Recentered Influence Function (RIF) Regression and Decomposition

Andrew Taeho Kim

Population Studies Center
University of Pennsylvania

Quantitative Methodology Working Group
April 12, 2024

Penn Population Studies Center

Budig and Hodges (2010) *vs.* Killewald and Bearak (2014)

- Does motherhood have varying effects on earnings across the full distribution of earners?
  - Budig and Hodges (2010) argue wage penalty for motherhood is proportionately largest for the lowest-paid workers
  - Killewald and Bearak (2014) argue one cannot infer the effect across the "full" distribution with conditional quantile regression
- Conditional Quantile Regression (CQR) *vs.* Unconditional Quantile Regression (UQR)

Conditional Quantile Regression (CQR)

OLS: Conditional Mean Function

$$y_i = \alpha + \beta x_i + \sum \gamma z_i + \varepsilon_i \tag{1}$$

- Compares the mean of the distribution of y conditional on z for unit change in x

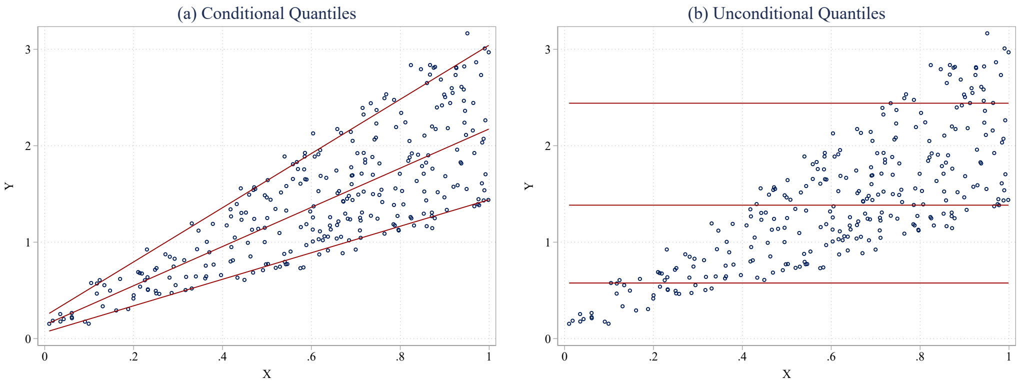
CQR: Conditional Quantile Function

$$Q_\tau(y|x,z) = \alpha(\tau) + \beta(\tau)x + \sum \gamma z + \varepsilon \tag{2}$$

- Compares quantile point Q_τ on the distribution of y conditional on z for unit change in x

Unconditional Quantile?

- Unconditional Quantiles: Quantiles of the overall distribution



Unconditional Quantile Regression (UQR)

- In CQR, low τ does not equate low value of y
  - Distribution of interest changes as a whole according to covariates
- Why UQR?
  - Effects of variables on the different parts of the “raw” or “original” distribution

(Recentered) Influence Function

- Firpo et al. (2009)
  - What do we do if we want to obtain partial effects of X on distributional statistics?
- Influence Function (IF) of a distributional statistic, $v(F_y)$
  - $IF(y_i, v, F_y) =$ influence of an individual observation on that distributional statistic
- Quantifies the changes in the distributional statistic by adding person i to the distribution

Recentered Influence Function

- Recentered Influence Function (RIF)
  - $RIF(y_i, v, F_y) = v(F_y) + IF(y_i, v, F_y)$
  - Linear approximation of the contribution of a single observation on the construction of the distributional statistic, $v(F_y)$
- $E[RIF(y_i, v, F_y)] = v(F_y)$
  - unconditional expectation of the RIF function equals $v(F_y)$
- $Var(v(F_y)) = \frac{1}{N} Var(RIF(y_i, v, F_y)) = \frac{1}{N} Var(IF(y_i, v, F_y))$
  - IF and RIF can be used to obtain the variance of distributional statistic, $v(F_y)$

RIF for Unconditional Quantiles

Firpo et al. (2009)

$$RIF(y_i, Q_\tau, F_y) = Q_\tau(y) + \frac{\tau - \Delta(y_i \leq Q_\tau(y))}{f_y(Q_\tau(y))} \tag{3}$$

- $Q_\tau(y)$: value of y at τth sample quantile
- $f_y(Q_\tau(y))$: density of y at $Q_\tau(y)$
- $\Delta(y_i \leq Q_\tau(y))$: indicator function equals 1 if y_i is below $Q_\tau(y)$

