

Quantile Regression and Structural Changes in the Italian Wage Equation

Marilena Furno*

October 2006[•]

Abstract

Quantile regressions are useful in investigating returns to education. They are more informative than OLS since they allow to look at the impact of the explanatory variables at different values of the conditional distribution. Furno (2006) defines a formal test for structural break with quantile regressions. By repeatedly implementing this test to a returns to education equation we control its stability both over time and with respect to its explanatory variables, in order to look for a deeper explanation of the break. We find that when we model returns to education in Italy by more than one equation, separating the sample by gender and region, the estimates turn out to be more stable, particularly the regression describing women wage. The gender and regional gaps enhance the changing coefficient problem.

Keywords: quantile regression; structural break; robustness; non-normality; test.

* Department of Economics, Università di Cassino.
Address for correspondence: Via Orazio 27/D, 80122, Napoli, Italy.
webpage: www.eco.unicas.it/docente/furno/
e-mail: furnoma@tin.it

[•] Preliminary draft, comments welcomed.

1. Introduction

The OLS estimates of a structural equation look at the mean of the conditional distribution, while quantile regressions (QR) allow to analyze more than one point of the conditional distribution, not only the center, i.e. the median, but also the tails, i.e. the lower and the upper quantiles. This characteristic makes QR extremely useful to evaluate returns to education. Indeed they allow to compute the impact of education on income at different points of the wage distribution. In addition, by computing the difference from one quantile to the other over time, it is possible to measure changes in income distribution and inequality.

Besides providing a more thorough and detailed analysis, the choice of QR is linked to a technical reason specific to the wage equation. The available data sets turn out to be truncated, since the lowest wages are most likely self-excluded from the interviews. Truncation at the lower tail of the distribution increases the mean of the dependent variable, which in turn causes a bias in the OLS estimates since OLS computes the line/plane passing through the conditional mean. Instead QR, which estimate the line/plane passing through the selected conditional quantile, are not affected by truncation since quantiles are more robust than the mean.

In this paper we look at a specific aspect of the returns to education equation. A test for structural break based on QR, discussed in Furno (2006), is here implemented to analyze the Italian wage structure. This test allows to verify the existence of a break and, by checking at different points of the conditional distribution, at what level of the dependent variable the break is more effective. Indeed, we find a case of break affecting only the lower or/and the upper tails of the distribution. Furthermore, we can verify which regressor has a greater influence on the break. The presence of a break is generally tested as a function of time.¹ The sample is split in two sub-periods and the test verifies if the coefficients estimated in the two sub-periods are significantly different from those estimated in the entire sample. By devising different sample splitting, specifically related to the characteristics of each explanatory variable, it is possible to verify if the break is linked, for instance, to a gender, a geographical or a generational gap. This allows to explore if and at what level of the dependent variable a regressor has an impact on the break.

¹ Time is seen as a proxy of technological changes.

The different ways in which we split the sample, in conjunction with the estimates of the equation at different quantiles, allow to extract information that could not be obtained otherwise and that may lead to improvements in the specification of the equation. Indeed they point out analogies and dissimilarities in the behavior of the coefficients which help analyzing if it is appropriate to estimate the equation in the presence of a break or if it is preferable to change the specification of the model. Indeed we find a more stable model when we consider returns to education separately by regions and genders. In particular, the equations describing working women are very close to stability while there is still a strong evidence of changing coefficients in the men sub-samples, although not as strong as at the beginning of the analysis. The implication is that gender and regional gaps widen the changing coefficients problem of the Italian wage equation.

The rest of the paper is organized as follows. Section 2 presents the test function. Section 3 estimates the Italian wage equation implementing OLS and QR. Sections 4 to 9 control the stability of the equation with respect to time and to other variables of the model. The final section is devoted to the conclusions.

2. The Quantile regression based test function

Consider the standard linear regression model $y_t = x_t \beta + \varepsilon_t$, where y_t is the dependent variable, x_t is the k -row vector of a single observation for the k explanatory variables, ε_t is the i.i.d. error term having continuous and strictly positive density $f(\cdot)$ at the median, in a sample of size n .

In the OLS framework, the test for structural break relies on the assumption of normally distributed errors, and is defined as follows:

$$C = \frac{[\tilde{u}'\tilde{u} - \hat{u}'\hat{u}] / d_1}{\hat{u}'\hat{u} / d_2} \quad (1)$$

The test compares the objective functions under the null and the alternative, and relies on the increase of the objective function and the worsening of the fit when unnecessary constraints are imposed. The numerator is the difference between the sum of squared residuals of the constrained and unconstrained models, $\tilde{u}'\tilde{u}$ and $\hat{u}'\hat{u}$ respectively, divided by the degrees of

freedom d_1 which are equal to the number of constraints. The denominator is the sum of squared residuals of the unconstrained model, adjusted by the degrees of freedom d_2 (the sample size less the number of estimated coefficients). The test function in (1) is asymptotically distributed as F_{d_1, d_2} .

Buchinsky (1994; 1995) first analyzes the U.S. labor income distribution using the QR estimator. Quantile regressions do not require distributional assumptions and allow to evaluate the same equation at different quantiles. The objective function of the quantile regression estimator at the selected quantile θ is:

$$V(b(\theta)) = \sum_{y_t > x_t b} \theta(y_t - x_t b) + \sum_{y_t < x_t b} (1 - \theta)(y_t - x_t b)$$

When $\theta=0.5$ the objective function simplifies into $V(b(0.5)) = \sum |y_t - x_t b|$, which computes the median conditional regression, or the least absolute deviation (LAD). Recently Angrist et al. (2006), analyzing the U.S. wage structure under incorrect model specification, point out an interesting rule of thumb to signal the presence of structural breaks in the quantile regression framework. They analyze weekly wages from 1980 to 2000, for U.S. born men aged 40-49, as a function of the years of schooling. They consider the QR estimates at the same quantile over different sample periods and notice that these regressions are far apart, the confidence intervals of the two estimated equations do not overlap and do not have any common region (Figure 1). They interpret this as an evidence of the existence of a structural break, but do not provide a formal statistical test.

Furno (2006) defines the test function for structural break based on quantile regressions:

$$C^1 = \frac{[\tilde{V}(b(\theta)) - \hat{V}(b(\theta))] / d_1}{\hat{V}(b(\theta)) / d_2} = \left(\frac{\tilde{V}(b(\theta))}{\hat{V}(b(\theta))} - 1 \right) \frac{d_2}{d_1} \quad (2)$$

The numerator is the difference between constrained and unconstrained objective functions of the quantile regression, adjusted by the number of constraints. The denominator is the estimated quantile regression objective function for the unconstrained model, adjusted by its degrees of freedom. The asymptotic distribution of C^1 relies on the results in Bai (1995)

on the consistency of QR estimators in the presence of structural break,² and on the results in Koenker and Bassett (1982) and Koenker and Machado (1999) on the asymptotic distribution of the likelihood ratio test in quantile regressions. C^1 is asymptotically distributed as F_{d_1, d_2} .

The advantage of this test over the standard OLS test of equation (1) is that it allows to check the existence of structural break at different quantiles, at the median as well as at the upper and lower quantiles, thus finding the impact of a structural break at different levels of the dependent variable. Indeed, it may be the case that the equation is stable at the mean of the dependent variable, but not at the low or/and at the upper quantiles. In the following sections there are cases which provide evidence in favor of a balancing effect, where the behavior in the tails to some extent cancels out at the middle of the wages conditional distribution.

Furthermore, there is another reason to prefer the QR test for structural break. Godfrey and Orme (2000) find that the OLS based test for structural break is undersized with uniform error distributions, while it is oversized with Student-t, lognormal and χ^2 distributions. Andrews (2003) shows that serial correlation causes over-rejection in the OLS-based test for structural break. These findings imply that in the OLS framework the normality assumption is crucial to define the asymptotic distribution of the test. On the contrary, normality is not required in quantile regressions and the behavior of the QR based test is not affected by the non-normality of the error distribution. The robustness of quantile regressions provides a pay off in the presence of non-normal and/or non-i.i.d. errors. Furno (2006) analyzes the behavior of the QR test for structural break in a Monte Carlo study, selecting different error distributions and various departures from the i.i.d. assumption: serial correlation, heteroskedasticity, and conditional heteroskedasticity. The simulations show that in many experiments the QR test outperforms the OLS based test, which instead tends to over-reject the true null.³

In what follows we implement the tests for structural break of equations (1) and (2) to the wage equation in Italy from 1989 to 2004. We compare the behavior of the two tests and we exploit the greater flexibility provided by the QR based test to point out the characteristics and the changes that the wage structure has undergone, not only over time but also with

² This result is quite general since consistency holds in case of known and of estimated break, for i.i.d. and for non-i.i.d. errors.

³ The over-rejection is particularly evident when there is serial correlation, with conditional heteroskedasticity, when the break is close to the end of the sample.

respect to other regressors, in order to verify if, by what extent and at which quantile they may have a role in the break. This may also suggest improvements in the model specification.

