D'Haultfœuille, Xavier; Maurel, Arnaud; Qiu, Xiaoyun; Zhang, Yichong

Working Paper

Estimating Selection Models without Instrument with Stata

IZA Discussion Papers, No. 12486

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: D'Haultfœuille, Xavier; Maurel, Arnaud; Qiu, Xiaoyun; Zhang, Yichong (2019) : Estimating Selection Models without Instrument with Stata, IZA Discussion Papers, No. 12486, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/202832

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12486

Estimating Selection Models without Instrument with Stata

Xavier D'Haultfœuille
Arnaud Maurel
Xiaoyun Qiu
Yichong Zhang

JULY 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12486

Estimating Selection Models without Instrument with Stata

Xavier D'Haultfœuille
CREST-ENSAE

Arnaud Maurel
Duke University, NBER and IZA

Xiaoyun Qiu
Northwestern University

Yichong Zhang
Singapore Management University

JULY 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Estimating Selection Models without Instrument with Stata*

This article presents the eqregsel command for implementing the estimation and bootstrap inference of sample selection models via extremal quantile regression. The command estimates a semiparametric sample selection model without instrument or large support regressor, and outputs the point estimates of the homogenous linear coefficients, their bootstrap standard errors, as well as the p-value for a specification test.

JEL Classification: C21, C24, C87, J31

Keywords: eqregsel, sample selection models, extremal quantile regressions

Corresponding author:
Arnaud Maurel
Duke University
Department of Economics
213 Social Sciences Building, Box 90097
Durham, NC 27708-0097
USA

E-mail: apm16@duke.edu

* Zhang acknowledges the financial support from Singapore Ministry of Education Tier 2 grant under grant no. MOE2018-T2-2-169.

1 Introduction

In this article, we present the command `eqregsel` for estimation and inference of endogenous sample selection models that implements the procedures developed in recent work by D'Haultfœuille et al. (2018).[1] Prior methods proposed in the econometric literature to estimate endogenous sample selection models rely on instruments and/or large support regressors. For the former, see, among others, Heckman (1974, 1979, 1990), Ahn and Powell (1993), Donald (1995), Buchinsky (1998), Chen and Khan (2003), Das et al. (2003), Newey (2009) and Vella (1998) for a survey. Chamberlain (1986) and Lewbel (2007) develop 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, as well as large support regressors are often difficult, if not impossible to find.

Instead, the method implemented in `eqregsel` does not require the presence of instruments or large support regressors.[2] Identification relies instead on the strategy initially proposed by D'Haultfoeuille and Maurel (2013), which is based on the idea that, provided that selection is endogenous, one can expect the effect of the outcome on selection to dominate those of the covariates, for large values of the outcome. `eqregsel` builds on the estimation method proposed by D'Haultfœuille et al. (2018) and implements a series of quantile regressions in the tails of the outcome distribution (extremal quantile regressions).[3] 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 et al. (2018).

Our command 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 to `heckman`. First, it does not require normality of the error term in the selection equation, nor linearity of the conditional expectation of the error term in the outcome equations. 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.

[1] The Stata command `eqregsel` can be downloaded from the following webpage: `http://www.amaurel.net/Packages`.

[2] 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 nor large support regressors. However, their approach is based on a different set of assumptions and, in contrast to our framework, delivers set- rather than point-identification.

[3] See Chernozhukov et al. (2017) for an overview of extremal quantile regression methods and recent applications.

The remainder of the paper is organized as follows. In Section 2, we recall the setup of the semiparametric endogenous sample selection model considered in D'Haultfœuille et al. (2018), and describe the data-driven procedure used to choose the quantile index for the extremal quantile regression. Section 3 describes how to implement the method in practice. Section 4 presents the `eqregsel` command. Section 5 illustrates the use of our command by estimating the black-white wage gap on US young males of the 1979 and 1997 National Longitudinal Surveys of Youth. Section 6 concludes.

