The Stata Journal (2020)
20, Number 2, pp. 297–308 DOI: 10.1177/1536867X20930998

Estimating selection models without an instrument with Stata

Xavier D'Haultfœuille
CREST-ENSAE
Paris, France
xavier.dhaultfoeuille@ensae.fr

Arnaud Maurel
Duke University, NBER, and IZA
Durham, NC
apm16@duke.edu

Xiaoyun Qiu
Northwestern University
Evanston, IL
xiaoyun.qiu@u.northwestern.edu

Yichong Zhang
Singapore Management University
Singapore
yczhang@smu.edu.sg

Abstract. In this article, we present the `eqregsel` command, which estimates and provides bootstrap inference for sample-selection models via extremal quantile regression. `eqregsel` estimates a semiparametric sample-selection model without an instrument or a large support regressor and outputs the point estimates of the homogeneous linear coefficients, their bootstrap standard errors, and the *p*-value for a specification test.

Keywords: st0598, eqregsel, sample-selection models, extremal quantile regressions

1 Introduction

In this article, we present the command `eqregsel`, which estimates and provides bootstrap inference of endogenous sample-selection models and implements the procedures developed in recent work by D'Haultfœuille, Maurel, and Zhang (2018). Prior methods to estimate endogenous sample-selection models proposed in the econometric literature rely on instruments, large support regressors, or both. For the former, see, among others, Heckman (1974, 1979, 1990); Ahn and Powell (1993); Donald (1995); Buchinsky (1998); Chen and Khan (2003); Das, Newey, and Vella (2003); Newey (2009); and Vella (1998) for a survey. Chamberlain (1986) and Lewbel (2007) developed identification strategies for sample-selection models in the absence of an instrument for selection. These alternative methods rely on the existence of a large support regressor. However, in practice, valid instruments and large support regressors are often difficult, if not impossible, to find.

Unlike prior methods, the method implemented in `eqregsel` does not require the presence of instruments or large support regressors.[1] Identification relies instead on the strategy initially proposed by D'Haultfœuille and Maurel (2013), which is based on

[1] See Honoré and Hu (2018) for a related recent work also motivated by the difficulty of finding instruments for sample selection. As is the case here, they do not require exclusion restrictions or large support regressors. However, their approach is based on a different set of assumptions and, in contrast with our framework, delivers set rather than point identification.

the idea that, provided that selection is endogenous, one can expect the effect of the outcome on selection to dominate that of the covariates for large values of the outcome. eqregsel builds on the estimation method proposed by D'Haultfœuille, Maurel, and Zhang (2018) and implements a series of quantile regressions in the tails of the outcome distribution (extremal quantile regressions).[2] The command outputs estimates for a set of user-specified coefficients of interest, their standard errors (estimated via bootstrap), and a p-value for the specification test described in D'Haultfœuille, Maurel, and Zhang (2018).

eqregsel complements the existing Stata command heckman for the estimation of sample-selection models. In terms of underlying assumptions, eqregsel has at least three distinctive features compared with heckman. First, it does not require normality of the error term in the selection equation or linearity of the conditional expectation of the error term in the outcome equation. Second, it does not restrict the selection process apart from an independence-at-infinity condition. Third, it allows for heterogeneous distributional effects of other control variables.

The remainder of the article is organized as follows. In section 2, we recall the setup of the semiparametric endogenous sample-selection model considered in D'Haultfœuille, Maurel, and Zhang (2018) and describe the data-driven procedure used to choose the quantile index for the extremal quantile regression. In section 3, we describe how to implement the method in practice. In section 4, we present the eqregsel command. In section 5, we illustrate the use of eqregsel by estimating the black–white wage gap on U.S. young males of the 1979 and 1997 National Longitudinal Surveys of Youth (NLSY79 and NLSY97). In section 6, we conclude.