3. Quantile regression and OLS estimates of the wage equation

To estimate returns to education, we use repeated cross-sectional data from eight waves of the Survey of Household Income and Wealth (SHIW), which is a representative sample of the Italian resident population. Sampling is in two stages, first municipalities and then households. Municipalities are divided into strata defined by regions and classes of population size (less than 20,000; from 20,000 to 40,000; more than 40,000). Households are then randomly selected from registry office records. The survey is conducted every other year and covers about 24,000 individuals and 8,000 households.⁴ Because of its sample design and its collection of detailed wealth statistics, the SHIW is similar to the Survey of Consumer Finances (SCF), which is representative of the U.S. population.⁵

From SHIW we select a sample of employees born between 1936 and 1965, with age between 20 and 65, which turns out to be a sample of $n=24174$ observations.⁶ The dependent variable is the log of yearly wages, net of taxes and social security contribution, expressed in 2002 euros using the CPI deflator. The explanatory variables are: age as a proxy for workers' experience; years of education, which takes the values 0, 5, 8, 13, 18, according to the completed school degree; a dummy variable for gender and a dummy for the south. The OLS estimated equation, with t-statistics in parenthesis, is the following

$$YL = 2.033 + .00798 AGE + .0465 EDUCATION - .355 WOMEN - .140 SOUTH \quad (3)$$

(121)
(24)
(68)
(-47)
(-23)

which shows a 4.6% increase in wage for each additional year of schooling; a 0.7% increase in wage for each additional year of experience as proxied by age. Income is 35% lower for women and 14% lower for residents in the south. Thus, a woman working in the south

⁴ The survey takes place in 1989, 1991, 1993, 1995, 1998, 2000, 2002, 2004.

⁵ The English version of the questionnaire and data can be downloaded from the Bank of Italy at http://www.bancaditalia.it/statistiche/ibf/statistiche/ibf/pubblicazioni/boll_stat/en_shiw00.pdf.

⁶ The years preceding 1936 present a comparatively smaller number of observations, due to the early Italian retirement rules.

experiences as much as a 49% reduction in her labor income. A worker with one additional year of both schooling and experience gets an increase in wage of 5.4%.

However OLS delivers biased estimates of returns to education. Indeed, the sample is truncated at the bottom of the wage distribution, because individuals with low productivity participate less to the labor market surveys than individuals with high productivity. This increases the mean of the wage distribution causing bias in the estimates of the conditional mean.⁷ The bias can be avoided by estimating the conditional median, which is little or not at all affected by truncation. Furthermore, returns to education may differ across the wage distribution: they are higher for high-productivity workers, and lower for low-productivity workers. This phenomenon, which can be explained by the influence on the dependent variable of an unobservable variable such as “ability”, is better analyzed by quantile regressions.

Accordingly, we estimate the same regression at different quantiles: 10%, 25%, 50%, 75%, 90%. The results, reported in Table 1, are summarized in Figure 2. The straight lines represent the OLS coefficients of equation (3), while the kinked lines show the variation of the quantile regression coefficients as estimated at the different quantiles. The kinks show that the impact on income of each determinant changes according to the different quantiles analyzed, i.e. to the different wage levels. For instance, the coefficients of experience and education are higher at the top quantiles of the wage distribution, while the negative effect of gender and regional area decreases around the median wages.

The comparison of the kinked and the straight line for each coefficient shows that the estimated coefficients of age and education at the median are smaller than at the mean, while the negative impact of gender and of the regional parameter is less effective at the median when compared to the mean. Indeed, working in the south is associated with a reduction in wages of 14% at the mean and 8% at the median. For a woman the reduction in wage is of 35% at the mean and 24% at the median. Thus, the wage cut of a woman working in the south is 49% at the mean but 32% at the median, while the increase in wage of a worker with one additional year of experience and of education goes from 5.4% at the mean to 4.1% at the median.

In the following sections we estimate this equation in different sub-samples, to check if there is any evidence of changes in the structure of the equation.

4. Structural break

To search for the presence of structural breaks, we estimate the regressions separately in three sub-samples, selecting respectively 1989-91, 1993-98 and 2000-04 data for the OLS estimates in equation (4), (5) and (6):

$$YL = 2.24 + .0059 AGE + .0394 EDUCATION - .231 WOMEN - .0788 SOUTH \quad (4)$$

(98.95) (13.81) (42.53) (17.47) (9.95)

$$YL = 1.84 + .0103 AGE + .0547 EDUCATION - .395 WOMEN - .176 SOUTH \quad (5)$$

(62.44) (18.13) (46.28) (28.17) (17.14)

$$YL = 1.899 + .0094 AGE + .0480 EDUCATION - .322 WOMEN - .195 SOUTH \quad (6)$$

(59.21) (15.53) (36.05) (27.24) (17.17)

Table 2 presents the results for the quantile regressions, while Figure 3 compares OLS and QR estimates. The graphs show that the behavior of the equation in the first sample is sizably different from its behavior in the second period, while in the third sample the equation tends to revert toward the first sample estimates assuming an intermediate position. This behavior occurs implementing both OLS and QR. The latter estimates, besides showing the different impact of each regressor at different wage quantiles, when computed over time can tell us about changes in income distribution and inequality as measured by quantiles. This is performed by controlling if each coefficient computed at different quantiles maintains the same distance from one quantile to the other as time elapses.⁸

To get further details, a regression for each year of the sample could be implemented. However, besides simply looking at the behavior over time, one has to test if the changes in the structure of the equation are statistically significant.

Starting with the OLS analysis, in order to implement the test function in (1) the sample has to be split in two sub-periods according to the occurrence of the break. The restricted

⁷ Since non-participation is more frequent among women, this implies that returns for women are particularly unreliable. This coefficient is likely to be overestimated with OLS.

⁸ For instance, in Table 2 the difference for the education coefficient estimated at the third with respect to the second quartile is $.0357 - .0309 = 0.0048$ in 1989-91 and $.0442 - .0357 = 0.0085$ in 2000-04. This shows an increase in the inequality of the distributions of returns to education from the median to the

model is given by equation (3). The unrestricted model allows the coefficients to change over time. This implies the estimation of the model in two different sub-samples. Since we do not really know the break point, we evenly split the years of the survey obtaining $n_a=13418$ observations for the 1989-95 period, and $n_b =10756$ for 1998-2004.⁹ The OLS results in the two sub-samples are, respectively:

$$YL = 2.060 + .00807 \text{ AGE} + .0467 \text{ EDUCATION} - .324 \text{ WOMEN} - .113 \text{ SOUTH} \quad (7a)$$

(100) (20) (56) (-29) (-15)

$$YL = 1.904 + .00912 \text{ AGE} + .0495 \text{ EDUCATION} - .339 \text{ WOMEN} - .188 \text{ SOUTH} \quad (7b)$$

(68) (17) (43) (-32) (-19)

In Figure 4 these results are reported as a straight line with diamonds for n_a , and with triangles for n_b . Triangles are higher than diamonds for the experience and the education coefficients, while they are lower for the gender and the regional ones. These estimates, when compared with the results in equation (3) yield the value of $C = 83.78$ for the OLS based test, rejecting the null hypothesis of constancy of the parameters.

We now consider the same problem in terms of QR and implement the test function in equation (2). We look at the constancy of the estimated coefficients over time at the median, at the first and third quartile (results for the 10th and 90th quantiles are available on request).¹⁰ The results are reported in Table 3, while their graphical representation is in Figure 4. Dots stand for the estimates in the n_a sample, going from 1989 to 1995, and squares for those computed in n_b , the 1998-2004 period. By comparing the QR coefficients over time we see that, at all quartiles, the gender and the regional coefficients in the most recent years (the kinked lines with squares) are below the dotted line, and particularly so at the first quartile. For the experience and the education coefficients the line with squares is instead above the dotted one. This confirms the OLS findings. By comparing these results with the estimates of the model in the entire sample of Table 1, that is by evaluating C^l at each quartile, we can verify if and where the change over time in the regression coefficients is significantly

third quartile, over time. However, a more general check needs to take into account the variations over time between quantiles of each regression coefficient.

⁹ The assumed break point occurs in 1996-1997.

¹⁰ Bai (1995) proves consistency of the QR estimator in case of known and unknown break point, with i.i.d. and non-i.i.d. error distributions.

different from zero. The test statistic yields the values of $C_{.25}^1 = 60.48$, $C_{.50}^1 = 51.83$, $C_{.75}^1 = 40.87$. Although they all reject the null hypothesis, they show that at the first quartile the changes are more sizable than at other points of the wage distribution. The structural break is thus affected by a great extent by the changes occurred in the lower wages over time.

5. Sample splits by year-of-birth

We can further analyze returns to education by asking if the basic relationship does change as a function of the age difference among workers. This would imply that i) young generations attain higher returns to education due, for instance, to a greater familiarity with new technologies; or alternatively that ii) the young generations get lower returns since the job market is saturated with over-qualified young workers. To verify if a generational gap is enhancing the break and if this gap increases or decreases returns to education, we separately estimate the wage equation for older workers, in a sample of size $n_c=10,834$ referring to employees born between 1936 and 1950, and for younger workers in a sample of $n_d=13,790$ observations for people born between 1951 and 1965.¹¹ The OLS estimates, respectively in samples n_c and n_d , are

$$YL = 2.642 - .0048 AGE + .0514 EDUCATION - .309 WOMEN - .119 SOUTH \quad (8c)$$

(62) (-6) (55) (-24) (-13)

$$YL = 2.056 + .0086 AGE + .0413 EDUCATION - .361 WOMEN - .163 SOUTH \quad (8d)$$

(86) (15) (41) (-39) (-20)

With respect to equation (3), the sample split produces a change of sign for the age coefficient, as shown in the top left graph of Figure 5. Therefore, in the older workers group experience has a negative coefficient, while the impact of education is larger. The opposite occurs in equation (8d), for the younger group, where the age coefficient is high, and the education coefficient is not so high. This evidence is in favor of the first alternative, (i), where young generations get larger returns due to their greater familiarity with technology. In the two top graphs of Figure 5, the above equations are represented by the straight line with

diamonds for the older workers, and the one with triangles for the younger group. The C test is $C=80.6$ and rejects the null of constant coefficients.