2 The framework and estimation method

2.1 Model and estimation

We consider the following outcome equation:

$$Y^* = X_1'\beta_1 + \varepsilon$$

where $Y^* \in \mathbb{R}$ and $X_1 \in \mathbb{R}^{d_1}$ are the outcome and covariates of interest, respectively. In the following, we seek to identify and estimate β_1. For that purpose, we rely on two key conditions. The first is that for any $\tau \in (0,1)$, the τ-th conditional quantile of ε satisfies

$$Q_{\varepsilon|X}(\tau|X) = \beta_0(\tau) + X_2'\beta_2(\tau), \tag{2.1}$$

where $X = (X_1', X_2')'$ and X_2 denotes other covariates. Then

$$Q_{Y^*|X}(\tau) = X_1'\beta_1 + \beta_0 + X_2'\beta_2(\tau). \tag{2.2}$$

The effect of X_1 is thus assumed to be homogenous across different quantile indices, while the effect of the other covariates 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 only observes D, $Y = DY^*$ and 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 $x \in \text{Supp}(X)$,

$$\lim_{y\to\infty} P(D = 1|X = x, Y^* = y) = h. \tag{2.3}$$

Combining (2.2) and (2.3), D'Haultfœuille et al. (2018, Theorem 1) shows that, under some

regularity conditions on the upper tail of ε, as $\tau \to 0$,

$$\begin{aligned} Q_{-Y|X}(\tau|X) &= Q_{-Y^*|X}(\tau/h|X) + o(1) \\ &= -X_1'\beta_1 - \beta_0(1-\tau/h) - X_2'\beta_2(1-\tau/h) + o(1). \end{aligned} \tag{2.4}$$

Therefore, (2.4) suggests that we can estimate β_1 by running a quantile regression of $-Y$ on $-X$ with a sufficiently small quantile index τ, i.e.,

$$\left(\widehat{\beta}_1', \widehat{\beta}_0(1-\tau/h), \widehat{\beta}_2'(1-\tau/h)\right)' = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(-Y_i + \overline{X}_i'\beta), \tag{2.5}$$

where $\rho_\tau(u) = (\tau - \mathbb{1}\{u < 0\})u$ is the check function used in quantile regressions and $\overline{X}_i = (X_{1i}', 1, X_{2i}')'$. Intuitively, for $\widehat{\beta}_1$ 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 et al. (2018) establish that if $\tau_n \to 0$ and $n\tau_n \to \infty$,[4], and under additional technical restrictions, $\widehat{\beta}_1$ is consistent and asymptotically normal.

As is standard with extremal quantile regressions (see Chernozhukov et al., 2017), 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^*, X).[5] Importantly, D'Haultfœuille et al. (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 $\widehat{\beta}_1^*$, assuming for simplicity that X_1 is a scalar ($d_1 = 1$). Then, Theorem 2 in D'Haultfœuille et al. (2018) implies that the percentile bootstrap confidence interval $[q_{\alpha/2}^*, q_{1-\alpha/2}^*]$ of β_1 has an asymptotic coverage 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 (2.1) and (2.3). Importantly, we can develop a specification test of these conditions, based on the implication that the coefficient β_1 in $Q_{-Y|X}(\tau_n|X)$ is the same across different extremal quantile indices τ_n (see (2.4)). Then, if the model is correctly specified, the two estimators $\widehat{\beta}_1(\ell\tau_n)$ (with $0 < \ell < 1$) and $\widehat{\beta}_1(\tau_n)$ of β_1, obtained respectively with $\tau = \ell\tau_n$ and $\tau = \tau_n$, should be close. Following this idea, consider the following J-test statistic:

$$T_J(\ell) = [(1/\ell) - 1]^2 (\widehat{\beta}_1(\tau_n) - \widehat{\beta}_1(\ell\tau_n))' \widehat{\Omega}^{-1} (\widehat{\beta}_1(\tau_n) - \widehat{\beta}_1(\ell\tau_n)), \tag{2.6}$$

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

[5] We refer to the definition of the rate above Theorem 2.2 in D'Haultfœuille et al. (2018) for more details.

where $\widehat{\Omega}$ is a (bootstrap) estimator of the asymptotic covariance of $\widehat{\beta}_1(\tau_n)$, 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 quantile of order $1-\alpha$ of a χ^2 distribution with d_1 degrees of freedom. Theorem 2.3 in D'Haultfœuille et al. (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 trade-off between bias and variance, which is governed by the quantile index τ_n used in the extremal quantile regression. We present in the following the algorithm outlined in D'Haultfœuille et al. (2018), which selects a suitable quantile index based on estimators of the bias and the variance of $\widehat{\beta}_1$.