RIF Regression

1. Calculate RIF on the distributional statistic for y
2. Use calculated RIF as a dependent variable instead of y in OLS

UQR utilizing RIF

RIF Regression on Unconditional Quantile Point τ

$$RIF(y_i, Q_\tau, F_y) = \alpha(\tau) + \beta(\tau)x_i + \sum \gamma z_i + \varepsilon_i \tag{4}$$

- $\beta(\tau)$: effect of a marginal change in x on the unconditional quantile τ of y

Caution

- Using RIF-based regression models to predict unconditional quantile levels
  - Risks assessing population-level effects, not individual-level treatment effects
- If y is wages, the coefficient for dummy variable X indicates the effect of the distribution of X variable has on the distribution of wages among high- (Q90), middle- (Q50), and low- (Q10) paid workers net of covariates
- The problem is that it is not an individual-level treatment effect
- Not an issue if the coefficient of interest is not FE

Quantile Treatment Effect using IPW

- Rios-Avila and Maroto (2022)
  - Quantile treatment effect (QTE) using inverse probability weighting produces treatment effect
  - RIF-regression based UQR with QTE can assess the gender wage gap among high- middle- and low-paid workers

UQR with QTE

$$RIF(y_i, Q_\tau, F_{y|x=1})x_i + RIF(y_i, Q_\tau, F_{y|x=0})(1 - x_i) = \alpha(\tau) + \beta(\tau)x_i + \sum \gamma z_i + \varepsilon_i \tag{5}$$

RIF Decomposition

Oaxaca-Blinder Decomposition

$$\bar{Y}_a - \bar{Y}_b = \underbrace{(\alpha_a - \alpha_b) + \sum(\beta_a - \beta_b)\bar{x}_a}_{\text{Coefficient Effect}} + \underbrace{\sum(\bar{x}_a - \bar{x}_b)\beta_b}_{\text{Composition Effect}} \tag{6}$$

- Firpo et al. (2018) show that the Oaxaca-Blinder decomposition of group mean wage differences is a particular instance of a more general decomposition of any distributional statistic
- $\bar{Y}_a - \bar{Y}_b$ can be extended to $RIF(y_i, Q_\tau, F_y)_a - RIF(y_i, Q_\tau, F_y)_b$

RIF Decomposition

1. Calculate RIF on the distributional statistic for y by groups of interest
2. Use calculated RIF as a dependent variable instead of y in Oaxaca-Blinder decomposition

RIF of Other Distributional Statistic (Rios-Avila 2020)

Mean

$$RIF(y, \mu_Y) = y \tag{7}$$

Variance

$$RIF(y, \sigma_Y^2) = (y - \mu_Y)^2 \tag{8}$$

Interquantile Range

$$RIF(y, IQR(\tau_1, \tau_2)) = RIF(y, Q_{\tau_1}, F_y) - RIF(y, Q_{\tau_2}, F_y) \tag{9}$$

- and many more

Summary

- RIF got popularized for the study on the distributions
- Provides a simple and computationally less complicated approach to exploring distributions

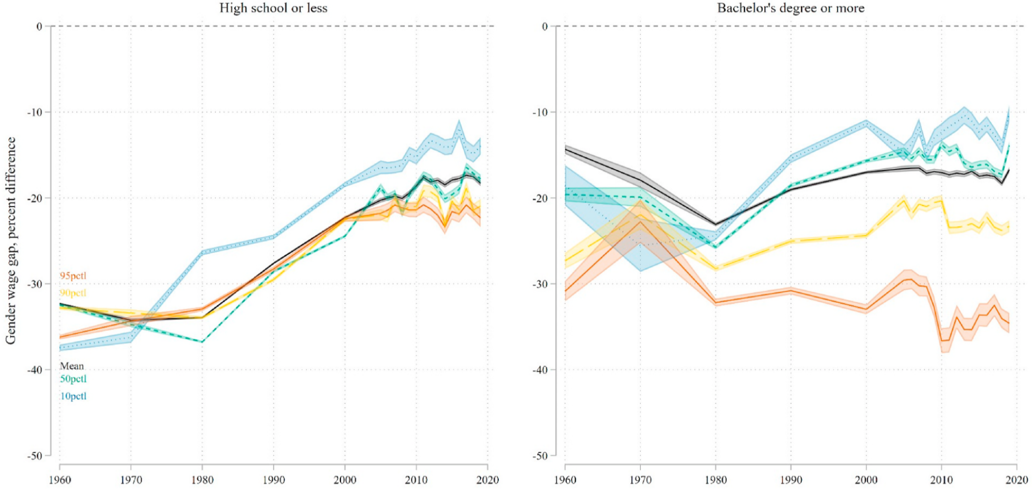
- RIFs are constructed to allow "any" distributional statistics to be assessed via OLS and its decomposition

