arhomme: A Stata implementation of the Arellano/Bonhomme (2017) estimator for quantile regression with selection correction

Martin Biewen (University of Tübingen, IZA Bonn)
Pascal Erhardt (University of Tübingen)

Swiss Stata Conference 2020
Virtual Bern, November 19, 2020

Outline

Sample selection bias

- Example: Female labor market participation and pay
  - How much *would* a woman with given characteristics be paid if she decided to work?
  - One cannot just look at women who *actually* work because these might be endogenously selected
  - A given woman might decide not to work if her potential pay is too low in comparison to alternative options
  - Basing a wage regression only on women who are observed working will lead to biased regression coefficients

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Selection correction models

- **Mean outcomes**
  - Heckman (1979), Ahn/Powell (1993), Andrews/Schafgans (1998), Chen/Khan (2003), Das/Newey/Vella (2003)
- **Distributional outcomes**
  - Buchinsky (1998, 2001), Albrecht et al. (2009)
  - Huber/Melly (2015) showed that this is too restrictive
  - First general solution: Arellano/Bonhomme (2017)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Arellano/Bonhomme (2017) method

- Model equations

$$Y^* = \mathbf{X}'\beta(U) \qquad \text{(Determinants of potential outcome)}$$

$$D = 1\{V \leq p(\mathbf{Z})\} \qquad \text{(Selection equation)}$$

$$Y = Y^* \text{ if } D = 1 \qquad \text{(Observable outcomes)}$$

- Model framed in terms of unobserved ranks

U (Rank of individual in conditional distribution $Y^*|\mathbf{X}$)

V (Rank in resistance towards selection)

- Ranks are jointly uniformly distributed

$$C_{U,V|\mathbf{X}=\mathbf{x}}(U,V) \qquad \text{(Copula function connecting ranks)}$$

Arellano/Bonhomme (2017) method

- Key insight

$$P[Y^* \leq \mathbf{X}'\beta(\tau) \ |D = 1, \mathbf{Z} = \mathbf{z}] = P[U \leq \tau | V \leq p(\mathbf{z}), \mathbf{Z} = \mathbf{z}]$$

$$= \frac{C_{U,V|\mathbf{X}=\mathbf{x}}(\tau, p(\mathbf{z}))}{p(\mathbf{z})} := G_{\mathbf{x}}(\tau, p(\mathbf{z}))$$

- Interpretation

  τ-quantiles in *overall* population correspond to
  $G_{\mathbf{x}}$-quantiles in *selected* population

➔ **'Rotated' quantile regression**

- For practical estimation, one has to assume a parametric model for copula (leading to a model for $G_{\mathbf{x}}(u, v)$)
- And for selection probability (e.g. probit)

Arellano/Bonhomme (2017) method

- Estimation (GMM + rotated quantile regression)

$$\widehat{\rho} = \underset{r \in \mathcal{R}}{argmin} \left\| \sum_{i=1}^{N} \sum_{l=1}^{L} \left[D_i \, \varphi(\mathbf{Z}_i) \left(1\left\{ Y_i < \mathbf{X}_i' \widehat{\beta}(\tau_l, r) \right\} - G(\tau_l, \Phi(\mathbf{Z}_i' \widehat{\gamma}); r) \right) \right] \right\|$$

$$\widehat{\beta}(\tau) = \underset{\mathbf{b}(\tau) \in \mathcal{B}}{argmin} \sum_{i=1}^{N} D_i \left[\widehat{G}_{\tau,i} \left(Y_i - \mathbf{X}_i' \mathbf{b}(\tau) \right)^{+} + \left(1 - \widehat{G}_{\tau,i} \right) \left(Y_i - \mathbf{X}_i' \mathbf{b}(\tau) \right)^{-} \right]$$

- Compare ‘unrotated’ (= ordinary) quantile regression

$$\tilde{\beta}(\tau) = \underset{\mathbf{b} \in \mathcal{B}}{argmin} \sum_{i=1}^{N} D_i \left[\tau \left(Y_i^* - \mathbf{X}_i' \mathbf{b} \right)^{+} + (1 - \tau) \left(Y_i^* - \mathbf{X}_i' \mathbf{b} \right)^{-} \right]$$

Algorithms and inference

- **Algorithms**
  - We use interior point algorithm by Morillo/Koenker/Eilers which we translated from Matlab to Mata
  - Often much faster than algorithm used in `qreg`
- **Inference**
  - Arellano/Bonhomme (2017) showed (pointwise) asymptotic normality but asymptotic variance matrix very complex
  - In practice they used ‘subsampling’ (Politis/Romano,1994)
  - But choice of subsample size is difficult issue
  - Bootstrap should be prefered if computationally realistic
  - We implement subsampling as well as conventional bootstrap

