantile instrumental-variable estimation with Stata

Victor Chernozhukov
Massachusetts Institute of Technology
Cambridge, MA
vchern@mit.edu

Ivan Fernández-Val
Boston University
Boston, MA
ivanf@bu.edu

Sukjin Han
University of Texas at Austin
Austin, TX
sukjin.han@austin.utexas.edu

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

Abstract. Many applications involve a censored dependent variable, an endogenous independent variable, or both. Chernozhukov, Fernández-Val, and Kowalski (2015, *Journal of Econometrics* 186: 201–221) introduced a censored quantile instrumental-variable (CQIV) estimator for use in those applications. The estimator has been applied by Kowalski (2016, *Journal of Business & Economic Statistics* 34: 107–117), among others. In this article, we introduce a command, `cqiv`, that simplifies application of the CQIV estimator in Stata. We summarize the CQIV estimator and algorithm, describe the use of `cqiv`, and provide empirical examples.

Keywords: st0576, cqiv, quantile regression, censored data, endogeneity, instrumental variable, control function

1 Introduction

In 2015, Chernozhukov, Fernández-Val, and Kowalski introduced a censored quantile instrumental-variables (CQIV) estimator. In this article, we introduce a command, `cqiv`, that implements the CQIV estimator in Stata. Our goal is to facilitate the use of `cqiv` in many applications.

Many applications involve censoring and 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 0, 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 addresses censoring and 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 `cqiv` 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 of the `cqiv` command, following Chernozhukov, Fernández-Val, and Kowalski (2015). In section 3, we introduce the use of `cqiv`. We provide an empirical application with examples that involve estimating Engel curves, as in Chernozhukov, Fernández-Val, and Kowalski (2015).

2 CQIV estimation

We first describe a model of triangular system for 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;[1] $\mathbf{w}$ be a vector of covariates, possibly containing c; and $\mathbf{z}$ be a vector of (possibly discrete) instrumental variables excluded from the equation for y^*. We observe $\{y_i, d_i, \mathbf{w}_i, \mathbf{z}_i, c_i\}_{i=1}^n$, a sample of size n of independent and identically distributed observations from the random vector $(y, d, \mathbf{w}, \mathbf{z}, c)$, which obeys

$$\begin{aligned} y &= \max(y^*, c) && (1) \\ y^* &= Q_{y^*}(u \mid d, \mathbf{w}, v) = \mathbf{x}'\boldsymbol{\beta}_0(u) && (2) \\ d &= Q_d(v \mid \mathbf{w}, \mathbf{z}) && (3) \end{aligned}$$

where v is a latent unobserved variable that accounts for the possible endogeneity of d, $\mathbf{x} = \mathcal{x}(d, \mathbf{w}, v)$ with $\mathcal{x}(d, \mathbf{w}, v)$ being a vector of transformations of $(d, \mathbf{w}, v)$, $Q_{y^*}(u \mid d, \mathbf{w}, v)$ is the u-quantile of y^* conditional on $(d, \mathbf{w}, v)$, $\boldsymbol{\beta}_0(u)$ is the vector of coefficients in the u-quantile function of y^* conditional on $(d, \mathbf{w}, v)$, $Q_d(v \mid \mathbf{w}, \mathbf{z})$ is the v-quantile of d conditional on $(\mathbf{w}, \mathbf{z})$, and

$$\begin{aligned} u &\sim U(0,1) \mid d, \mathbf{w}, \mathbf{z}, v, c \\ v &\sim U(0,1) \mid \mathbf{w}, \mathbf{z}, c \end{aligned}$$

This CQIV regression model nests the uncensored case of the quantile instrumental-variable (QIV) regression by making c arbitrarily small. As an example for the CQIV model, in the Engel curve application of Chernozhukov, Fernández-Val, and Kowalski (2015), y is the expenditure share in alcohol (bounded from below at $c = 0$), d is total expenditure on nondurables and services, $\mathbf{w}$ are household demographic characteristics, and $\mathbf{z}$ 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 "nonessential" goods, such as alcohol, tend to expend a higher proportion of their incomes, and therefore they tend 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 the 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 or labor and then decides

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

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 or labor, then the consumption budget shares do not depend on labor income given the consumption expenditure (see, for example, Deaton and Muellbauer [1980]). This justifies the use of labor income as an exclusion restriction.