The QR analysis leads to the estimates reported in Table 4 and summarized in Figure 5. The kinked lines with dots are the QR estimates for older workers while those with squares are the QR coefficients for younger employees. These results confirm the presence of a larger value of the education coefficient and a negative sign in the experience coefficient for the older workers. The latter, however, becomes very small and non-significantly different from zero at the upper quartile, so that we can conclude that the variable experience, for the older group, has a sizable impact only at the low quartile, diminishes at the median and disappear at the higher wages. The gender and the regional variable become equal in the two sub-samples at the higher incomes. The test function assumes the values $C_{.25}^1 = 49.82$, $C_{.50}^1 = 42.91$, $C_{.75}^1 = 42.68$, and once again the null of constant coefficients is rejected. In a sample splitting looking at generational gap as a component of structural change, the test C^1 assumes approximately the same value at each quartile, with no clear prevalence of one quartile over the others.

6. Sample split by education

Then we investigate the impact of the variable education on the instability of the wage structure. We split the sample into workers with at most 8 years and workers with more than 8 years of schooling. The sample sizes are respectively of $n_e=12271$ and $n_f=11903$. We keep the variable education in the model since it comprises more than two categories so that, after the sample splitting, the variable is still meaningful in both restricted and unrestricted models. The OLS estimates are:

$$YL = 2.179 + .0043 AGE + .0521 EDUCATION - .407 WOMEN - .174 SOUTH \quad (9e)$$

(71) (9) (22) (-34) (-20)

$$YL = 1.843 + .0122 AGE + .0452 EDUCATION - .307 WOMEN - .109 SOUTH \quad (9f)$$

(59) (27) (24) (-32) (-13)

¹¹ The variables age, which is a proxy for experience, and year-of-birth do not necessarily refer to the same interviewed individuals. We keep the variable age in the model, since the sample splitting

Experience, as proxied by age, is extremely relevant in the highly educated workers, and the straight line with triangles in Figure 6 is at the top of the graph, but for them one additional year of schooling has a smaller impact on income with respect to workers with a lower degree of sample n_e , represented by the line with diamonds. The test for changing coefficients is $C=48.9$.

The QR analysis provides the estimates reported in Table 5. At all quartiles experience has a larger impact in the more than 8 years of schooling group, while does not change much for the less well educated group. This is shown in the top left graph by a line with squares always above an almost flat dotted line. At the first quartile, for the less well educated workers of group n_e , education is very relevant and the negative impact of gender and region is strong, as shown by the dotted line. The larger impact of education for group n_e at the first quartile decreases at the median and more so at third quartile. Therefore, one additional year of education is most rewarding for workers with more than 8 years of schooling earning the highest wages. This can be seen in Figure 6. The graphs are characterized by the intersection between the QR lines depicting the coefficients for the two groups, dots for workers with at most 8 years of schooling and squares for the others. The intersection, which can not be detected using OLS, is in the graphs of education, gender and region. Each of these variables reverses its effect in going from low to high incomes. For instance, in the southern regions the highest incomes and highly educated workers are more penalized than the workers with at most 8 years of schooling. Furthermore, at the upper quartile highly educated women are slightly more penalized than women with less schooling. The test for structural change assumes values $C_{.25}^1 = 57.58$, $C_{.50}^1 = 17.17$, $C_{.75}^1 = 37.27$, which allow to reject the null. The very low value of the test at the median is due to the reversion phenomenon. This is actually a case where the structural change tends to cancel at the center of the conditional distribution while being effective in the tails.

7. Sample split by gender

We consider now the gender variable, in order to measure its contribution to the changing coefficient and, if any is found, at what income level it occurs. The sample is split into men, $n_g=19965$, and women, $n_h=4209$. This time the constrained model, estimated in the entire

according to the year-of-birth differs from the age variable in the equation.

sample, does not include the gender variable. Indeed, since this is a dummy variable, it becomes meaningless in the unconstrained model. The OLS estimates in the entire sample, i.e. in the constrained model, in the sample of working men and in the sample of working women,¹² which together provide the unrestricted model, are respectively the following:

$$YL = 1.954 + .0088 \text{ AGE} + .0437 \text{ EDUCATION} - .115 \text{ SOUTH} \quad (10)$$

(112) (26) (61) (-18)

$$YL = 2.055 + .0078 \text{ AGE} + .0451 \text{ EDUCATION} - .140 \text{ SOUTH} \quad (10g)$$

(118) (23) (63) (-23)

$$YL = 1.560 + .0089 \text{ AGE} + .0533 \text{ EDUCATION} - .142 \text{ SOUTH} \quad (10h)$$

(32) (9) (27) (-7)

In the above comparison the variables age and education have a different impact in the two sub-samples, much higher for women than for men, triangles versus diamonds in Figure 7. The test function in the OLS case is $C=571.2$ and the null of constant coefficients across gender is strongly rejected. The QR results reported in Table 6 are less homogeneous. At the first quartile, for the working women, the negative impact of the regional area is very large, while the variables age and education present very large values. At the median age, region and education do not differ much between men and women. At the third quartile the behavior is reversed. Experience, education, and the absolute value of the regional area become larger in the working men sample and smaller for women. In Figure 7 this results in intersecting QR lines. In this picture all the coefficients, experience, education and region, reverse their impact in going from low to high income when the sample is split with respect to the gender variable. Once again, this can not be detected with OLS. The test for changing coefficients assumes the values of $C_{.25}^1 = 309.25$, $C_{.50}^1 = 262.81$, $C_{.75}^1 = 257.17$, which strongly reject the null and which show that the widest change occurs at the first quartile. If we compare the results of equation (3) and Table (1) with the results discussed in this section, we can see that the earlier estimates, computed in the entire sample, are closer to the equation computed in n_g , the men sub-sample. This is possibly due to the comparatively greater number of observations in n_g .

¹²We recall the warning about the possible bias of the OLS estimates in the sub-sample of women.

8. Sample split by region

Next we analyze the impact of the geographical variable. The sample split is north-center, $n_i=15419$, versus south, $n_j=8755$. In this model the gender variable is reintroduced, while the regional dummy is dropped since it coincides with the sample splitting rule. The OLS estimates of the restricted model, the equation estimated using the north-center data n_i , and the equation considering the n_j sample for the south, respectively, are:

$$YL = 1.977 + .0077 AGE + .0476 EDUCATION - .339 WOMEN \quad (11)$$

(118) (24) (69) (-45)

$$YL = 2.080 + .0077 AGE + .0431 EDUCATION - .351 WOMEN \quad (11i)$$

(104) (20) (52) (-42)

$$YL = 1.833 + .0082 AGE + .0516 EDUCATION - .364 WOMEN \quad (11j)$$

(62) (14) (43) (-23)

Age and education present slightly larger coefficients in the southern regions. The computed test is $C=152.84$. The quantile regression estimates are reported in Table 7. At the lower quartile experience, education and the negative impact of gender is larger in the south. At the median there is little difference between the two sets of estimates, while at the third quartile the estimated coefficients have a larger impact for the center-north regions. Once again this results in the intersection of the QR lines of Figure 8. The computed tests are $C_{.25}^1 = 95.6$, $C_{.50}^1 = 47.5$, $C_{.75}^1 = 50.7$, the null is rejected at all quartiles but the estimated value of the test at the first quartile is much larger than elsewhere.

9. Sample split by gender and region

Table 8 reports the results of the tests implemented in each of the sample splitting so far analyzed. In the upper part of this table the OLS based test is always larger than the QR counterpart while, within the different quantile regressions, the first quartile yields larger

values of C^1 .¹³ Thus the large value of C at the mean, as computed by OLS, is mostly driven by what occurs at the first quartile, i.e. at the low wages.

The different sample splitting and tests, besides providing useful insights on the behavior of the coefficients at different points of the conditional distribution, can be used to verify if returns to education are well described by the regression in equation (3), which accounts for the gender and regional gaps by introducing dummy variables.

Table 8 shows that gender is the variable which causes the strongest rejection of the hypothesis of stability, with both OLS and QR estimators. Therefore we reconsider the analysis of the gender sample splitting in equations (10g) and (10h) of Section 7, together with the corresponding QR estimates of Table 6, and verify for each equation if the coefficients change over time. We take into account the 1989-1995 and 1998-2004 time periods with a break at 1996-1997. The corresponding C and C^1 tests are reported in Table 8 in the rows labeled “gender over time”. They show that the women equation, although rejecting the null of constant coefficients, is more stable than the working men sector. The latter has changed more at the first and second quartile than at the third one.

The second high value of C and C^1 , as shown in the top section of Table 8, is given by a structural change depending on the regional variable. Thus we check if, by separately modeling the two regions as in equations (11i) and (11j) of Section 8 and in the corresponding QR estimates of Table 7, the coefficients change over time. Once again we consider as break point the interval 1996-97. The results are in Table 8, rows labeled “region over time”. The results do not show stability in any of the two regions, and although OLS signals a greater dynamism in the south, we do not really find much of a difference between the north-center and the southern regions, at least in the first two quartiles.

Finally, we separately model returns to education with respect to gender and region. We divide the sample into four parts: women working in the south, $n_h=1096$; women working in the north-center, $n_k=3113$; men working in the south, $n_l=7659$; men working in the north-center, $n_m=12306$. In each sample we regress wage as a function of age and schooling by OLS and quantile regressions. The results are in Table 9 and in Figure 9. The latter shows how the behavior of the coefficients across quantiles is quite different in the four sub-samples. For instance both age and education are directly related to the wage quartile for men working in the north-center region, while are inversely related for women working in the south. The

¹³ Although in case of the year-of-birth sample splitting the difference across quartiles is really minor.

behavior of the coefficients in the regressions of men working in the south, instead, is comparable to the behavior of the regression coefficients of women working in the north-center. In the four sub-samples we can further check the stability of the coefficients over time. The computed values of the tests C and C^1 are reported in the last rows of Table 8 labeled “gender and region over time”. These results confirm that the regressions for women yield very small values of the test function, supporting the evidence of few changes and little dynamic over time in these two sub-groups of the sample, more so in the north-center than in the south. The coefficients in the sub-samples for the working men, instead, do change over time and are characterized by a greater dynamism. In particular C^1 assumes a quite large value for men working in the south at the upper quartile, which drives upward the OLS based C test.

