Sources of Wage Inequality:

Decomposing the Conditional Gini Coefficient

Carlos Hurtado♣

Jun 5, 2025

Abstract

This paper introduces a new econometric method to identify factors influencing the disparities within the distribution of a positive random variable, focusing on US wages between 1986 and 2015. I relate the conditional Lorenz curve to the conditional quantile function to additively decompose the conditional Gini index. Moreover, this paper presents a technique to disentangle temporal changes in the distribution. The analysis shows that despite reduced impacts of race and gender on wages, persistent disparities require ongoing intervention, while higher education, especially college degrees, significantly reduces wage inequality.

Keywords: Gini Index, Wage Structure, Inequality, Quantile Regression

JEL Classification Numbers: C13, C21, C43, D63

♣ Economics Department, Robins School of Business, 102 UR Drive, University of Richmond, VA 23173, USA. Email: churtado@richmond.edu. Ph; 713-933-9027. The author would like to thank Roger Koenker for his advice and guidance throughout this research project. I thank comments and suggestions from Mark Borgschulte, George Deltas, Bhash Mazumder, Elizabeth Powers, and seminar participants at the Midwest Graduate Student Summit, LACEA-LAMES 2016, and the HCEO 2017 Summer School of Socioeconomic Inequality

1 Introduction

Since the 1980s, increasing attention has been given to economic inequality in the US. [1] Complex interactions of sociodemographic factors, individual attributes, and their returns shape income and wage distributions. Recognizing that these characteristics and endowments are interconnected is crucial, and their combined influence on economic inequality is complex and elusive. Advanced statistical methods are essential to unravel this complexity and accurately measure their impact on inequality. Despite extensive research on economic disparities through income or wage distribution estimates, a substantial knowledge gap persists in understanding how *inequality measures* respond to these multifaceted influencing factors.

In this study, I present a new econometric method to evaluate the impact of various factors on the disparities within the distribution of a positive random variable, explicitly focusing on wages. This method decomposes the conditional Gini index by utilizing conditional quantile regressions and the relationship between the Lorenz curve and quantile function, allowing for precise identification and quantification of influential factors on inequality. [2] I demonstrate this methodology with data from the Ongoing Rotation Group (ORG) of the Current Population Survey (CPS) for 1986 and 2015, introducing an approach to decompose wage distribution changes using counterfactual scenarios based on estimated conditional Gini coefficients.

Building on seminal contributions by Oaxaca (1973) and Blinder (1973), recent research has moved beyond classical labor models to analyze wage inequality throughout the entire

[1] See Levy and Murnane (1992), Katz (1999), Autor et al. (2008), Guvenen et al. (2014), and Abel and Deitz (2019) for a review of the literature.

[2] The Gini coefficient, ranging from 0 ("perfect equality") to 1 ("perfect inequality"), is preferred for its simplicity and comparability cross populations. It avoids parametric assumptions, unlike measures like the Atkinson Index or Theil and Generalized Entropy measures, which depend on societal inequality aversion parameters. For further details on modern inequality measurements, see Hufe et al. (2022).

distribution. [3] This shift is first exemplified by Buchinsky (1994) using quantile regression to demonstrate a stronger impact of education on higher earners' wages [4]. More recently, Bayer and Charles (2018) find significant earnings improvements for college-educated black men at higher income percentiles, indicating significant positional gains within this group.

Advancements in modeling the entire wage distribution, notably by Machado and Mata (2005), have deepened our understanding of the factors shaping wage distribution. [5] They developed a counterfactual decomposition technique using conditional quantile regression to estimate marginal (log) wage distributions, aligned with a conditional distribution derived from these regressions. [6] Their work enables constructing counterfactual scenarios to compare marginal wage distributions and explore wage differentials across various quantiles, assessing the dynamics of the entire wage distribution.

Traditional inequality analysis methods, like kernel density estimates, focusing on isolated quantiles, or quantile-based ratios, can miss intricate dynamics in wage distributions. My proposed method overcomes this by analyzing all distribution quantiles and linking the conditional quantile function with the conditional Lorenz curve. [7] This

[3] The classical model inspired literature focusing on average wage differences, accounting for individual and institutional characteristics, as exemplified by Katz and Murphy (1992), Bound and Johnson (1992), Blau and Kahn (1996), and Card and Lemieux (2001). For a review of many of the decomposition methods, refer to Fortin et al. (2011).

[4] Angrist et al. (2006) observed comparable results in a recent US subsample, echoed by Arellano and Bonhomme (2017) in the UK context.

[5] DiNardo et al. (1996) pioneered wage distribution modeling by developing an estimation procedure to analyze counterfactual (log) wage distributions using kernel density methods on weighted samples.

[6] Recently, Firpo, Fortin, and Lemieux (2009) introduced unconditional quantile regressions, which estimates the impact of covariates on unconditional quantiles. Using weighted samples, Firpo, Fortin, and Lemieux (2018) extend this approach by proposing a decomposition method that examines changes across the entire unconditional distribution.

[7] A conditional quantile function estimates percentile values in data subgroups under specific conditions, like median earnings for college graduates. This function is integral to conditional quantile regression, which uses predictor variables to predict these quantiles, providing insights into varying relationships across a distribution. Alternatively, unconditional quantile methods, such as those proposed by Firpo,et al. (2009), capture the total effect of covariate changes on the entire wage distribution. When their assumptions hold, they offer valuable insights into inequality trends. However, policy discussions often focus on within-group disparities, where conditional quantile regression better isolates how covariates shape inequality among similar workers.

comprehensive approach evaluates the entire wage distribution, linking (log) wage determinants directly to inequality through the conditional Gini index, thus offering a precise inequality measure without relying on density modeling.

I employ CPS ORG hourly wage data from 1986 and 2015 to exemplify my methodology and contrast it with the approach of Machado and Mata (2005), hereafter referred to as MM. My model, which includes a range of individual, job-related, and demographic factors, uncovers several key findings. [8] The manufacturing sector shifted to increase wages by 2015; race and gender impacts on wages, though lessened, still demand action; and crucially, higher education, particularly college degrees, noticeably reduces wage inequality, reflected in a decreased conditional Gini index, highlighting the importance of college education in addressing wage disparities.

I implement the MM method alongside my proposed technique to showcase the advantages of the later. My method reveals a notable decrease in lower-end wage inequality, almost balancing the higher-end increase, and elucidates wage inequality shifts over time. My approach highlights the substantial effects of unionization, manufacturing, and urbanization on wage levels, reflecting economic changes. It also shows that higher education, particularly college degrees, significantly impacts reducing inequality. This method surpasses MM by offering a comprehensive analysis of wage dynamics across income levels and socio-economic factors, especially emphasizing the pivotal role of education in mitigating wage disparities.

The structure of this paper unfolds as follows. Section 2 establishing the theoretical link between the conditional Lorenz curve and the conditional Gini index; Section 3 detailing the proposed estimation method; Section 4 offers an account of the US hourly wage data employed in this analysis; Section 5 discussing the empirical application and insights of the method; and Section 6 concludes.

[8] The model also incorporates state and industry fixed effects, essential for capturing macroeconomic shifts, sectoral changes, and global trends such as trade liberalization and technological advancements.

2 Conditional Lorenz Curve and Gini Index

The Lorenz curve is a powerful instrument for illustrating the inequality present in the distribution of a positive random variable. In the context of wage inequality, it represents the cumulative share of wages against the cumulative percentage of earners, from the lowest to the highest. Adhering to the framework in Koenker (2005), I define the Lorenz curve as:

$$L(\tau) = \frac{\int_0^\tau Q_Y(t)dt}{\int_0^1 Q_Y(t)dt} = \frac{1}{\mu}\int_0^\tau Q_Y(t)dt, \tag{1}$$

where Y is a continuous and positive random variable with a cumulative density function $F_Y(y)$, quantile function denoted as $Q_Y(t) = \inf\{y: F_Y(y) \geq t\} = F^{-1}(t)$, with $y_\tau = Q_Y(\tau)$ and mean μ satisfying $0 < \mu < \infty$. As show in Appendix A, the application of a monotonic transformation, denoted $h(\cdot)$, which satisfies $h(Y) \geq 0$ and $0 < \mu_h < \infty$, where $\mu_h = E[h(y)]$, culminates in a Lorenz curve of the *transformed* variable as given by

$$L_h(\tau) = \frac{1}{\mu_h}\int_0^\tau Q_{h(Y)}(t)dt = \frac{\tau E[h(y)|h(y) \leq h(y_\tau)]}{\mu_h}. \tag{2}$$

Drawing upon the fact that $0 \leq E[h(y)|h(y) \leq h(y_\tau)] \leq E[h(y)] = \mu_h$, and considering $\tau \in (0,1)$, it is clear that the Lorenz curve of the transformed variable lies between zero and one.

Let us consider $Q_{h(Y)}(t|x)$, with $t \in (0,1)$, to represent the t-th conditional quantile of the distribution of $h(Y)$, given a vector of covariates denoted by $x \in R^P$. I propose modeling this conditional quantile function as a linear combination of the covariates, illustrated as:

$$Q_{h(Y)}(t|x) = x^T\beta(t) = \sum_{j=1}^{P} x_j\,\beta_j(t), \tag{3}$$

where each $\beta_j(t)$ is the coefficient aligned with j-th covariate at the t-th quantile. Next, let $\lambda_h(\tau) \in R^P$ be a vector where its j -th element is defined as $\lambda_{h,j}(\tau) = \frac{1}{\tau}\int_0^\tau \beta_j(t)dt$, representing, in essence, the mean of the j-th coefficient within the interval $(0, \tau)$ [9]. From equations (2) and (3), the conditional Lorenz curve of the transformed variable is expressed as:

$$\mathrm{L_h}(\tau|x) = \frac{1}{\mu_h}\int_0^\tau Q_{h(Y)}(t|x)dt = \frac{1}{\mu_h}\sum_{j=1}^{P} x_j \int_0^\tau \beta_j(t)dt = \frac{\tau x^T \lambda_h(\tau)}{\mu_h}. \quad (4)$$

By comparing equations (2) and (4), it becomes clear that $E\big[h(y)|x \wedge \big(h(y) \leq h(y_\tau)\big)\big]$ equates to $x^T\lambda_h(\tau)$. By taking the limit when τ goes to one, I deduce that $E[h(y)|x]$ is given by $x^T\lambda_h(1) = x^T\int_0^1 \beta(t)dt$, provided that the integral exists for each characteristic j.

The Gini coefficient, derived from the Lorenz curve, summarizes the distribution disparity of a positive random variable. Its relationship with the Lorenz curve is:

$$G = 1 - 2\int_0^1 L(\tau)\, d\tau, \quad (5)$$

where G represents the Gini index; this index quantifies the degree of deviation of a given random variable's Lorenz curve from the line that indicates perfect equality.[10]

I compute the conditional Gini coefficient, considering a vector of covariates, by integrating the conditional Lorenz curve described in Equation (4), into the Gini index's definition:

[9] This convention implies that $\int_0^\tau \beta_j\,(t)dt = \tau\lambda_{h,j}(\tau)$, and $\lambda_h(\tau) = \frac{1}{\tau}\big(\int_0^\tau \beta_1(t)dt\,, \cdots, \int_0^\tau \beta_P(t)dt\big) = \frac{1}{\tau}\int_0^\tau \beta(t)dt$.

[10] The line of perfect equitability is the Lorenz curve of a degenerate random variable δ_μ, which only takes the single value μ.

$$G_h(x) = 1 - 2\int_0^1 L_h(\tau|x)d\tau$$
$$= 1 - \frac{1}{\mu_h}\sum_{j=1}^{P} x_j \int_0^1 \int_0^\tau 2\beta_j(t)dt\, d\tau, \quad (6)$$

where $x \in R^P$. This Equation (6) constitutes an additive decomposition of the conditional Gini index. This analytical tool is invaluable in scrutinizing the progression of variations in the distribution of $h(Y)$, contingent on the factor endowments and sociodemographic characteristics, x_j, as well as the returns (prices) associated with these endowments and characteristics, $\frac{1}{\mu_h}\int_0^1 \int_0^\tau 2\beta_j(t)dt\, d\tau$.

Notice that the coefficient can be reformulated by partitioning the interval (0,1) into n equally spaced sub-intervals:

$$G = 1 - \sum_{i=0}^{n-1} 2\int_{\tau_i}^{\tau_{i+1}} L(\tau)d\tau, \quad (7)$$

where $\tau_i = \frac{i}{n}$, for $i = 0, \cdots, n-1$. By definition, $\tau_{i+1} - \tau_i = \frac{1}{n}$, and noting that the area beneath the line of perfect equitability can be expressed in terms of rectangles and triangles, it becomes evident that

$$G = 2\left[\sum_{i=0}^{n-1}\left(\frac{i}{n^2} + \frac{1}{2n^2} - \int_{\tau_i}^{\tau_{i+1}} L(\tau)d\tau\right)\right]. \quad (8)$$

This connection offers a direct numerical approximation of the Gini coefficient, making it a flexible tool in inequality analysis.

Equation (8) uniquely identifies sub-intervals significantly affecting the Gini index, spotlighting quantiles mainly increasing distribution inequality. Combining equations (6) and (8), the coefficient, given a vector of covariates, is redefined as:

$$G_h(x) = 1 - \sum_{i=0}^{n-1}\sum_{j=1}^{P} x_j \frac{1}{\mu_h}\int_{\tau_i}^{\tau_{i+1}}\int_{0}^{\tau} 2\beta_j(t)dt\,d\tau$$
$$= 2\left[\sum_{i=0}^{n-1}\left(\frac{i}{n^2}+\frac{1}{2n^2}-\sum_{j=1}^{P} x_j \frac{1}{\mu_h}\int_{\tau_i}^{\tau_{i+1}}\int_{0}^{\tau}\beta_j(t)dt\,d\tau\right)\right]. \quad (9)$$

Equation (9) decomposes the conditional coefficient additively in terms of both endowments and characteristics, x_j, and the key sub-intervals influencing distribution inequality. This approach enhances our understanding of factors behind income or wealth disparities in diverse economic contexts.

2.1 Impact of Individual Characteristics on the Conditional Gini Index

I use Equation (6) to compute the variation in the conditional Gini index, resulting from a *small* positive change in a characteristic j from x_j to x_j':

$$\frac{\Delta G_h(x)}{\Delta x_j} = \frac{G_h(x_j', x_{-j}) - G_h(x_j\ , x_{-j})}{x_j' - x_j} = -\frac{1}{\mu_h}\int_0^1\int_0^{\tau} 2\beta_j(t)dt\,d\tau \stackrel{\text{def}}{=} -\frac{\Pi_j}{\mu_h}, \quad (10)$$

where $x = (x_j, x_{-j}) = (x_1, \cdots, x_j, \cdots, x_p) \in R^P$. I assume $\mu_h > 0$, which indicates that the direction of the change in the Gini coefficient exclusively relies on the sign of

$$\Pi_j = \int_0^1\int_0^{\tau} 2\beta_j(t)dt\,d\tau.$$

A negative Π_j indicates that a marginal positive change in covariate j leads to an increase in the conditional Gini index, suggesting heightened inequality in the distribution of $h(Y)$. Conversely, a positive shift in covariate j with a positive Π_j results in a decrease in the inequality found in the distribution of $h(Y)$.