The arhomme command

`arhomme` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] ,

`select`([*depvar*$_s$ [=]]*varlist*$_s$) [`rhopoints(`#`)` `taupoints(`#`)` `meshsize(`#`)` `centergrid(`#`)` `frank` `gaussian` `plackett` `joema` `nostderrors` `subsample(`#`)` `repetitions(`#`)` `instrument(`*varname*`)` `copulaparameter(`*varname*`)` `quantiles` (#[#[#...]]) `graph` `output(`[*normal*] [*bootstrap*]`)`]

- `arhomme` is byable
- `pweights` are allowed
- Postestimation: `predict`, `test` etc.

Empirical illustration 1: `heckman` data set

```
. webuse womenwk

. /* estimate median regression with selection correction */

. arhomme wage educ age, select(educ age married children) gaussian q(.5) tau(7) rep(250)

option subsample left unspecified: subsample automatically set to 2000 (bootstrap)
use option nostderrors to disable estimation of covariance matrix

First step estimation (probit model) successfully completed.

Second step (gaussian copula parameter estimation) successfully completed.
Found objective function minimum 8.993e-07 for rho = -0.6517

Third step (minimization of rotated check function) successfully completed.

Initialising standard error estimation by 2000 out of 2000 bootstrap method:
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
```

(Output continued on next page)

Empirical illustration 1: `heckman` data set

```
------------------------------------------------------------------------------
Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)
------------------------------------------------------------------------------
                                                 Number of obs.    =      2,000
                                                 Num. of selected =      1,343
                                                 Rho points        =         19
                                                 Tau points        =          7
                                                 Meshsize          =     1.0000
                                                 Spearman's rho    =    -0.6339
                                                 Kendall's tau     =    -0.4519
                                                 Blomqvist's beta =    -0.4519
                                                 Minimum Fval      =  8.993e-07
                                                 Replications      =        250
                                                 Subsample Size    =      2,000
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
select       |
   education |   .0583645   .0111586     5.23   0.000      .036494    .0802351
         age |   .0347211   .0042541     8.16   0.000     .0263832    .0430591
     married |   .4308575   .0745429     5.78   0.000     .2847561    .5769589
    children |   .4473249   .0279817    15.99   0.000     .3924817    .5021681
       _cons |  -2.467365   .1915084   -12.88   0.000    -2.842715   -2.092015
-------------+----------------------------------------------------------------
.5_quantile  |
       _cons |   .5695862   1.392844     0.41   0.683    -2.160337     3.29951
   education |   1.016767   .0760082    13.38   0.000      .867794    1.165741
         age |    .203274   .0259338     7.84   0.000     .1524446    .2541034
-------------+----------------------------------------------------------------
_anc         |
         rho |  -.6516752   .0759974    -8.57   0.000    -.8006273    -.502723
------------------------------------------------------------------------------
```

Empirical illustration 2: Arellano/Bonhomme (2017b)

```
. /* Replicates empirical application in Arellano/Bonhomme (2017b),
> Handbook of Quantile regression based on Huber/Melly (2015) data */
.
. arhomme lwage $X [pw=wgt], sel(ft = $X $B) tau(4) rho(39) gauss subsample(1000) rep(500) quant(.25 .5 .75)

------------------------------------------------------------------------------
Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)
------------------------------------------------------------------------------
                                                 Number of obs.   =     44,562
                                                 Num. of selected =     20,055
                                                 Rho points       =         39
                                                 Tau points       =          4
                                                 Meshsize         =     1.0000
                                                 Spearman's rho   =    -0.0945
                                                 Kendall's tau    =    -0.0631
                                                 Blomqvist's beta =    -0.0631
                                                 Minimum Fval     =  1.473e-08
                                                 Replications     =        500
                                                 Subsample Size   =      1,000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ft           |
      educ_7 |   .5869417   .0428666    13.69   0.000     .5029247    .6709586
      educ_8 |    .073392   .0226713     3.24   0.001     .0289571    .1178269
      educ_9 |   .2325266   .0261318     8.90   0.000     .1813092    .2837441
     educ_11 |   .0598427   .0287012     2.09   0.037     .0035894    .1160959
     educ_13 |   .1910608   .0310857     6.15   0.000      .130134    .2519876
         exp |   .0036565   .0044806     0.82   0.414    -.0051252    .0124382
        exp2 |  -.0003162   .0001053    -3.00   0.003    -.0005225   -.0001098
...
```

