Selection effects and the distribution of the gender pay gap - a novel approach

Christoph Breunig [1] Peter Haan [2, 3] Boryana Ilieva [3,4]

[1]Bonn University

[2]FU Berlin

[3]DIW Berlin

[4]HU Berlin

Wage distributions, selection patterns, and gender pay gaps

- In today's labor markets, women continue to earn less than men. (Blau & Kahn 2017).
  - In Germany, overall raw gender wage gap is about 20%. The raw gender wage gap in full time work is about 14%.
- The gender gap is not only present at the mean but over the full wage distribution.
- Challenges to measure the gender pay gap correctly: gender specific selection pattern into employment might bias gender gaps.
  - Selection effects might be different for men and women
  - Direction of selection bias for gender gap not clear

Methods to account for selection

- Standard approach: Selection correction based on an instrument as proposed by Heckman (1974, 1979):
  - i) Instrument needs to be a driver of selection, ii) and strictly exogenous to the earnings process.
  - Framework extended to the entire wage distribution (Arellano & Bonhomme 2017).
  - Exclusion restriction sometimes hard to justify in practice - in particular for men.
- Alternative approach: Imputation only based on observables, e.g. Olivetti and Petrongolo (2008), Blau et al. (2021)

This Paper I

- Building on Breunig et al. (2018), D'Haultfoeuille (2011), we provide a novel set of assumptions for non-parametric identification of wage distributions and moments given selectively observed data.
- Instead of estimating the selection process using an exclusion restriction, we estimate the latent wage process as a weighted function of the non-random observations of wages.
- Identification of the inverse probability weights relies on an instrument.
    - Instrument needs to shift the outcome variable.
    - Instrument is unrelated to selection conditional on the latent outcome process.

This Paper II

- We use the estimation framework to analyse the distribution of gender wage gaps in full-time employment in Germany.
- Instrument based on the initial earnings information from administrative data of the German Institute for Employment Research.
  - Initial earnings are correlated with current latent wage process
  - 1) Selection into employment depends on current reservation and market wages. 2) Conditional on the this latent wage process initial earnings are not informative for selection process
- We estimate selection free full-time earnings distributions by gender and gender pay gap distributions in full-time for year 2017.
- We analyze the conditional distribution of gender gaps by experience and education.

Literature

- Gender gaps
  - Blau & Kahn (2006, 2017), Blundell et al. (2007), Mulligan & Rubinstein (2008)
- Selection correction
  - Ahn & Powell (1993), Das et al. (2003), Heckman (1974, 1979)
  - Arellano & Bonhomme (2017)
- Imputation based on observables
  - Olivetti and Petrongolo (2008), Blau et al. (2021)

1. Econometric model
   - Identification assumptions
   - Estimator intuition
2. Data
   - Sample
   - Instrument
3. Cross-section estimation results
   - Earnings distributions
   - Distribution of the gender pay gap
4. Conclusion

Econometric model

Latent process framework applied to labor market participation and earnings

$$
\begin{aligned}
Y^* &= h(X, W, U) \\
D &= p(Y^*, X) + V \\
Y &= Y^* \qquad \text{if} \qquad D = 1,
\end{aligned}
$$

where unobserved variables are denoted by a star

- Y^* - potential labor earnings
- X - covariates
- W - instrument
- D - participation indicator
- U and V - error terms
- $p(y, x) = \mathbb{P}(D = 1 | Y^* = y, X = x)$ - selection probability.

Intuition for identification

$$
\begin{aligned}
E[Y^*|W] &= E[Y^* \frac{P(D=1|Y^*, W)}{P(D=1|Y^*, W)}|W] \\
&= E[Y^* \frac{D}{P(D=1|Y^* W)}|W] \\
&= E[Y^* \frac{D}{P(D=1|Y^*)}|W] \\
&= E[Yg(y)|W]
\end{aligned}
$$

where $g(y) = 1/\mathbb{P}(D=1|Y^*=y)$ and $Y = Y^* D$

Estimation

Estimation: Two-step inverse probability weighting procedure.