- Interpretation of coefficients requires caution, as they indicate changes in the distribution as well
- Interpretation of FE is complicated; most suggest using QTE

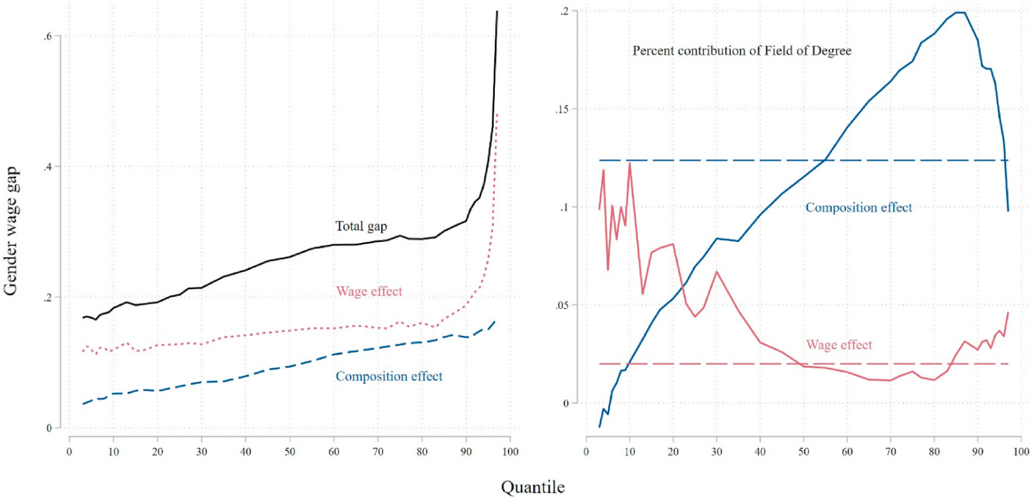
RIF on Unconditional Quantiles

- Quadlin et al. (2023)
  - Does gender differences in educational credentials contribute to the high-wage earnings gap by gender?
- UQR and Oaxaca-Blinder decomposition of the gender wage gap