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

$$y^* \quad = \quad \beta_{00} + \beta_{01} d + \boldsymbol{\beta}_{02} \mathbf{w} + \Phi^{-1}(\epsilon) \quad \epsilon \sim U(0,1) \tag{4}$$

where Φ^{-1} denotes the quantile function of the standard normal distribution. 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^*,

$$\begin{aligned} Q_{y^*}(u \mid d, \mathbf{w}, v) = \mathbf{x}'\boldsymbol{\beta}_0(u) = \beta_{00} + \beta_{01} d + \boldsymbol{\beta}_{02}\mathbf{w} + \rho_0 \Phi^{-1}(v) \\ + (1-\rho_0^2)^{1/2}\Phi^{-1}(u) \end{aligned} \tag{5}$$

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

Given this model, Chernozhukov, Fernández-Val, and Kowalski (2015) introduce the estimator for the parameter $\boldsymbol{\beta}_0(u)$ as

$$\widehat{\boldsymbol{\beta}}(u) = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{\dim(\mathbf{x})}} \frac{1}{n} \sum_{i=1}^{n} 1\left(\widehat{\mathbf{s}}_i' \widehat{\boldsymbol{\gamma}} > \varsigma\right) \rho_u \left(y_i - \widehat{\mathbf{x}}_i' \boldsymbol{\beta}\right) \tag{6}$$

where $\rho_u(z) = \{u - 1(z<0)\}z$ is the asymmetric absolute loss function of Koenker and Bassett (1978), $\widehat{\mathbf{x}}_i = \mathcal{x}(d_i, \mathbf{w}_i, \widehat{v}_i)$, $\widehat{\mathbf{s}}_i = \mathcal{s}(\widehat{\mathbf{x}}_i, c_i)$, $\mathcal{s}(\mathbf{x}, c)$ is a vector of transformations of $(\mathbf{x}, c)$, ς is a positive cutoff, 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 CQR estimator to a setting where there is possible endogeneity. As described in Chernozhukov, Fernández-Val, and Kowalski (2015), this algorithm is based on the following implication of the model:

$$P\left\{y \le \mathbf{x}'\boldsymbol{\beta}_0(u) \mid \mathbf{x}, c, \mathbf{x}'\boldsymbol{\beta}_0(u) > c\right\} = P\left\{y^* \le \mathbf{x}'\boldsymbol{\beta}_0(u) \mid \mathbf{x}, c, \mathbf{x}'\boldsymbol{\beta}_0(u) > c\right\} = u$$

provided that $P\{\mathbf{x}'\boldsymbol{\beta}_0(u) > c\} > 0$. In other words, $\mathbf{x}'\boldsymbol{\beta}_0(u)$ is the conditional u-quantile of the observed outcome for the observations for which $\mathbf{x}'\boldsymbol{\beta}_0(u) > c$; that is, 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, Fernández-Val, and Kowalski (2015) refer to them as the "quantile-uncensored" observations. The multiplier $1(\widehat{\mathbf{s}}_i'\widehat{\boldsymbol{\gamma}} > \varsigma)$ 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, Fernández-Val, and Kowalski (2015).

`cqiv` implements the CQIV estimator, which is computed using an iterative procedure where each step takes the form specified in equation (6) with a particular choice of $1(\widehat{\mathbf{s}}_i'\widehat{\boldsymbol{\gamma}} > \varsigma)$. 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. Because $\{\mathbf{x}'\boldsymbol{\beta}_0(u) > c\} \equiv \{P(y^* \leq c \mid \mathbf{x}, c) < u\}$, quantile-uncensored observations have a conditional probability of censoring that is lower than the quantile index u. The linear part of the conditional quantile function, $\mathbf{x}_i'\boldsymbol{\beta}_0(u)$, 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, $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}(u) > c_i$, and iterate.