Furthermore, μ_h serves as a scaling parameter that normalizes the Lorenz curve and the Gini coefficient within the range of zero to one. Consequently, the magnitude of Π_j indicates the extent of change in the Gini index following an adjustment in covariate j. Larger absolute values of Π_j correspond with more pronounced shifts in the absolute Gini coefficient. I label the absolute value of $\frac{\Pi_j}{\mu_h}$ as the impact of covariate j on the distribution of $h(Y)$. Under this premise, certain covariates exert a more substantial impact on the distribution of $h(Y)$ compared to others.

2.2 Temporal Changes in the Distribution of $h(Y)$

I explore the complexities of the distribution of $h(Y)$ to assess how various factors influence *distributional changes over time*. This decomposition helps distinguish effects from shifts in individual traits versus changes in returns to these attributes. Similar decomposition analyses appear in prior research: DiNardo et al. (1996) use kernel density estimates on reweighted samples for counterfactual wage distributions, Firpo, Fortin, and Lemieux (2018), hereafter FFL, apply unconditional quantile regressions to decompose wage changes across the entire distribution, and MM develop a technique using conditional quantile regression. In all scenarios, including mine, the decomposition broadens the Oaxaca (1973) method, initially forged to investigate counterfactual disparities in average earnings.

I aim to explore changes in the distribution of $h(Y)$ across two years, denoted by $\Psi \in \{0,1\}$, through two counterfactual scenarios. First, I assess the inequality in the distribution of $h(Y)$ for $\Psi = 1$, using the distribution of covariates in year $\Psi = 0$. Second, I evaluate the disparity in the distribution of $h(Y)$ in year $\Psi = 1$, assuming only one covariate follows the

distribution in year $\Psi = 0$. This approach helps understand the impacts on $h(Y)$ due to changes in covariates and their returns.

Let us model the conditional quantile function in year Ψ as

$$Q_{h(Y)}(t|x;\Psi) = x^T \beta_\Psi(t), \tag{11}$$

where $\beta_\Psi(t)$ represents the coefficients of the covariates in year Ψ at quantile t, and x is a vector of covariates. Now, let $X(\Psi)$ denote an $N_\Psi \times P$ matrix of data on these covariates in year Ψ with N_Ψ denoting the number of observations and P the number of covariates. Also, denote by $\bar{X}_j(\Psi)$ the average of column j of the matrix $X(\Psi)$. Using the additive decomposition of the Gini coefficient, Equation (6), I propose an estimate for the conditional Gini index in year Ψ:

$$\hat{G}_h^\Psi = 1 - \sum_{j=1}^{P} \bar{X}_j(\Psi) \frac{\widehat{\Pi}_j^\Psi}{\hat{\mu}_h^\Psi}, \tag{12}$$

where $\hat{\mu}_h^\Psi$ and $\widehat{\Pi}_h^\Psi$ are the respective estimates for μ_h and Π_j in year Ψ.

Assuming for simplicity that changes in covariates don't modify their returns, despite potential general equilibrium effects, I estimate the conditional Gini index for year $\Psi = 1$ with covariate distribution from year $\Psi = 0$ as follows: [11]

$$\hat{G}_h^1\big(X(0)\big) = 1 - \sum_{j=1}^{P} \bar{X}_j(0) \frac{\widehat{\Pi}_j^1}{\hat{\mu}_h^1}. \tag{13}$$

For this discussion, let G_h^Ψ denote the Gini index computed from a sample in year Ψ. I calculate changes in the Gini coefficient to capture shifts in the distribution of $h(Y)$:

[11] This assumption is inherent in the Oaxaca (1973) decomposition and present in DiNardo *et al.* (1996) and MM.

$$G_h^1 - G_h^0 = \hat{G}_h^1 - \hat{G}_h^0 + \text{residual}$$
$$= \underbrace{\hat{G}_h^1 - \hat{G}_h^1\big(X(0)\big)}_{\text{change in covariates}} + \underbrace{\hat{G}_h^1\big(X(0)\big) - \hat{G}_h^0}_{\text{change in returns}} + \text{residual}. \quad (14)$$

The change in the distribution tied to shifts in individual characteristics is measured by $\hat{G}_h^1 - \hat{G}_h^1\big(X(0)\big)$, where returns remain constant and only the covariates vary. On the other hand, the shift in the distribution of $h(Y)$ triggered by changes in returns to individual traits is encapsulated by $\hat{G}_h^1\big(X(0)\big) - \hat{G}_h^0$, where covariates stay the same and only returns change.

I define $X_{-j}^1(0) = \big(\bar{X}_1(1), \cdots, \bar{X}_j(0), \cdots, \bar{X}_P(1)\big)$ as a vector in R^P, where the j-th entry represents the average of characteristic j in year $\Psi = 0$, while all other entries are the averages of the respective covariates in year $\Psi = 1$. To pinpoint the effect of a single covariate changing from year $\Psi = 0$ to $\Psi = 1$, I introduce the impact on the change in the distribution of $h(Y)$:

$$\hat{G}_h^1 - \hat{G}_h^1\left(X_{-j}^1(0)\right) = -\left(\bar{X}_j(1) - \bar{X}_j(0)\right)\frac{\widehat{\Pi}_j^1}{\hat{\mu}_h^1}. \quad (15)$$

The Equation assumes a particular sequence of changes from $\Psi = 0$ to $\Psi = 1$, which is arbitrary. Exploring what would occur at $\Psi = 0$ with covariates of $\Psi = 1$ offers alternate insights into the effects of changes in characteristics and their returns.

2.2.1 Alternative Methods

I also apply the MM method to understand changes in the distribution of $h(Y)$, estimating the entire distribution to identify factors influencing temporal shifts.[12] First, I model the conditional quantile of $h(Y)$ in year Ψ as given by equation (11):

$$Q_{h(Y)}(t|x;\Psi) = x^T \beta_\Psi(t).$$

Second, I employ the following steps to estimate the implied marginal densities:

1. Generate a random sample of size m from a uniform random variable on $[0,1]$: $u_1, \cdots, u_m$
2. Estimate $Q_{h(Y)}(t|x;\Psi)$ yielding m estimates $\hat{\beta}_\Psi(u_i)$.
3. Generate a random sample of size m with replacement from $X(\Psi)$, the $N_\Psi \times P$ matrix of data on covariates, denoted by$\{x_i^*(\Psi)\}_{i=1}^m$.
4. Generate a random sample of $h(Y)$ that is consistent with the conditional distribution defined by the model: $\left\{\eta_i^*(\Psi) \stackrel{\text{def}}{=} x_i^*(\Psi)^T \hat{\beta}_\Psi(u_i)\right\}_{i=1}^m$.

For generating a random sample from the marginal distribution of $h(Y)$ as it would have been in $\Psi = 1$ —assuming all covariates had been as in $\Psi = 0$—I use $X(0)$ in the third step above, assuming covariate changes do not modify their returns.

To construct a counterfactual where only one covariate, $x_i(1)$, is distributed as in year $\Psi = 0$, I introduce additional steps based on the method by MM. The authors defined a partition of the covariate $x_i(1)$ in J classes, $C_j(1)$, with relative frequencies $f_j(\cdot)$, for $j = 1, \cdots, J$, and propose the following:

1. Generate $\{\eta_i^*(1)\}_{i=1}^m$, a random sample of $h(Y)$, with size m, that is consistent with

[12] This method uses the probability integral transformation theorem, which states that if U is uniformly distributed on [0,1], then $F^{-1}(U)$ follows the distribution F.

the conditional distribution defined by the model.

2. Take the first class, $C_1(1)$, and select all elements of $\{\eta_i^*(1)\}_{i=1}^m$ that are generated using this class, $I_1 = \{i | x_i(1) \in C_1(1)\}$, that is $\{\eta_i^*(1)\}_{i \in I_1}$. Generate a random sample of size $m \times f_1(0)$ with replacement from $\{\eta_i^*(1)\}_{i \in I_1}$.

3. Repeat step 2 for $j = 2, \cdots, J$.

These approaches enable me to decompose changes in the density of $h(Y)$ based on these generated samples.

Let $\hat{f}\big(\eta(\Psi)\big)$ be an estimator of the marginal density of an observed sample of $h(Y)$ in year Ψ, and $\hat{f}\big(\eta^*(\Psi)\big)$ an estimator of the density of $h(Y)$ based on the generated sample $\{\eta_i^*(\Psi)\}_{i=1}^m$. I denote $\hat{f}(\eta^*(1); X(0))$ as an estimate of the counterfactual density in $\Psi = 1$ if the covariates had been distributed as in $\Psi = 0$. Similarly, $\hat{f}(\eta^*(1); x_i(0))$ is an estimate of the density in $\Psi = 1$ if only the i-th covariate is distributed as in $\Psi = 0$. For a summary statistic $\alpha(\cdot)$ (e.g., a quantile or scale measure), the decomposition of changes in α can be written as:

$$\alpha\big(\hat{f}(\eta(1))\big) - \alpha\big(\hat{f}(\eta(0))\big) = \underbrace{\big(\hat{f}(\eta^*(1))\big) - \alpha\left(\hat{f}\big(\eta^*(1); X(0)\big)\right)}_{\text{change in covariates}} \tag{16}$$

$$+ \underbrace{\alpha\left(\hat{f}\big(\eta^*(1); X(0)\big)\right) - \alpha\left(\hat{f}\big(\eta^*(0)\big)\right)}_{\text{change in returns}} \tag{17}$$

$$+ \text{residual} \tag{18}$$

Likewise, the individual contribution of a covariate is:

$$\alpha\big(\hat{f}(\eta^*(1))\big) - \alpha\big(\hat{f}\big(\eta^*(1)\big); x_i(0)\big). \tag{19}$$

This approach estimates the full distribution to pinpoint factors influencing shifts in specific quantiles, unlike the proposed method, which assesses inequality across the whole distribution directly.

Another approach to decomposing distributional changes is the method proposed by FFL, which relies on Recentered Influence Function (RIF) regressions to estimate the effect of covariates on unconditional quantiles or other statistics, such as the Gini index. FFL provides insights into how changes in returns and covariates (they call those changes in composition and structure) shape inequality over time without estimating the conditional distribution. [13] Appendix F deeply explains the FFL method, its scope and limitations.

To clarify, Unconditional Quantile Regression (UQR), introduced by Firpo, Fortin, and Lemieux (2009), differs fundamentally from Conditional Quantile Regression (CQR) in both purpose and interpretation. UQR estimates how changes in covariates influence the unconditional distribution of the outcome by regressing the RIF of a quantile on those covariates. In contrast, CQR focuses on how the quantiles of the outcome vary within subpopulations defined by specific characteristics. For example, it estimates how the 90th percentile of wages differs across individuals with similar education and experience. While UQR reflects aggregate distributional shifts, CQR captures within-group disparities by holding covariates constant.

Indeed, FFL emphasizes that CQR captures only within-group effects by isolating dispersion among observably similar individuals, while unconditional approaches combine both within-group and between-group variation. This difference is not merely technical; it reflects fundamentally different research goals. The choice between UQR and CQR should depend on whether the aim is to understand overall distributional changes or disparities among individuals with similar characteristics.

From a policy perspective, conditional measures of inequality are especially valuable when the focus is on disparities among observably similar individuals. Policy-oriented research often seeks to understand wage gaps among workers with the same education, experience, or industry. In such cases, unconditional measures may obscure these within-

[13] An essential distinction between the FFL method and my proposed approach is that the RIF transformation depends on the entire wage distribution and the distributions of covariates through its effects on wages, making it sensitive to changes in the distribution of characteristics. See Appendix F for details.

group disparities by blending them with between-group composition effects. The conditional Gini coefficient addresses this issue directly by capturing the residual inequality that remains after accounting for covariates. By conditioning on characteristics, researchers can answer questions such as: how unequal are wages among similar workers, and which factors drive that remaining inequality?

By contrast, unconditional decomposition methods such as the UQR/RIF-regression method of FFL decompose total inequality without conditioning on covariates, which comes with two conceptual trade-offs.

First, RIF-based regressions rely on the full outcome distribution to construct the dependent variable, which makes the estimates sensitive to large shifts in that distribution and less capable of capturing nonlinear interactions. The influence function offers an exact approximation only for infinitesimal changes. When covariate changes are substantial or the inequality measure is highly nonlinear, the linear RIF regression may introduce significant approximation errors. Rothe (2015) demonstrates that for some nonlinear functionals, like the Gini, the effects of covariates may be inaccurately estimated. FFL also note that the RIF transformation's dependence on the entire distribution can lead to specification problems when covariate distributions change considerably.[14]

Second, unconditional approaches do not separate within-group from between-group inequality. This limits their usefulness for policy questions that target disparities among observably similar individuals. For example, the impact of a minimum wage on low-educated workers may be obscured by broader composition effects in an unconditional framework.

In contrast, my method focuses on the conditional Gini, allowing a decomposition of residual inequality within comparable groups. This makes it better suited for evaluating

[14]. Appendix F provides empirical evidence of this issue using my comprehensive hourly wage data from the CPS ORG for 1986 and 2015. Panel B of Figure F.3 shows that the composition effect estimated via RIF regressions deviates from the effect obtained through reweighting, particularly across the middle of the distribution and above the 80th percentile, signaling specification error. Table F.2 confirms that these discrepancies are statistically significant, highlighting how substantial shifts in the covariate distribution distort RIF regression estimates.

targeted policies and understanding the drivers of wage disparities among similar individuals.

3 Estimation Procedure

In the previous section, I measure the impacts and decompose the temporal changes in the distribution of $h(Y)$ by estimating $\Pi_j = \int_0^1 \int_0^\tau 2\beta_j(t)dt\,d\tau$ as detailed in Equation (6). A straightforward estimate would be $\widehat{\Pi}_j = \int_0^1 \int_0^\tau 2\hat{\beta}_j(t)dt\,d\tau$, where $\hat{\beta}_j$ represents the estimated quantile regression coefficient.

To clarify the estimation approach, I consider $Q_{h(Y)}(t|x)$ for $t \in (0,1)$ to be the t-th conditional quantile function of $h(Y)$, given a vector of covariates $x \in R^P$. I posit that the conditional quantile function can be model as:

$$Q_{h(Y)}(t|x) = x^T\beta(t), \tag{20}$$

Here, $\beta(t)$ is a vector in R^P, with its entries being the quantile regression coefficients. From Koenker and Bassett (1978), we know that the quantile regression estimates exist and minimize a weighted sum of absolute residuals.

My goal is to estimate the impact using quantile regression coefficients estimates:

$$\widehat{\Pi}_j = \int_0^1 \int_0^\tau 2\hat{\beta}_j(t)dt\,d\tau = \sum_{i=0}^{n-1} \int_{\tau_i}^{\tau_{i+1}} \int_0^\tau 2\hat{\beta}_j(t)dt\,d\tau. \tag{21}$$

One method to determine the double integrals in Equation (21) is to perform a numerical computation on a grid for evaluations of $\hat{\beta}_j(t)$. However, a challenge emerges when extending one-dimensional integration methods to multiple dimensions: the required

function evaluations increase exponentially. For instance, using m evaluation points would require estimations proportional to m^2.