- Step 1 - Weight: Obtain some instrumental variables estimator for the inverse selection probability function $g(y) = 1/\mathbb{P}(D = 1|Y^* = y)$
  - g(y)=1/P(D=1|Y*=y) is unknown and cannot be estimated directly since we do not observe Y*
  - But we can estimate the function g using the exclusion restriction
- Step 2 - Weighted estimation: Construct an estimator for e.g. mean, median or distribution using the weight which is equal to the inverse selection probability (Step 1).

Data

- Administrative data from the German Social Security (SIAB) records.
- Regional File of the Sample of Integrated Labour market Biographies (SIAB-R), a factually anonymous 2 % random sample of the population of the Integrated Employment Biographies (IEB), IAB.
- We observe daily earnings, age, education, and detailed records of each (un)employment episode.
- Sample
  - Cross-section year 2017
  - Ages 25-50
  - German by nationality throughout recorded labor history
  - Wage distribution trimmed at 3% from below and at 15% from above
  - at least information about three years

Sample statistics: employment

	Women			Men		
	Mean/Share	Min	Max	Mean/Share	Min	Max
Age	37.34	25	50	36.06	25	50
Education						
No Vocational Training (%)	6.86			8.63		
Vocational Training (%)	77.85			80.40		
University (%)	13.88			9.45		
Employment Status						
Non-Employed (%)	7.37			9.57		
Part-Time (%)	51.33			13.32		
Full-Time (%)	41.30			77.11		
Experience						
Non-Employed	1.85	0	30	2.20	0	32
Part-Time	5.96	0	33	1.95	0	27
Full-Time	8.09	0	33	10.48	0	33
Individuals	94974			86484		

Note: Mean values and percentage shares.

Source: SIAB Estimation sample

Sample statistics: wages

Table 1: Wages and raw gender gaps

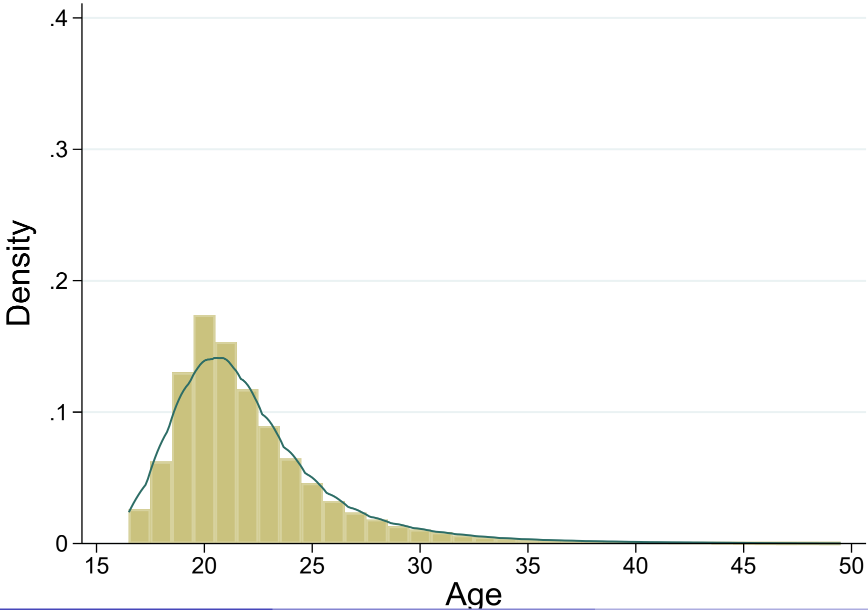
	Min	Q25	Median	Mean	Q75	Max
2017 Wages All	6.86	75.49	98.04	101.99	123.53	195.10
Wages Female	6.86	63.73	85.29	87.64	103.92	127.45
Wages Male	6.86	81.37	105.88	109.88	135.29	195.10
Raw wage gap	0.00	17.64	20.59	26.79	31.37	67.65
Wage gap in %	0.00	21.68	19.45	24.38114	23.19	34.67453