`cqiv` provides different ways of estimating the control variable v, which can be chosen with the option `firststage(`*string*`)`. If $Q_d(v \mid \mathbf{w}, \mathbf{z})$ is invertible in v, the control variable has several equivalent representations:

$$v = \vartheta_0(d, \mathbf{w}, \mathbf{z}) \equiv F_d(d \mid \mathbf{w}, \mathbf{z}) \equiv Q_d^{-1}(d \mid \mathbf{w}, \mathbf{z}) \equiv \int_0^1 1\{Q_d(v \mid \mathbf{w}, \mathbf{z}) \leq d\} dv$$

$F_d(d \mid \mathbf{w}, \mathbf{z})$ is the distribution of d conditional on $(\mathbf{w}, \mathbf{z})$. Different estimators of v can be constructed based on parametric or semiparametric models for $F_d(d \mid \mathbf{w}, \mathbf{z})$ and $Q_d(v \mid \mathbf{w}, \mathbf{z})$. Let $\mathbf{r} = r(\mathbf{w}, \mathbf{z})$, with $r(\mathbf{w}, \mathbf{z})$ being a vector of collecting transformations of $(\mathbf{w}, \mathbf{z})$ specified by the researcher. When *string* is `quantile`, a quantile regression model is assumed, where $Q_d(v \mid \mathbf{w}, \mathbf{z}) = \mathbf{r}'\boldsymbol{\pi}_0(v)$, $\boldsymbol{\pi}_0(v)$ is the vector of coefficients in the v-quantile function of d conditional on $(\mathbf{w}, \mathbf{z})$, and

$$v = \int_0^1 1\{\mathbf{r}'\boldsymbol{\pi}_0(v) \leq d\} dv$$

The estimator of v then takes the form

$$\widehat{v} = \tau + \int_\tau^{1-\tau} 1\{\mathbf{r}'\widehat{\boldsymbol{\pi}}(v) \leq d\} dv \tag{7}$$

where $\widehat{\boldsymbol{\pi}}(v)$ 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 (7) can be approximated numerically using a finite grid of quantiles.[2] Specifically, the fitted values for prespecified quantile indices (whose number n_q is controlled by the option `nquant(`#`)`) are calculated, which then yields

$$\widehat{v}_i = \frac{1}{n_q} \sum_{j=1}^{n_q} 1\{\mathbf{r}_i'\widehat{\boldsymbol{\pi}}(v_j) \leq d_i\}$$

[2] The use of the integral to obtain a generalized inverse is convenient to avoid monotonicity problems in $v \mapsto \mathbf{r}'\widehat{\boldsymbol{\pi}}(v)$ that are due to misspecification or sampling error. Chernozhukov, Fernández-Val, and Galichon (2010) developed asymptotic theory for this estimator.

For other related quantile regression models that can alternatively be used, see Chernozhukov, Fernández-Val, and Kowalski (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,

$$v = F_d(d \mid \mathbf{w}, \mathbf{z}) = \Lambda \left\{ \mathbf{r}' \boldsymbol{\pi}_0(d) \right\}$$

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

$$\widehat{v} = \Lambda \left\{ \mathbf{r}' \widehat{\boldsymbol{\pi}}(d) \right\} \tag{8}$$

where $\widehat{\boldsymbol{\pi}}(d)$ is the maximum likelihood estimator of $\boldsymbol{\pi}_0(d)$ at each d (see, for example, Foresi and Peracchi [1995], and Chernozhukov, Fernández-Val, and Melly [2013]).[3] The expression (8) 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 the option `nthresh(`#`)`. Concretely, for threshold d_j with $j = 1, \ldots, n_t$,

$$\widehat{v}_i = \Lambda \left\{ \mathbf{r}_i' \widehat{\boldsymbol{\pi}}(d_j) \right\} \qquad \text{for } i\text{'s s.t. } d_{j-1} \leq d_i < d_j \text{ with } d_0 = -\infty \text{ and } d_{n_t} = \infty$$

where $\widehat{\boldsymbol{\pi}}(d_j)$ is a probit or logit estimate with $\widetilde{d}_i(d_j) = 1\{d_i \leq d_j\}$ as a dependent variable and $\mathbf{r}_i$ as regressors.