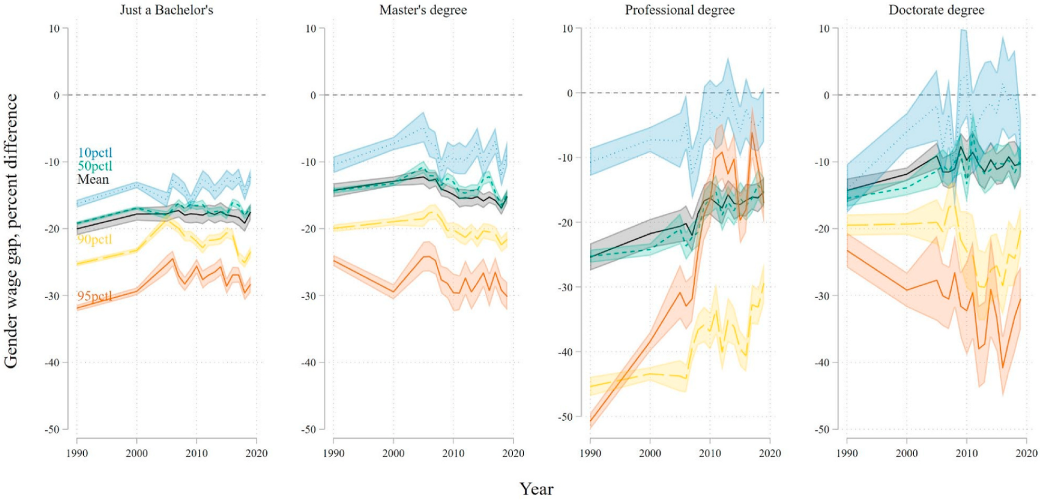
Gender Wage Gap Over Time



Decomposition of Gender Wage Gap



Gender Wage Gap by Degree Type

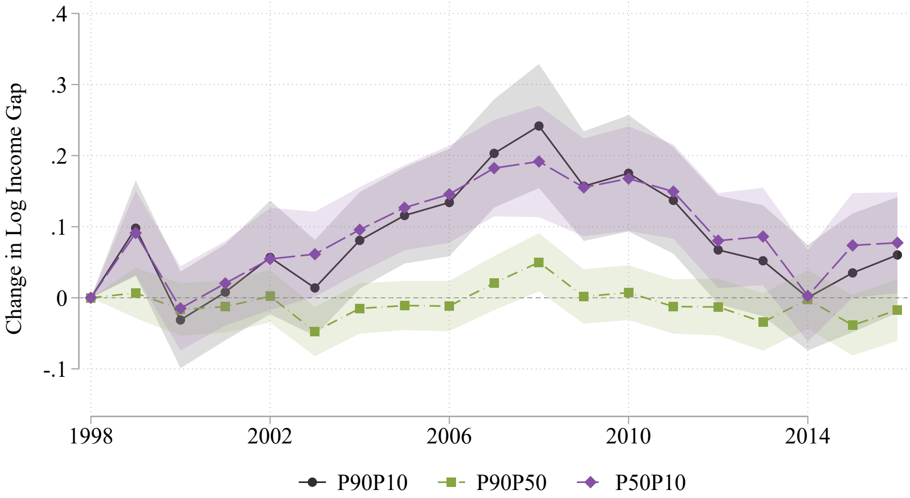


RIF on Interquantile Range

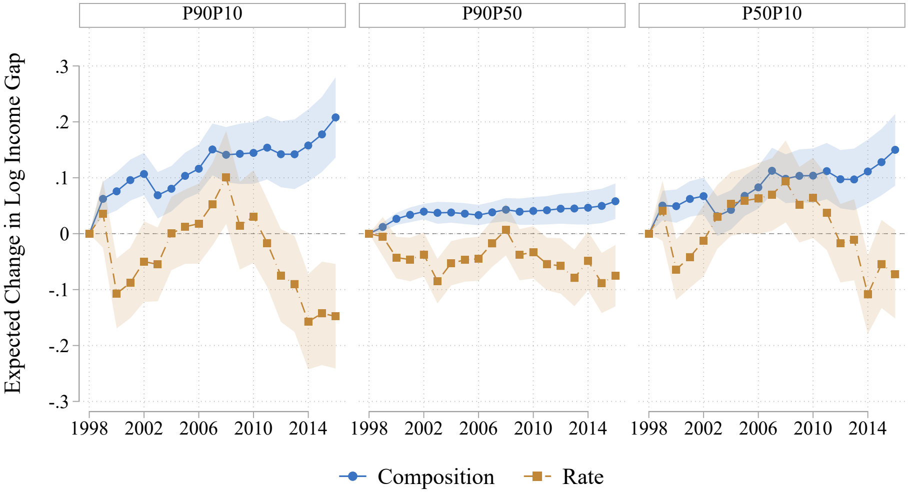
- Kim and Kim (2024)
  - What factors account for the change in bottom income inequality in Korea?
  - Does the aging population matter?

- $\Delta(Q_{90} - Q_{10}) = \Delta(Q_{90} - Q_{50}) + \Delta(Q_{50} - Q_{10})$
- Oaxaca-Blinder decomposition of $\Delta(Q_{90} - Q_{10})$, $\Delta(Q_{90} - Q_{50})$, and $\Delta(Q_{50} - Q_{10})$

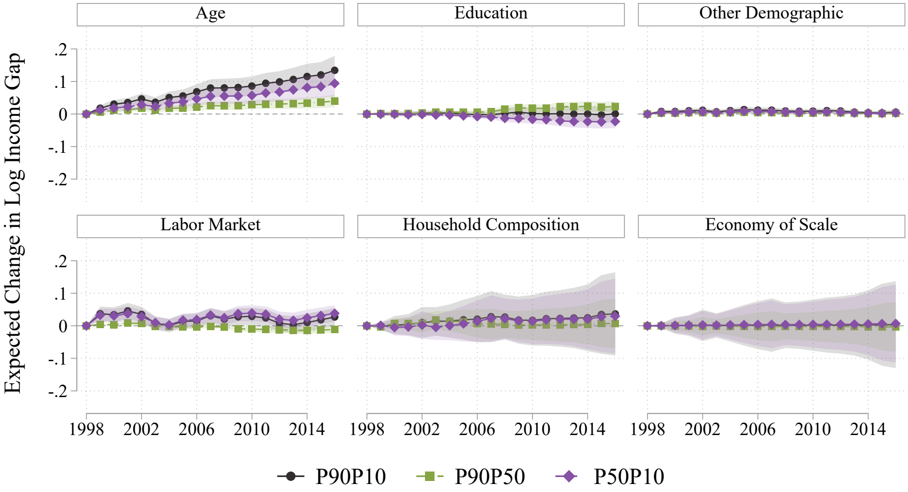
Changes in Inequality of Log Equivalized Market Income