Instrument

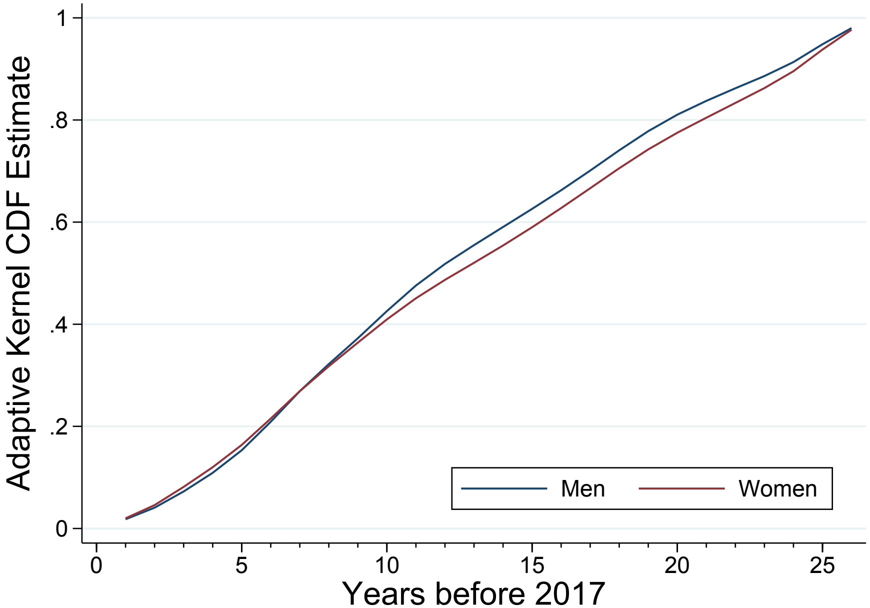
Instrument: Earnings history

- Potential earnings in current period is correlated with earnings history, e.g individual effects.
- According to the labor supply model, the earnings history has no direct effect on the selection process in the current period, **conditional** on the current wage:
  - Individuals select into employment by comparing their current reservation wage with the current period market offer.
  - Earnings from far back affect selection only through effects on current wage.
- Potential threads to identification from earnings history
  - Unemployment insurance payments
  - Persistent economic shocks
- We use earnings history from as far in the past as possible.

Instrument



Instrument



Adaptive Kernel CDF Estimate
0
.2
.4
.6
.8
1
0
5
10
15
20
25
Years before 2017
Men
Women

Instrument relevance

[OLD!!!]