To simplify calculations, I suggest deriving a smooth approximation for $\hat{\beta}_j(t)$ using a known functional form with an identifiable antiderivative. This allows for analytical calculation of $\widehat{\Pi}_j$ using the functional form, keeping the number of function evaluations to m, in line with the initial set of evaluation points.[15]

I use orthogonal polynomials to approximate these continuous functions.[16] This technique produces a singular polynomial of order K that minimizes the squared error between the smoothing polynomial and the observed function values. One limitation when using orthogonal polynomials to compute Π_j is the need to define the polynomial's order, K. However, a notable advantage is the ease of computation for Π_j's estimate. I can determine the double integrals of Equation (21) in a single step.

I assume that $\hat{\beta}_j(t)$ is continuous over the interval [0,1]. As Judd (1998) describes, I can approximate this function using orthogonal polynomials on this interval.[17] Several orthogonal polynomial families, like Legendre, Chebyshev, Laguerre, and Hermite, vary in their weighting functions and domains. For bounded domain functions, the simplest weighting function is $w(x) = 1$, matching Legendre polynomials. Therefore, for simplicity, I use Legendre polynomials to approximate $\hat{\beta}_j(t)$ on [0,1].

---

[15] Several smoothing techniques for continuous functions exist. Splines, a well-known technique, approximate functions with polynomials segments that joint at knots. However, estimating Π_j using splines is challenging due to the need for selecting knots and the requirement for multiple piecewise integrations based on these knots.

[16] A weighting function, $w(x)$, defined on $[a, b]$, is positive almost everywhere and has a finite integral on $[a, b]$. Given a weighting function, the inner product between the polynomials f and g is $\langle f, g\rangle = \int_a^b f(x)g(x)w(x)dx$. A family of polynomials $\{p_n(x)\}$ is orthogonal relative to $w(x)$ if and only if $\langle p_m, p_n\rangle = 0$ for all $m \neq n$.

[17] The Weierstrass Approximation Theorem allows for uniform approximation of $\hat{\beta}_j(t)$ over [0,1] to any precision with these polynomials.

I derive the least-square approximation of $\hat{\beta}_j(t)$ using a polynomial of order K on [0,1].[18] Specifically, let $\tilde{\beta}_{j,K}(t)$ represent a polynomial described as:

$$\tilde{\beta}_{j,K}(t) = \alpha_{0,j}p_0(t) + \alpha_{1,j}p_1(t) + \cdots + \alpha_{K,j}p_K(t).$$

Here, $\{p_k\}_{k=0}^{K}$ denotes the first $K+1$ Legendre polynomials. The objective is to minimize the sum of the squared errors between $\hat{\beta}_j(t)$ and $\tilde{\beta}_{j,K}(t)$, defined by

$$e\big(\alpha_{0,j}, \cdots, \alpha_{K,j}\big) = \int_0^1 \big[\hat{\beta}_j(t) - \tilde{\beta}_{j,K}(t)\big]^2 dt.$$

For a given K, I define

$$\tilde{\tilde{\beta}}_{j,K}(t) = \underset{\alpha_{0,j},\cdots,\alpha_{K,j}}{argmin}\ e\big(\alpha_{0,j}, \cdots, \alpha_{K,j}\big);$$

the polynomial $\tilde{\tilde{\beta}}_{j,K}(t)$ is a smoothed approximation of $\hat{\beta}_j(t)$ based on $K+1$ known polynomials. One of the major benefits of this approximation is its closed-form antiderivatives, which streamline computation. By leveraging this smoothed approximation, I can efficiently determine the influence of the j-th covariate on the inequality of the distribution of $h(Y)$, as described in Equation (21). [19]

I use bootstrapping to assess the significance of $\widehat{\Pi}_j$ and establish its confidence intervals. The process involves N observations resampled $\mathbb{R}$ times with replacement. Each iteration estimates the quantile regression coefficients, $\hat{\beta}(t)$, and their smooth polynomial approximations, $\tilde{\tilde{\beta}}_{j,K}(t)$, from the resampled data. Using this smooth approximation, I compute an estimate of $\widehat{\Pi}_j$ based on Equation (21). I compute the point estimate for the

[18] Appendix B provides details of the approximation process.

[19] Appendix C features robustness tests confirming the proposed method's high accuracy and precision across various polynomial orders.

impacts, Π_j, by averaging these $\mathbb{R}$ estimates. The 95% bootstrap confidence intervals are constructed using the 2.5-th and 97.5-th quantiles from these estimates. [20]

4 Hourly Wage Series From the CPS ORG

In this study, I scrutinize data from the CPS to analyze the shifts in wage distribution in the US from 1980 to 2015. Since 1979, the CPS ORG has actively engaged workers in a comprehensive survey, collecting detailed earnings data. This data facilitates the accurate estimation of hourly wages, which can be derived directly from reported hourly earnings or calculated by dividing weekly earnings by the corresponding hours worked per week.

Nevertheless, the task of utilizing the ORG data comes with its set of challenges, as highlighted by Acemoglu and Autor (2011).[21] To ensure I overcome all these challenges, I utilize programs available under the GNU General Public License created by the Center for Economic and Policy Research (CEPR). This resource integrates data from the National Bureau of Economic Research Annual Earnings Files with the CPS basic monthly files, establishing a consistent wage series utilizing the data from the CPS ORG.[22]

For the final sample, I adjust wages to 2020's monetary value using the Consumer Price Index issued by the Bureau of Labor Statistics. [23] I focus on hourly workers aged 16 to 65, earning $1 to $100 (in 1979 dollars). Following Acemoglu and Autor (2011), I calculate

[20] The accuracy of my estimation hinges on precisely modeling the conditional quantile function, especially the linearity assumption. For a large sample, as detailed in Bantli and Hallin (1999) and Koenker (2005), we know that $\hat{\beta}_n(t) \to \beta(t)$, and therefore $\widehat{\Pi}_j \approx \Pi_j$. Appendix C confirms the estimation's accuracy across different polynomial orders.

[21] To ensure consistency in hourly wage data across years, I adjust for the CPS's methodological changes over three decades. These include modifications in top coding weekly earnings, categorizing overtime, tips, and commissions for hourly workers, and variations in reporting 'usual weekly hours.'

[22] My development of a consistent hourly wage series from 1980 to 2015 owes much to the CEPR programs. For details, see Schmitt (2003) on the series and Acemoglu and Autor (2011) for the methodology. A replication package is available at https://doi.org/10.7910/DVN/HJEVTW and my website.

[23] I use the seasonally adjusted US city average index for all items (Series Id: CUSR0000SA0).

potential experience by subtracting the number of years of education from individuals' ages and deducting five additional years to account for elementary schooling. Finally, I include indicators for female, nonwhite and unionized workers and maintain uniform classifications for twenty industries, including a consistent category for manufacturing employees.

In this study, I meticulously categorize educational attainment, recognizing its crucial impact on earnings as established by seminal works in economics. I draw upon the insights of Card (1999), Autor, et al. (2003), Goldin and Katz (2007), and Acemoglu and Autor (2011), all of whom underscore education's pivotal role in shaping labor market outcomes and highlight its position as a key determinant of earnings and employment opportunities.

I include various education levels in my analysis, from non-school attendees to bachelor's degree holders and above. Specifically, I include the associate degree category to acknowledge its proven positive impact on earnings. Recent studies, including those by Jepsen et al. (2014), Bahr et al. (2015), Stevens et al. (2019), and Grosz (2020), emphasize the significant earnings benefits of community college programs, validating the inclusion of associate degrees for a comprehensive wage impact assessment.

My analysis covers a large sample of around 165,000 workers annually from 1980 to 2015. Table 1 presents summary statistics for men and women, highlighting the gender wage gap. Men's average real wages remained steady at $24.4 (3.2), while women experienced a systematic wage increase, reducing the gender wage gap over the years. The table shows rising educational attainment for both genders, with women surpassing men in average education levels. Additionally, Table 1 reveals a decline in unionization and manufacturing employment, alongside an increase in nonwhite workers' participation during this period.

5 Sources of Wage Inequality in the US

Building on the relationship between the conditional Gini coefficient and conditional quantile regression, I employ the comprehensive hourly wage data from the CPS ORG for my empirical application. I set $h(\cdot)$ as the natural logarithm and choose 1986 ($\Psi = 0$) as the

start year, marking the beginning of growing wage inequality in the US during the early 1980s. The analysis extends to 2015 ($\Psi = 1$), covering three decades of rising wage disparity.

I model the conditional quantile function of the logarithm of wages as:

$$Q_{\ln w_{isj}}(t|x_{isj};\Psi) = x_{isj}^{T}\beta_{\Psi}(t) + \eta_s(t) + \gamma_j(t) + \varepsilon_{isj}(t) \qquad (22)$$

Here, x_{isj} represents the characteristic for individual i, in state s, working in industry j. This vector encompasses job-related attributes such as unionization status, public sector employment, manufacturing job, and part-time work status. Demographic attributes include indicators for nonwhite, female, and marital status. Additional controls cover a quadratic in potential experience, urban living indicators, education categories, decade-based experience indicators, and their interactions with education classes. Finally, $\eta_s(t)$ and $\gamma_j(t)$ are state and industry fixed effects, respectively.

In my model, I include independent variables as robust controls to account for macroeconomic and structural changes. The industry fixed effects capture the shift from manufacturing-heavy sectors to service- and tech-oriented sectors (Blum, 2008). Furthermore, job-related attributes and state fixed effects capture international trade impacts and regional disturbances, including the 'China shock' (Autor et al., 2013) and other nuances of trade liberalization (Chongvilaivan and Hur, 2011). Urban living and education indicators underline the influence of technological advances on wages, showing how technology benefits specific regions and skill sets (Hühne and Herzer, 2017).

Figure 1 displays the quantile regression coefficients estimated of Equation (22) for a grid of 69 equally spaced points over the (0,1) interval [24]. In each panel, the solid line corresponds to the estimates in 1986, while the dashed line depicts those from 2015. In each

[24] The 69 equally spaced points create a grid that has constant step of around 1.4%.

case, the shaded regions around the lines correspond to the 95% confidence interval obtained by computing a Huber sandwich estimate using a local estimate of the sparsity.

Figure 1 offers a comprehensive view of the factors that affect wages in 1986 and 2015 across the entire distribution. Starting with union status, the positive wage premium associated with union membership was more pronounced in 1986 across all quantiles but has reduced by 2015. For public sector employees, there is a premium for lower wage workers, and a penalty for upper wage employees, and these premium and penalty remains similar in magnitude for both 1986 and 2015.

Figure 1 depicts the evolving impact of manufacturing jobs on wages across three decades. Initially, in 1986, manufacturing jobs negatively affected wages, but by 2015, they became a positive determinant for wages across most quantiles, with diminishing impact at the highest income levels. This transition from negative to positive impact reflects the evolving role of the manufacturing sector in wage dynamics, likely due to structural changes, economic shifts, and external factors like the 'China shock' and trade agreements, which contributed to a more productive manufacturing sector in the US over this period.

Looking at demographic factors, nonwhite and women workers both show wage penalties, with the latter having a slightly more pronounced difference, especially in higher quantiles. Notably, this disparity has narrowed for both groups over the 30-year span, indicating progress, albeit limited, in wage equality. Moreover, married workers enjoy a wage premium, particularly in 1986 in the lower end of the distribution. However, the advantage at the lower wages diminishes by 2015.

Figure 1 underscores a shift in the urban wage premium between 1986 and 2015. In 1986, the figure shows an evident urban wage premium, especially prominent in the middle to higher quantiles. However, by 2015 this premium appears to diminish, suggesting a decrease in the urban wage advantage. This trend complements the findings of De la Roca and Puga (2017), emphasizing the dynamic nature of knowledge spillover in urban environments and its potential impact on wages over time.

Figure 1 shows different wage trends from 1986 to 2015 across education levels. In 1986, high school and some college-educated individuals had higher wages in lower quantiles than their 2015 peers, indicating better pay for less advanced degrees. Associate degree earners had similar wages in both years, with the 2015 cohort slightly ahead in higher earnings. Notably, college graduates in 2015 consistently outearned their 1986 counterparts, highlighting the growing value of higher education, particularly for top earners, reflecting changing labor market dynamics over three decades.

5.1 Impact of Individual Characteristics

I implement the procedure described in section 3 to estimate the sources of wage inequality in the US. I choose the order of the polynomial approximation to be 7 for 1985 ($\Psi = 0$) and 2015 ($\Psi = 1$). [25] Additionally, I set 1,000 bootstrap repetitions for estimation, compute $\hat{\mu}_{ln}$ as the weighted average of the logarithm of 2020 real wages, and estimate $\widehat{\Pi}_j$ from Equation (20) for both years to assess the impact estimate $\frac{\widehat{\Pi}_j}{\widehat{\mu}_{ln}}$.

Table 2 shows estimation results for selected covariates. As discussed in section 2.1, given a *small* positive change in covariate j, a positive sign of $\widehat{\Pi}_j$ is associated with a reduction in inequality of the (log) wage distribution. Each cell in Table 2 shows the impact estimation and its 95% bootstrap confidence interval, with columns representing different years. Notably, all covariates in Table 2 are binary, except potential experience, an important consideration when thinking about small positive changes.

Evaluating the results in Table 2, one can discern the relative impacts of various factors on wage inequality between 1986 and 2015. Unionization consistently decreased wage disparities; a rising proportion of unionized workers led to notable wage gap reductions, although its influence waned over time, pointing to evolving workplace dynamics and the role of unions. Public sector employment had a marginal positive effect in 1986, but this

[25] Appendix C confirms that the results are robust to the choice of the polynomial order.

effect disappeared by 2015. In contrast, the manufacturing sector, initially worsening wage inequality in 1986, began to mitigate it by 2015.

Table 2 shows that potential experience has a modest but steady role in reducing wage disparities. Marital status steadily helps narrow the wage gap, while urban residency's impact on inequality slightly diminished over time. Race and gender factors, although negative, have seen reduced adverse impacts. However, their ongoing effects highlight the necessity for persistent interventions.

Higher education levels, particularly college education, have a significant impact on reducing wage inequality. An increase in college-educated workers notably lowers the conditional Gini index, more so than high school education, reflecting the premium on college education in wage determination. The results in Table 2 not only highlight the variables impacting wage inequality in the US during the study period but also reveal the changing magnitude and direction of these impacts. Some factors actively reduce the wage gap, while others highlight ongoing areas for intervention. The next subsection explores these temporal shifts in wage distribution.

5.2 Temporal Changes in the Distribution of (Log) Wages

In this subsection, I explore wage distribution changes over time, contrasting my method with the MM algorithm. I split the interval (0,1) into four sub-intervals (Q1, Q2, Q3, Q4) and use a seventh-degree polynomial approximation, alongside a repetition count of 1,000. For the MM method, I follow the guidelines from subsection 2.2.1, setting m to 4,500 and using $\alpha(\cdot)$ as the quantile statistic. This comparison aims to highlight the advantages of my proposed approach.

Table 3a presents MM decomposition results for wage data from 1986 and 2015, showing quantiles from the 1st to the 99th percentile of the estimated log-transformed wage distributions, capturing a wide spectrum of the wage distribution. The third column

specifically reports the changes in these quantiles, including both point estimates and 95% bootstrap confidence intervals derived from 1,000 bootstrap samples.

Column 3 of Table 3a reflects overall wage changes from 1986 to 2015, revealing the MM method's limitations in capturing nuanced inequality changes. It shows significant wage increases at the distribution's extremes, with 30.5% growth at the lowest quantile and 29% at the highest. However, the MM method's focus on discrete quantiles falls short in providing a complete picture of inequality changes across the wage spectrum, as indicated by the marginal increase in the Gini coefficient from 11.04 to 11.09.