Lastly, when *string* is `ols`, a linear regression model $d = \mathbf{r}'\boldsymbol{\pi}_0 + v$ is assumed, and $\widehat{v}$ is a transformation of the ordinary least-squares (OLS) residual:

$$\widehat{v}_i = \Phi \left\{ \left(d_i - \mathbf{r}_i' \widehat{\boldsymbol{\pi}} \right) / \widehat{\sigma} \right\} \tag{9}$$

where Φ is the standard normal distribution, $\widehat{\boldsymbol{\pi}}$ is the OLS estimator of $\boldsymbol{\pi}_0$, 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).

2.1 CQIV algorithm

The algorithm recommended in Chernozhukov, Fernández-Val, and Kowalski (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) three-step CQR algorithm.

[3] Chernozhukov, Fernández-Val, and Melly (2013) developed asymptotic theory for this estimator.

For each desired quantile u, perform the following steps:

0. Obtain $\widehat{v}_i = \widehat{\vartheta}(d_i, \mathbf{w}_i, \mathbf{z}_i)$ from (7), (8), or (9), and construct $\widehat{\mathbf{x}}_i = x(d_i, \mathbf{w}_i, \widehat{v}_i)$.

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

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

3. Obtain the three-step CQIV coefficient estimates $\widehat{\boldsymbol{\beta}}^1(u)$, 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, one can use a probit, logit, or any other model that fits the data well. `cqiv` provides the option `ldv2(`*string*`)`, where *string* can be `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 $\mathbf{x}_i'\boldsymbol{\beta}_0(u) > c_i$. To choose the value of k_0, it is advisable that a constant fraction of observations satisfying $\Lambda(\widehat{\mathbf{s}}_i'\widehat{\delta}) > 1 - u$ is excluded from J_0 for each quantile. To do so, one needs to set k_0 as the q_0th quantile of $\Lambda(\widehat{\mathbf{s}}_i'\widehat{\delta})$ conditional on $\Lambda(\widehat{\mathbf{s}}_i'\widehat{\delta}) > 1 - u$, 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 the option `drop1(`*#*`)`.

Remark 2 (Step 2). To choose the cutoff ς_1, it is advisable that a constant fraction of observations satisfying $\widehat{\mathbf{x}}_i'\widehat{\boldsymbol{\beta}}^0(u) > c_i$ is excluded from J_1 for each quantile. To do so, one needs to set ς_1 to be the q_1th quantile of $\widehat{\mathbf{x}}_i'\widehat{\boldsymbol{\beta}}^0(u) - c_i$ conditional on $\widehat{\mathbf{x}}_i'\widehat{\boldsymbol{\beta}}^0(u) > c_i$, 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 the option `drop2(`*#*`)`.[5]

[4] As an optional fourth step, one can update the set of quantile-uncensored observations J_2 by replacing $\widehat{\boldsymbol{\beta}}^0(u)$ with $\widehat{\boldsymbol{\beta}}^1(u)$ 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 the `cqiv` command, because Chernozhukov, Fernández-Val, and Kowalski (2015) find little gain of iterating in terms of bias, root mean squared 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, Fernández-Val, and Kowalski (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 $\widehat{\boldsymbol{\beta}}^0(u)$ 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 $q_0 > q_1$.

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

2.2 Weighted bootstrap algorithm

Chernozhukov, Fernández-Val, and Kowalski (2015) recommend obtaining standard errors and confidence intervals through either weighted bootstrap or nonparametric bootstrap procedures. We focus on the weighted bootstrap here. To speed up the computation, we propose a procedure 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 $(e_{1b}, \dots, e_{nb})$ independent and identically distributed from the standard exponential distribution.
2. Reestimate the control variable in the weighted sample, $\widehat{v}_{ib}^e = \widehat{\vartheta}_b^e(d_i, \mathbf{w}_i, \mathbf{z}_i)$, and construct $\widehat{\mathbf{x}}_{ib}^e = x(d_i, \mathbf{w}_i, \widehat{v}_{ib}^e)$.
3. Estimate the weighted quantile regression

$$\widehat{\boldsymbol{\beta}}_b^e(u) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{\dim(\mathbf{x})}} \sum_{i \in J_{1b}} e_{ib}\rho_u\left(y_i - \boldsymbol{\beta}'\widehat{\mathbf{x}}_{ib}^e\right)$$

where $J_{1b} = \{i : \widehat{\boldsymbol{\beta}}(u)'\widehat{\mathbf{x}}_{ib}^e > c_i + \varsigma_1\}$ and $\widehat{\boldsymbol{\beta}}(u)$ is a consistent estimator of $\boldsymbol{\beta}_0(u)$, for example, the three-stage CQIV estimator $\widehat{\boldsymbol{\beta}}^1(u)$.