2 The framework and estimation method

2.1 Model and estimation

We consider the outcome equation

$$Y^* = \mathbf{X}_1'\boldsymbol{\beta}_1 + \varepsilon$$

where $Y^* \in \mathbb{R}$ and $\mathbf{X}_1 \in \mathbb{R}^{d_1}$ are the outcome and covariates of interest, respectively. In the following, we seek to identify and estimate $\boldsymbol{\beta}_1$. For that purpose, we rely on two key conditions. The first is that for any $\tau \in (0,1)$, the τth conditional quantile of ε satisfies

$$Q_{\varepsilon|X}(\tau|\mathbf{X}) = \beta_0(\tau) + \mathbf{X}_2'\boldsymbol{\beta}_2(\tau) \tag{1}$$

where $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$ and $\mathbf{X}_2$ denotes other covariates. Then

$$Q_{Y^*|\mathbf{X}}(\tau|\mathbf{X}) = \mathbf{X}_1'\boldsymbol{\beta}_1 + \beta_0(\tau) + \mathbf{X}_2'\boldsymbol{\beta}_2(\tau) \tag{2}$$

[2] See Chernozhukov, Fernández-Val, and Kaji (2018) for an overview of extremal quantile regression methods and recent applications.

The effect of $\mathbf{X}_1$ is thus assumed to be homogeneous across different quantile indices, while the effect of the other covariates $\mathbf{X}_2$ is allowed to be heterogeneous across the distribution of Y^*.

Y^* is not directly observed. Instead, and denoting by D the selection dummy, the econometrician observes only D, $Y = DY^*$, and $\mathbf{X}$. The second key condition is that, conditional on having "large" outcomes, selection is independent of the covariates. More precisely, we assume that there exists a constant $h \in (0,1]$ such that for all $x \in \text{Supp}(\mathbf{X})$,

$$\lim_{y\to\infty} P(D=1|\mathbf{X}=\mathbf{x}, Y^*=y) = h \tag{3}$$

In some cases, it may be more plausible to impose that, conditional on having "small" outcomes ($Y^* \to -\infty$), selection is independent of the covariates. This case can be handled simply by replacing Y with $-Y$ and $\mathbf{X}$ with $-\mathbf{X}$ hereafter.

Combining (2) and (3), D'Haultfœuille, Maurel, and Zhang (2018, theorem 2.1) show that, under some regularity conditions on the upper tail of ε, as $\tau \to 0$,

$$\begin{aligned} Q_{-Y|\mathbf{X}}(\tau|\mathbf{X}) &= Q_{-Y^*|\mathbf{X}}(\tau/h|\mathbf{X}) + o(1) \\ &= -\mathbf{X}_1'\boldsymbol{\beta}_1 - \beta_0(1-\tau/h) - \mathbf{X}_2'\boldsymbol{\beta}_2(1-\tau/h) + o(1) \end{aligned} \tag{4}$$

Therefore, (4) suggests that we can estimate $\boldsymbol{\beta}_1$ by running a quantile regression of $-Y$ on $-\mathbf{X}$ with a sufficiently small quantile index τ; that is,

$$\left\{\widehat{\boldsymbol{\beta}}_1', \widehat{\beta}_0(1-\tau/h), \widehat{\boldsymbol{\beta}}_2'(1-\tau/h)\right\}' = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho_\tau\left(-Y_i + \overline{\mathbf{X}}_i'\boldsymbol{\beta}\right)$$

where $\rho_\tau(u) = (\tau - \mathbb{1}\{u<0\})u$ is the check function used in quantile regressions, n denotes the sample size, and $\overline{\mathbf{X}}_i = (\mathbf{X}_{1i}', 1, \mathbf{X}_{2i}')'$. Intuitively, for $\widehat{\boldsymbol{\beta}}_1$ to be consistent, τ should depend on n and tend to 0 as n tends to infinity. However, it should not tend too quickly to 0; otherwise, the extremal quantile regression would be unstable. Formally, and letting τ_n denote the quantile index, D'Haultfœuille, Maurel, and Zhang (2018) establish that if $\tau_n \to 0$ and $n\tau_n \to \infty$,[3], and under additional technical restrictions, $\widehat{\boldsymbol{\beta}}_1$ is consistent and asymptotically normal.