(Output omitted, continued on next page)

Empirical illustration 2: Arellano/Bonhomme (2017b)

```
-------------+----------------------------------------------------------------
.25_quantile |
       _cons |    1.95851    .0880627    22.24   0.000      1.78591    2.13111
      educ_7 |   .2042428    .0716669     2.85   0.004     .0637782   .3447073
      educ_8 |   .1047957     .018662     5.62   0.000     .0682188   .1413727
      educ_9 |   .0759379    .0213025     3.56   0.000     .0341859     .11769
     educ_11 |   .2806021    .0230099    12.19   0.000     .2355035   .3257007
     educ_13 |   .1891199    .0264449     7.15   0.000     .1372889    .240951
         exp |   .0163292     .003148     5.19   0.000     .0101592   .0224992
...
```

(Output omitted)

```
-------------+----------------------------------------------------------------
.75_quantile |
...
```

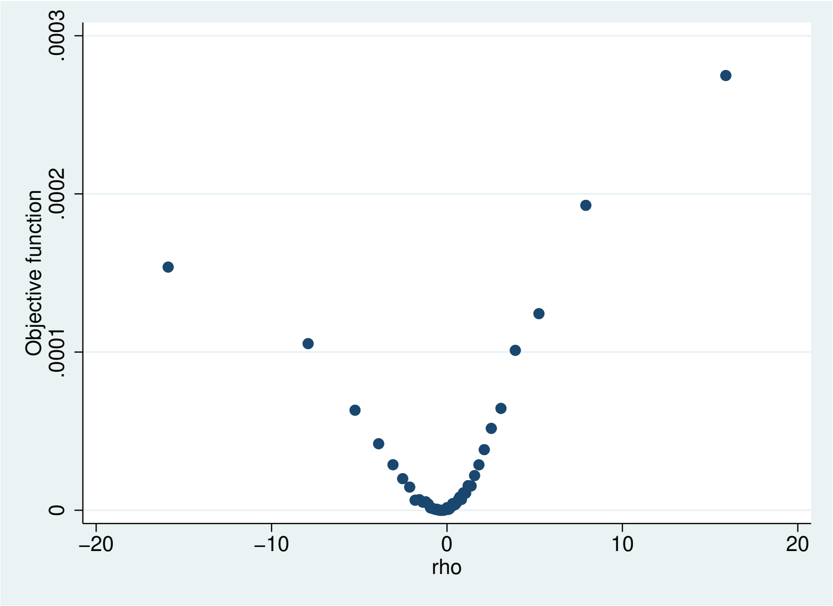
(Output omitted)

```
         exp |   .0301231    .0032301     9.33   0.000     .0237923   .0364539
        exp2 |  -.0004632    .0000761    -6.09   0.000    -.0006123  -.0003141
     exp_edu |   .0032962    .0007618     4.33   0.000      .001803   .0047893
    exp2_edu |    -.00007    .0000194    -3.61   0.000     -.000108   -.000032
     midwest |  -.1216595    .0160573    -7.58   0.000    -.1531313  -.0901878
       south |  -.0978834    .0164904    -5.94   0.000     -.130204  -.0655627
        west |  -.0157402    .0168777    -0.93   0.351    -.0488198   .0173394
     married |   .0210093    .0123316     1.70   0.088    -.0031602   .0451788
-------------+----------------------------------------------------------------
_anc         |
         rho |  -.0989229    .0535576    -1.85   0.065    -.2038938    .006048
------------------------------------------------------------------------------
```

Empirical illustration 3: Arellano/Bonhomme (2017)

```
. /* Partly replicates empirical application in original article
> Arellano/Bonhomme (Ectra, 2017) and illustrates grid search options */
.
. /* estimate on subsample single women */
. // first, crude estimation
.
. arhomme lw $X, sel(work = $X s_zero) frank graph rho(49) tau(4) q(.5) nostd
...
```

(Output omitted)