Columns 4 to 6 of Table 3a dissect the total wage distribution changes into three parts: covariate changes (Equation 16), variations in returns (Equation 17), and a residual component (Equation 18). This breakdown analyzes the factors affecting wage changes at different income levels. However, interpreting these results in terms of inequality is complex. The analysis shows mixed outcomes, with some quantiles experiencing positive effects and others negative, making it difficult to ascertain a clear overall impact on wage inequality.

The analysis reveals contrasting effects of covariates on wages across income levels. Changes in covariates have little effect on wages at lower percentiles (1st and 10th), but significantly increase wages at the 25th percentile and above, benefiting higher-income individuals. The impact of returns on characteristics varies: positive at both distribution ends, indicating the value of certain attributes or skills, but negative at the median, implying average earners benefit less or other factors are at play.

Finally, Columns 7 to 15 in Table 3a indicate varied wage impacts of individual covariates across income levels. Unionization negatively affects higher wage percentiles (75th and 90th), while manufacturing shows minimal impact. Nonwhite workers experience wage disadvantages across a broad range (10th to 90th percentiles), and women's wage impact is mixed, with disadvantages at higher wages but gains at the lowest quantile. Urban living significantly impacts only the 90th percentile. Crucially, higher education, particularly college degrees, positively affects wages across most percentiles, underscoring its role in wage growth.

Table 3b provides a detailed breakdown of the shifts in the wage distribution, employing the additive decomposition approach to the Gini index as outlined in Equation (14). The format of this table mirrors that of Table 3a, where each cell offers two pieces of information: the initial entry denotes the point estimate, while the subsequent entry indicates the 95% bootstrap confidence intervals derived from 1,000 bootstrap samples. Within the context of this table, a *negative* point estimate suggests a *decrease* in the Gini coefficient, signaling a *decline* in wage inequality. Hence, any negative figures within the table can be interpreted as factors contributing to a more equitable wage distribution.[26]

Table 3b analyzes wage inequality changes from 1986 to 2015 across quartiles. It reveals decreased inequality for lower wage earners (Q1 and Q2), with a significant reduction in Q1. Conversely, the top earners (Q4) experienced increased inequality, indicating growing disparities at higher income levels. The proposed method's advantage is evident, as it shows a significant decrease in inequality at the lower end, almost balancing the increase at the higher end, thus providing a clearer understanding of how wage inequality has evolved over time.

Table 3b shows the effects of covariates and the associated returns on wage inequality. Covariates have no significant impact in the lowest quartile (Q1) but are influential in the second to fourth quartiles (Q2-Q4), affecting middle to upper income brackets. There is a consistent, significant increase in wage inequality across all quartiles due to changes in returns to certain characteristics, more so in the higher wage quartiles (Q3 and Q4). This indicates that certain skills or attributes are increasingly valuable, especially in the higher wage sectors, leading to a widening wage gap.[27]

[26] Tables F.1 and F.2 in Appendix F present the results obtained by using the FFL method.

[27] A detailed comparison between my method and FFL is presented in Section F.4 of Appendix F. The FFL approach primarily identifies an increase in inequality in the middle of the distribution but finds little change at the lower and upper tails or in the overall Gini index. In contrast, my method not only captures the same increase in mid-distribution inequality but also uncovers a significant reduction in inequality at the lower quartile and a notable rise at the upper quartile. Moreover, while the FFL method does not detect statistically significant covariate effects on the Gini index, my approach identifies meaningful contributions of individual factors to both the overall Gini and specific quartiles. These differences

Table 3b reveals that unionization and manufacturing significantly impact wage inequality across all quartiles, with changes in these characteristics leading to increased disparities. This reflects broader economic shifts, including the decline in unionized jobs and the transformation of manufacturing due to globalization and automation. These changes result in divergent wage outcomes, with higher-skilled workers benefiting while those with fewer skills may face stagnation. My methodology, unlike the MM method, precisely isolates these changes in returns, offering a clearer understanding of how these sectors contribute to growing wage disparities.

Regarding demographics, the analysis separates the impact of race and gender. Table 3b shows that racial disparities, represented by the Nonwhite variable in Column 9, contribute to increasing wage inequality consistently across all quartiles, suggesting that wages are diverging along racial lines. However, the gender factor, analyzed through the Women variable in Column 10, does not manifest a significant effect on wage inequality, indicating that gender by itself may not be a dominant factor in wage disparity within the scope of this analysis.

Column 11 of Table 3b shows that urbanization consistently reduces wage inequality across all quartiles, suggesting that it acts as an equalizer in wage distribution. Urban areas, offering diverse job opportunities and higher earning potential, appear to enable more equitable wage dispersion. This finding highlights urbanization as a key factor in reducing wage disparities, a nuance uniquely captured by my proposed method, unlike the MM method.

Lastly, the influence of educational attainment, particularly the possession of a College Degree shown in Column 15, displays a robust negative and significant impact on inequality, with the most substantial effects observable in the uppermost quartile and the aggregate. The importance of a college education in the contemporary job market is underscored by

arise because the FFL method relies on unconditional quantiles and is sensitive to shifts in the wage distribution, whereas my method, based on conditional quantile regression, provides a more precise decomposition of inequality within groups, making it better suited for analyzing within-group wage disparities.

this trend; as college-educated workers are increasingly favored, the resulting wage premium compresses the upper tail of the wage distribution, culminating in a reduction of inequality at the higher wage levels. This underscores the critical role of higher education in the quest to diminish wage inequality.

In summary, my method provides a detailed analysis of wage distribution changes, offering deeper insights than the MM method. It effectively captures wage dynamics across income levels and assesses the impact of socioeconomic factors like unionization, manufacturing shifts, reflecting the complex interplay of globalization, technological advancement, and skill levels within the workforce. The analysis reveals that urbanization reduces wage inequality and higher education, particularly for higher earners, plays a crucial role in equalizing wages. This comprehensive approach uncovers various factors driving wage inequality, providing valuable evidence for policymakers to address wage disparities.

5.2.1 Discussion

As I conclude my analysis, I find it crucial to integrate my research outcomes with the widely held views on how education affects wage inequality. Consider an economy divided into two types of workers, the low-skilled and the high-skilled, with education measuring their skill level. The average wages for low-skilled workers, which I will call w_L, contrast with the w_H earned by their high-skilled counterparts. The educational premium, reflected by the w_H/w_L ratio, is something one can roughly estimate by regressing log wages on years of education.

As the proportion of high-skilled workers rises, it triggers a significant drop in wage inequality, and this happens through two main forces. The price effect kicks in first: the more these high-skilled workers flood the market, the more their relative wages start to fall. Then there is the composition effect: a larger slice of our workforce is climbing up to the high-skill, better-paid ranks, which naturally compresses the wage gap. The results show that educational progress is not just theoretical but a tangible lever for lessening wage disparities.

6 Conclusion

This study presents a new method to examine US wage disparities (1986-2015), decomposing the conditional Gini index through the conditional Lorenz curve and quantile function. Utilizing data from the CPS ORG, the key findings include the evolving role of manufacturing in wage dynamics, reduced but persistent race and gender wage impacts, and significant inequality reduction through higher education, especially college degrees. This methodological advancement offers new insights into wage distribution and inequality.

This study enhances the literature on decomposition methods by using the conditional Gini coefficient to measure inequality in log wage distribution directly, bypassing density modeling. Traditional methods relied on kernel estimates or selected quantile analyses, often overlooking comprehensive inequality measures. The proposed approach, assuming a linear conditional quantile function for log wages, reveals how the impacts of different factors on wages have evolved. Notably, when compared with the MM algorithm, this method not only aligns with MM's conclusions but also uncovers aspects of wage inequality that the MM method did not distinctly identify.

A limitation of this study, as an analog to Oaxaca (1973) decomposition, is the presumption that changes in the characteristics do not modify the returns of those characteristics. Moreover, the analysis only accounts for changes in the covariates from 1986 to 2015, but the proposed decomposition technique could have considered counterfactual scenarios in reverse order. More importantly, the linear decomposition works for a particular transformation of wages for which the conditional quantile functions are assumed to be linear in parameters (i.e., log wages), but this may not be a natural scale to analyze the distribution disparity.

Further research could usefully explore how to account for the general equilibrium effects given changes in the distribution of the covariates, because those changes will also affect the returns to the characteristics. Moreover, a future study investigating different counterfactual scenarios and more recent years of analysis would be very interesting. A natural progression of this work is to extend the proposed method to the untransformed

variable (i.e., wages) to address questions related to the inequality of the distribution of the variable in levels.

Data Availability Statement

The data and replication package supporting the findings of this study are available at https://doi.org/10.7910/DVN/HJEVTW and on my website. The replication package includes code to reproduce all tables and figures using my proposed method, as well as the Machado and Mata (2005) and Firpo, Fortin, and Lemieux (2018) methods.

The MORG data used in this study is publicly accessible at http://www.nber.org/morg/annual/, and CPS basic monthly files are available at https://www.census.gov/data/datasets/time-series/demo/cps/cps-basic.html.
Additionally, the CEPR code utilized in this research can be downloaded from https://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/cps-org-data/.

This comprehensive data and code availability ensure that the analysis can be fully replicated and facilitate transparency in research.

Disclosure of Interest:

The author declares no competing interests.

Funding:

No funding was received for this research.

References

Abel, J. R., & Deitz, R. (2019). Why are some places so much more unequal than others? Federal Reserve Bank of New York Economic Policy Review, 25(1), 58-75.

Acemoglu, D., & Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Handbook of labor economics (Vol. 4, pp. 1043-1171). Elsevier.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the US wage structure. Econometrica, 74(2):539–563.

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

Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The China syndrome: Local labor market effects of import competition in the United States. American economic review, 103(6), 2121-2168.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. The Quarterly journal of economics, 118(4), 1279-1333.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in us wage inequality: Revising the revisionists. The Review of economics and statistics, 90(2):300–323.

Bahr, P. R., Dynarski, S., Jacob, B., Kreisman, D., Sosa, A., & Wiederspan, M. (2015). Labor Market Returns to Community College Awards: Evidence from Michigan. A CAPSEE Working Paper. Center for Analysis of Postsecondary Education and Employment.

Bayer, P., & Charles, K. K. (2018). Divergent paths: A new perspective on earnings differences between black and white men since 1940. The Quarterly Journal of Economics, 133(3), 1459-1501.

Blau, F. and Kahn, L. (1996). International differences in male wage inequality: Institutions versus market forces. Journal of Political Economy, 104(4):791–837.

Blinder, A. (1973). Wage discrimination: Reduced form and structural estimates. Journal of Human Resources, 8:436–455.

Blum, B. S. (2008). Trade, technology, and the rise of the service sector: The effects on US wage inequality. Journal of International Economics, 74(2), 441-458.

Bound, J. and Johnson, G. (1992). Changes in the Structure of Wages in the 1980's: An Evaluation of Alternative Explanations. American Economic Review, 82(3):371–92.

Buchinsky, M. (1994). Changes in the us wage structure 1963-1987: Application of quantile regression. Econometrica: Journal of the Econometric Society, pages 405–458.

Card, D. (1999). The causal effect of education on earnings. Handbook of labor economics, 3, 1801-1863.

Card, D. and Lemieux, T. (2001). Can falling supply explain the rising return to college for younger men? a cohort-based analysis. The Quarterly Journal of Economics, 116(2):705–746.

Chongvilaivan, A., & Hur, J. (2011). Outsourcing, labour productivity and wage inequality in the US: a primal approach. Applied Economics, 43(4), 487-502.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Econometrica, 64:1001–1044.

El Bantli, F. and Hallin, M. (1999). L 1-estimation in linear models with heterogeneous white noise. Statistics & probability letters, 45(4):305–315.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional quantile regressions. Econometrica, 77(3), 953-973.

Firpo, S. P., Fortin, N. M., & Lemieux, T. (2018). Decomposing wage distributions using recentered influence function regressions. Econometrics, 6(2), 28.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. Handbook of labor economics, 4:1–102.

Goldin, C., & Katz, L. F. (2007). The race between education and technology: The evolution of US educational wage differentials, 1890 to 2005.

Grosz, M. (2020). The returns to a large community college program: Evidence from admissions lotteries. American Economic Journal: Economic Policy, 12(1), 226-253.

Guvenen, F., Kuruscu, B., & Ozkan, S. (2014). Taxation of human capital and wage inequality: A cross-country analysis. Review of Economic Studies, 81(2), 818-850.

Hufe, P., Kanbur, R., & Peichl, A. (2022). Measuring unfair inequality: Reconciling equality of opportunity and freedom from poverty. The Review of Economic Studies, 89(6), 3345-3380.

Hühne, P., & Herzer, D. (2017). Is inequality an inevitable by-product of skill-biased technical change?. Applied Economics Letters, 24(18), 1346-1350.

Jepsen, C., Troske, K., & Coomes, P. (2014). The labor-market returns to community college degrees, diplomas, and certificates. Journal of Labor Economics, 32(1), 95-121.

Judd, K. (1998). Approximation methods. In Press, T. M., editor, Numerical Methods in Economics, volume 1. MIT Press Books, 1 edition.

Katz, L. F. (1999). Changes in the wage structure and earnings inequality. Handbook of labor economics, 3:1463–1555.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963-1987: Supply and demand factors. Quarterly Journal of Economics, 107:35–78.

Koenker, R. (2005). Inequality measures and their decomposition. In Quantile regression. Cambridge University Press, Cambridge.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. Econometrica: journal of the Econometric Society, pages 33–50.

Levy, F. and Murnane, R. J. (1992). Us earnings levels and earnings inequality: A review of recent trends and proposed explanations. Journal of economic literature, 30(3):1333–1381.

Machado, J. A. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of Applied Econometrics, 20:445–465.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. International Economic Review, 14:693–709.

Roca, J. D. L., & Puga, D. (2017). Learning by working in big cities. The Review of Economic Studies, 84(1), 106-142.

Rothe, C. (2015). Decomposing the composition effect: the role of covariates in determining between-group differences in economic outcomes. Journal of Business & Economic Statistics, 33(3), 323-337.