As is standard with extremal quantile regressions (see Chernozhukov, Fernández-Val, and Kaji [2018]), the rate of convergence is not the usual parametric root-n rate. Moreover, in this case, this rate depends on unknown features of the distribution of $(D, Y^*, \mathbf{X})$.[4] Importantly, D'Haultfœuille, Maurel, and Zhang (2018) show that the bootstrap is consistent for inference and does not require the knowledge of the rate of convergence. To illustrate this, let q_γ^* denote the quantile of order γ of the bootstrap estimator $\widehat{\boldsymbol{\beta}}_1^*$, assuming for simplicity that X_1 is a scalar ($d_1 = 1$). Then theorem 2 in

[3] This corresponds to the so-called intermediate order case in extreme value theory, in contrast with the extreme order case, where one would have $n\tau_n \to k$ for some $k > 0$.

[4] We refer to the definition of the rate above theorem 2.2 in D'Haultfœuille, Maurel, and Zhang (2018).

D'Haultfœuille, Maurel, and Zhang (2018) implies that the percentile bootstrap confidence interval (CI) $[q^*_{\alpha/2}, q^*_{1-\alpha/2}]$ of β_1 has an asymptotic coverage rate of $1-\alpha$. Such an interval does not require the knowledge of the rate of convergence.

The results above rely on two main conditions, namely, (2) and (3). Importantly, we can develop a specification test of these conditions based on the implication that the coefficients $\boldsymbol{\beta}_1$ are the same across different extremal quantile indices τ_n [see (4)]. Then, if the model is correctly specified, the two estimators $\widehat{\boldsymbol{\beta}}_1(\ell\tau_n)$ (with $0 < \ell < 1$) and $\widehat{\boldsymbol{\beta}}_1(\tau_n)$ of $\boldsymbol{\beta}_1$, obtained respectively with $\tau = \ell\tau_n$ and $\tau = \tau_n$, should be close. Following this idea, consider the J-test statistic

$$T_J(\ell) = \{(1/\ell) - 1\}^2 \left\{\widehat{\boldsymbol{\beta}}_1(\tau_n) - \widehat{\boldsymbol{\beta}}_1(\ell\tau_n)\right\}' \widehat{\boldsymbol{\Omega}}^{-1} \left\{\widehat{\boldsymbol{\beta}}_1(\tau_n) - \widehat{\boldsymbol{\beta}}_1(\ell\tau_n)\right\} \tag{5}$$

where $\widehat{\boldsymbol{\Omega}}$ is a (bootstrap) estimator of the asymptotic covariance of $\widehat{\boldsymbol{\beta}}_1(\tau_n)$, properly normalized by the rate of convergence in view of the discussion above. Then we reject the test at the nominal level α whenever $T_J(\ell) > q_{d_1}(1-\alpha)$, where $q_{d_1}(1-\alpha)$ is the $(1-\alpha)$th quantile of a χ^2 distribution with d_1 degrees of freedom. Theorem 2.3 in D'Haultfœuille, Maurel, and Zhang (2018) establishes that for any $0 < \ell < 1$, the test has an asymptotic level of α. It also proves that under some local alternatives, the local power is maximized at $\ell^* = \arg\max_{\ell \in [0,1]} \ell\{\ln(l)\}^2/(1-\ell) \simeq 0.2$.

2.2 Choice of the quantile index

The performance of extremal quantile estimators depends on a tradeoff between bias and variance, which is governed by the quantile index τ_n used in the extremal quantile regression. In the following, we present the algorithm outlined in D'Haultfœuille, Maurel, and Zhang (2018), which selects a suitable quantile index based on estimators of the bias and the variance of $\widehat{\boldsymbol{\beta}}_1$.