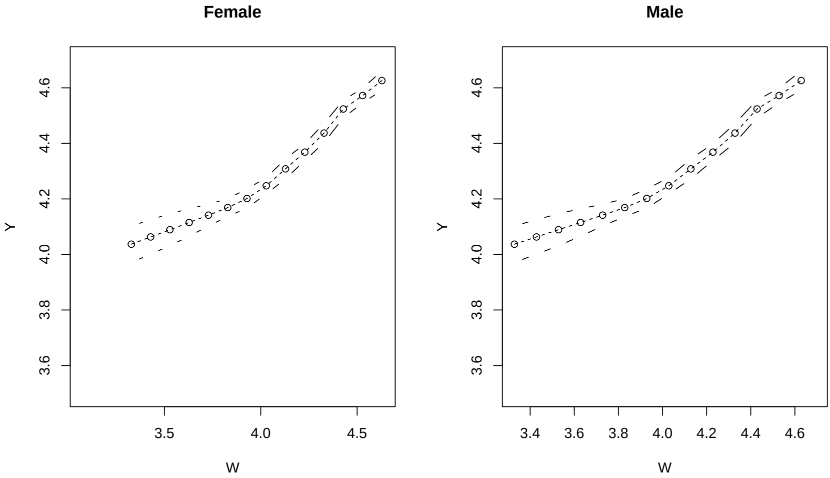


Figure 3: Association of instrument and observed outcome

Results

Estimation results: Distributions

figures/raw_vs_corrected_cdfs_male_and_female_v5_c2.pdf

Estimation results: Gender pay gaps

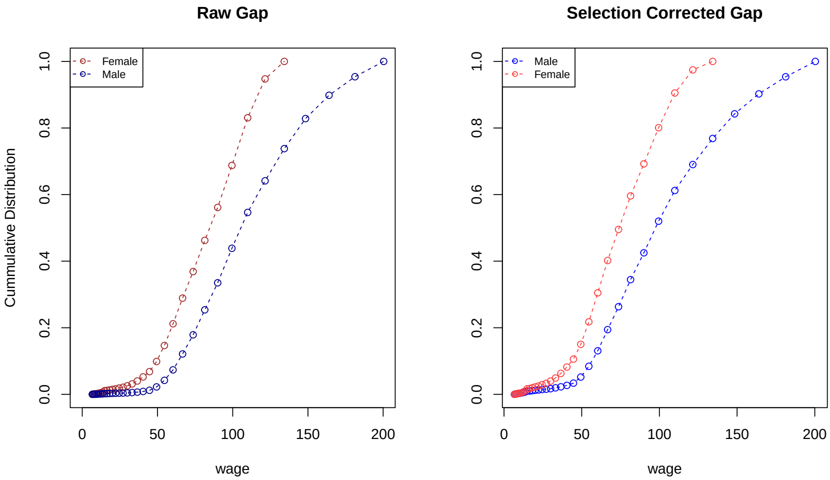


Figure 5: Gender pay gaps

Estimation results: Distributions, High Exp

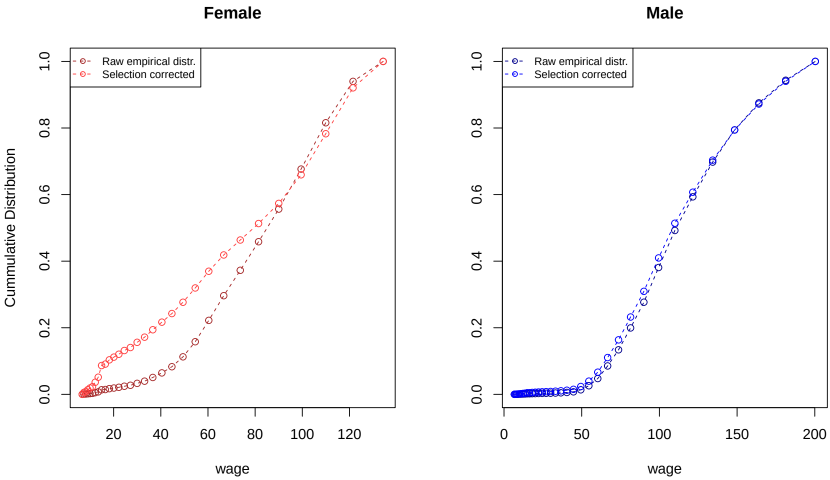


Figure 6: Selection corrected earnings distributions by gender

Estimation results: Distributions, Middle Exp

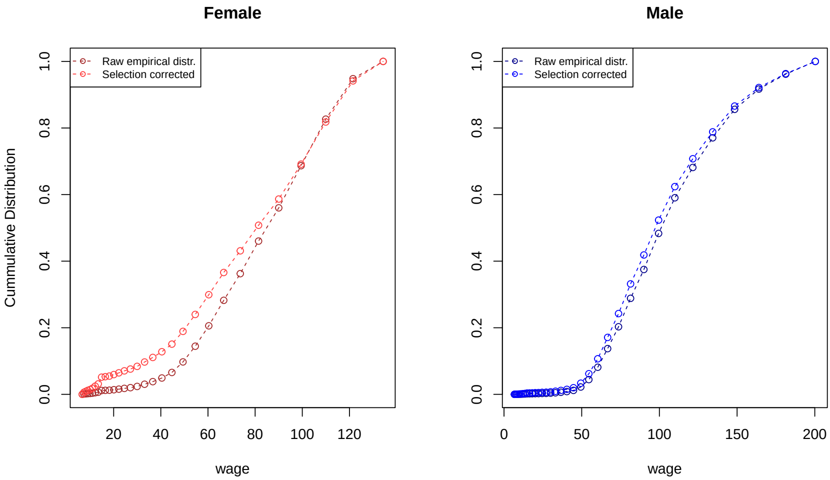


Figure 7: Selection corrected earnings distributions by gender

Estimation results: Distributions, Low Exp

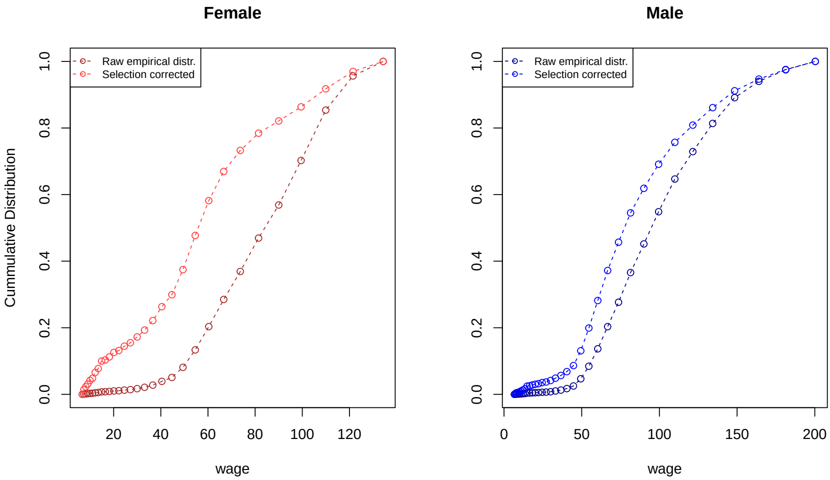


Figure 8: Selection corrected earnings distributions by gender

Estimation results: Distributions, High Educ

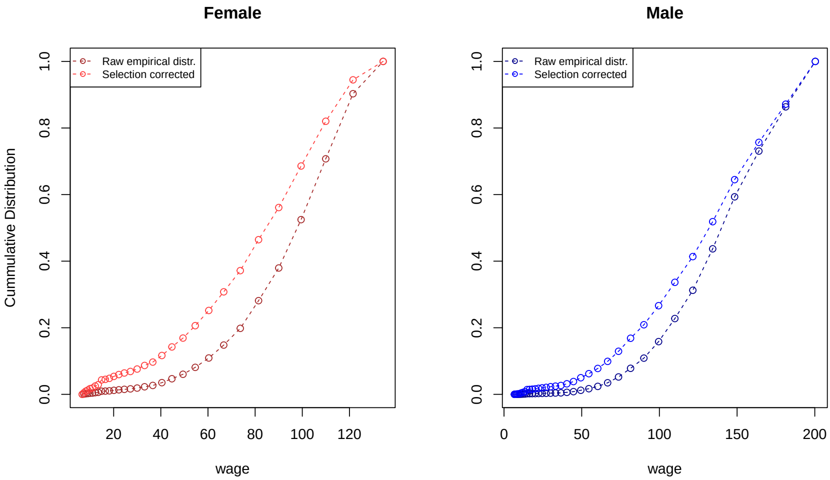


Figure 9: Selection corrected earnings distributions by gender

Estimation results: Distributions, Middle Educ

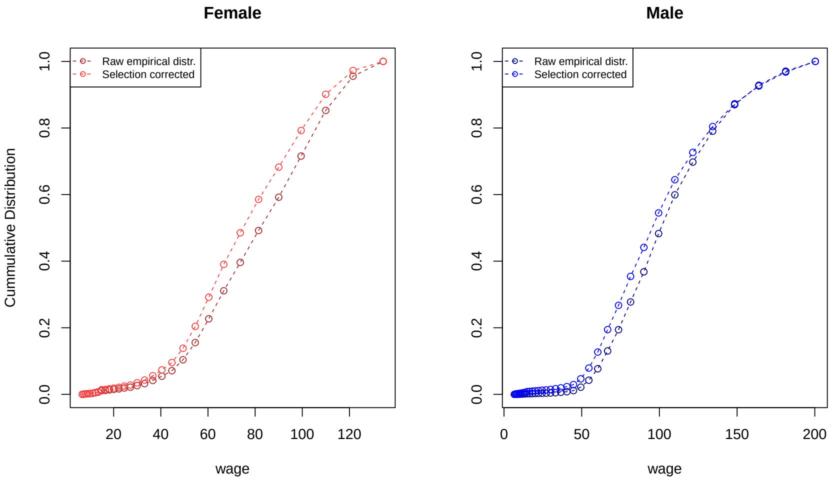


Figure 10: Selection corrected earnings distributions by gender

Estimation results: Distributions, Low Educ

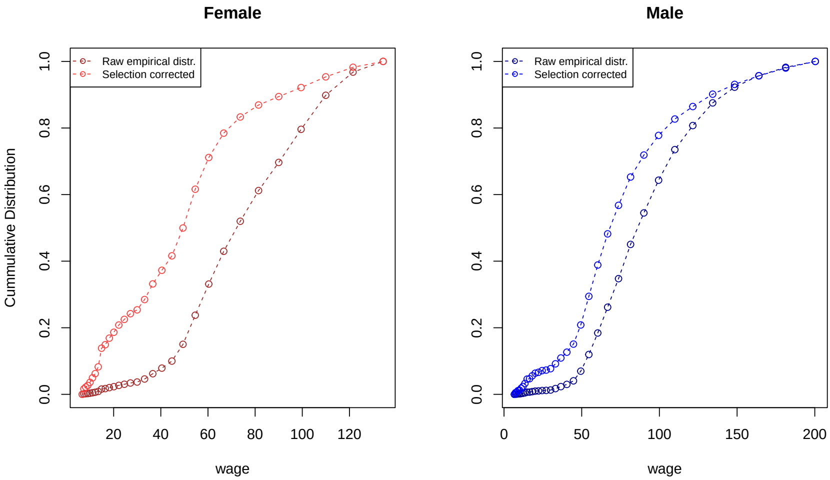


Figure 11: Selection corrected earnings distributions by gender

Conclusion

- Extend the econometric strategy of Breunig et al. (2018), D'Haultfoeuille (2011) and provide a framework for identification and estimation of a quantile regression function on selectively observed data.