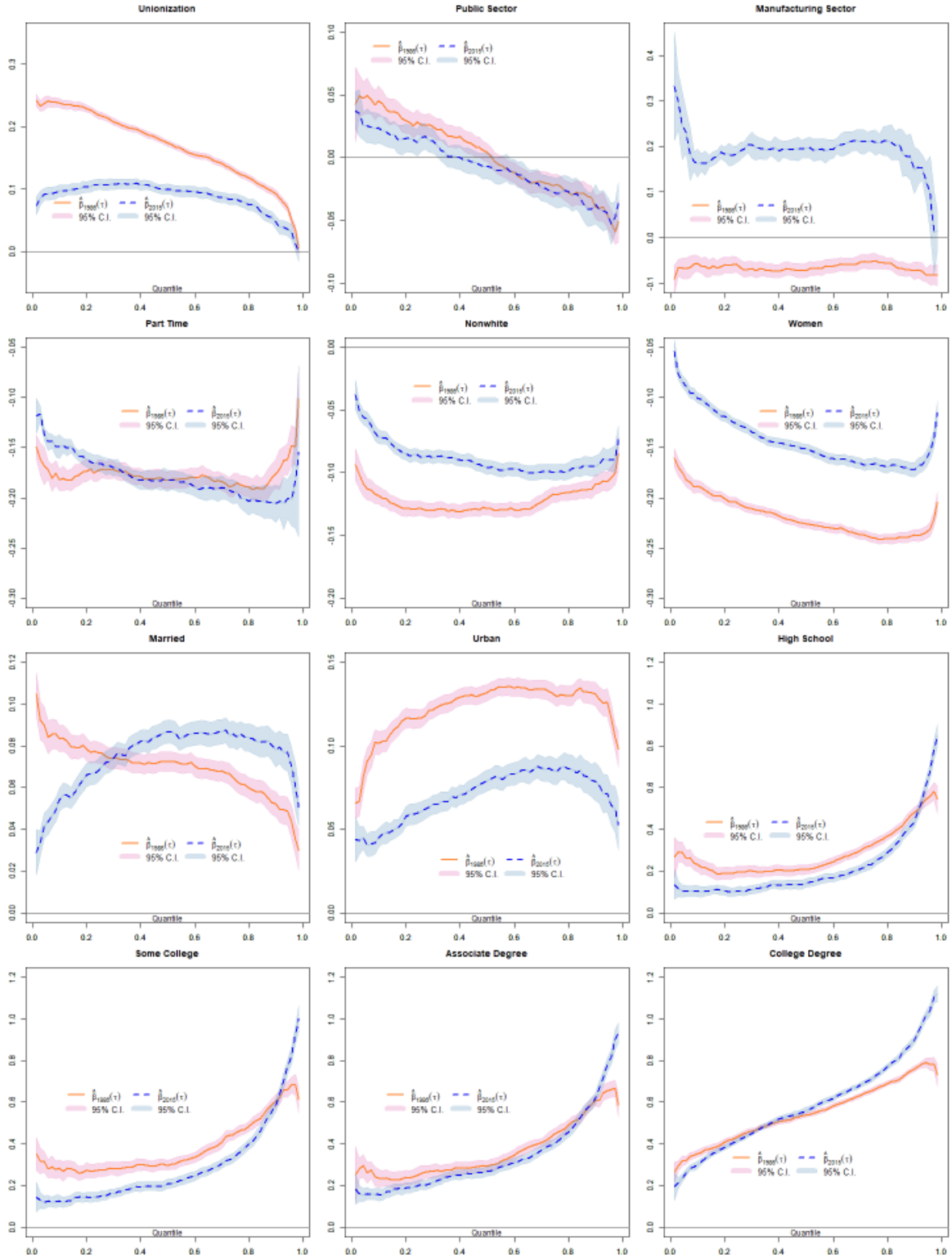
Schmitt, J. (2003). Creating a consistent hourly wage series from the Current Population Survey's Outgoing Rotation Group, 1979-2002. Version 0.9 (August). Washington DC: Center for Economic and Policy Research.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological), pages 683–690.

Stevens, A. H., Kurlaender, M., & Grosz, M. (2019). Career technical education and labor market outcomes: Evidence from California community colleges. Journal of Human Resources, 54(4), 986-1036.

Figures

Figure 1: Selected Coefficients Estimates From the Quantile Regression



Note: This figure compares 1986 and 2015 quantile regression coefficient estimates from Equation (22), calculated for 69 points in the (0,1) interval. It features solid lines for 1986 and dashed lines for 2015, with shaded areas indicating 95% confidence intervals using a Huber sandwich estimate and local sparsity estimation. The regression, utilizing CPS ORG hourly wage data, includes individual, job-related, demographic, and macroeconomic factors, plus state and industry fixed effects.

Tables

Table 1: Summary Statistics for the CPS 1980-2015

			Men					Women				
Year	Log Real Minimum Wage[A]	Manufact. Employ.	Log Real Wage[A]	Union[B]	Nonwhite	Education	Experience[C]	No. Obs.	Log Real Wage[A]	Union[B]	Nonwhite	Education	Experience[C]	No. Obs.
1980	2.28	0.26	3.20		0.17	12.75	18.26	106,932	2.79		0.18	12.76	17.64	87,888
1981	2.26	0.25	3.18		0.18	12.82	18.19	99,527	2.78		0.19	12.80	17.68	83,229
1982	2.20	0.24	3.18		0.18	12.92	18.23	92,249	2.80		0.19	12.91	17.68	79,594
1983	2.16	0.23	3.18	0.28	0.18	12.99	18.11	91,053	2.81	0.18	0.19	12.99	17.59	79,017
1984	2.12	0.23	3.17	0.26	0.18	13.02	17.91	92,732	2.82	0.17	0.19	13.03	17.53	80,796
1985	2.09	0.23	3.17	0.25	0.20	13.03	18.02	93,768	2.83	0.16	0.20	13.07	17.58	82,833
1986	2.07	0.22	3.19	0.24	0.20	13.07	17.94	92,085	2.86	0.16	0.20	13.12	17.64	83,459
1987	2.03	0.22	3.18	0.23	0.21	13.08	18.00	92,014	2.87	0.15	0.21	13.15	17.69	84,693
1988	1.99	0.21	3.17	0.23	0.22	13.11	17.99	88,096	2.87	0.15	0.21	13.19	17.78	81,205
1989	1.95	0.21	3.19	0.22	0.22	13.14	18.09	89,459	2.88	0.15	0.22	13.23	18.02	82,998
1990	2.02	0.20	3.17	0.21	0.23	13.12	18.05	93,498	2.88	0.15	0.23	13.27	18.01	87,323
1991	2.09	0.20	3.16	0.21	0.24	13.18	18.24	90,154	2.89	0.15	0.23	13.32	18.26	85,381
1992	2.06	0.19	3.15	0.21	0.24	13.00	18.55	88,364	2.90	0.15	0.23	13.14	18.64	84,530
1993	2.03	0.19	3.14	0.20	0.24	13.06	18.59	86,820	2.91	0.15	0.23	13.20	18.78	83,925
1994	2.00	0.19	3.14	0.20	0.24	13.09	18.62	82,362	2.92	0.15	0.24	13.24	18.81	80,375
1995	1.98	0.19	3.13	0.19	0.24	13.12	18.74	82,515	2.90	0.14	0.24	13.26	18.96	79,910
1996	2.06	0.18	3.12	0.19	0.26	13.12	18.98	73,043	2.90	0.14	0.25	13.29	19.08	71,486
1997	2.12	0.18	3.14	0.18	0.27	13.10	19.09	74,587	2.92	0.14	0.26	13.30	19.27	72,771
1998	2.10	0.18	3.18	0.18	0.27	13.13	19.22	75,596	2.95	0.13	0.27	13.32	19.34	73,483
1999	2.08	0.17	3.21	0.18	0.27	13.18	19.34	76,758	2.97	0.13	0.27	13.35	19.45	74,452
2000	2.05	0.17	3.22	0.17	0.28	13.18	19.47	77,723	2.98	0.13	0.28	13.36	19.60	75,175
2001	2.02	0.16	3.23	0.17	0.28	13.23	19.73	82,354	3.00	0.13	0.28	13.41	19.85	80,052
2002	2.00	0.15	3.24	0.16	0.28	13.26	20.01	87,802	3.03	0.13	0.28	13.46	20.09	86,481
2003	1.98	0.14	3.24	0.16	0.31	13.23	20.16	85,508	3.03	0.13	0.30	13.50	20.46	85,314
2004	1.95	0.13	3.24	0.15	0.32	13.25	20.25	84,298	3.03	0.13	0.30	13.54	20.55	83,302
2005	1.92	0.13	3.22	0.15	0.32	13.24	20.46	84,895	3.03	0.13	0.31	13.58	20.66	83,789
2006	1.89	0.13	3.22	0.14	0.33	13.26	20.51	85,018	3.03	0.12	0.31	13.60	20.79	82,926
2007	1.99	0.13	3.23	0.14	0.33	13.31	20.60	83,742	3.04	0.13	0.32	13.67	20.86	82,382
2008	2.06	0.13	3.23	0.15	0.33	13.39	20.81	82,316	3.05	0.13	0.32	13.75	21.03	81,683
2009	2.17	0.12	3.26	0.15	0.33	13.46	21.15	78,900	3.07	0.13	0.32	13.80	21.29	80,037
2010	2.15	0.12	3.24	0.14	0.33	13.49	21.26	78,101	3.07	0.13	0.32	13.85	21.43	78,973
2011	2.12	0.12	3.22	0.14	0.34	13.52	21.19	77,791	3.06	0.13	0.32	13.89	21.50	77,681
2012	2.10	0.12	3.22	0.13	0.35	13.56	21.35	78,104	3.05	0.12	0.35	13.92	21.51	76,849
2013	2.09	0.12	3.22	0.13	0.36	13.58	21.34	78,071	3.06	0.12	0.35	13.99	21.44	76,426
2014	2.07	0.12	3.21	0.13	0.37	13.59	21.35	78,762	3.06	0.12	0.36	14.02	21.34	76,565
2015	2.07	0.12	3.24	0.13	0.37	13.63	21.29	77,812	3.08	0.12	0.37	14.05	21.28	75,642

[A] 2020 Constant Dollars

[B] Union status of workers was not collected in the outgoing rotation group supplements from 1980 to 1982. However, using the May pension supplement it may be possible estimate this summary statistic for a subsample of the population

[C] Potential experience is computed as age - years of education - 5

Note: The table presents summary statistics for my refined sample from the CPS ORG, utilizing the CEPR Uniform Extracts to generate a consistent series of hourly wages. All wages have been adjusted to 2020 USD using the CPI series CUSR0000SA0. My focus is on workers aged 16 to 65, earning hourly wages between $1 and $100, adjusted to 1979 dollars. Potential experience is calculated by subtracting the number of years of education and an additional five years for elementary schooling from each individual's age. Both education and potential experience are expressed in years. The columns labeled "manufacturing," "union," and "nonwhite" represent the proportion of workers in manufacturing jobs, unionized positions, or those who did not identify as white, respectively. All summary statistics are weighted by the CPS sample weights.

Table 2: Impact Estimates of Selected Covariates

	1986	2015
Unionization status	0.071	0.032
	0.069; 0.073	0.029; 0.035
Public sector job	0.006	0.001
	0.003; 0.010	-0.002; 0.005
Manufacturing job	-0.022	0.066
	-0.028;-0.017	0.059; 0.074
Potential Experience	0.0078	0.0052
	0.0072;0.0084	0.0047;0.0058
Part time employee	-0.059	-0.056
	-0.062;-0.055	-0.059;-0.052
Nonwhite	-0.044	-0.028
	-0.046;-0.042	-0.030;-0.026
Female	-0.07	-0.041
	-0.072;-0.068	-0.043;-0.040
Married	0.027	0.023
	0.025; 0.028	0.021; 0.025
Urban area	0.041	0.021
	0.039; 0.043	0.018; 0.023
High school	0.081	0.04
	0.072; 0.090	0.034; 0.047
Some college	0.109	0.058
	0.098; 0.119	0.051; 0.065
Associate degree	0.099	0.076
	0.088; 0.111	0.069; 0.084
College degree	0.161	0.153
	0.157; 0.166	0.149; 0.158

Note: The table presents the impact of individual characteristics on wage inequality for years 1986 and 2017. A positive sign of the reported impact is associated with a reduction in the inequality of the distribution of (log) wages given a small increase in the corresponding characteristic. The first entry of each cell in the table presents the impact estimation, whereas the second reports the 95% bootstrap confidence interval. Each column exhibits the results for the corresponding year. I computed the figures using the estimation procedure described in section 3. The polynomial approximation order is 7 for both years. The bootstrap uses 1,000 repetitions for each year. For the bootstrap, in each iteration, I calculated the $\hat{\mu}_{ln}$ using the weighted average of the logarithm of real wages, adjusted to 2020 dollar value; Then, I compute the estimate $\widehat{\Pi}_j$ from Equation (21) and compute the impact estimate $\widehat{\Pi}_j/\hat{\mu}_{ln}$

Table 3a: Decomposition of Wage Changes (1986-2015) Using the MM Method

	Marginals			Aggregate Contributions			Individual Covariates								
	1986 (1)	2015 (2)	Change (3)	Covariates (4)	Returns (5)	Residual (6)	Unionization (7)	Manufacturing (8)	Nonwhite (9)	Women (10)	Urban (11)	High School (12)	Some College (13)	Associate Degree (14)	College Degree (15)
1st quant.	1.56	1.86	0.305	-0.027	0.288	0.044	0.030	0.038	-0.012	0.043	0.042	0.050	0.040	0.029	0.087
			0.284;0.328	-0.075;0.009	0.242;0.347		-0.014;0.083	-0.033;0.078	-0.057;0.058	0.001;0.092	-0.020;0.084	-0.004;0.086	-0.019;0.089	-0.030;0.071	0.017;0.146
10th quant.	2.15	2.29	0.132	0.015	0.082	0.035	-0.007	-0.010	-0.034	0.026	0.006	0.031	0.008	0.017	0.051
			0.132;0.136	-0.008;0.035	0.043;0.104		-0.034;0.017	-0.032;0.014	-0.063;-0.005	-0.003;0.051	-0.022;0.027	-0.001;0.053	-0.020;0.038	-0.009;0.040	0.027;0.073
25th quant.	2.47	2.53	0.066	0.029	0.005	0.032	-0.020	-0.016	-0.046	0.004	0.014	0.014	0.012	-0.004	0.046
			0.063;0.071	0.006;0.047	-0.020;0.028		-0.053;0.005	-0.043;0.009	-0.074;-0.021	-0.024;0.031	-0.011;0.038	-0.012;0.041	-0.016;0.035	-0.031;0.025	0.019;0.070
Median	2.87	2.94	0.065	0.087	-0.041	0.019	-0.019	-0.004	-0.040	0.004	0.025	0.032	0.005	-0.001	0.067
			0.053;0.066	0.062;0.107	-0.063;-0.013		-0.049;0.006	-0.033;0.022	-0.075;-0.015	-0.025;0.029	-0.004;0.051	0.006;0.058	-0.028;0.034	-0.031;0.026	0.039;0.098
75th quant.	3.30	3.39	0.089	0.131	-0.025	-0.016	-0.050	-0.034	-0.068	-0.022	-0.027	0.040	-0.023	-0.001	0.092
			0.079;0.098	0.101;0.157	-0.055;0.003		-0.091;-0.013	-0.068;0.002	-0.105;-0.036	-0.056;0.015	-0.065;0.009	0.004;0.077	-0.063;0.013	-0.039;0.033	0.051;0.125
90th quant.	3.63	3.81	0.179	0.130	0.023	0.026	-0.061	-0.045	-0.096	-0.047	-0.041	-0.010	-0.039	-0.016	0.067
			0.173;0.195	0.086;0.178	-0.026;0.066		-0.107;-0.026	-0.080;0.003	-0.139;-0.052	-0.085;-0.013	-0.077; 0.000	-0.056;0.038	-0.075;0.006	-0.059;0.028	0.023;0.121
99th quant.	4.08	4.37	0.290	0.134	0.105	0.052	0.004	0.009	-0.033	0.016	0.024	0.021	0.002	0.022	0.090
			0.290;0.290	0.058;0.197	0.040;0.186		-0.048;0.058	-0.041;0.081	-0.079;0.024	-0.038;0.078	-0.034;0.094	-0.034;0.107	-0.057;0.060	-0.036;0.080	0.025;0.150
Gini of LogW	11.04	11.09	0.05	-1.7019	2.8591	-1.106	-0.231	-0.096	-0.173	-0.321	-0.246	-0.176	-0.305	-0.088	0.013
			-0.554;0.664	-2.767;-0.577	2.394;3.327		-0.494;0.033	-0.356;0.169	-0.441;0.093	-0.590;-0.053	-0.511;0.020	-0.446;0.097	-0.567;-0.041	-0.349;0.173	-0.250;0.276

Note: The table presents the results of the MM decomposition applied to wage data. The first two columns list the estimated log-transformed wages in 1986 and 2015 for selected quantiles. The third column reports the changes in these quantiles over the period. Columns four to six break down the total changes in the wage distribution into segments associated with covariates, variations in returns, and the residual component. The final columns, seven to fifteen, illustrate the effects of specific individual covariates. All point estimates come with 95% bootstrap confidence intervals, derived from 1,000 bootstrap samples and displayed beneath each computed value.