Specifically, consider the same test statistic as in (5), but where $(\ell\tau_n, \tau_n)$ are replaced by $(\ell_1\tau_n, \ell_2\tau_n)$, with $\ell_1 < 1 < \ell_2$:

$$T_J(\tau) = (1/\ell_1 - 1/\ell_2)^2 \left\{\widehat{\boldsymbol{\beta}}_1(\ell_2\tau) - \widehat{\boldsymbol{\beta}}_1(\ell_1\tau)\right\}' \widehat{\boldsymbol{\Omega}}^{-1} \left\{\widehat{\boldsymbol{\beta}}_1(\ell_2\tau) - \widehat{\boldsymbol{\beta}}_1(\ell_1\tau)\right\}$$

D'Haultfœuille, Maurel, and Zhang (2018) show that the difference between the median of $T_J(\tau)$ and the median of a chi-squared distribution with d_1 degrees of freedom can serve as a proxy for the bias of the estimator.

The idea, then, is to estimate this difference using subsampling.[5] For each subsample and each quantile index τ within a grid $\mathcal{G}$, one can compute $T_J(\tau)$. Let $M_{\text{sub}}(\tau)$ denote the median of these test statistics over different subsamples for a given τ, and let M_{d_1} denote the median of the chi-squared distribution with d_1 degrees of freedom. Then, the proxy of the bias is defined as

[5] We recall that subsampling corresponds to a bootstrap without replacement of size $b_n < n$. Though often less accurate than the standard bootstrap, subsampling has the advantage of being consistent under much weaker conditions. See Politis, Romano, and Wolf (1999) for an introduction.

$$\widehat{\mathrm{diff}}_n(\tau) = \frac{|M_{\mathrm{sub}}(\tau) - M_{d_1}|}{\sqrt{b_n \tau}}$$

where b_n denotes the subsample size.

Similarly, the asymptotic covariance matrix is estimated by the covariance matrix of the subsampling estimator of $\boldsymbol{\beta}_1$, multiplied by the normalizing factor b_n/n. Denote by $\widehat{\mathrm{Var}}_n(\tau)$ the sum of the diagonal elements of this covariance matrix. The quantile index is selected to optimize the bias-variance tradeoff,

$$\widehat{\tau}_n = \arg\min_{\tau \in \mathcal{G}} \widehat{\mathrm{Var}}_n(\tau) + \widehat{\mathrm{diff}}_n(\tau)$$

where $\mathcal{G}$ denotes a finite grid within $(0,1)$. This procedure results in undersmoothing compared with a more standard tradeoff between variance and squared bias. As with the case of nonparametric regressions, this is needed to control the asymptotic bias that would otherwise affect the limiting distribution of the estimator. We refer to D'Haultfœuille, Maurel, and Zhang (2018) for simulation-based evidence that this choice leads to estimators that are both accurate and only very mildly biased, thus leading to reliable inference on $\boldsymbol{\beta}_1$.

3 Implementation

We summarize how we implement the method described above in `eqregsel`.

1. Draw B bootstrap samples and B subsamples of size b_n.
2. For each $\tau \in \mathcal{G}$:

    a. Compute the estimator of $\boldsymbol{\beta}(\tau) = \{\boldsymbol{\beta}_1', \beta_0(1-\tau/h), \boldsymbol{\beta}_2'(1-\tau/h)\}'$:

    $$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau \left(-Y_i + \overline{\mathbf{X}}_i' \boldsymbol{\beta}\right)$$

    Let $\widehat{\boldsymbol{\beta}}_1(\tau)$ denote the vector comprising the first d_1 components of $\widehat{\boldsymbol{\beta}}(\tau)$.

    b. Compute

    $$\widehat{\boldsymbol{\Omega}}(\tau) = \frac{1}{B} \sum_{b=1}^{B} \left\{\widehat{\boldsymbol{\beta}}_1^b(\tau) - \widehat{\boldsymbol{\beta}}_1(\tau)\right\} \left\{\widehat{\boldsymbol{\beta}}_1^b(\tau) - \widehat{\boldsymbol{\beta}}_1(\tau)\right\}'$$

    with $\widehat{\boldsymbol{\beta}}_1^b(\tau)$ being the bootstrap estimator of $\boldsymbol{\beta}_1$ on the bth bootstrap sample.

