SUPPLEMENT TO "QUANTILE REGRESSION UNDER MISSPECIFICATION, WITH AN APPLICATION TO THE U.S. WAGE STRUCTURE": VARIABLE DEFINITIONS, DATA, AND PROGRAMS

(*Econometrica*, Vol. 74, No. 2, March 2006, 539–563)

BY JOSHUA ANGRIST, VICTOR CHERNOZHUKOV, AND IVÁN FERNÁNDEZ-VAL

THIS SUPPLEMENT provides added technical details related to the data, variable definitions, and estimation. The paper has two empirical components: estimation of quantile regression weighting schemes and robust inference on the quantile regression process for earnings equations. Both rely on Census microdata for 1980, 1990, and 2000. The original raw data are available from the Integrated Public Use Microdata Series (IPUMS) web site and our Stata extracts are available here. In addition to a description of the data and variables, this supplement includes all Stata and R (version 2.0.1) command files used to construct Figures 1 and 2, and Table I.

1. DATA

The data used in the empirical application were drawn from the 1% self-weighting 1980 and 1990 samples, and the 1% weighted 2000 sample, all from the IPUMS web site (Ruggles et al. (2003)). The sample consists of U.S.-born black and white men aged 40–49 with at least five years of education, with positive annual earnings and hours worked in the year preceding the census, and with a nonzero sampling weight. Individuals with imputed values for age, education, earnings, or weeks worked were also excluded from the sample. After this selection process, the final sample sizes were 65,023, 86,785, and 97,397 for 1980, 1990, and 2000.

The log-earnings variable is the average log weekly wage, calculated as the log of reported annual income from work divided by weeks worked in the previous year. Annual income is expressed in 1989 dollars using the Personal Consumption Expenditures Price Index, extracted from the Bureau of Economic Analysis web site *http://www.bea.gov*.

The education variable for 1980 corresponds to the highest grade of school completed, coded in Table SI. For the purposes of Figure 2 and most of the empirical work, years of schooling for 1990 and 2000 censuses were imputed from categorical schooling variables in Table SII. This is similar to the imputation used by Angrist and Krueger (1999) and many others, when working with 1990 and later schooling variables.

2. VARIABLES

The selection process described in the previous section creates the Stata data files *census80.dta*, *census90.dta*, and *census00.dta* that contain the data for the

TABLE SI

Years of Schooling	Highest Grade of School Completed
5	5th grade of elementary school
6	6th grade of elementary school
7	7th grade of elementary school
8	8th grade of elementary school
9	9th grade of high school
10	10th grade of high school
11	11th grade of high school
12	12th grade of high school
13	1st year of college
14	2nd year of college
15	3rd year of college
16	4th year of college
17	5th year of college
18	6th year of college
19	7th year of college
20	8th or more year of college

census years 1980, 1990, and 2000, respectively. These files include the following variables:

perw: individual sampling weights,
logw: average log weekly wage in 1989 dollars,
educ: years of schooling,
black: indicator variable for race that takes the value 1 for blacks,

TABLE SII

Years of Schooling	Educational Attainment
8	5th, 6th, 7th, or 8th grade
9	9th grade
10	10th grade
11	11th or 12th grade, no diploma
12	High school graduate, diploma or GED
13	Some college, but no degree
14	Completed associate degree in college, occupational program
15	Completed associate degree in college, academic program
16	Completed bachelor's degree, not attending school
17	Completed bachelor's degree, but now enrolled
18	Completed master's degree
19	Completed professional degree
20	Completed doctorate

age: age in years,
exper: potential experience, calculated as $age - educ - 6$,
exper2: square of *exper*.

All calculations involving the 2000 sample use the sampling weights *perw*.

3. DESCRIPTION OF STATA AND R COMMAND FILES

The files are divided into three groups. The first contains the files used to generate Figure 1. These include programs that nonparametrically estimate conditional quantiles, run quantile regressions, estimate the importance weights, and obtain the density weights. The second group contains the R command file that generates Figure 2. The third group contains the Stata do file that produces Table I.

For replication purposes, note that all the command files assume that the data sets *census80.dta*, *census90.dta*, and *census00.dta* are located in a folder with path *h:\quantiles\supplement\data*. The programs produce output files and auxiliary data files located in a folder *h:\quantiles\supplement\results*. These folders should be created before running the command files in the order that they are described below. Alternatively, the folder paths can be changed in the programs.

3.1. *Files Used for Figure 1*

1. *cq.do*: Obtains nonparametric estimates of the conditional quantiles of log earnings given schooling. These estimates are just the sample quantiles of log earnings for each level of schooling.

2. *qr.do*: Quantile regressions and Chamberlain's minimum distance estimates fitting conditional quantiles to schooling across cells. The outcomes of this programs are the Stata data files *census80g.dta*, which contains estimates of the conditional quantiles, and the quantile regression and Chamberlain fitted values for each level of schooling; and *census80qr.dta*, which contains the quantile regression residuals and specification errors for each individual.