- We use the proposed methodolody to correct for selection into full-time employment and estimate labor earnings distributions for both genders as well as the gender gap along the distribution.
- SIAB data on daily wages and earliest available wage as an instrument.
- Agenda:
  - Evaluate the gender pay gap simultaneously controlling for key characteristics such as education, experience, sector of employment, etc., and for selection - quantile regressions (RIF).

Thank you!

References I

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

Arellano, M. & Bonhomme, S. (2017), 'Quantile selection models with an application to understanding changes in wage inequality', *Econometrica* **85**(1), 1–28.

Blau, F. D. & Kahn, L. M. (2006), 'The us gender pay gap in the 1990s: Slowing convergence', *Ilr Review* **60**(1), 45–66.

Blau, F. D. & Kahn, L. M. (2017), 'The gender wage gap: Extent, trends, and explanations', *Journal of economic literature* **55**(3), 789–865.

Blundell, R., Gosling, A., Ichimura, H. & Meghir, C. (2007), 'Changes in the distribution of male and female wages accounting for employment composition using bounds', *Econometrica* **75**(2), 323–363.

References II

Breunig, C., Mammen, E. & Simoni, A. (2018), 'Nonparametric estimation in case of endogenous selection', *Journal of Econometrics* **202**(2), 268–285.

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

D'Haultfoeuille, X. (2011), 'On the completeness condition in nonparametric instrumental problems', *Econometric Theory* **27**(03), 460–471.

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica: Journal of the Econometric Society* pp. 679–694.

Heckman, J. J. (1979), 'Sample selection bias as a specification error', *Econometrica: Journal of the econometric society* pp. 153–161.

Mulligan, C. B. & Rubinstein, Y. (2008), 'Selection, investment, and women's relative wages over time', *The Quarterly Journal of Economics* **123**(3), 1061–1110.

Non-parametric identification proof

Theorem

The quantile regression function $h(\cdot, q)$ is identified for each $q \in (0,1)$.

$$\mathbb{P}(Y^* \leq h(X,W,q)|X = x, W = w) = q$$

for some $q \in (0,1)$. The iterated law of expectations and the exclusion restriction imposed on W, i.e., $\mathbb{P}(D = 1|Y^*, X, W) = p(Y^*, X)$ implies

$$\begin{aligned} q &= \mathbb{E}[\mathbb{1}\{Y^* \leq h(X,W,q)\}|X = x, W = w] \\ &= \mathbb{E}\Big[\mathbb{1}\{Y^* \leq h(X,W,q)\}\frac{p(Y^*,X)}{p(Y^*,X)}\Big|X = x, W = w\Big] \\ &= \mathbb{E}\Big[\mathbb{1}\{Y \leq h(X,W,q)\}\frac{D}{p(Y,X)}\Big|X = x, W = w\Big] \end{aligned}$$

where the right hand side only depends on observable variables. Now due to strict monotonicty in $h(\cdot, q)$ identification of the quantile regression function follows.