Table 3b: Wage Distribution Shifts (1986-2015) Using an Additive Decomposition of Gini Index

	Marginals			Aggregate Contributions			Individual Covariates								
	Gini 1986 (1)	Gini 2015 (2)	Change (3)	Covariates (4)	Returns (5)	Residual (6)	Unionization (7)	Manufacturing (8)	Nonwhite (9)	Women (10)	Urban (11)	High School (12)	Some College (13)	Associate Degree (14)	College Degree (15)
Q1	1.82	1.68	-0.1460	-0.0127	0.0756	-0.2088	0.0143	0.0462	0.0223	0.0034	-0.0064	0.0237	-0.0155	0.0012	-0.0685
			-0.278:-0.021	-0.062:0.039	0.041:0.108		0.008:0.021	0.030:0.063	0.017:0.028	-0.006:0.013	-0.009:-0.003	0.015:0.032	-0.024:-0.007	-0.010:0.011	-0.094:-0.042
Q2	3.70	3.61	-0.0867	-0.2067	0.4840	-0.3640	0.0466	0.1263	0.0814	0.0127	-0.0236	0.0735	-0.0543	0.0042	-0.2783
			-0.301:0.126	-0.380:-0.023	0.394:0.573		0.026:0.068	0.083:0.171	0.061:0.104	-0.021:0.048	-0.035:-0.011	0.047:0.099	-0.085:-0.026	-0.034:0.039	-0.380:-0.173
Q3	3.69	3.81	0.1212	-0.5305	0.9510	-0.2993	0.0787	0.2128	0.1471	0.0241	-0.0468	0.1439	-0.1101	0.0084	-0.5660
			-0.113:0.355	-0.863:-0.179	0.807:1.096		0.044:0.115	0.139:0.288	0.111:0.188	-0.040:0.092	-0.068:-0.023	0.093:0.195	-0.172:-0.052	-0.068:0.077	-0.773:-0.351
Q4	1.83	1.99	0.1629	-0.9520	1.3486	-0.2336	0.1042	0.3024	0.2155	0.0366	-0.0730	0.2862	-0.2123	0.0152	-0.9412
			0.041:0.280	-1.470:-0.421	1.151:1.547		0.058:0.153	0.198:0.410	0.162:0.275	-0.060:0.139	-0.107:-0.035	0.185:0.387	-0.331:-0.100	-0.122:0.139	-1.285:-0.583
Total	11.04	11.09	0.0514	-1.7019	2.8591	-1.1058	0.2438	0.6877	0.4663	0.0768	-0.1499	0.5274	-0.3922	0.0291	-1.8540
			-0.554:0.664	-2.767:-0.577	2.394:3.327		0.136:0.357	0.450:0.932	0.351:0.596	-0.126:0.293	-0.219:-0.073	0.340:0.713	-0.611:-0.184	-0.234:0.266	-2.532:-1.149

Note: The table details the shifts in wage distribution from 1986 to 2015, employing an additive decomposition of the Gini index as detailed in Equation (14). Each entry in Table 3b comprises two parts: the point estimate followed by the 95% bootstrap confidence intervals, obtained from 1,000 bootstrap samples. Negative point estimates in this table imply a reduction in the Gini coefficient, indicating a decline in wage inequality. The table spans four quartiles and the entire wage distribution, offering insights into the dynamics of wage changes and the impact of diverse socioeconomic factors.

Appendices: For online publication only

A. The Lorenz Curve as an Expected Value

Consider Z, a continuous random variable with support in R. Let its cumulative distribution function be $F_Z(z)$ and its probability density function be $f_Z(z)$. Assume that the conditional expectation $E[z|z \leq a]$ exists and is finite for every $a \in R$. For a given τ in the interval (0,1), define the quantile function $Q_Z(t) = \inf\{z: F_Z(z) \geq t\} = F_Z^{-1}(t)$, and denote $z_\tau = Q_Z(\tau)$.

Then, the integral of $Q_Z(t)$ from 0 to τ can be expressed as follows:

$$\begin{aligned}\int_0^\tau Q_Z(t)dt &= \int_{-\infty}^{z_\tau} z f_Z(z)dz \\ &= \tau \int_{-\infty}^{z_\tau} z \frac{f_Z(z)}{F_Z(z_\tau)} dz \\ &= \tau \int_{-\infty}^{z_\tau} z f_{Z<z_\tau}(z)\, dz \\ &= \tau E[z|z \leq z_\tau].\end{aligned} \tag{23}$$

From Equation (23), it becomes evident that the integral of $Q_Z(t)$ from 0 to 1 is equal to the expected value of z.

Moreover, $\forall \tau \in (0,1)$

$$E[z|z \leq z_\tau] = \int_{-\infty}^{z_\tau} z \frac{f_Z(z)}{F_Z(z_\tau)} dz \leq z_\tau \int_{-\infty}^{z_\tau} \frac{f_Z(z)}{F_Z(z_\tau)} = z_\tau$$

and,

$$z_\tau = z_\tau \int_{z_\tau}^{\infty} \frac{f_Z(z)}{1 - F_Z(z_\tau)} dz \leq \int_{z_\tau}^{\infty} z \frac{f_Z(z)}{1 - F_Z(z_\tau)} dz = E[z|z \geq z_\tau].$$

Then, $\forall \tau \in (0,1)$

$$0 \leq (1 - \tau)(E[z|z \geq z_\tau] - E\,[z|z \leq z_\tau]),$$

which implies

$$E[z|z \leq z_\tau] \leq \tau E[z|z \leq z_\tau] + (1-\tau)E[z|z \geq z_\tau] = E[z]. \quad (24)$$

Let Y be a continuous and positive random variable with a cumulative density function $F_Y(y)$, quantile function denoted by $Q_Y(t) = \inf\{y: F_Y(y) \geq t\} = F_Y^{-1}(t)$, and $y_\tau = Q_Y(\tau)$. Assume that $0 < E[y] < \infty$. Let $h(\cdot)$ be a continuous and monotone function. Define $Z = h(Y)$ and $\mu_h = E[h(y)] = E[z]$. Assume that $h(\cdot)$ is such that $h(Y) \geq 0$ and $0 < \mu_h < \infty$. By the properties of the quantile function, $Q_{h(Y)}(t) = h\big(Q_Y(t)\big)$. Then, using Equation (23), the Lorenz curve of the transformed variable is given by

$$L_h(\tau) = \frac{1}{\mu_h}\int_0^\tau Q_{h(Y)}(t)dt = \frac{\tau E[h(t)|h(t) \leq h(y_\tau)]}{E[h(y)]}.$$

Using the inequality in (24), the transformed Lorenz curve takes values between 0 and 1.

By the definition of the Gini coefficient, we have

$$\begin{aligned} G_h &= 1 - 2\int_0^1 L_h(\tau)d\tau \\ &= 1 - \frac{2}{\mu_h}\int_0^1 \tau E[h(y)|h(y) \leq h(y_\tau)]\, d\tau. \end{aligned}$$

The Gini index, G_h, is always less than or equal to 1 because $h(Y) \geq 0$. Furthermore, based on the inequality presented in Equation (24), we can deduce that

$$E[h(y)|h(y) \leq h(y_\tau)] \leq E[h(y)],$$

which implies

$$\int_0^1 \tau E[h(y)|h(y) \leq h(y_\tau)]d\tau \leq \int_0^1 \tau E[h(y)]d\tau = \frac{1}{2}E[h(y)] = \frac{\mu_h}{2}.$$

In other words, the Gini index, G_h, is always positive.

B. Integral Approximation

The Gini index takes the form

$$G_h(x) = 1 - \frac{1}{\mu_h}\sum_{j=1}^{P} x_j \int_0^1 \int_0^\tau 2\beta_j(\tau)dt\, d\tau.$$

I use the family of Legendre polyonomy of degree K to approximate each quantile regression coefficient estimate, $\hat{\beta}_j(t)$:

$$\hat{\beta}_j(t) \approx \tilde{\hat{\beta}}_{j,K}(t) = \hat{\alpha}_0 p_0(t) + \cdots + \hat{\alpha}_K p_K(t) = \sum_{i=0}^{K} \hat{\alpha}_i p_i(t).$$

Finally, I approximate the impact of characteristic j as

$$\frac{2}{\mu_h}\sum_{i=0}^{K} \hat{\alpha}_i \int_0^1 \int_0^\tau p_i(t)\, d\tau.$$

To compute the vector $\hat{\alpha} = (\hat{\alpha}_0, \cdots, \hat{\alpha}_K)$, first create a fixed grid t_{n_g} with n_g equally-spaced points on (0,1). Using statistical software, compute n_g quantile regression coefficient estimates, one for each point on the grid. Represent these estimates as a $1 \times n_g$ vector $\hat{\beta}_j\left(t_{n_g}\right)$. In a similar fashion, define $p_i\left(t_{n_g}\right)$ as the $1 \times n_g$ vector that computes the i-th Legendre polynomial at each grid point. Subsequently, define the $n_g \times (K+1)$ matrix $P\left(t_{n_g}\right)$ with each $p_i\left(t_{n_g}\right)$ as its columns, resulting in $P\left(t_{n_g}\right) = \left[p_0\left(t_{n_g}\right), \cdots, p_K\left(t_{n_g}\right)\right]$. The values of $\hat{\alpha}_i$ are then computed as the scalars that minimize the squared error between $\hat{\beta}_j\left(t_{n_g}\right)$ and $P\left(t_{n_g}\right) \times \hat{\alpha}$. In other words, the entries of the vector $\hat{\alpha}$ are the OLS coefficient estimates from the model with $\hat{\beta}_j\left(t_{n_g}\right)$ as the dependent variable and the design matrix $P\left(t_{n_g}\right)$.

C. Accuracy of the Estimating Procedure

To better grasp the implications of the polynomial approximation's order, K, consider a hypothetical scenario. Let's say a researcher postulates a model for the conditional quantile function of wage logarithms, using the transformation $h(\cdot) = \ln(\cdot)$. Further, let's assume the researcher gauges a quantile regression coefficient, $\hat{\beta}_j(t)$, across a grid of $m = 69$ quantiles.

Figure C.1 illustrates a representative smooth least-square approximation of this assumed quantile regression coefficient, $\hat{\beta}_j(t)$. In each of the figure's panels, the dashed line represents $\tilde{\tilde{\beta}}_{j,K}(t)$. Panel (a) showcases the outcomes of employing a smoothed polynomial approximation of second degree. If our focus remains solely on the quantile regression's point estimate, a second-degree polynomial appears to inadequately represent the estimate. Yet, when accounting for the 95% confidence interval, it's evident that a second-degree polynomial might offer a reasonable approximation. On the other hand, Panel (b) depicts the outcome using a sixth-degree polynomial. Increasing the polynomial's degree enhances the approximation fidelity to the quantile regression coefficient's point estimate. Such an enhancement might be a preferable approach, depending on the research objectives.

The panels in Figure C.1 elucidate that utilizing a higher polynomial degree improves the approximation's adherence to the quantile regression's point estimate. However, even with a higher degree, the polynomial might not completely smooth out abrupt variations in the estimates. Such variations could likely arise from data scarcity for specific quantiles—typically at the extreme top or bottom 1% of wage distributions. Selecting the polynomial's degree presents a balancing act. While a greater degree enhances the approximation's precision—something that may be sought after—it doesn't always align with the primary objective of the approximation. Additionally, as demonstrated in Figure 1, if the smoothing polynomial resides within the confidence bands, then the approximation might be deemed '*satisfactory*', despite any minor discrepancies.

Moreover, let's keep the assumption that $h(\cdot) = \ln(\cdot)$ and assume further that the conditional quantile function of the logarithm of y can be modeled by a simple linear relation:

$$Q_{\ln(y)}(t|x) = \beta_0(t) + x\beta_1(t) + \varepsilon(t).$$

For simplicity, let's assume that the intercept is constant, e.g., $\beta_0 = 0.5$, and the slope parameter increases linearly with quantiles, e.g., $\beta_1(t) = 0.2 + 0.05t$. Under these simplifying assumptions, the numerator of the impact estimates would be:

$$\Pi_0 = 2\int_0^1 \int_0^\tau 0.5dt\, d\tau = 0.5$$

and

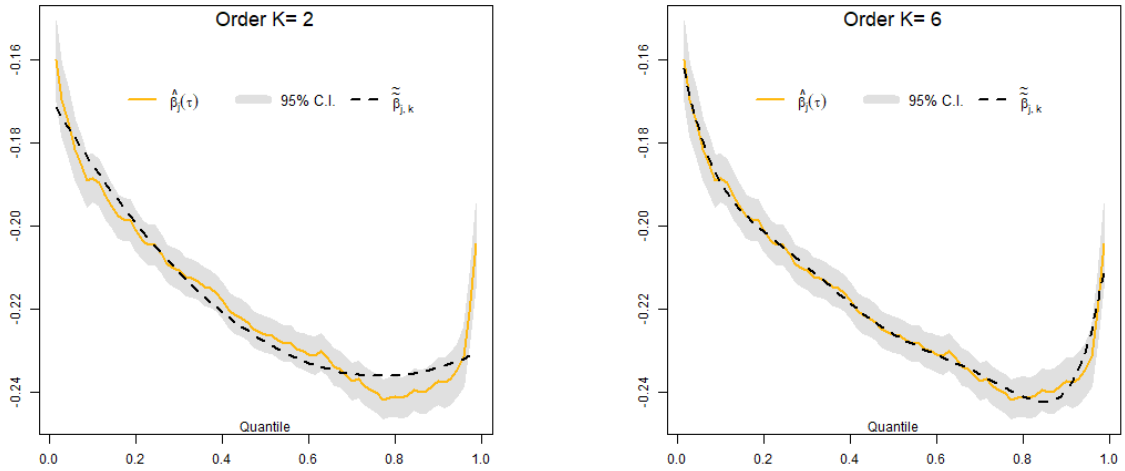
$$\Pi_1 = 2\int_0^1 \int_0^\tau 0.2 + 0.05tdt\, d\tau = 0.21667$$

To evaluate the precision of the estimation procedure, I generated a simulated dataset comprising $N = 1{,}000$ observations. For this dataset, the variable x is normally distributed, truncated at 20, and characterized by a mean of 35 and a standard deviation of 8. I set the error term to be uniformly distributed between zero and one, and to increase proportionally with x by 0.05 to achieve a slope that linearly rises with quantiles. Panel (a) of Figure C.1 showcases the simulated dataset, featuring selected estimated lines representing the conditional quantile functions. Panel (b) of the same figure displays the estimated conditional quantile regression coefficients marked with dots, in contrast to the actual conditional quantile regression coefficient, depicted as a solid line.