c. Compute, for each subsample $s = 1 \dots S$, the estimator of $\boldsymbol{\beta}_1$ $[\widehat{\boldsymbol{\beta}}_1^s(\tau)]$, and the J-test statistic:[6]

$$T_J^s(\tau) = (b_n/n)(1/\ell_1 - 1/\ell_2)^2 \left\{\widehat{\boldsymbol{\beta}}_1^s(\ell_2\tau) - \widehat{\boldsymbol{\beta}}_1^s(\ell_1\tau)\right\}' \\ \widehat{\boldsymbol{\Omega}}(\tau)^{-1} \left\{\widehat{\boldsymbol{\beta}}_1^s(\ell_2\tau) - \widehat{\boldsymbol{\beta}}_1^s(\ell_1\tau)\right\}$$

d. Compute $\widehat{\mathrm{diff}}_n(\tau) = \{|M_{\mathrm{sub}}(\tau) - M_{d_1}|\}/(\sqrt{b_n\tau})$, where $M_{\mathrm{sub}}(\tau)$ denotes the median of $\{T_J^1(\tau), \dots, T_J^B(\tau)\}$.

e. Compute $\widehat{\mathrm{Var}}_n(\tau) = (b_n/n)\sum_{k=1}^{d_1} \widehat{\Sigma}(\tau)_{kk}$, where $\widehat{\Sigma}(\tau)_{kk}$ is the kth diagonal term of

$$\widehat{\boldsymbol{\Sigma}}(\tau) = \frac{1}{S}\sum_{s=1}^{S} \left\{\widehat{\boldsymbol{\beta}}_1^s(\tau) - \overline{\boldsymbol{\beta}}_1(\tau)\right\} \left\{\widehat{\boldsymbol{\beta}}_1^s(\tau) - \overline{\boldsymbol{\beta}}_1(\tau)\right\}'$$

with

$$\overline{\boldsymbol{\beta}}_1(\tau) = \frac{1}{S}\sum_{s=1}^{S} \widehat{\boldsymbol{\beta}}_1^s(\tau)$$

3. Compute $\widehat{\tau}_n = \arg\min_{\tau\in\mathcal{G}} \widehat{\mathrm{Var}}_n(\tau) + \widehat{\mathrm{diff}}_n(\tau)$.

4. Define $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_1(\widehat{\tau}_n)$ and $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}(\widehat{\tau}_n)$. $\mathrm{CI}_{1-\alpha}(\beta_{1k})$ of level $1-\alpha$ on the kth component of $\boldsymbol{\beta}_1$ are then equal to

$$\mathrm{CI}_{1-\alpha}(\beta_{1k}) = \left[\widehat{\beta}_{1k} - z_{1-\alpha/2}\sqrt{\widehat{\Omega}_{kk}}, \widehat{\beta}_{1k} + z_{1-\alpha/2}\sqrt{\widehat{\Omega}_{kk}}\right]$$

where $\widehat{\Omega}_{kk}$ is the kth diagonal term of $\widehat{\boldsymbol{\Omega}}$ and $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of a standard normal variable.

5. Compute $\widehat{\boldsymbol{\beta}}_1(0.2\widehat{\tau}_n)$ and then $T_J(0.2)$, as defined in (5), to perform the specification test of the model.

In practice, we consider an equally spaced grid $\mathcal{G}$ with lower bound min(0.1, $80/b_n$), upper bound 0.3, and a number of points equal to $n_{\mathcal{G}}$. The lower bound is motivated by the fact that if the effective subsampling size τb_n becomes too small, then the intermediate order asymptotic theory is likely to be a poor approximation (see Chernozhukov and Fernández-Val [2011] for a related discussion). To compute $T_J^s(\tau)$ in step 2c above, we use $(\ell_1, \ell_2) = (0.9, 1.1)$.