Specifically, consider the same test statistic as in (2.6), but where $(\ell\tau_n, \tau_n)$ are replaced by $(\ell_1\tau_n, \ell_2\tau_n)$, with $\ell_1 < 1 < \ell_2$:

$$T_J(\tau) = [1/\ell_1 - 1/\ell_2]^2(\widehat{\beta}_1(\ell_2\tau) - \widehat{\beta}_1(\ell_1\tau))'\widehat{\Omega}^{-1}(\widehat{\beta}_1(\ell_2\tau) - \widehat{\beta}_1(\ell_1\tau)).$$

D'Haultfœuille et al. (2018) shows 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.[6] 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

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

where b_n denotes the subsample size.

Similarly, the asymptotic covariance matrix is estimated by the covariance matrix of the subsampling estimator of β_1, multiplied by the normalizing factor b_n/n. Denote by $\widehat{\text{Var}}_n(\tau)$ the sum of the diagonal elements of this covariance matrix. The quantile index is selected to

[6] 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 et al. (1999) for an introduction.

optimize the bias-variance trade-off:

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

where $\mathcal{G}$ denotes a finite grid within $(0, 1)$. This procedure results in undersmoothing in comparison with a more standard trade-off between variance and squared bias. Similarly to 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 et al. (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 β_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 $\beta(\tau) = (\beta_1', \beta_0(1 - \tau/h), \beta_2'(1 - \tau/h))'$:

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

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

   (b) Compute

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

   with $\widehat{\beta}_1^b(\tau)$ the bootstrap estimator of β_1 on the b-th bootstrap sample.

   (c) Compute, for each subsample $s = 1 \dots B$, the estimator of β_1 ($\widehat{\beta}_1^s(\tau)$), and the J-test statistic:[7]

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

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

[7] 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.

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

$$\widehat{\Sigma}(\tau) = \frac{1}{B}\sum_{s=1}^{B}(\widehat{\beta}_1^s(\tau) - \overline{\beta}_1(\tau))(\widehat{\beta}_1^s(\tau) - \overline{\beta}_1(\tau))' \quad \text{with} \quad \overline{\beta}_1(\tau) = \frac{1}{B}\sum_{s=1}^{B}\widehat{\beta}_1^s(\tau).$$

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

4. Define $\widehat{\beta}_1 = \widehat{\beta}_1(\widehat{\tau}_n)$ and $\widehat{\Omega} = \widehat{\Omega}(\widehat{\tau}_n)$. Confidence intervals $\text{CI}_{1-\alpha}(\beta_{1k})$ of level $1-\alpha$ on the k-th component of β_1 are then equal to

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

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

5. Compute $\widehat{\beta}_1(0.2\widehat{\tau}_n)$ and then $T_J(0.2)$, as defined in (2.6), 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 Fernandez-Val, 2011 for a related discussion). To compute $T_J^s(\tau)$ in Step 2.(c) above, we use $(\ell_1, \ell_2) = (0.9, 1.1)$.

4 The `eqregsel` command

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

4.1 Syntax

The syntax of `eqregsel` is as follows:

`eqregsel Y X1 X2` [*if*] [*in*] [`, hom(#) subs(#) grid(#) rep(#)`]

4.2 Description

`eqregsel` computes $\widehat{\beta}_1$ in(2.2) based on the data-driven τ_n detailed in Section 2.2 above. It also reports its standard errors and 95% confidence intervals. Finally, it computes the p-value of this specification test using $\ell = 0.2$.

`X1` is the list of variable entering in X_1 in Model (2.2).

`X2` is the list of variable entering in X_2 in Model (2.2).

4.3 Options

`hom(#)` specifies d_1, the number of variables in X_1. The code then returns their estimated effects and standard errors. The default value is 1.

`subs(#)` specifies the subsample size b_n. Following D'Haultfœuille et al. (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 value is 40.