10. Conclusions

Quantile regressions provide a useful tool to investigate the wage structure of the Italian economy. They have been very fruitful in showing the different impact of the explanatory variables on the different levels of the dependent variable. In particular, we found a reversion phenomenon in some coefficients which can not be made manifest with OLS. At the lower quantiles some variables have a large (small) impact on wage but this impact becomes instead small (large) at the upper quartiles. We also found a case where one coefficient becomes close to zero and non-significant at the upper quartile, while being very significant at the first quartile and at the median.

The hypothesis of constant coefficients has been verified and rejected implementing an F test for structural break based on quantile regressions. The test has been implemented at different quantiles to check if the break has the same relevance at different points of the conditional distribution. The standard way to test for a break is with respect to time, according to the idea that time is a proxy for technological changes. We have explored all the explanatory variables in order to verify if one or more of them had a key role in the break. We divide the sample according to different criteria: young versus old, highly educated versus less well educated, women versus men, working in the north-center regions versus south. The idea is to test if the wages structure is stable across the various distinctions or if returns to education are better modeled with separate equations.

By sorting out younger from older workers, we found that the experience coefficient is negative for older workers, which can be explained by their lack of familiarity with new technologies. However, this phenomenon becomes non-significant at the third quartile, i.e. at the higher wages. Separating less from highly educated workers, we found that the coefficients for the two groups reverse their impact going from the first to the last quartile. Indeed, education is very relevant at low wages for the less well educated group, while it drops at the high wages. The vice versa occurs for highly educated workers. Same phenomenon has been found when sorting out men from women: education is large at the first quartile for women while becomes small at the third quartile, and the vice versa occurs in the working men equation. The sample split by regions shows a large value of education in the south at the lower quartile, which decreases approaching the upper quartile, and the converse occurs in the north-center sub-sample. The reverse impact of the coefficients across quantiles is an example of how a structural change may affect only the tails and cancel out at the middle values of the dependent variable.

On the overall, the gender and the regional gaps are by far the largest contributors to the break in the wage equation. This means that the gender and the regional gaps have affected the structure of the Italian wage equation more than time, or education or generational gaps. Then we separately model returns splitting the sample with respect to region, gender, region and gender. When we implement once more the test for structural change within each of these sub-samples, we can see that the regressions for the working women yield very small values in both regional areas, which can be interpreted as a tendency to stability for women returns to education. The tests implemented in the working men sub-samples yield instead larger values, still signaling that coefficients do change over time.

Figure 1. Angrist et al (2006) impact of schooling coefficient on weekly wage, U.S. data

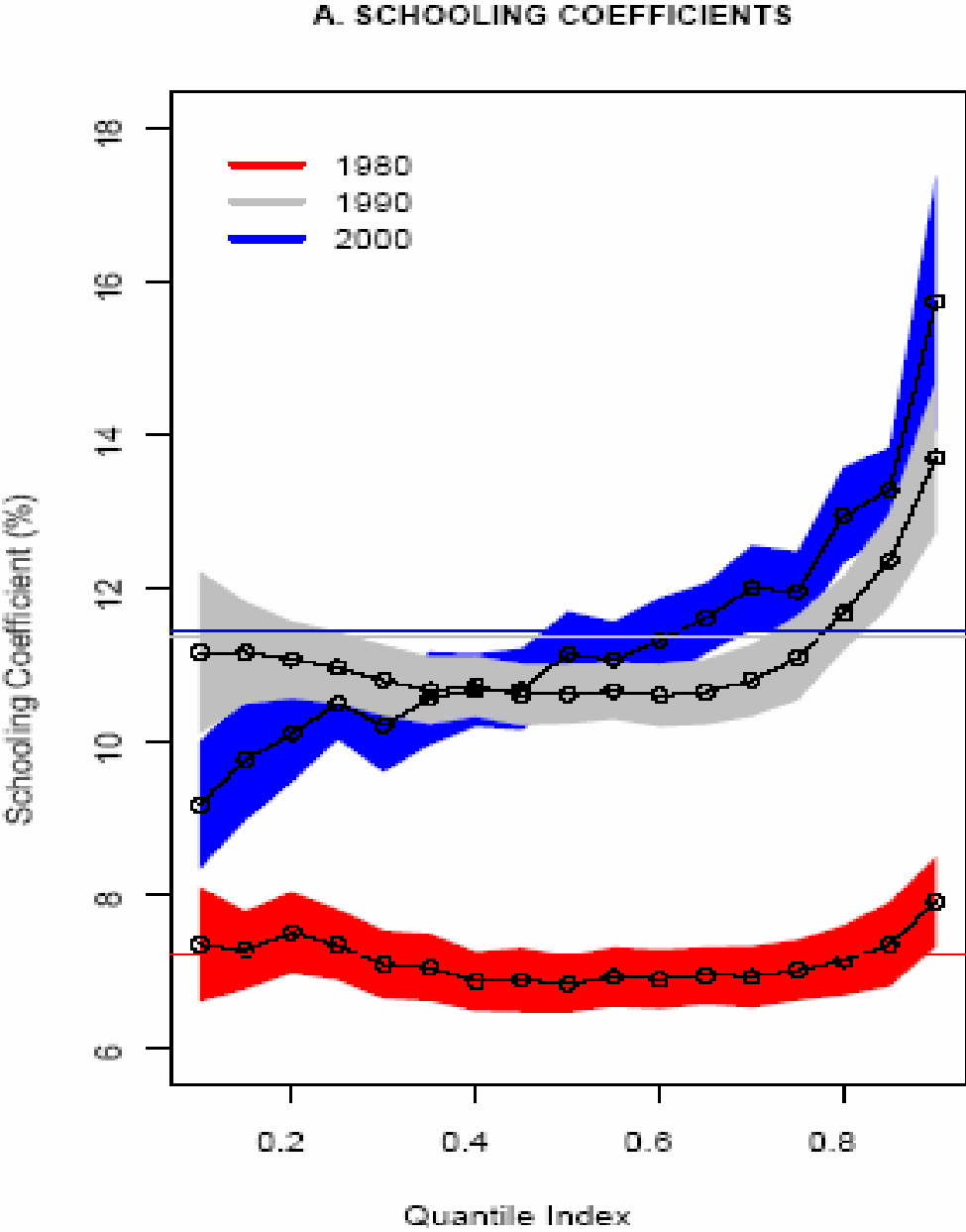


Figure 2. Estimates of the regression coefficients in the entire sample

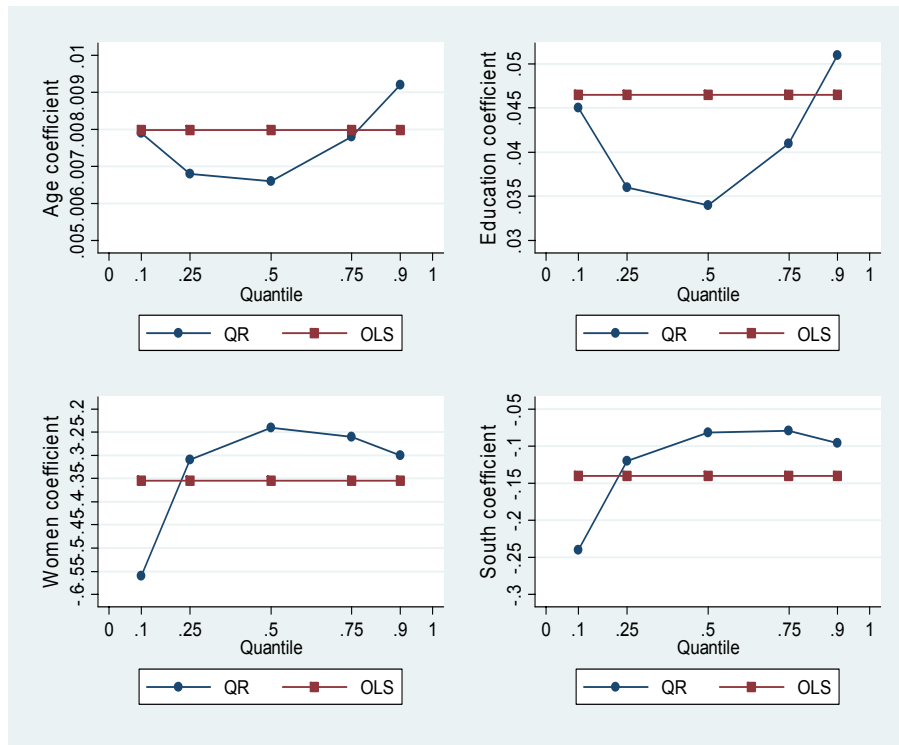


Table 1. Estimates in the entire sample, 1989-2004

	OLS	0.10 quantile	0.25 quantile	0.50 quantile	0.75 quantile	0.90 quantile
Age	0.0079 (24.93)	0.0079 (14.51)	0.0068 (20.80)	0.0066 (28.07)	0.0078 (27.43)	0.0092 (23.83)
Education	0.0465 (68.38)	0.0453 (39.03)	0.0360 (52.32)	0.0349 (69.06)	0.0419 (65.45)	0.0516 (55.52)
Women	-0.355 (47.56)	-0.567 (46.93)	-0.317 (42.66)	-0.247 (44.55)	-0.269 (39.49)	-0.300 (32.16)
South	-0.140 (23.94)	-0.247 (26.87)	-0.122 (21.08)	-0.0829 (19.02)	-0.0799 (14.97)	-0.0966 (13.28)
Constant	2.033 (121.94)	1.686 (63.66)	2.019 (122.65)	2.209 (178.56)	2.285 (148.17)	2.337 (106.93)
Observations	24174	24174	24174	24174	24174	24174

Note. Student-t statistics in parenthesis.