I estimate the numerator of the impact using the integral approximation detailed in Appendix B. Varying the order of polynomial approximation from K =2 to 10, I compute the 95% bootstrap confidence intervals using the 2.5th and 97.5th quantiles from 1,000 repetitions. Table C.1 presents the results of this performance test. The first column of the table displays the exact numerators of the impact estimates. Columns two through ten reveal the estimated results using Legendre polynomials of the corresponding orders. Each estimate's cell includes the 95% bootstrap confidence interval. The final two rows of the table feature the results of the hypothesis tests $\widehat{\Pi}_i \neq \Pi_i$ for i =0,1. The key insight from this table is the accuracy of the procedure and the minimal effect of the polynomial approximation's order on the accuracy of the estimation. This exercise validates the estimation procedure's effectiveness in quantifying the covariates' impact.

Figure C.1: Example of approximation using Legendre polynomials

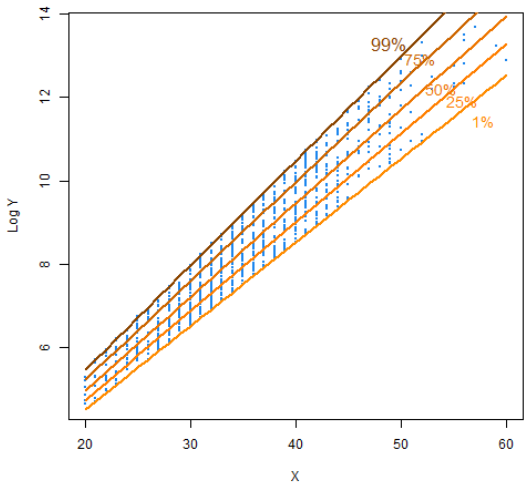
(a) Polynomial of degree 2 **(b)** Polynomial of degree 6



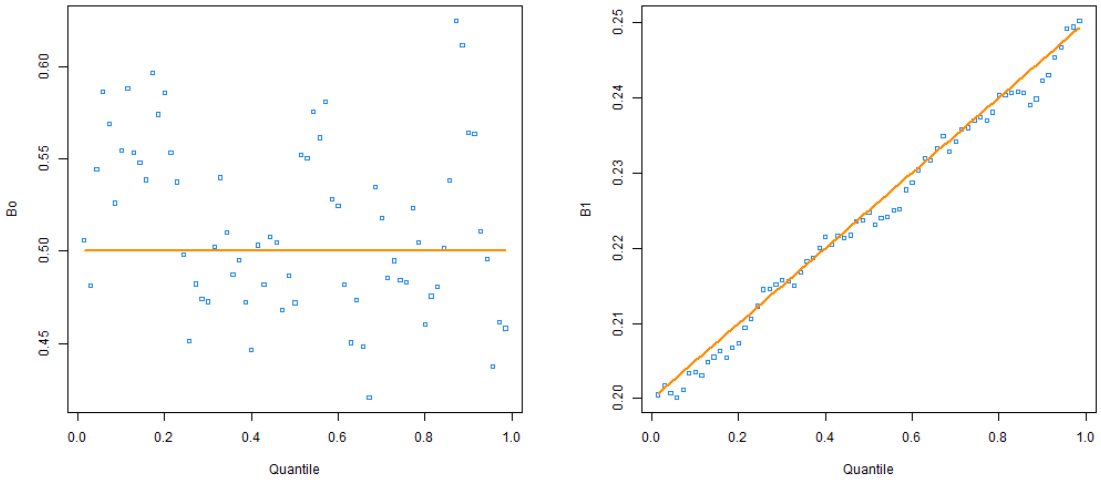
Note: The figure presents panels (a) and (b) to illustrative examples of least-square approximations on coefficients derived from a quantile regression, using Legendre polynomials of order two and six, respectively. In both panels, I showcase a generic quantile regression coefficient estimate, $\hat{\beta}_j(t)$, alongside its 95% confidence interval. The dashed line in each panel represents the least-square approximation using Legendre polynomials, $\widetilde{\widetilde{\beta}}_{j,K}(t)$, for $K = 2, 6$, respectively

Figure C.2:

(a) Simulated dataset and selected conditional quantile relationships



(b) Coefficient estimates and actual parameters



Note: Panel (a) showcases the simulated dataset, and panel (b) displays the estimated quantile regression coefficients. The dataset in panel (a) comprises $N = 1{,}000$ observations with the variable x following a normal distribution, truncated at 20. This distribution has a mean of 35 and a standard deviation of 8. The simulation's error term is uniformly distributed between zero and one, incrementally increasing in proportion to x by 0.05 to replicate a slope that linearly ascends with the quantiles. Panel (b) compares the estimated conditional quantile regression coefficients marked with dots, against the actual conditional quantile regression coefficient, shown as a solid line.

Table C.1: Performance Test Results Using Various Orders of Polynomial Approximation

	Exact Impact (1)	Order 2 (2)	Order 3 (3)	Order 4 (4)	Order 5 (5)	Order 6 (6)	Order 7 (7)	Order 8 (8)	Order 9 (9)	Order 10 (10)
Intercept	0.50000	0.51978	0.51994	0.51938	0.51909	0.51782	0.51833	0.51898	0.51822	0.51843
		0.3712;0.6683	0.3715;0.6684	0.3708;0.668	0.3712;0.667	0.3701;0.6655	0.3698;0.6669	0.3705;0.6675	0.3701;0.6663	0.3699;0.667
x	0.21667	0.21576	0.21575	0.21576	0.21577	0.21582	0.21580	0.21577	0.21580	0.21580
		0.21132;0.2202	0.21131;0.22019	0.21132;0.2202	0.21135;0.22019	0.2114;0.22023	0.21136;0.22025	0.21133;0.22022	0.21137;0.22023	0.21136;0.22024
$\widehat{\Pi}_0 \neq \Pi_0$		Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject
$\widehat{\Pi}_1 \neq \Pi_1$		Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject

Note: The table shows the results from the performance test. The numerator of the impact is estimated using various orders of polynomial approximation, ranging from $K = 2$ to 10. The first column provides the exact numerators of the impact estimates, while columns two through ten display the estimated results using Legendre polynomials of each specified order. Below each estimate, the corresponding 95% bootstrap confidence intervals are provided. The bootstrap confidence intervals are derived from 1,000 repetitions. The last two rows of the table show the results of the hypothesis test $\widehat{\Pi}_i \neq \Pi_i$ for $i = 0,1$, underlining the accuracy of the procedure and the minimal impact of the polynomial approximation's order on this accuracy.

D. Inequality in the Distribution of Y

I compute the linear decomposition of the Gini index in Equation (6) for the transformed variable, $h(Y)$. However, I also find it compelling to gauge the influence of individual characteristics on the distribution of the positive random variable, Y. Studying the transformed variable instead of the variable in its original scale might pose a conflict between the statistical and economic objectives of this study. But, considering the assumed properties of the transformation $h(\cdot)$ and using the properties of the quantile function, I derive:

$$\begin{aligned} Q_Y(t|x) &= h^{-1}\left(Q_{h(Y)}(t|x)\right) \\ &= h^{-1}(x^T\beta(t)), \end{aligned} \tag{25}$$

This implies:

$$L(\tau|x) = \frac{1}{\mu}\int_0^{\tau} h^{-1}(x^T\beta(t))dt \tag{26}$$

and

$$G(x) = 1 - \frac{2}{\mu}\int_0^1\int_0^{\tau} h^{-1}\left(x^T\beta(t)\right)dt\,d\tau. \tag{27}$$

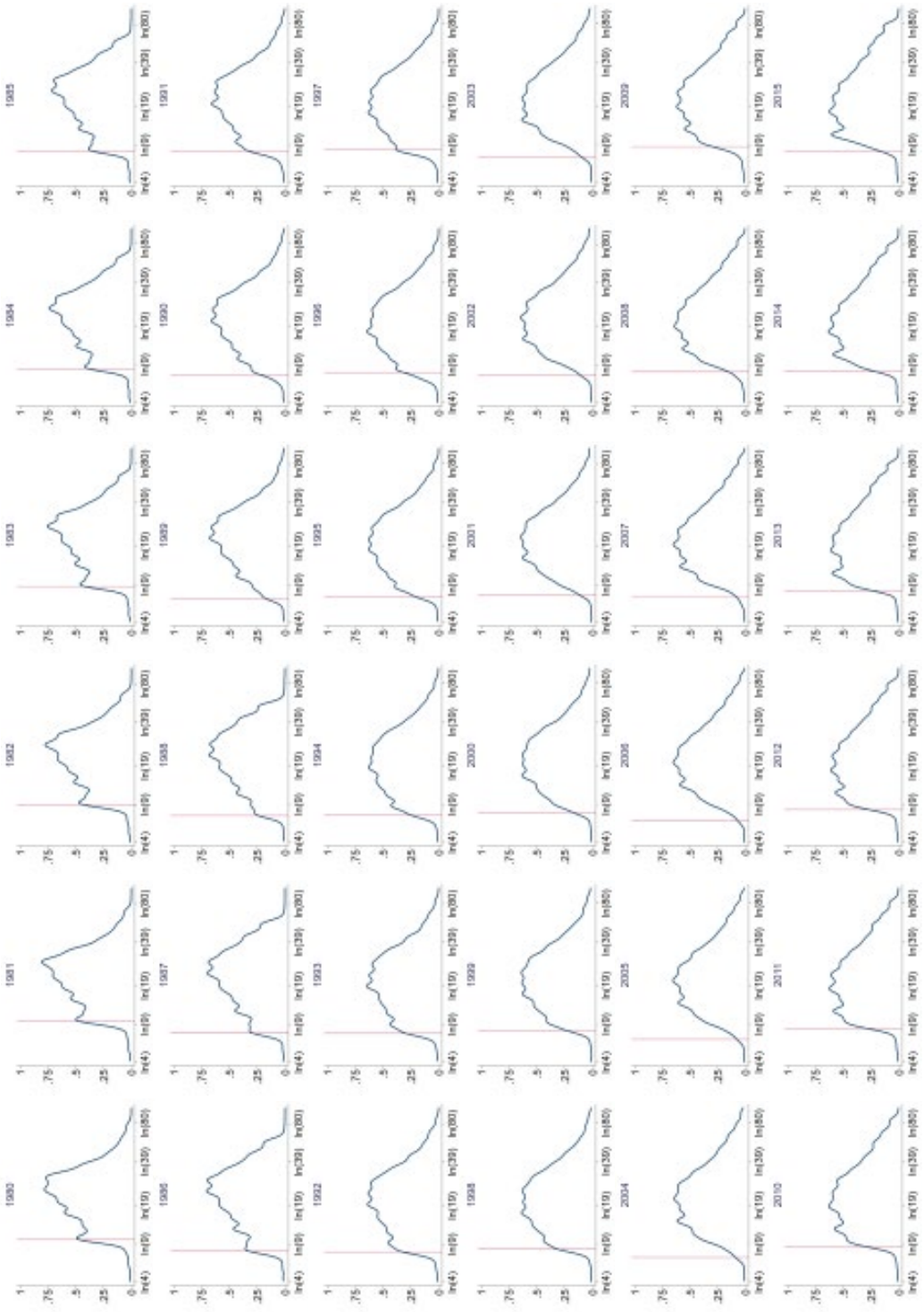
While the previous relationship is not inherently linear because $h^{-1}(\cdot)$ isn't linear, Equation (27) establishes a connection between the Gini coefficient of Y and a transformation of a linear combination of the quantile regression coefficients. I acknowledge that this link requires further exploration in future research.

E. Wage Distribution Disparities

To elucidate the wage distribution disparity both intra- and inter-gender, I present weighted kernel density estimates of the hourly wages for men and women spanning from 1980 to 2015 in figures E.1a and E.1b.[28] These graphs feature a vertical line indicating the respective (log) real minimum wage as referenced in column two of Table 1, which shows the concentration of wage distributions at the lower range. These representations clearly illustrate a significant expansion in the upper tail of the distribution in recent years compared to preceding periods. Moreover, I point out an evident broadening in the spread of hourly wages relative to the mean over time for both sexes, more pronounced in women's data. This visual representation aligns with previous findings by Levy and Murnane (1992), DiNardo et al. (1996), Katz (1999), and Autor et al. (2008).

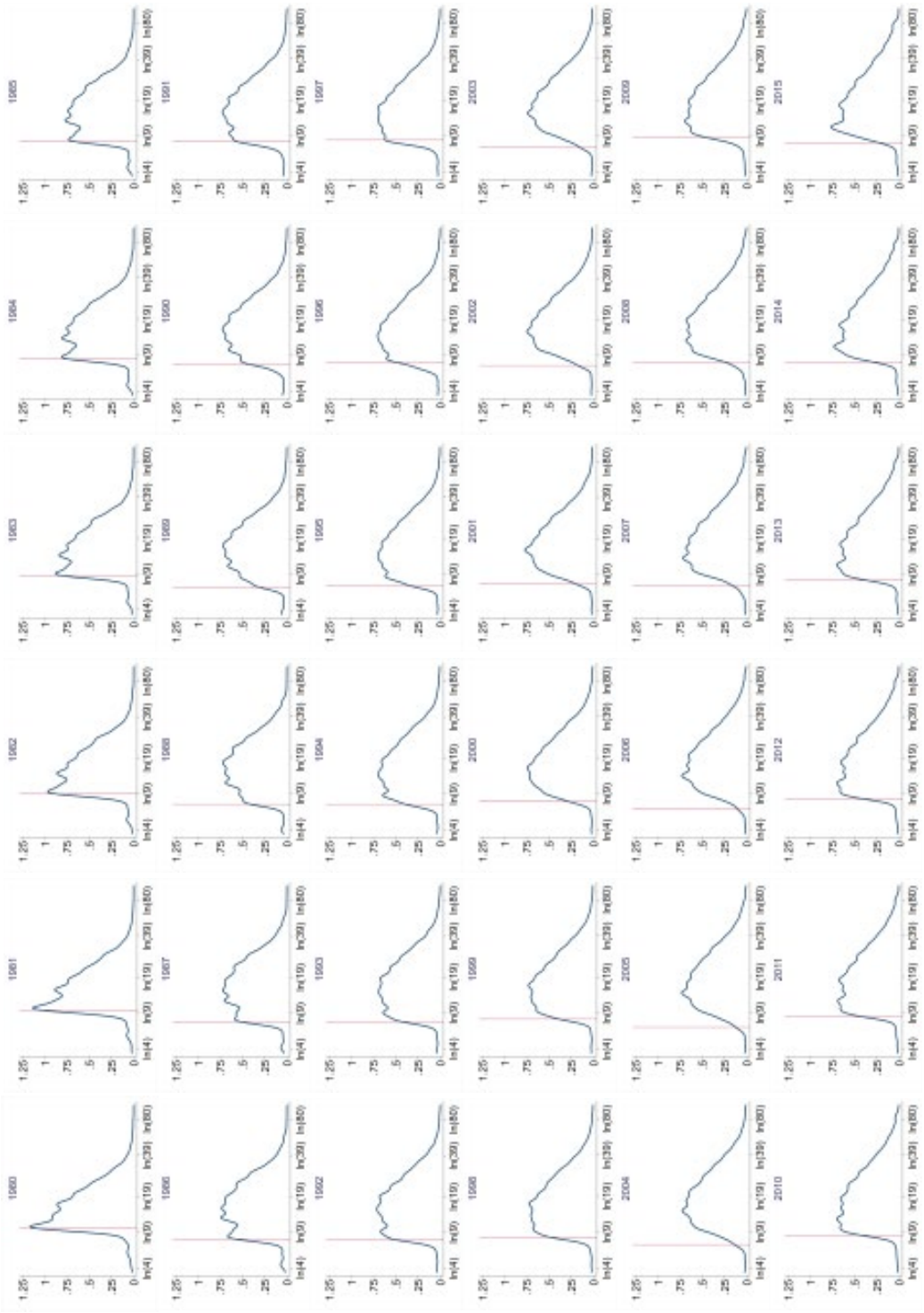
[28] These figures bear resemblance to those found in DiNardo et al. (1996), though a key distinction lies in my application of the CPS sample weights, in contrast to their implementation of hours-weighted kernel estimates. As in DiNardo et al. (1996), I determine the bandwidth for this estimation employing the Sheather and Jones (1991) method.