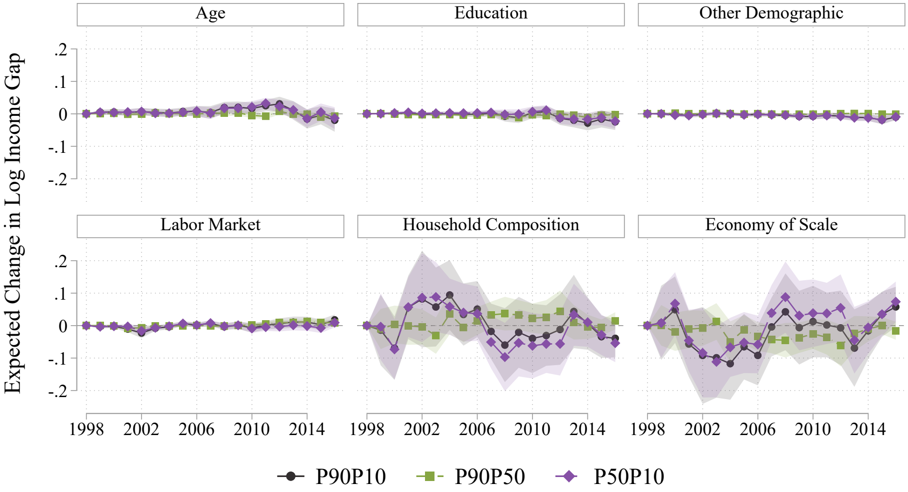
Decomposition of the Changes



Detailed Decomposition of the Changes: Composition Effect



Detailed Decomposition of the Changes: Rate Effect



Packages

- Stata
  - `ssc install rif`
  - https://journals.sagepub.com/doi/full/10.1177/1536867X20909690
- R
  - `install.packages("dineq")`
  - https://cran.r-project.org/web/packages/dineq/index.html

Data

- Data

```
. use "../workingdata/QM0412", clear

. desc

Contains data from ../workingdata/QM0412.dta
 Observations:        44,917
    Variables:             9                  10 Apr 2024 20:19
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
hrwage          float   %9.0g                 Log Hourly Wage
age             float   %9.0g                 Age
age2            float   %9.0g                 Age Squared
wt              float   %9.0g                 Person Weight
red             float   %9.0g      red        Levels of Education
mst             float   %9.0g      mst        Marital Status
region          float   %9.0g      REGION     Census Region
nchild          float   %9.0g      NCHILD     Number of Children
fem             float   %9.0g      fem        Women
--------------------------------------------------------------------------------
Sorted by:
```

Data

- Variables

```
. sum [aw=wt]
```

Variable	Obs	Weight	Mean	Std. dev.	Min	Max
hrwage	44,917	80350973.5	3.114268	.7138432	0	9.959576
age	44,917	80350973.5	38.63798	8.800404	25	54
age2	44,917	80350973.5	1570.339	694.0618	625	2916
wt	44,917	80350973.5	2528.909	1226.684	120.2	11599.43
red	44,917	80350973.5	3.341025	1.221831	1	7
mst	44,917	80350973.5	2.560073	.9962262	1	4
region	44,917	80350973.5	28.26111	10.05346	11	42
nchild	44,917	80350973.5	1.001234	1.201521	0	9
fem	44,917	80350973.5	.4817279	.4996716	0	1