Figure 3. Regression coefficients over time, 3 sub-samples: 1989-91, 1993-98, 2000-04

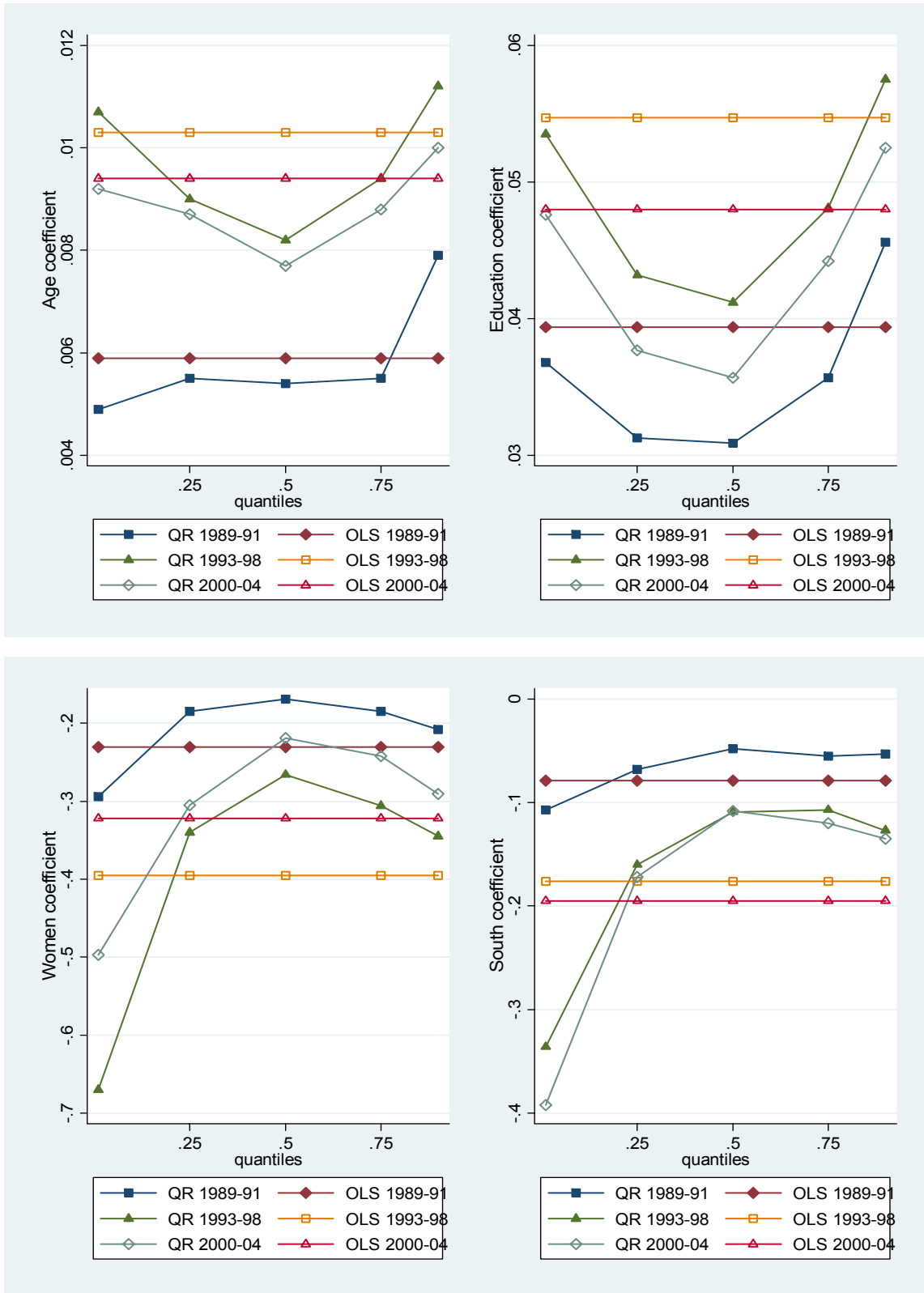


Table 2. Quantile regression estimates in three sub-samples: 1989-91, 1993-98, 2000-04

	0.10 quantile			0.25 quantile			0.50 quantile			0.75 quantile			0.90 quantile		
	1989-91	1993-98	2000-04	1989-91	1993-98	2000-04	1989-91	1993-98	2000-04	89-91	1993-98	2000-04	1989-91	1993-98	2000-04
Age	0.005 (5.62)	0.011 (7.76)	0.009 (7.88)	0.006 (11.66)	0.009 (16.92)	0.009 (14.37)	0.005 (15.90)	0.008 (18.05)	0.008 (15.88)	0.006 (13.36)	0.009 (20.79)	0.009 (15.96)	0.008 (11.85)	0.011 (16.55)	0.010 (11.46)
Education	0.037 (19.38)	0.054 (17.99)	0.048 (19.75)	0.031 (31.57)	0.043 (38.92)	0.038 (29.46)	0.031 (41.80)	0.041 (43.63)	0.036 (33.55)	0.036 (37.38)	0.048 (49.47)	0.044 (34.21)	0.046 (27.07)	0.058 (37.41)	0.053 (23.90)
Women	-0.295 (11.90)	-0.670 (20.89)	-0.498 (23.92)	-0.186 (13.37)	-0.340 (26.85)	-0.305 (27.09)	-0.170 (16.08)	-0.266 (23.73)	-0.220 (23.18)	-0.186 (14.02)	-0.307 (26.99)	-0.242 (21.89)	-0.208 (9.65)	-0.345 (19.91)	-0.292 (16.47)
South	-0.107 (7.32)	-0.336 (14.56)	-0.393 (20.42)	-0.068 (8.23)	-0.160 (17.35)	-0.172 (16.18)	-0.048 (7.66)	-0.109 (13.30)	-0.109 (11.94)	-0.056 (7.02)	-0.107 (12.87)	-0.120 (11.35)	-0.053 (4.14)	-0.128 (10.12)	-0.135 (8.00)
Constant	1.993 (45.33)	1.451 (21.91)	1.549 (27.76)	2.176 (90.59)	1.842 (69.14)	1.866 (61.90)	2.340 (129.52)	2.084 (88.39)	2.099 (81.85)	2.464 (109.16)	2.167 (88.83)	2.167 (71.14)	2.450 (64.32)	2.207 (58.37)	2.263 (45.52)
Observations	7304	8834	8036	7304	8834	8036	7304	8834	8036	7304	8834	8036	7304	8834	8036

Note. Student-t statistics in parenthesis.