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 the 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, Fernández-Val, and Kowalski (2015), we focus on weighted bootstrap, partly because it has practical advantages over nonparametric bootstrap to deal with discrete regressors with small cell sizes, because 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. It can implement both censored and uncensored QIV estimation under either exogeneity or endogeneity. The estimators proposed by Chernozhukov, Fernández-Val, and Kowalski (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.

All the variables in the parentheses of the syntax are those involved in the first-stage estimation of CQIV and QIV.

3.3 Options

Model

quantiles(*numlist*) specifies the quantiles at which the model is fit 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. The default is censorpt(0). An inappropriately specified censoring point will generate errors in estimation.

censorvar(*varname*) specifies the censoring variable (that is, the random censoring point) of the dependent variable.

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

`uncensored` selects uncensored QIV estimation.

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

`firststage(`*string*`)` determines the first-stage estimation procedure, where *string* may be specified as `quantile` for quantile regression (the default), `distribution` for distribution regression (either probit or logit), or `ols` for OLS estimation. Be aware that `firststage(distribution)` can take a long time to execute.

`firstvar(`*varlist*`)` specifies the list of variables other than instruments that are included in the first-stage estimation. The 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 `firststage(quantile)`. The default is `nquant(50)`; that is, 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 specified as `firststage(distribution)`. The default is `nthresh(50)`; that is, 50 evenly spaced thresholds (that is, 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 limited dependent variable model used in the first-stage estimation when the estimation procedure is `firststage(distribution)`, where *string* is either `probit` for probit estimation (the default) or `logit` for logit estimation.

`ldv2(`*string*`)` determines the limited dependent variable 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 the censored dependent variable when the censoring is due to economic reasons, such as 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; that is, the average of

$$1\{Q_{y^*}(u \mid d, \mathbf{w}, v) > c\}\partial_d Q_{y^*}(u \mid d, \mathbf{w}, v) = 1\{\mathcal{x}(d, \mathbf{w}, v)'\boldsymbol{\beta}_0(u) > c\}\partial_d\, \mathcal{x}(d, \mathbf{w}, v)'\boldsymbol{\beta}_0(u)$$

over all observations. If the underlying function is nonlinear in the endogenous variable, average marginal effects must be calculated directly from the coefficients without the `corner` option. For details of the related concepts, see section 2.1 of Chernozhukov, Fernández-Val, and Kowalski (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). The default is `drop1(10)`.

`drop2(#)` sets the proportion of observations q_1 with estimates 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). The default is `drop2(3)`.

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

Inference

`confidence(`*string*`)` specifies the type of confidence intervals. If *string* is specified as `no`, which is the default, then no confidence intervals are calculated. If *string* is specified as `boot` or `weightboot`, then either nonparametric bootstrap or weighted bootstrap (respectively) t-percentile symmetric confidence intervals are calculated. The weights of the weighted bootstrap are generated from the standard exponential distribution. Be aware that `confidence(boot)` and `confidence(weightboot)` can take a long 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 `confidence(boot)` or `confidence(weightboot)` is also specified. The default is `bootreps(100)`.

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

`level(#)` sets the confidence level. The default is `level(95)`.

Robust check

`norobust` suppresses the robustness diagnostic test results. There are no diagnostic test results to suppress when `uncensored` is used.

3.4 Stored results

`cqiv` stores 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)`	`cqiv`
`e(depvar)`	name of dependent variable
`e(regression)`	name of the implemented regression: either `cqiv`, `qiv`, or `cqr`
`e(endogvar)`	name of endogenous regressor
`e(instrument)`	names of instrumental variables
`e(censorvar)`	name of censoring variable
`e(regressors)`	names of the regressors
`e(firststage)`	type of first-stage estimation
`e(confidence)`	type of confidence intervals
Matrices	
`e(results)`	matrix containing the estimated coefficients, means, standard errors, and lower and upper bounds of confidence intervals
`e(quantiles)`	row vector containing the quantiles at which CQIV has been estimated
`e(robustcheck)`	matrix containing the results for the robustness diagnostic test results; see table 1 below

In the following table, we present the CQIV robustness diagnostic tests suggested in Chernozhukov, Fernández-Val, and Kowalski (2015) for the CQIV estimator with an OLS estimate of the control variable. See section 2.1 of that article for the definitions of k_0, ς_1, J_0, and 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.