Control Variables Setup

```
. qui tab red, gen(educ)
. qui tab mst, gen(mrst)
. qui tab region, gen(rgnn)
.
. isvar educ2-educ7 mrst2-mrst4 rgnn2-rgnn9
variables: educ2 educ3 educ4 educ5 educ6 educ7 mrst2 mrst3 mrst4 rgnn2 rgnn3 rgnn4 rgnn5 rgnn6 rgnn7 rgnn8 rgnn9
. local ctrl age age2 nchild `r(varlist)´
. center `ctrl´, inplace
(modified variables: age age2 nchild educ2 educ3 educ4 educ5 educ6 educ7 mrst2 mrst3 mrst4 rgnn2 rgnn3 rgnn4 rgnn5
```

UQR using RIF-Regression

(1) Calculate RIF for quantiles of log hourly wage, then OLS

```
. forvalues q = 10(10)90 {
  2.          egen hrwage_q`q´ = rifvar(hrwage), q(`q´) weight(wt)
  3. }

.
. forvalues q = 10(10)90 {
  2.          qui reg hrwage_q`q´ i.fem `ctrl´ [pw=wt]
  3.          eststo ols_q`q´
  4. }
```

(2) Use `rifhdreg`

```
. forvalues q = 10(10)90 {
  2.          qui rifhdreg hrwage i.fem `ctrl´ [pw=wt], rif(q(`q´))
  3.          eststo rif_q`q´
  4. }
```

Results

```
. esttab ols_q10 ols_q50 ols_q90 rif_q10 rif_q50 rif_q90, ///
>         mtitle(ols_q10 ols_q50 ols_q90 rif_q10 rif_q50 rif_q90) ///
>         b(3) se(3) varwidth(10) lab noobs keep(*.fem)
```

	(1) ols_q10	(2) ols_q50	(3) ols_q90	(4) rif_q10	(5) rif_q50	(6) rif_q90
men	0.000	0.000	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)	(.)
women	-0.229***	-0.240***	-0.291***	-0.229***	-0.240***	-0.291***
	(0.013)	(0.008)	(0.013)	(0.013)	(0.008)	(0.013)

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- Gender wage gap is larger at upper quantiles
- Is it gender wage gap?

Interpretation

- we would say gender wage gap at Q10 is -20.5% $(= 100 \times (e^{-.229} - 1))$ and at Q90 is -25.2%
- however, UQR provides linear approximations of changes in how unconditional quantiles of the dependent variable change when there is a small change in the distribution of independent characteristics
- at Q10, 10 percentage point increase in the share of women may decrease wages by 2.0%
- at Q90, 10 percentage point increase in the share of women may decrease wages by 2.5%

Obtain QTE (1)

Calculate IPW → Obtain RIF → OLS

```
. qui logit fem `ctrl´ [pw=wt]
. qui predict IPW0
. qui gen     IPW = .
. qui replace IPW = 1/IPW0     if fem == 1
. qui replace IPW = 1/(1-IPW0) if fem == 0
. qui gen     IPWwt = wt * IPW
.
. forvalues q = 10(10)90 {
  2.          egen hrwage_q`q´_ipw = rifvar(hrwage), q(`q´) weight(IPWwt) by(fem)
  3. }
.
. forvalues q = 10(10)90 {
  2.          qui reg hrwage_q`q´_ipw i.fem `ctrl´ [pw=IPWwt]
  3.          eststo olsIPW_q`q´
  4. }
```

Obtain QTE (2)

Use rifhdreg

```
. forvalues q = 10(10)90 {
  2.          qui rifhdreg hrwage i.fem `ctrl´ [pw=wt], rif(q(`q´)) over(fem) rwlogit(`ctrl´) ate
  3.          eststo ate_q`q´
  4. }
```

QTE Results

```
. esttab olsIPW_q10 olsIPW_q50 olsIPW_q90 ate_q10 ate_q50 ate_q90, ///
>         mtitle(olsIPW_q10 olsIPW_q50 olsIPW_q90 ate_q10 ate_q50 ate_q90) ///
>         b(3) se(3) varwidth(10) lab noobs keep(*.fem)
```