Figure 4. Estimates of changes over time in two sub-samples: 1989-95, 1998-04

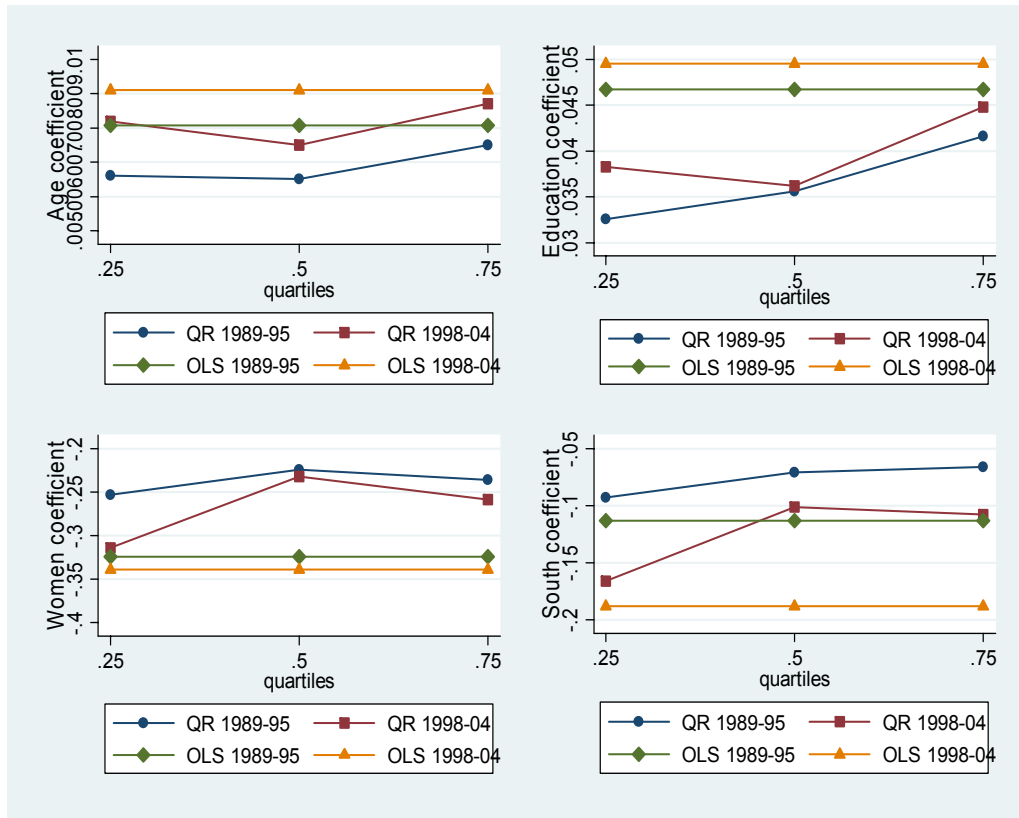


Table 3. Estimates in two sub-samples, sample split over time

	OLS		0.25		0.50		0.75	
	1989-95	1998-04	1989-95	1998-04	1989-95	1998-04	1989-95	1998-04
Age	0.00807 (20.58)	0.00912 (17.28)	0.0066 (17.70)	0.0082 (14.07)	0.0065 (20.57)	0.0075 (17.51)	0.0073 (19.63)	0.0087 (18.86)
Education	0.0467 (56.36)	0.0495 (43.46)	0.0326 (45.45)	0.0383 (30.99)	0.0356 (52.70)	0.0362 (39.08)	0.0416 (49.53)	0.0448 (42.10)
Women	-0.324 (29.72)	-0.339 (32.12)	-0.253 (24.99)	-0.314 (27.91)	-0.224 (25.26)	-0.232 (26.99)	-0.236 (22.04)	-0.259 (27.17)
South	-0.113 (15.87)	-0.188 (19.47)	-0.093 (14.09)	-0.166 (16.32)	-0.071 (12.23)	-0.101 (12.84)	-0.066 (9.44)	-0.108 (12.45)
Constant	2.060 (100.88)	1.904 (68.85)	2.056 (107.91)	1.883 (64.52)	2.238 (134.51)	2.109 (93.63)	2.327 (114.06)	2.173 (85.65)
Observations	13418	10756	13418	10756	13418	10756	13418	10756

Note. Student-t statistics in parenthesis.

Figure 5. Comparison of younger versus older workers, sample split by year of birth

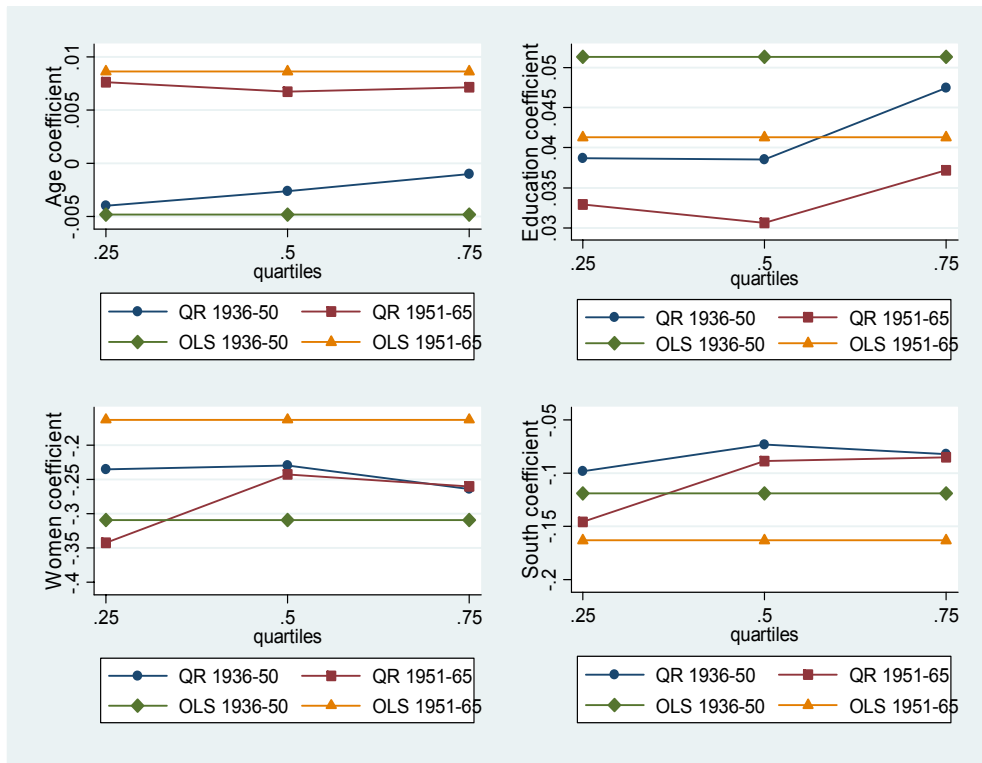


Table 4. Estimates in two sub-samples, sample split by year of birth

	OLS		0.25		0.50		0.75	
	1936-50	1951-65	1936-50	1951-65	1936-50	1951-65	1936-50	1951-65
Age	-0.0048 (6.02)	0.0086 (15.70)	-0.004 (6.07)	0.0076 (12.06)	-0.0026 (3.68)	0.0067 (16.54)	-0.001 (1.41)	0.0071 (15.04)
Education	0.0514 (55.19)	0.0413 (41.96)	0.0387 (48.35)	0.0329 (29.08)	0.0385 (46.83)	0.0306 (41.78)	0.0475 (53.27)	0.0372 (40.83)
Women	-0.309 (24.01)	-0.361 (39.41)	-0.235 (21.91)	-0.342 (32.64)	-0.230 (20.26)	-0.243 (35.65)	-0.264 (21.96)	-0.260 (31.95)
South	-0.119 (13.91)	-0.163 (20.61)	-0.098 (13.74)	-0.146 (16.49)	-0.073 (9.61)	-0.089 (15.15)	-0.082 (10.32)	-0.085 (12.20)
Constant	2.642 (62.48)	2.056 (86.50)	2.548 (72.75)	2.014 (75.38)	2.656 (71.07)	2.239 (126.78)	2.697 (67.33)	2.341 (110.17)
Observations	10384	13790	10384	13790	10384	13790	10384	13790

Note. Student-t statistics in parenthesis.

Figure 6: Comparing different levels of education, 0-8 versus 13-18 years of school

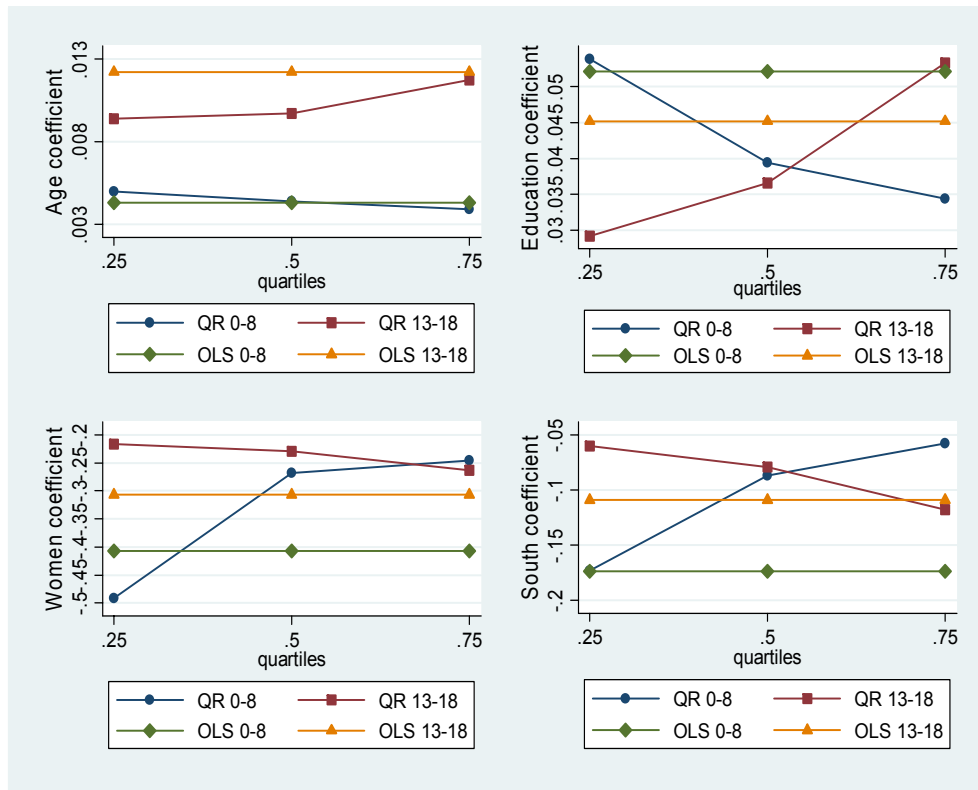


Table 5. Estimates in two sub-samples, sample split by education

	OLS		0.25		0.50		0.75	
	0-8	13-18	0-8	13-18	0-8	13-18	0-8	13-18
Age	0.0043 (9.24)	0.0122 (27.37)	0.005 (10.90)	0.0094 (20.50)	0.0044 (14.18)	0.0097 (22.17)	0.0039 (10.73)	0.0117 (23.04)
Education	0.0521 (22.14)	0.0452 (24.91)	0.0539 (24.28)	0.0292 (16.18)	0.0394 (25.26)	0.0366 (20.56)	0.0344 (17.57)	0.0533 (25.72)
South	-0.174 (20.82)	-0.109 (13.30)	-0.173 (22.00)	-0.0601 (7.39)	-0.0873 (15.71)	-0.0799 (9.91)	-0.058 (8.55)	-0.118 (12.56)
Women	-0.407 (34.53)	-0.307 (32.40)	-0.492 (43.86)	-0.217 (23.22)	-0.268 (34.34)	-0.230 (24.80)	-0.246 (25.82)	-0.263 (24.39)
Constant	2.179 (71.29)	1.843 (59.76)	2.001 (69.12)	1.962 (63.75)	2.284 (112.66)	2.045 (67.50)	2.497 (100.01)	1.971 (55.55)
Observations	12271	11903	12271	11903	12271	11903	12271	11903

Note. Student-t statistics in parenthesis.