4 The eqregsel command

We describe below the syntax, options, and stored results associated with the `eqregsel` command. Note that it relies on the `moremata` package (Jann 2005). If the latter is not already installed, one must type `ssc install moremata` in the Stata Command window. The `eqregsel` command is compatible with Stata 14 and later versions.

[6] The term b_n/n accounts for the fact that the rate of convergence of the J statistic on the subsample is b_n/n times the rate of convergence on the whole sample.

4.1 Syntax

The syntax of `eqregsel` is as follows:

`eqregsel` *Y X*1 *X*2 [*if*] [*in*] [, `hom(`#`)` `subs(`#`)` `grid(`#`)` `rep(`#`)` `small`]

4.2 Description

`eqregsel` computes $\widehat{\boldsymbol{\beta}}_1$ in (2) based on the data-driven τ_n detailed in section 2.2 above. It also reports its standard errors and 95% CIs. Finally, it computes the p-value of this specification test using $\ell = \ell^*$.

*X*1 is the list of variables entering in $\mathbf{X}_1$ in (2).

*X*2 is the list of variables entering in $\mathbf{X}_2$ in (2).

4.3 Options

`hom(`#`)` specifies d_1, the number of variables in $\mathbf{X}_1$. The code then returns their estimated effects and standard errors. The default is `hom(1)`.

`subs(`#`)` specifies the subsample size b_n. Following D'Haultfœuille, Maurel, and Zhang (2018), and letting $x^+ = \max(0, x)$, the default value is set to

$$b_n = 0.6n - 0.2(n-500)^+ - 0.2(n-1000)^+ - 0.2\left\{1 - \frac{\ln(2000)}{\ln(n)}\right\}(n-2000)^+$$

`grid(`#`)` specifies $n_{\mathcal{G}}$, the number of grid points. The default is `grid(40)`.

`rep(`#`)` specifies B, the number of bootstrap and subsampling replications. The default is `rep(150)`.

`small` specifies that (3) holds when $Y^* \to -\infty$ rather than when $Y^* \to \infty$.

4.4 Stored results

`eqregsel` stores the following in `e()`:

Scalars

`e(tau0)`	quantile index $\widehat{\tau}_n$
`e(specificationtest)`	p-value of the specification test
`e(subs)`	subsample size when selecting the quantile index
`e(homvar)`	number of variables with homogeneous effects on the outcome

Matrices

`e(beta_hom)`	a $d_1 \times 1$ matrix containing the estimated coefficient or coefficients of interest
`e(std_b)`	a $d_1 \times 1$ matrix containing the standard error of the estimator or estimators

5 Example

We use the command **eqregsel** to estimate the black–white wage gap among young males from NLSY79 and NLSY97, revisiting the work of D'Haultfœuille, Maurel, and Zhang (2018) on this question. We are particularly interested in the evolution of the gap between these two cohorts.

We use the same samples and definitions of variables as D'Haultfœuille, Maurel, and Zhang (2018). In particular, we consider that an individual in the NLSY79 is a nonparticipant if he did not work in 1990 or in 1991. The outcome of interest is the (potential) log-wage, which is defined as the log of the mean real wages in 1990 and 1991 for workers who worked both years and the log of the real wage in the year of employment for those who worked only one year. We apply the same rules with the years 2007 and 2008 for individuals in the NLSY97.

In our specification, we estimate for the two samples the effect of the black dummy on the log of wages (**log_wage**), controlling for the Hispanic dummy (**hispanic**), age (**age**), Armed Forces Qualification Test (AFQT) score (**afqt**), and AFQT squared (**afqt2**). The AFQT scores cannot be directly compared across both NLSY cohorts, partly because of changes in how the test was administered. To handle this issue, we use a modified version of the AFQT constructed using the equipercentile mapping proposed by Altonji, Bharadwaj, and Lange (2012). We also restrict the samples to the respondents who took the test when they were 16 or 17, to address the issue that the rank within the AFQT distribution may vary with the age of the respondent at the time of the test. The final sample sizes are equal to 1,077 and 1,123 for the NLSY79 and NLSY97 cohorts, respectively. The overall labor force participation rates for the two corresponding samples are equal to 95.1% and 89.7%. However, they reach only 90.6% and 81.4% for black males.