3. *delta.do*: Auxiliary program for estimating the importance weights. This program obtains, for each level of schooling, the grid of values for log earnings where the density is estimated. Using the data file *census80qr.dta*, the outcome of this program is the file *census80delta.dta* that contains the grid of values for each level of schooling.

4. *importance_weights.do*: Derives the importance weights by estimating kernel densities in the grid of points obtained from *delta.do* and weighted-averaging these densities. See Section 4 for details. The results, together with the quantile regression weights (importance weights × histogram of schooling), are added to *census80g.dta*.

5. *density_weights.do*: Obtains the density weights by estimating kernel densities at the nonparametric estimates of the conditional quantiles. This program uses the data file *census80qr.dta* created by *qr.do* and adds the results to *census80g.dta*.

6. *histogram.do*: Saves individual levels of schooling in the data file *census80i.dta* to generate the histogram.

7. *figure1.R*: R command file that generates Figure 1, based on the information created by the previous Stata do files. In particular, this program uses the data files *census80g.dta*, which contains nonparametric estimates of the conditional quantiles, quantile regression fits, Chamberlain fits, quantile regression weights, importance weights, and density weights, and *census80i.dta*, which contains the data needed to construct the histogram of education. This program uses the library *foreign*, which needs to be installed in R before running the program.

3.2. *File Used for Figure 2*

The R command file *figure2.R* generates the two panels of Figure 2. This program uses the libraries *foreign* and *quantreg*, which need to be installed in R before running the program. The uniform bands were obtained by subsampling using $B = 500$ repetitions with subsample size $b = 5n^{2/5}$ and a grid of quantiles $\mathcal{T}_{K_n} = \{0.10, 0.11, \dots, 0.90\}$. The main text and Chernozhukov and Fernández-Val (2005) discuss subsampling for QR inference in greater detail.

3.3. *File Used for Table I*

The Stata do file *table1.do* makes all the calculations needed to construct Table I. Here, the covariates in the estimation of the conditional quantiles and quantile regressions include years of schooling, race, and a quadratic function of experience. All summary measures were calculated using the distribution of the covariates in each year.

4. ESTIMATING THE QR WEIGHTING FUNCTION

We calculate the importance weights using the expression from Theorem 1:

$$w_\tau(X, \beta) = \int_0^1 (1-u) \cdot f_Y\big(u \cdot X'\beta + (1-u) \cdot Q_\tau(Y|X)|X\big). \tag{1}$$

The integral was estimated with a grid of 101 points between the nonparametric estimates of the CQF ($\widehat{Q}_\tau(Y|X)$) and the QR approximation ($X'\widehat{\beta}(\tau)$) for each value of the discrete covariates, X. (Programs use the change of variable $\epsilon_\tau = Y - Q_\tau(Y|X)$ with $f_{\epsilon_\tau}(e|X) = f_Y(e + Q_\tau(Y|X)|X)$.) This gives rise to the discrete approximation formula for the importance weights:

$$\begin{aligned}(2)\qquad &\widehat{w}_\tau(x, \widehat{\beta}(\tau))\\ &= \frac{1}{101}\sum_{u=1}^{101}\left(1-\frac{u-1}{100}\right)\\ &\quad \times \widehat{f}_Y\left(\frac{u-1}{100}\cdot x'\widehat{\beta}(\tau) + \left(1-\frac{u-1}{100}\right)\cdot \widehat{Q}_\tau(Y|X=x)\middle|X=x\right).\end{aligned}$$

We used kernel density estimates of $f_Y(y|X=x)$ with a Gaussian kernel and bandwidth (h) determined by

$$\begin{aligned}(3)\qquad m &= \min\left[\sqrt{\mathrm{Var}\left[Y-\widehat{Q}_\tau(Y|X=x)|X=x\right]},\right.\\ &\qquad\left.\frac{IQR_{0.25,0.75}[Y-\widehat{Q}_\tau(Y|X=x)|X=x]}{1.349}\right],\end{aligned}$$

$$(4)\qquad h = \frac{0.9\cdot m}{n^{1/5}}.$$

This bandwidth choice is optimal in the sense that it minimizes mean integrated square error with Gaussian data and a Gaussian kernel (Silverman (1986)). The density weights were calculated similarly.

REFERENCES

ANGRIST, J., AND A. KRUEGER (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, Vol. 3, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier Science, 1277–1366.

CHERNOZHUKOV, V., AND I. FERNÁNDEZ-VAL (2005): "Subsampling Inference on Quantile Regression Processes," *Sankhyā*, 67, 253–276.

RUGGLES, S., M. SOBEK, et al. (2003): "Integrated Public Use Microdata Series," Version 3.0. Historical Census Project, University of Minesota, Minneapolis.

SILVERMAN, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.