Figure 7. Comparing returns of men versus women, sample split by gender

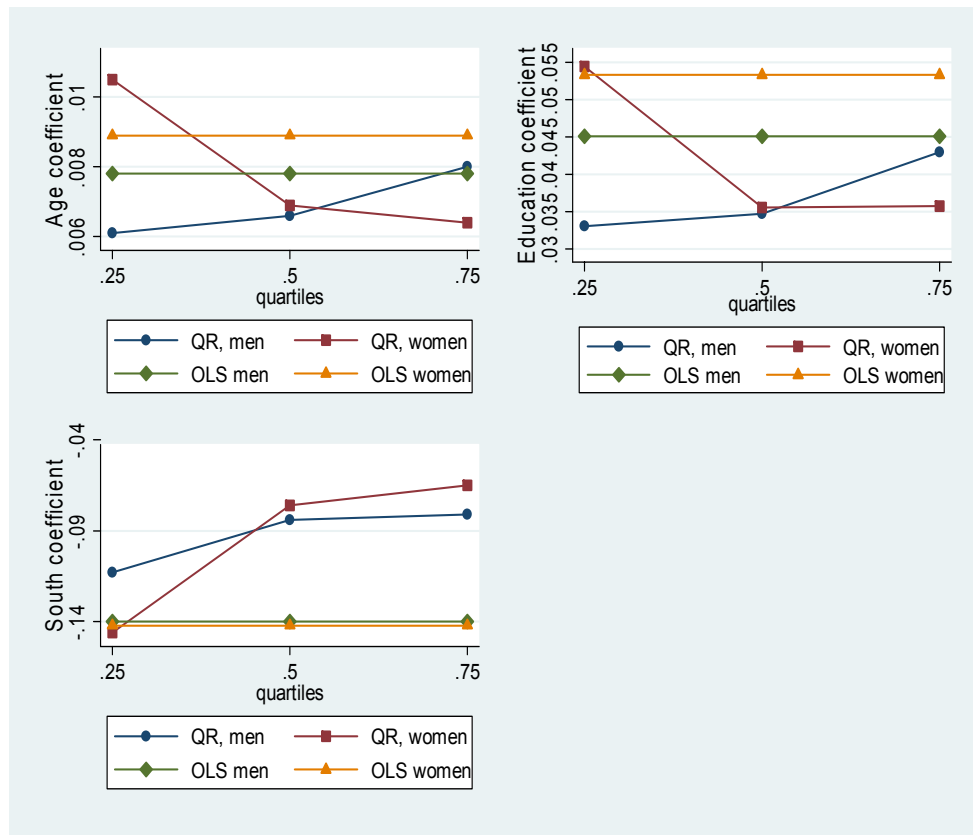


Table 6. Estimates in two sub-samples, sample split by gender

	0.25			0.50			0.75		
	full sample	women	men	full sample	women	men	full sample	women	men
Age	0.007 (23.29)	0.011 (8.31)	0.006 (18.92)	0.007 (23.88)	0.007 (10.96)	0.007 (25.30)	0.008 (28.04)	0.006 (10.01)	0.008 (22.86)
Education	0.032 (47.89)	0.054 (18.83)	0.033 (49.14)	0.032 (52.96)	0.036 (25.88)	0.035 (62.78)	0.040 (62.00)	0.036 (25.25)	0.043 (54.04)
South	-0.093 (16.47)	-0.146 (5.66)	-0.114 (20.40)	-0.061 (11.75)	-0.077 (5.96)	-0.084 (17.88)	-0.068 (12.65)	-0.065 (4.88)	-0.082 (12.64)
Constant	1.968 (122.43)	1.312 (19.91)	2.077 (127.64)	2.188 (148.68)	1.942 (58.66)	2.214 (163.63)	2.256 (144.63)	2.134 (62.01)	2.265 (118.47)
Observations	24174	4209	19965	24174	4209	19965	24174	4209	19965

Note. Student-t statistics in parenthesis.

Figure 8. Comparing returns of north-center versus south, sample split by regions

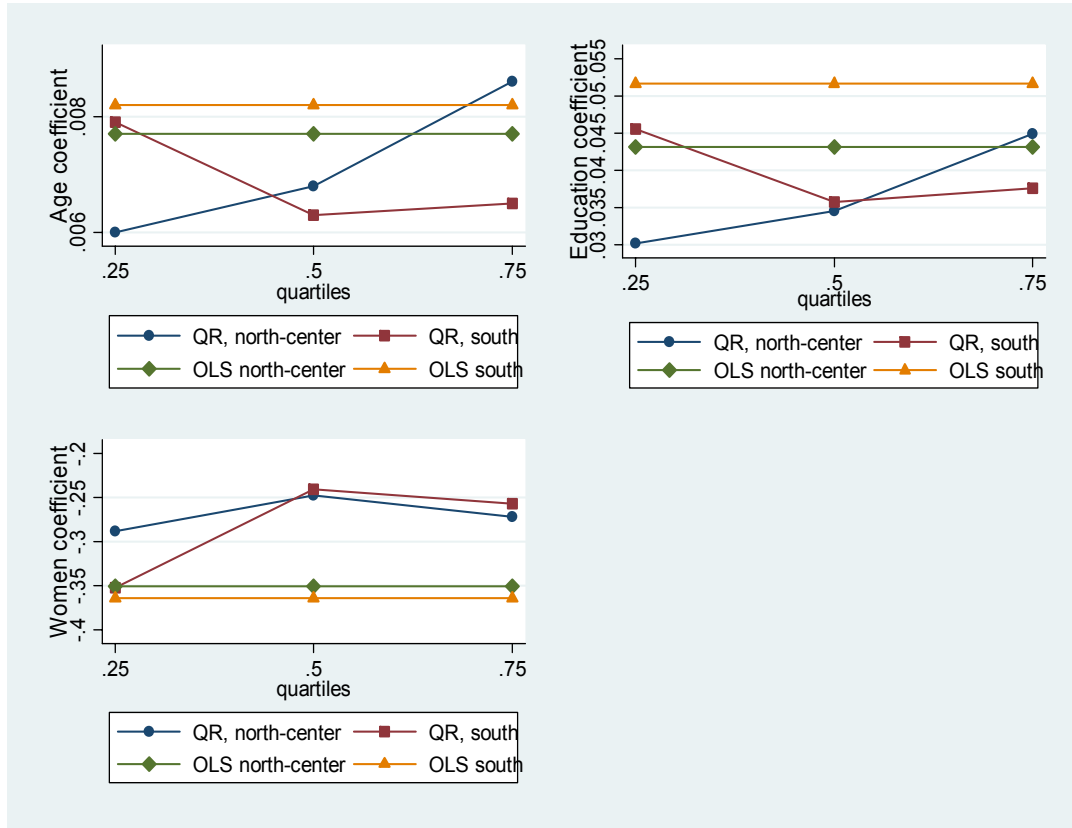


Table 7. Estimates in two sub-samples, sample split by region

	0.25			0.50			0.75		
	full sample	south	north-center	full sample	south	north-center	full sample	south	north-center
Age	0.007 (20.39)	0.008 (12.71)	0.006 (16.25)	0.006 (23.56)	0.006 (14.68)	0.007 (24.16)	0.007 (23.69)	0.007 (14.14)	0.009 (24.39)
Education	0.035 (51.88)	0.046 (34.54)	0.030 (38.98)	0.035 (61.05)	0.036 (40.30)	0.035 (56.17)	0.042 (60.28)	0.038 (37.57)	0.045 (54.51)
Women	-0.296 (40.86)	-0.352 (21.43)	-0.288 (37.04)	-0.239 (38.01)	-0.242 (20.72)	-0.249 (40.53)	-0.261 (35.12)	-0.258 (19.98)	-0.273 (34.61)
Constant	1.994 (123.30)	1.745 (56.25)	2.110 (113.26)	2.190 (157.08)	2.132 (98.53)	2.207 (148.48)	2.274 (136.06)	2.299 (95.49)	2.222 (112.99)
Observations	24174	8755	15419	24174	8755	15419	24174	8755	15419

Note. Student-t statistics in parenthesis.

Table 8. Values of the tests for structural change

	OLS	Quantile regressions		
		.25	.50	.75
Sample splitting				
time	83.78	60.48	51.83	40.87
year of birth	80.6	49.82	42.91	42.68
education	48.9	57.58	17.17	37.27
gender	571.2	309.2	262.8	257.1
region	<u>152.84</u>	<u>95.6</u>	47.5	<u>50.7</u>
gender over time	16.96	15.89	11.73	11.95
	89.58	59.31	54.37	42.23
region over time	63.02	42.29	31.28	29.42
	36.47	46.36	34.31	20.88
gender and region over time	11.54	8.81	8.53	10.17
	8.68	10.80	6.70	4.11
	69.23	42.96	32.07	96.32
	42.65	36.01	39.67	25.06

Note. In bold are the highest values, underlined are the second highest values of the test for each column.