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

4.4 Saved results

The `eqregsel` command saves the following in `e()`:

1. `e(tau0)`, a scalar containing the quantile index $\widehat{\tau}_n$.
2. `e(specificationtest)`, a scalar containing the p-value of the specification test.
3. `e(subs)`, a scalar containing the subsample size b_n.
4. `e(homvar)`, a scalar containing d_1, the number of variable(s) with homogenous effect(s) on the outcome.
5. `e(beta_hom)`, a $d_1 \times 1$ matrix containing the estimated coefficient(s) of interest.
6. `e(std_b)`, a $d_1 \times 1$ matrix containing the standard error of the estimator(s).

5 Example

We use the command `eqregsel` to estimate the black-white wage gap among young males from the National Longitudinal Surveys of Youth 1979 and 1997 (NLSY79 and NLSY97), revisiting the work of D'Haultfœuille et al. (2018) on this question. We are in particular interested in the evolution of the gap between these two cohorts. We use the same samples and definitions of variables as in D'Haultfœuille et al. (2018). In particular, we consider that an individual in the NLSY79 is a nonparticipant if he did not work in 1990 nor 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 Hispanic dummy (`hispanic`), age (`age`), AFQT (Armed Forces Qualification Test score) and AFQT squared (`afqt` and `afqt2`). The AFQT scores cannot be directly compared across both NLSY cohorts, in part 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 et al. (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%. They only reach 90.6% and 81.4% for Black males, however.

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.[8]

[8]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",clear

.
. gen 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 as well as 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 OLS 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 OLS 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 dropout from the labor market (Juhn, 2003). Since 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.[9]

6 Conclusion

In this paper we have discussed how to use the `eqregsel` command to estimate and conduct inference on sample selection models, following D'Haultfœuille et al. (2018). Unlike alternative

[9] We also estimate the wage gap using the Heckman two-step estimator, without any instrument. We obtain very 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 instrument, this estimator strongly relies on functional form restrictions and is often unstable.

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. The Stata command `eqregsel` makes it possible to easily use this procedure.

References

Ahn, H., and J. L. Powell. 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58(1-2): 3–29.

Altonji, J., P. Bharadwaj, and F. Lange. 2012. Changes in the characteristics of American youth: Implications for adult outcomes. *Journal of Labor Economics* 30(4): 783–828.

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): 1–30.

Chamberlain, G. 1986. Asymptotic efficiency in semiparametric model with censoring. *Journal of Econometrics* 32(2): 189–218.

Chen, S., and S. Khan. 2003. Semiparametric estimation of a heteroskedastic sample selection model. *Econometric Theory* 19(6): 1040–1064.

Chernozhukov, V., and I. Fernandez-Val. 2011. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Review of Economic Studies* 78(2): 559–589.

Chernozhukov, V., I. Fernandez-Val, and T. Kaji. 2017. Extremal quantile regression: An overview. In *Handbook of Quantile Regression*, ed. R. Koenker, V. Chernozhukov, X. He, and L. Peng. Chapman and Hall/CRC.

Das, M., W. Newey, and F. Vella. 2003. Nonparametric estimation of sample selection models. *Review of Economic Studies* 70(1): 33–58.

D'Haultfoeuille, X., and A. Maurel. 2013. Another look at the identification at infinity of sample selection models. *Econometric Theory* 29(1): 213–224.

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(1): 129–142.

Donald, S. 1995. Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics* 65(2): 347–380.

Heckman, J. J. 1974. Shadow prices, market wages, and labor supply. *Econometrica* 42(4): 679–694.

———. 1979. Sample selection bias as a specification error. *Econometrica* 47(1): 153–161.

———. 1990. Varieties of selection bias. *The American Economic Review* 80(2): 313–318.

Honoré, B., and L. Hu. 2018. Selection without exclusion. Working paper.

Juhn, C. 2003. Labor market dropouts and trends in the wages of black and white men. *Industrial and Labor Relations Review* 56(4): 643–662.

Lewbel, A. 2007. Endogenous selection or treatment model estimation. *Journal of Econometrics* 141(2): 777–806.

Newey, W. 2009. Two step series estimation of sample selection models. *The Econometrics Journal* 12(1): 217–229.

Politis, D. N., J. P. Romano, and M. Wolf. 1999. *Subsampling*. Springer Science & Business Media.

Vella, F. 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* 33(1): 127–169.