We report below the output of the **eqregsel** procedure applied to the NLSY79 and NLSY97 samples, respectively. We use the default parameters. We can see from the estimation output that the default subsample sizes used in bootstrapping are 515 and 524, given the total sample size of 1,077 and 1,123. The procedure also displays the estimated computing time along with a progress bar. Although in this example estimation is performed at a limited computational cost, this feature makes it possible for the user to stop the execution of the command. If needed, one can then save on execution time by setting a lower number of bootstrap and subsampling replications or a lower number of grid points.[7]

[7] The computation times reported in these examples are obtained on an Intel Xeon CPU 2.40 GHz processor with 128 GB of RAM, using Stata/MP 14.2.

```
. use "bw_nlsy7997.dta"

. generate afqt2= afqt^2

. eqregsel log_wage black hispanic age afqt afqt2 if cohort79
The estimation will take about 5.333333 minutes.
|---------------|---------------|---------------|---------------|--------------|
0              20              40              60              80            100
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Number of observations =       1077
Optimal quantile index =       .245
J test(p-value) =  .81287468
Subsampling size used in bootstrapping =        515
Number of variables of interest =         1

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.1185019   .0431142    -2.75   0.006    -.2030043   -.0339996
------------------------------------------------------------------------------

. eqregsel log_wage black hispanic age afqt afqt2 if cohort97
The estimation will take about 5.333333 minutes.
|---------------|---------------|---------------|---------------|--------------|
0              20              40              60              80            100
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Number of observations =       1123
Optimal quantile index =        .29
J test(p-value) =  .77565885
Subsampling size used in bootstrapping =        524
Number of variables of interest =         1

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.1588783   .0406563    -3.91   0.000    -.2385632   -.0791935
------------------------------------------------------------------------------
```

The estimation results point to statistically and economically significant black–white wage gaps for the two cohorts. We also observe a wider black–white wage gap for the 1997 cohort relative to the 1979 cohort, with an increase in the estimated gap from about 11.9% to 15.9%. Note, however, that the difference is not significant at usual levels (p-value $= 0.51$). Interestingly, the p-values of the specification tests imply that one cannot reject our specification for either cohort at any standard statistical level.

It is interesting to compare the estimated black–white wage gap with the results of a simple ordinary least-squares regression of the log of hourly wages on a black dummy and the same set of controls. The estimated black–white wage gap drops from 11.9% and 15.9%, for our specifications, to 8.1% and 9.7% (with standard errors equal to 0.035 and 0.041), for the ordinary least-squares specification that ignores selection. That the estimated wage gap is larger in magnitude when we use our method is consistent with the underlying sample-selection issue. Indeed, among males, blacks are significantly more likely to drop out from the labor market (Juhn 2003). Because dropouts tend to have lower potential wages, one can expect that not controlling for endogenous labor market participation will result in underestimating the black–white wage differential.[8]

6 Conclusion

In this article, we have discussed how to use the `eqregsel` command to estimate and conduct inference on sample-selection models, following D'Haultfœuille, Maurel, and Zhang (2018). Unlike alternative estimation methods that have been proposed in the literature, the method does not require the presence of instruments or large support regressors. The estimator is simply based on a quantile regression in the tail but with a quantile index chosen in a data-driven fashion. `eqregsel` makes it possible to easily use this procedure.

7 Acknowledgments

Yichong Zhang acknowledges the financial support from the Singapore Ministry of Education Tier 2 grant under grant no. MOE2018-T2-2-169 and the Lee Kong Chian fellowship.