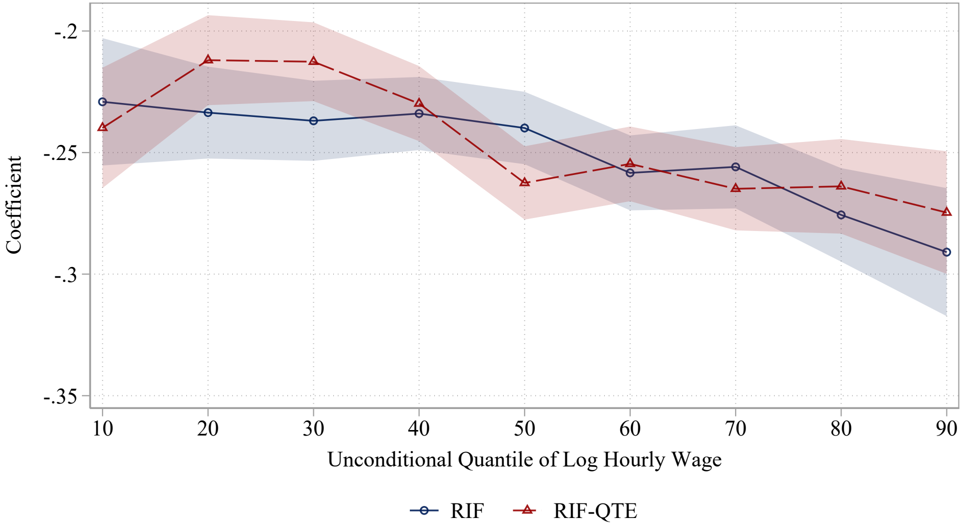
	(1) olsIPW_q10	(2) olsIPW_q50	(3) olsIPW_q90	(4) ate_q10	(5) ate_q50	(6) ate_q90
men	0.000	0.000	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)	(.)
women	-0.240***	-0.262***	-0.275***	-0.240***	-0.262***	-0.275***
	(0.013)	(0.008)	(0.013)	(0.013)	(0.008)	(0.013)

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Interpretation

- being women is associated with 21.3% $(= 100 \times (e^{-0.240} - 1))$ lower wages at Q10 and 24.0% lower wages at Q90

QTE Results Comparison



QTE Results Comparison

- Difference is “negligible” but...
- Interpretation of RIF regression coefficient require caution

RIF Based Oaxaca-Blinder Decomposition

```
.
. local ctro (age: age age2) (educ: educ2-educ7) (mrst: mrst2-mrst4) (rgnn: rgnn2-rgnn9) nchild
.
. forvalues q = 10(10)90 {
  2.          qui oaxaca_rif hrwage `ctro´ [pw=wt], rif(q(`q´)) by(fem) swap wgt(1) rwlogit(`ctrl´)
  3.          eststo decomp_q`q´
  4. }
```

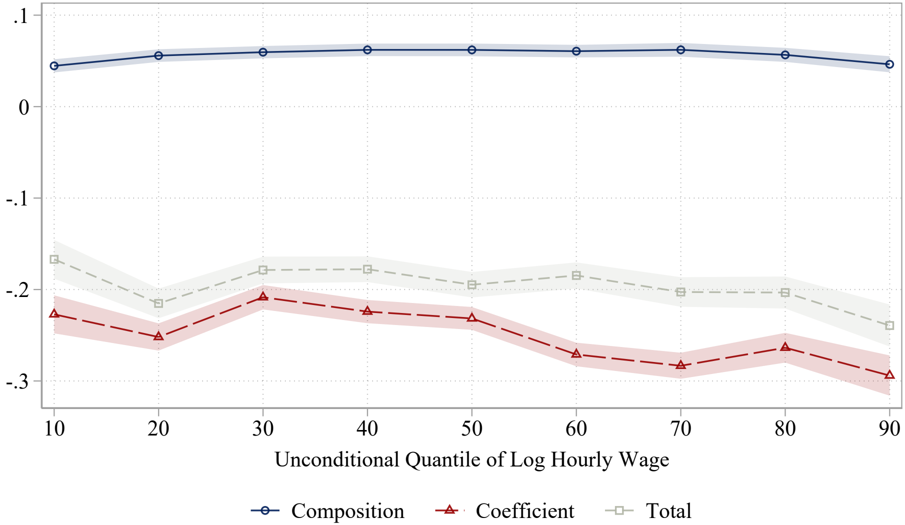
Decomposition Results

```
. esttab decomp_q10 decomp_q30 decomp_q50 decomp_q70 decomp_q90, mtitle(q10 q30 q50 q70 q90) ///
>         b(3) nose not keep(Overall:* explained:* unexplained:*) noobs
```