Figure 9. Behavior of the coefficients, sample split by gender and region

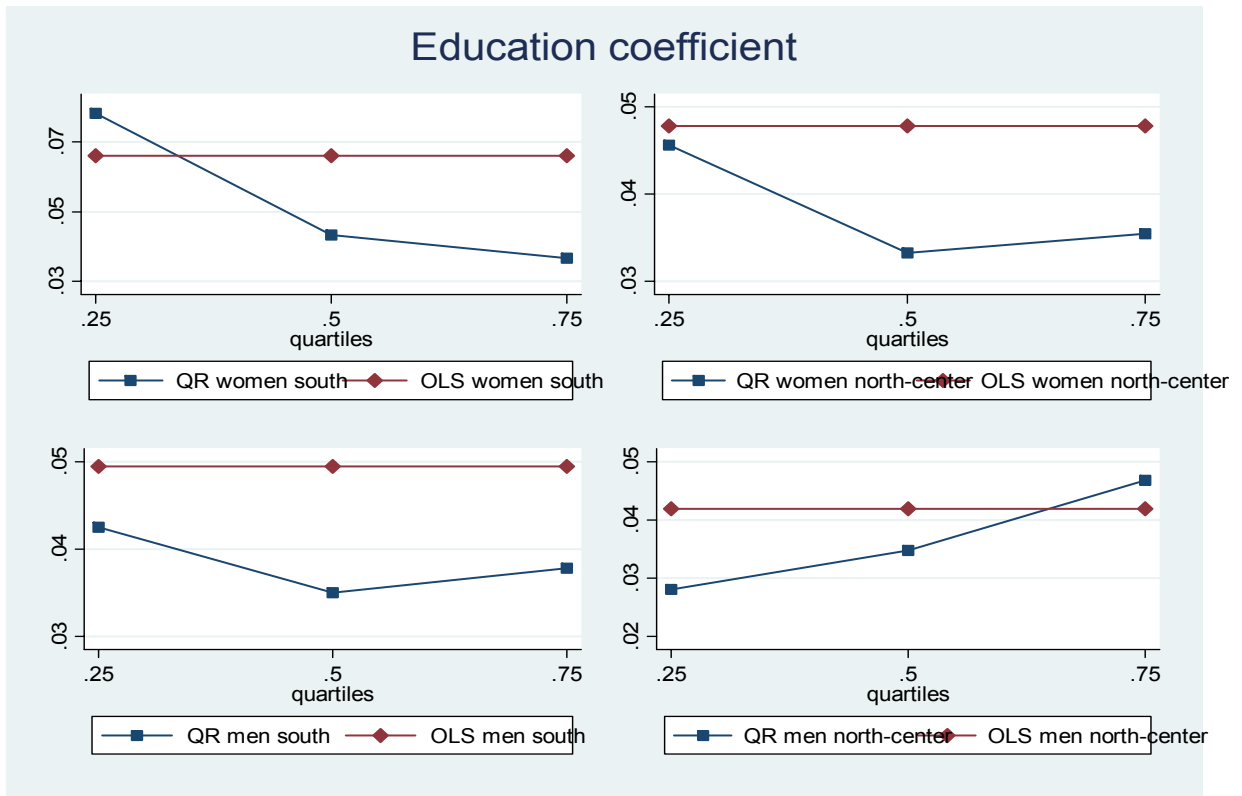
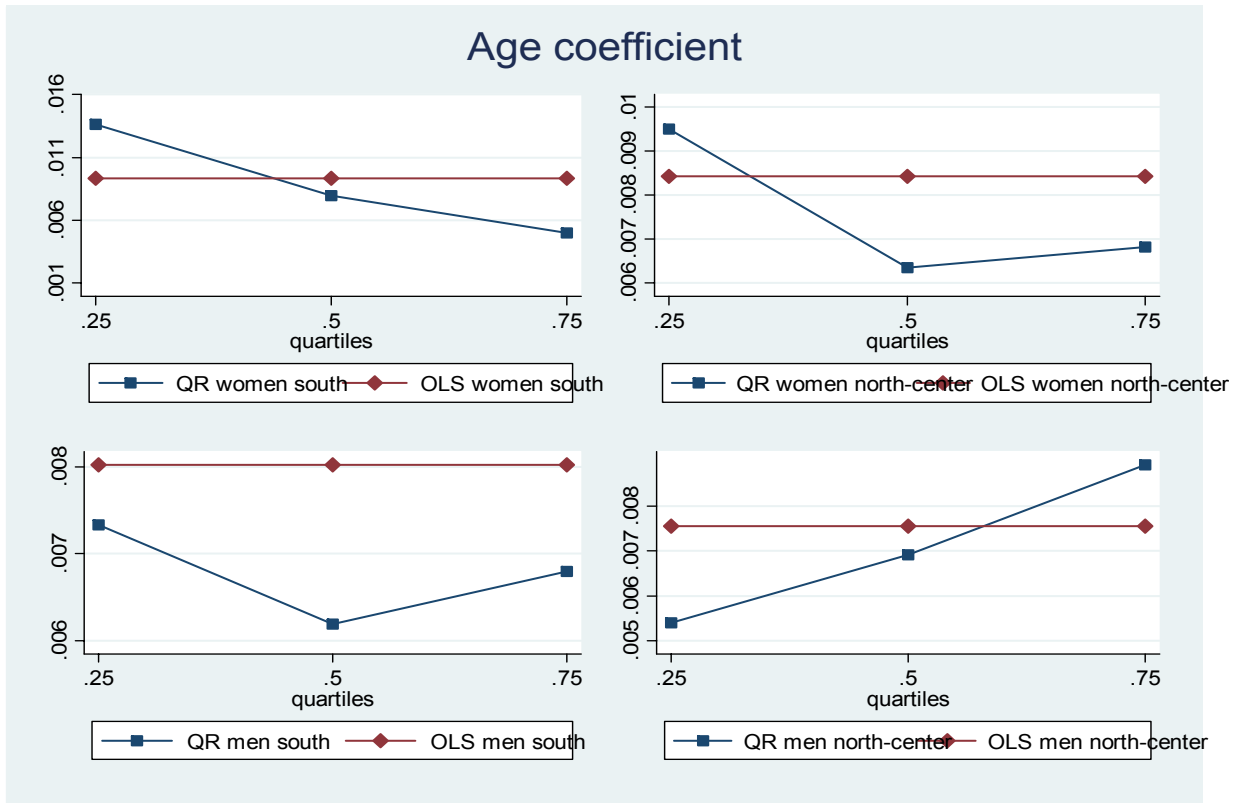


Table 9. Estimates of age and education, sample split by gender and region

	0.25						0.50						0.75						OLS					
	women			men			women			men			women			men			women			men		
	south	north-center	north-center	south	north-center	north-center	south	north-center	north-center	south	north-center	north-center	south	north-center	north-center	south	north-center	north-center	south	north-center	north-center	south	north-center	north-center
Age	0.014 (6.06)	0.010 (7.48)	0.005 (15.61)	0.007 (12.12)	0.006 (7.86)	0.006 (12.84)	0.005 (3.87)	0.007 (8.62)	0.007 (21.02)	0.006 (12.84)	0.006 (7.86)	0.005 (3.87)	0.007 (8.62)	0.007 (22.05)	0.009 (4.67)	0.009 (8.35)	0.009 (22.05)	0.007 (13.35)	0.007 (13.35)	0.007 (22.05)	0.009 (4.67)	0.009 (8.35)	0.008 (13.40)	0.008 (19.01)
Education	0.078 (15.59)	0.046 (15.80)	0.028 (39.48)	0.043 (33.78)	0.033 (18.84)	0.035 (35.58)	0.037 (13.53)	0.036 (20.06)	0.035 (48.79)	0.035 (35.58)	0.033 (18.84)	0.037 (13.53)	0.036 (20.06)	0.038 (34.25)	0.047 (15.81)	0.048 (21.59)	0.047 (49.68)	0.038 (34.25)	0.038 (34.25)	0.047 (49.68)	0.066 (15.81)	0.050 (40.49)	0.050 (48.67)	0.042 (48.67)
Constant	0.702 (6.05)	1.476 (21.97)	2.159 (125.85)	1.804 (59.92)	1.991 (46.58)	2.145 (89.23)	2.126 (31.69)	2.122 (49.39)	2.200 (128.00)	2.145 (89.23)	1.991 (46.58)	2.126 (31.69)	2.122 (49.39)	2.288 (86.32)	1.262 (12.45)	1.646 (30.78)	2.193 (97.84)	2.288 (86.32)	2.288 (86.32)	2.193 (97.84)	1.262 (12.45)	1.646 (30.78)	1.862 (62.40)	2.100 (101.23)
Observatio	1096	3113	12306	7659	3113	7659	1096	3113	12306	7659	3113	1096	3113	12306	1096	3113	12306	7659	7659	12306	1096	3113	7659	12306

References

- Andrews, D., 2003, "End-of-sample instability tests", *Econometrica*, vol. 71, 1661-1694.
- Angrist, J., Chernozhukov, V., Fernandez-Val, I. (2006), "Quantile regression under misspecification, with an application to the U.S. wage structure", *Econometrica*, vol. 74, 539-563.
- Bai, J. (1995), "Least absolute deviation estimation of a shift", *Econometric Theory*, vol. 11, 403-436.
- Buchinsky, M. (1994), "Changes in the US wage structure 1963-1987: Application of quantile regression", *Econometrica*, vol. 62, 405-458.
- Buchinsky, M. (1995), "Quantile regression, Box-Cox transformation model, and the US wage structure, 1963-1987", *Journal of Econometrics*, vol. 65, 109-154.
- Godfrey, L., Orme, C. (2000), "Controlling the significance levels of prediction error tests for linear regression models", *Econometrics Journal*, vol. 3, 66-83.
- Furno, M. (2006), "Misspecification and parameter instability in quantile regression", mimeo, University of Cassino.
- Koenker, R., Bassett, G. (1982), "Tests of linear hypotheses and l_1 -estimation", *Econometrica*, vol. 50, 1577-1583.
- Koenker, R., Machado, J. (1999), "Goodness of fit and related inference processes for quantile regression", *Journal of the American Statistical Association*, vol. 94, 1296-1310.