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	1.7	1.45	2.01	50.7	46.7	54.9	52.3	48.2	56.7
0.1	1.71	1.44	1.96	52.8	49.5	55.5	54.5	51.1	57.3
0.25	1.71	1.46	1.98	56.3	53.6	58.7	58.1	55.3	60.6
0.5	1.72	1.44	2.02	60.1	57.6	63.4	62	59.4	65.4
0.75	1.73	1.47	1.99	64	61.2	66.8	66	63.1	68.9
0.9	1.75	1.44	2.01	67.4	64.6	70.6	69.5	66.6	72.8
0.95	1.76	1.49	2.02	69.3	65.6	72.8	71.5	67.7	75.1
				Percent J0 in J1			Count in J1 not in J0		
Quantile	Median	Min	Max	Median	Min	Max	Median	Min	Max
0.05	1.6	1.33	1.85	100	97.7	100	36	0	81
0.1	1.6	1.33	1.85	100	99	100	37	7	74
0.25	1.6	1.33	1.85	100	99.6	100	40	15	68
0.5	1.6	1.33	1.85	100	99.6	100	43	23	78
0.75	1.6	1.33	1.85	100	99.7	100	47	17	74
0.9	1.6	1.33	1.85	100	99.7	100	50	15	88
0.95	1.6	1.33	1.85	100	99.1	100	51	16	97

Comparison of Objective Functions

	Objective Step 3			0bjective Step 2			Objective Step 3<0bjective 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 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, Fernández-Val, and Kowalski (2015) for more discussion.

3.5 Examples

We illustrate how to use `cqiv` with some examples. For the dataset, we use a household expenditure dataset for alcohol consumption drawn from the British Family Expenditure Survey; see Blundell, Chen, and Kristensen (2007) and Chernozhukov, Fernández-Val, and Kowalski (2015) for a detailed description of the data. We are interested in learning how the share of total expenditure on alcohol (`alcohol`) is affected by (the logarithm of) total expenditure (`logexp`), controlling for the number of children (`nkids`). For the endogenous expenditure, we use disposable income, that is, (the logarithm of) gross earnings of the head of the household (`logwages`), as an excluded instrument.

```
. use alcoholengel
```

Given this dataset, we can generate part of the empirical results of Chernozhukov, Fernández-Val, and Kowalski (2015):

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

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

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

And the CQR estimation can be implemented with the `exogenous` option:

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

Here are other 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 (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 last example, `logexp2` cannot be included in the first-stage regression when distribution regression is implemented, because `logexp2` is a monotone transformation of `logexp`. Thus, the distribution estimation yields a perfect fit.

4 Acknowledgments

We thank Simon Jäger, Blaise Melly, and Sanna Nivakoski for helpful comments.

5 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 19-4
. net install st0576      (to install program files, if available)
. net get st0576          (to install ancillary files, if available)
```

6 References

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

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

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

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

Chernozhukov, V., and H. Hong. 2002. Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association* 97: 872–882.

Deaton, A. S., and J. Muellbauer. 1980. *Economics and Consumer Behavior*. Cambridge: 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: 451–466.

Gorman, W. M. 1959. Separable utility and aggregation. *Econometrica* 27: 469–481.

Koenker, R., and G. Bassett, Jr. 1978. Regression quantiles. *Econometrica* 46: 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: 107–117.

Powell, J. L. 1986. Censored regression quantiles. *Journal of Econometrics* 32: 143–155.

About the authors

Victor Chernozhukov is Ford International Professor at the Department of Economics and the Center for Statistics and Data Science at MIT.

Ivan Fernández-Val is a Professor in Economics at Boston University.

Sukjin Han is an Assistant Professor in Economics at University of Texas at Austin.

Amanda Kowalski is Gail Wilensky Professor of Applied Economics and Public Policy at University of Michigan.