Documentation of MATLAB code for "Extremal Quantile Regressions for Selection Models and the Black-White Wage Gap"

Xavier D'Haultfœuille* Arnaud Maurel† Yichong Zhang‡

June 2014

The batch file **mytry.m** loads the data **cooked79_2** and produces the results reported in Table 5 of section 4 of our paper "Extremal Quantile Regressions for Selection Models and the Black-White Wage Gap". In particular, the code conducts an automatic pretest for homoskedasiticity and yields the point estimates and its standard deviation of β and δ corresponding to the homoskedastic variable(s) along with the CLR bound of the τ-th marginal quantile treatment effect for the heteroskedasitic variable(s) where the quantile index τ is user-specified. In a special note, the grid generated in this function should satisfy the data restriction. For example, in our case of using NLSY79 data, **black** and **hispanic** cannot be both 1 while $\mathbf{AFQT}^2$ is fully determined on **AFQT**. For the detail, please check the program. The main function it calls is **myfun_combined.m**. In **myfun_combined.m**, the default spacing parameters we use are $m = 1.2$, $l = (0.65, 0.85, 1.15, 1.45)$. The number of subsamples is set to be 500. The subsample size is (150,300,500,600) for sample size (250, 500, 1,000 and 2,000) and the corresponding linear interpolation for sample size in between. For sample size N larger than 2000, the subsample size is set to be $600 + 0.2(N - 2000)$. Also, we use the rule of thumb tuning parameter and Gaussian kernel to estimate the nonparametric pieces when computing the CLR bound. In the following, we describe the function **myfun_combined.m** as well as the additional functions this function calls.

*CREST. E-mail address: xavier.dhaultfoeuille@ensae.fr.
†Duke University, NBER and IZA. E-mail address: apm16@duke.edu.
‡Duke University. E-mail address: yz98@duke.edu.

myfun_combined.m

Input

- Xd: a matrix of discrete dependent variables
- Xc: a matrix of discrete dependent variables
- Y: a vector of independent variables
- Xdnp: a matrix of free discrete dependent variables (free means it is not determined by other components of X = (Xd,Xc))
- Xcnp: a matrix of free continuous dependent variables
- quant: the quantile of interest of the CLR bound
- grid: the grid of X used to compute the CLR bound
- gridnp: the grid of free X used in nonparametric estimation when computing CLR bound

Output

- delta_hetero: the estimator of delta without imposing partial homoskedasticity
- beta_hetero: the estimator of beta without imposing partial homoskedasticity
- V_hetero_star: the asymptotic variance covariance matrix for delta_hetero (and also beta_hetero). The convergence rate for delta_hetero is $\sqrt{n\tau}$ where n is the total sample size and τ is the quantile index used to compute delta_hetero
- Nb_hetero_star: the convergence for beta_hetero
- qstard: the optimal quantile index used to estimate both delta_hetero and beta_hetero
- phi: a 1 by d vector contains the results of pretest where d is the dimension of X (exclude intercept). If j-th entry is 1, then it indicates the j-th variable is homoskedastic, i.e. $\delta_j = 0$.
- delta_hom: the point estimates of delta under partial homoskedasticity for those variables who do not pass the pretest
- beta_hom: the point estimates of beta under partial homoskedasticity for those variables who pass the pretest
- Nb_hom_star: the convergence rate of beta_hom

- V_homd_star: the variance covariance matrix of delta_hom
- V_homb_star: the variance covariance matrix of beta_hom
- qstarbd: the optimal quantile index used to estimate delta_hom
- qstarbb: the optimal quantile index used to estimate beta_hom
- theta0: the CLR bound for quant-th marginal quantile treatment effect for those variables who do not pass the pretest (where quant is the input argument)

Function it calls

- **myfun_hetero.m**: produce point estimates of δ and β without assuming partial homoskedasticity.
- **myfun_hom.m**: produce point estimates of δ and β under partial homoskedasticity.
- **bound.m**: produce upper and lower bound of $Q_{\varepsilon}(\tau)$.

myfun_hetero.m:

Input

- tau: quantile
- m: a tuning parameter that is used to normalize our estimator. should be different from but very close to 1. In simulation and applications, m is set to be 1.2
- b0: the location normalizing factor
- d0: the scale normalizing factor
- X: dependent variables
- Y: independent variables
- l: equations we will explore by minimum distance estimations are indexed by l. In general, we will use quantile level tau, tau*l to estimate beta and delta
- delta_hetero: the user supplied starting value of delta. If it is left unspecified, the program will use OLS estimator

Output

- out: $2(d-1)$ by 1 vector collects (d-1) estimator of delta and (d-1) estimator of beta, where d is the dimension of x (including intercept).
- V: The asymptotic variance for delta
- dis: minimum distance of delta computed by plugging in the extremal estimator of delta
- Nb1: the convergence rate for beta

Functions it calls

- **rq_fnm.m**: It conducts quantile regression of Y on X at user specified quantile index τ. The code is programmed by Roger Koenker and we download it from his personal webpage. For further remarks, please check the comments in the code.

myfun_hom.m

Input

- tau: quantile
- m: a tuning parameter that is used to normalize our estimator. should be different from but very close to 1. In simulation and applications, m is set to be 1.2
- X: dependent variables
- Y: independent variables
- l: equations we will explore by minimum distance estimations are indexed by l. In general, we will use quantile level tau, tau*l to estimate beta and delta
- phi: a 1 by d vector contains the results of pretest where d is the dimension of X (exclude intercept). If j-th entry is 1, then it indicates the j-th variable is homoskedastic, i.e. $\delta_j = 0$.
- delta_MSE: the user supplied starting value of delta. If it is left unspecified, the program will use OLS estimator

Output

- par: a (d-1) by 1 vector collects estimators of heteroskedastic delta and estimators of homoskedastic beta, where d is the dimension of x (including intercept).
- Vd: the asymptotic variance matrix for delta
- Vb: the asymptotic variance matrix for beta
- dis_d: minimum distance for delta evaluated by plugging in the extremal estimator of delta under partial homoskedasticity
- dis_b: minimum distance for beta evaluated by plugging in the extremal estimator of beta under partial homoskedasticity

Functions it calls

- **rq_fnm.m**

bound.m

Input

- Y: the independent variable
- D: $(Y > 0)$
- Xdnp: a matrix of free discrete dependent variables (free means it is not determined by other components of X = (Xd,Xc))
- Xcnp: a matrix of free continuous dependent variables
- tau: the quantile of interest of the CLR bound
- gridnp: the grid of free X used in nonparametric estimation when computing CLR bound

Output

- y1: the lower bound
- y2: the upper bound

Functions it calls

- **condQ.m**: It computes the conditional probability $P(Y < y|D = 1, X = x)$

condQ.m

Input

- y: the value at which the conditional probability is evaluated
- xc: the continuous variable value at which the conditional probability is conditional on
- xd: the discrete variable value at which the conditional probability is conditional on
- Y: the independent variable
- Xc: the continuous elements of X
- Xd: the discrete elements of X
- h: the tunning parameter

Output

- q: estimator of $P(Y < y|D = 1, X = x)$