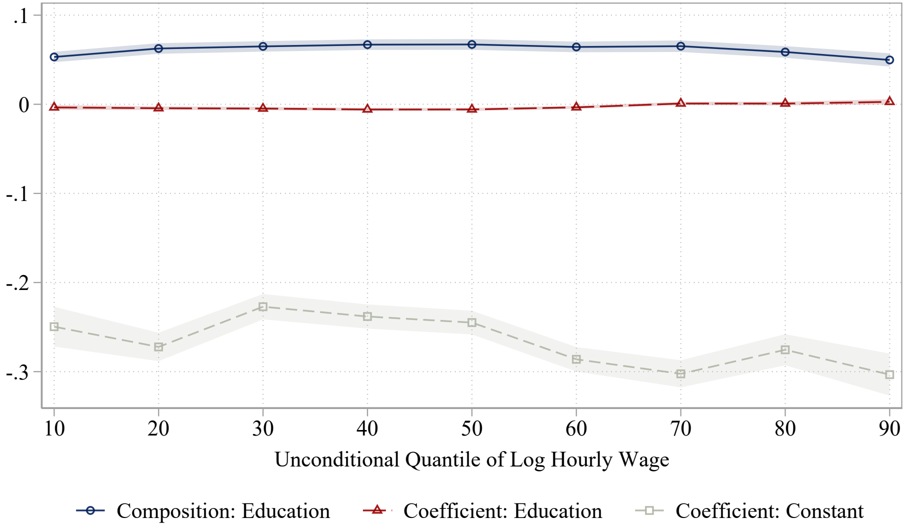
	(1) q10	(2) q30	(3) q50	(4) q70	(5) q90
Overall					
group_1	2.240***	2.703***	3.015***	3.334***	3.836***
group_c	2.184***	2.677***	2.982***	3.258***	3.787***
group_2	2.407***	2.882***	3.209***	3.537***	4.075***
tdifference	-0.167***	-0.179***	-0.195***	-0.203***	-0.239***
t_explained	0.056***	0.026***	0.033***	0.076***	0.049***
t_unexplai~d	-0.223***	-0.205***	-0.228***	-0.279***	-0.288***
explained					
total	0.056***	0.026***	0.033***	0.076***	0.049***
p_explained	0.045***	0.060***	0.062***	0.062***	0.046***
specif_err	0.012	-0.033***	-0.029***	0.014	0.003
unexplained					
total	-0.223***	-0.205***	-0.228***	-0.279***	-0.288***
rwg_error	0.004	0.003	0.004	0.004	0.006
p_unexplai~d	-0.227***	-0.208***	-0.232***	-0.283***	-0.294***

* p<0.05, ** p<0.01, *** p<0.001

OB Decomposition Results



Contribution of Education?



Thank you!

Questions or Comments?

atkim@sas.upenn.edu

References: CQR vs. UQR

- Borah, Bijan J. and Anirban Basu. 2013. "Highlighting Differences between Conditional and Unconditional Quantile Regression Approaches through an Application to Assess Medication Adherence." *Health Economics* 22:1052–1070
- Budig, Michelle J. and Melissa J. Hodges. 2010. "Differences in Disadvantage: Variation in the Motherhood Penalty across White Women's Earnings Distribution." *American Sociological Review* 75:705–728
- Killewald, Alexandra and Jonathan Bearak. 2014. "Is the Motherhood Penalty Larger for Low-Wage Women? A Comment on Quantile Regression." *American Sociological Review* 79:350–357

References: RIF

- Firpo, Sergio P., Nicole M. Fortin, and Thomas Lemieux. 2009. "Unconditional Quantile Regressions." *Econometrica* 77:953-–973
- Firpo, Sergio P., Nicole M. Fortin, and Thomas Lemieux. 2018. "Decomposing Wage Distributions Using Recentered Influence Function Regressions." *Econometrics* 6:28
- Rios-Avila, Fernando. 2020. "Recentered Influence Functions (RIFs) in Stata: RIF Regression and RIF Decomposition." *Stata Journal* 20:51–94
- Rios-Avila, Fernando and Michelle Lee Maroto. 2022. "Moving Beyond Linear Regression: Implementing and Interpreting Quantile Regression Models With Fixed Effects." *Sociological Methods & Research*

References: Examples

- Quadlin, Natasha, Tom VanHeuvelen, and Caitlin E. Ahearn. 2023. "Higher Education and High-Wage Gender Inequality." *Social Science Research* 112:102873
- Kim, ChangHwan and Andrew Taeho Kim. 2024. "Aging and the Rise in Bottom Income Inequality in Korea." *Research in Social Stratification and Mobility* 89:100882