8 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 20-2
. net install st0598      (to install program files, if available)
. net get st0598          (to install ancillary files, if available)
```

[8] We also estimate the wage gap using the Heckman two-step estimator without any instrument. We obtain imprecisely estimated gaps of 24.2% and −21.2%, with standard errors of 0.48 and 0.68. This could be expected: in the absence of an instrument, this estimator strongly relies on functional form restrictions and is often unstable.

9 References

Ahn, H., and J. L. Powell. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58: 3–29. https://doi.org/10.1016/0304-4076(93)90111-H.

Altonji, J. G., P. Bharadwaj, and F. Lange. 2012. Changes in the characteristics of American youth: Implications for adult outcomes. *Journal of Labor Economics* 30: 783–828. https://doi.org/10.1086/666536.

Buchinsky, M. 1998. The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics* 13: 1–30. https://doi.org/10.1002/(SICI)1099-1255(199801/02)13:1⟨1::AID-JAE474⟩3.0.CO;2-A.

Chamberlain, G. 1986. Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics* 32: 189–218. https://doi.org/10.1016/0304-4076(86)90038-2.

Chen, S., and S. Khan. 2003. Semiparametric estimation of a heteroskedastic sample selection model. *Econometric Theory* 19: 1040–1064. https://doi.org/10.1017/S0266466603196077.

Chernozhukov, V., and I. Fernández-Val. 2011. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Review of Economic Studies* 78: 559–589. https://doi.org/10.1093/restud/rdq020.

Chernozhukov, V., I. Fernández-Val, and T. Kaji. 2018. Extremal quantile regression. In *Handbook of Quantile Regression*, ed. R. Koenker, V. Chernozhukov, X. He, and L. Peng, chap. 18, chap. 18. Handbooks of Modern Statistical Methods, Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781315120256-18.

Das, M., W. K. Newey, and F. Vella. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* 70: 33–58. https://doi.org/10.1111/1467-937X.00236.

D'Haultfœuille, X., and A. Maurel. 2013. Another look at the identification at infinity of sample selection models. *Econometric Theory* 29: 213–224. https://doi.org/10.1017/S026646661200028X.

D'Haultfœuille, X., A. Maurel, and Y. Zhang. 2018. Extremal quantile regressions for selection models and the black–white wage gap. *Journal of Econometrics* 203: 129–142. https://doi.org/10.1016/j.jeconom.2017.11.004.

Donald, S. G. 1995. Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics* 65: 347–380. https://doi.org/10.1016/0304-4076(93)01590-I.

Heckman, J. J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42: 679–694. https://doi.org/10.2307/1913937.

———. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161. https://doi.org/10.2307/1912352.

———. 1990. Varieties of selection bias. *American Economic Review* 80: 313–318.

Honoré, B., and L. Hu. 2018. Selection without exclusion. FRB of Chicago Working Paper No. WP-2018-10. http://doi.org/10.21033/wp-2018-10.

Jann, B. 2005. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s455001.html.

Juhn, C. 2003. Labor market dropouts and trends in the wages of black and white men. *ILR Review* 56: 643–662. https://doi.org/10.1177/001979390305600406.

Lewbel, A. 2007. Endogenous selection or treatment model estimation. *Journal of Econometrics* 141: 777–806. https://doi.org/10.1016/j.jeconom.2006.11.004.

Newey, W. K. 2009. Two-step series estimation of sample selection models. *Econometrics Journal* 12: S217–S229. https://doi.org/10.1111/j.1368-423X.2008.00263.x.

Politis, D. N., J. P. Romano, and M. Wolf. 1999. *Subsampling*. New York: Springer.

Vella, F. 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* 33: 127–169. https://doi.org/10.2307/146317.

About the authors

Xavier D'Haultfœuille is a professor at CREST-ENSAE.

Arnaud Maurel is an assistant professor in the Department of Economics at Duke University and a faculty research fellow at the NBER and IZA.

Xiaoyun Qiu is a doctoral student in the Department of Economics at Northwestern University.

Yichong Zhang is an assistant professor of economics in the School of Economics at Singapore Management University.