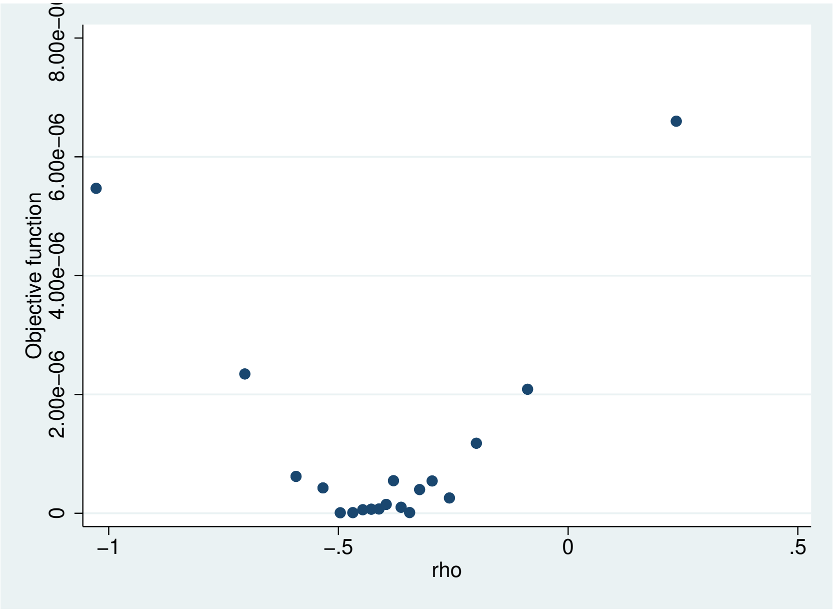
Empirical illustration 3: Arellano/Bonhomme (2017)

```
. local c = e(rho)

. graph save "Graph" "H:\_Pascal\soepdata\stata journal example_3\female single objective function.gph"
(file H:\_Pascal\soepdata\stata journal example_3\female single objective function.gph saved)

.
. // now a more detailed search
. arhomme lw $X, sel(work = $X s_zero) frank graph rho(19) tau(7) q(.5) center(‘c’) mesh(0.1) nostd
...
```

(Output omitted)



Empirical illustration 3: Arellano/Bonhomme (2017)

```
. /* next, estimate standard errors by subsampling */
. local s = 1000 + ceil(sqrt(_N))

. arhomme lw $X, sel(work = $X s_zero) fra rho(39) tau(7) q(.5) center(`c') rep(250) sub(`s')
...
```

(Output omitted)

```
Initialising standard error estimation by 1154 out of 23583 bootstrap method:
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................     50
..........................................
numerical derivatives are approximate nearby values are missing
x.......    100
..................................................    150
..................................................    200
..................................................    250
.    251
Probit model failed to converge for 1 subsample(s).
```

(Output omitted)

```
------------------------------------------------------------------------------
Arellano & Bonhomme (2017) selection model
(conditional quantile regression with sample selection)
------------------------------------------------------------------------------
                                                 Number of obs.   =     23,583
                                                 Num. of selected =     15,185
```

(Output omitted)

```
...
_anc         |
         rho |   -.495928   .5468089    -0.91   0.364    -1.567654    .5757978
------------------------------------------------------------------------------
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Summary

- `arhomme` implements Arellano/Bonhomme (2017) quantile regression with sample selection correction
- `arhomme` is fast and comfortable
- Potentially applicable in many fields in which there is need for correcting conditional distributions for sample selection
- *Unconditional* distributions corrected for sample selection can be obtained by aggregation (Chernozhukov et al., 2013)

Thank you!

Contact:

Martin Biewen

Mohlstraße 36, 72074 Tübingen
Telefon: +49 7071 29-75438
Telefax: +49 7071 29-5013
martin.biewen@uni-tuebingen.de

Main references

Albrecht, J., A. van Vuuren, and S. Vroman. 2009. Counterfactual distributions with sample selection adjustments: Econometric theory and an application to the Netherlands. *Labour Economics* 16: 383-396.

Arellano, M. and S. Bonhomme. 2017. Quantile Selection Models with an Application to Understanding Changes in Wage Inequality. *Econometrica* 85: 1-28

Arellano, M. and S. Bonhomme. 2017b. Sample Selection in Quantile Regression: A Survey. In *Handbook of Quantile Regression*, eds. Koenker, R., V. Chernozhukov, X. He, and L. Peng. New York: Chapman and Hall., Chapter 13.

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.

Chernozhukov, V., I. Fernandez-Val, and B. Melly. 2013. Inference on Counterfactual Distributions. *Econometrica* 81: 2205-2268.

Heckman, J.J. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47: 153-161.

Huber, M., and B. Melly. 2015. A Test of the Conditional Independence Assumption in Sample Selection Models. *Journal of Applied Econometrics* 30: 1144-1168.