Figure E.1a: Kernel density estimates of men's (log) real wages 1980 - 2015 ($2020)



Note: The figure presents kernel density estimates for hourly (log) wages of men, covering the years 1980 through 2015. In each panel, a vertical line indicates the federal minimum wage for that specific year. I have converted all wage values to 2020 USD using the CPI series CUSR0000SA0. This figure draws on my sample from the CPS ORG, employing the CEPR Uniform Extracts to create a consistent hourly wage series. I have narrowed my focus to include workers aged 16 to 65, with hourly wages ranging from $1 to $100 in 1979 dollars.

Figure E.1b: Kernel density estimates of women's (log) real wages 1980 - 2015 ($2020)



Note: The figure presents kernel density estimates for hourly (log) wages of women, covering the years 1980 through 2015. In each panel, a vertical line indicates the federal minimum wage for that specific year. I have converted all wage values to 2020 USD using the CPI series CUSR0000SA0. This figure draws on my sample from the CPS ORG, employing the CEPR Uniform Extracts to create a consistent hourly wage series. I have narrowed my focus to include workers aged 16 to 65, with hourly wages ranging from $1 to $100 in 1979 dollars.

F. Decomposing Wage Distributions Using Recentered Influence Function Regressions

Recent research by Firpo, Fortin, and Lemieux (2009) introduces Unconditional Quantile Regression (UQR), a regression approach that examines how changes in the distribution of explanatory variables affect the unconditional quantiles of an outcome variable. Their method relies on regressing the Recentered Influence Function (RIF) of a given quantile on explanatory variables, allowing researchers to extend the Oaxaca-Blinder decomposition to study distributional effects beyond mean differences. This framework provides insight into how covariates contribute to wage dispersion at different points of the distribution, making it particularly valuable for analyzing inequality dynamics.

Building on this work, Firpo, Fortin, and Lemieux (2018), hereafter FFL, refine the UQR decomposition by introducing a reweighting procedure that decomposes wage changes into composition effects (shifts in the distribution of covariates) and wage structure effects (changes in how covariates relate to wages). Their approach enables a more detailed examination of wage inequality across quantiles. However, a key limitation of their method is that the RIF transformation depends on the full distribution of wages, making results sensitive to shifts in the distribution of covariates. This sensitivity can lead to specification errors due to nonlinearities.

The following sections detail the FFL decomposition method, discuss its identification strategy, and present my implementation using CPS ORG data from 1986 and 2015. I compare the FFL approach to my proposed method, which models conditional quantiles using conditional quantile regression. This comparison highlights key differences in how each method captures the role of covariates in shaping wage inequality, emphasizing the advantages of focusing on within-group variation.

F.1 Wage Decomposition

This section presents the decomposition method implemented by FFL. To keep the exposition consistent with the application, I will focus on the case where the outcome variable, Y, represents wages. The decomposition compares similar individuals across two points in time, t= 0 and t = 1.

Suppose we observe a random sample of $N = N_0 + N_1$ individuals, where N_0 and N_1 are the number of individuals at times t= 0 and t = 1, respectively. Let's index individuals by $i = 1, \cdots, N$. Let's denote by $y_{1,i}$ the wage that individual i was paid at time 1, and $y_{0,i}$ the wage paid at time 0. Therefore, for each individual i, the observed wage, y_i, can be defined as $y_i = y_{1,i} \cdot \Psi_i + \mathrm{y}_{0,i} \cdot (1 - \Psi_i)$, where $\Psi_i = 1$ if individual i was observed at time 1 and 0 otherwise.

The probability that an individual i is observed at time t = 1 is p. The conditional probability that an individual i is observed at t = 1, given their observed characteristics $x_i \in \mathbb{R}^K$, is defined as the propensity score, $p(X) = \Pr(\Psi = 1 \mid X = x)$. Wage determination depends on observed components, x_i, and unobserved components, $\epsilon_i \in \mathbb{R}$, through the wage structure functions:

$$y_{\mathrm{t},i} = g_{\mathrm{t}}(x_i, \epsilon_i), \qquad \mathrm{t} = 0,1,$$

where $g_{\mathrm{t}}(\cdot,\cdot)$ are unknown real-valued mappings such that $g_{\mathrm{t}}: \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}_+ \cup \{0\}$. Also, assume the joint distribution of (Y, Ψ, X) is unknown.

From observed data on (Y, Ψ, X), it is possible to identify the distributions $Y_1|\Psi = 1 \sim dF_1$ and $Y_0|\Psi = 0 \sim dF_0$ non-parametrically; estimating the counterfactual distribution $Y_0|\Psi = 1 \sim dF_C$ is essential for the decomposition. The counterfactual distribution F_C represents the wages that would have prevailed under the wage structure when $\Psi = 0$, but with the observed and unobserved characteristics for $\Psi = 1$.

To formalize the decomposition, let ν represent a functional of the conditional joint distribution of $(Y_1, Y_0)|\Psi$. Specifically, $\nu: \mathcal{F} \to \mathbb{R}$, where $\mathcal{F}$ is a class of distribution functions such that $\forall F \in \mathcal{F}$, $\|\nu(F)\| < +\infty$. Examples of ν include means, quantiles, variances, or Gini indices. Notice how these functionals use the entire distribution $F \in \mathcal{F}$.

The overall wage gap, measured in terms of the functional ν, is defined as:

$$\Delta\nu_O = \nu(F_1) - \nu(F_0) = \nu_1 - \nu_0.$$

Adding and subtracting the counterfactual distribution F_C, we can express the wage gap as:

$$\Delta\nu_O = \underbrace{\nu_1 - \nu_C}_{\text{change in covariates}} + \underbrace{\nu_C - \nu_0}_{\text{change in returns}} = \underbrace{\Delta\nu_R}_{\text{structure}} + \underbrace{\Delta\nu_X}_{\text{composition}},$$

Where, Δv_X reflects the effect of changes in the returns due to changes in the distribution of X, and Δv_R reflects changes in the wage structure functions $g_t(\cdot,\cdot)$, provided we can hold the distribution of observables and unobservables fixed for $\Psi = 1$.

F.2 Identification and Estimation of the Composition and Structure Effects

FFL proves the identification of the counterfactual distribution, F_C, based on two assumptions. First, the ignorability assumption: Conditional on X, the unobserved components, ϵ, are independent of Ψ. Second, the overlapping support assumption: $\forall X \in \mathbb{R}^K$, $0 < p(X) = Pr(\Psi = 1 \mid X) < 1$.

Under these assumptions, FFL shows that one can identify the parameters of the counterfactual distribution $Y_0|\Psi = 1 \sim dF_C$. The identification relies on three weighting functions: $\omega_1(\Psi) = \frac{\Psi}{p}$, which transforms features of the marginal distribution of Y into features of the conditional distribution of $Y_1 \mid \Psi = 1$; $\omega_0(\Psi) = \frac{1-\Psi}{1-p}$, which transforms features of the marginal distribution of Y into features of the conditional distribution of $Y_0 \mid \Psi = 0$; and $\omega_C(\Psi, X) = \left(\frac{p(X)}{1-p(X)}\right) \cdot \left(\frac{1-\Psi}{p}\right)$, which transforms features of the marginal distribution of Y into features of the counterfactual distribution of $Y_0 \mid \Psi = 1$.

FFL shows that, under the ignorability and overlapping support assumptions, the distribution function $F_t(y)$ for t = 0,1 can be expressed as

$$F_t(y) = \mathbb{E}[\omega_t(\Psi) \cdot \mathbb{I}\{Y \leq y\}], \qquad t = 0, 1$$

and the counterfactual distribution as

$$F_C(y) = \mathbb{E}[\omega_C(\Psi, X) \cdot \mathbb{I}\{Y \leq y\}].$$

The identification of Δv_R and Δv_X follows from the fact that these quantities can be expressed as functionals of the distributions obtained by weighting the observations with the weighting functions described before.

F.2.1 The RIF regressions

The Influence Function (IF) of a distributional statistic, $\nu(F)$, measures the sensitivity of the statistic to small changes in the underlying distribution F. The IF is defined as:

$$IF(Y;\nu,F) = \lim_{\epsilon\to 0} \frac{\nu\big((1-\epsilon)F + \epsilon\delta_Y\big) - \nu(F)}{\epsilon},$$

where δ_Y is a point mass at Y. Intuitively, the IF captures how much the statistic $\nu(F)$ would change if an infinitesimal fraction of the distribution F were replaced by a point mass at Y. By definition $\mathbb{E}[(y;\nu,F)] = 0$.

For a given outcome Y, the RIF is defined as:

$$RIF(Y;\nu,F) = \nu(F) + IF(Y;\nu,F),$$

which ensures that the expectation of the RIF equals $\nu(F)$. Different summary statistics would have different RIF transformations, and the RIF regression is the OLS estimate of the conditional expectation of the RIF transformation.

In the current context, one can assume linearity of the RIF as

$$m_{\mathrm{t}}^{\nu}(x) = X\gamma_{\mathrm{t}}^{\nu}, \qquad \mathrm{t} = 0,1,$$

and

$$m_C^{\nu}(x) = X\gamma_C^{\nu}.$$

where,

$$\gamma_t^{\nu} = (\mathbb{E}[XX'|\Psi = t])^{-1} \cdot \mathbb{E}[RIF(Y_t;\nu_t,F_t)X|\Psi = t], \qquad t = 0,1,$$

and

$$\gamma_C^{\nu} = (\mathbb{E}[XX'|\Psi = 1])^{-1} \cdot \mathbb{E}[RIF(Y_0;\nu_C,F_C)X|\Psi = 1].$$

F.2.2 Estimates of the Structure and Composition Effects

In this setup, the structure effect is estimated as

$$\Delta\nu_R = \mathbb{E}[X|\Psi = 1]'(\gamma_1^\nu - \gamma_C^\nu),$$

and the composition effect is

$$\Delta\nu_X = \mathbb{E}[X|\Psi = 1]'\gamma_C^\nu - \mathbb{E}[X|\Psi = 0]'\gamma_0^\nu = (\mathbb{E}[X|\Psi = 1]' - \mathbb{E}[X|\Psi = 0]')\gamma_0^\nu + Err^\nu,$$

where, $Err^\nu = \mathbb{E}[X|\Psi = 1]'(\gamma_C^\nu - \gamma_0^\nu)$ is the approximation error. The error arises because the FFL's regression-based procedure provides only a first-order approximation of the composition effect. In practice, this error can be estimated as the difference between the reweighting-based estimate of the composition effect ($\nu_C - \nu_0$) and the estimate obtained using the RIF regression approach, which is $(\mathbb{E}[X|\Psi = 1] - \mathbb{E}[X|\Psi = 0])'\gamma_0^\nu$. When the RIF regression approach accurately approximates the composition effect, the error should be close to zero. Therefore, the magnitude of this error serves as a specification test for the validity of FFL's regression-based procedure.

In the context of estimation using the RIF regression, a key challenge is the sensitivity of the method to changes in the distribution of X. In conventional regression analysis, OLS estimates can depend on the distribution of X when the conditional expectation of Y given X is nonlinear. The issue becomes more complex for RIF regressions, particularly for distributional statistics beyond the mean.

The RIF transformation depends not only on the outcome variable Y but also on the overall distribution of Y. When the distribution of X changes, it can shift the distribution of Y, which in turn affects the value of $RIF(Y; \nu, F)$ for a given Y. This dependency on F has direct implications for RIF regressions. Since the left-hand side of the regression is no longer the same transformation of Y, the coefficients in the RIF regression are also affected. In essence, changing the distribution of X indirectly alters the estimated relationship between X and Y by modifying the underlying RIF values.

Examining the approximation error provides a practical specification test for assessing the validity of FFL's regression-based procedure. While their method may perform well for specific statistics, it can encounter challenges for others, particularly those affected by significant non-linearities. These non-linearities arise because changes in the distribution of X can alter the distribution of Y, which in turn affects the RIF values used in the regression. This dependency can lead to

inaccuracies in the estimated coefficients of the RIF regression, highlighting the importance of careful evaluation when applying their method.

F.3 RIF Transformations for Unconditional Quantiles and the Gini Index

Quantiles are key distributional statistics that partition a distribution into intervals with equal probabilities. For a quantile q_τ at level τ, the RIF is defined as:

$$RIF(y; q_\tau, F) = q_\tau + \frac{\tau - \mathbb{I}(y \leq q_\tau)}{f_Y(q_\tau)},$$

where

τ is the chosen quantile level (e.g., $\tau = 0.5$ for the median), $\mathbb{I}(y \leq q_\tau)$ is an indicator function that equals one if y is smaller or equal to q_τ, and $f_Y(\cdot)$ is the probability density of Y, which must be estimated.

The RIF transformation adjusts Y to reflect its contribution to the quantile q_τ. The first term anchors the transformation at the quantile value, while the second term measures deviations from the quantile based on the density. This transformation allows researchers to analyze how covariates affect the unconditional quantile by regressing $RIF(y; q_\tau, F)$ on the covariates X. The resulting coefficients estimate the marginal effect of each covariate on the selected quantile, reflecting its contribution to shifts in that part of the distribution.

It is also possible to compute the RIF transformation for the Gini index. In Firpo, Fortin, and Lemieux (2009), they define the Generalized Lorenz Curve as

$$GL(p, F_Y) = \int_{-\infty}^{F^{-1}(p)} z dF_Y(z)$$

where $F^{-1}(p)$ is the quantile function associated with F_Y. The Gini index is related to the area under the Generalized Lorenz Curve as

$$G(F_Y) = 1 - \frac{2}{\mu}\int_0^1 GL(p, F_Y)dp,$$

Where $\mu = E(Y)$. Define:

$$R(F_Y) = \int_0^1 GL(p, F_Y)dp.$$

Then, the RIF for the Gini index evaluated at y is:

$$RIF(y; G, F_Y) = 1 + B(F_Y)y + C(y; F_Y)$$

where:

$B(F_Y) = 2\mu^{-2}R(F_Y)$ and $C(y; F_Y) = -2\mu^{-1}\{y[1 - p(y)] + GL(p(y), F_Y)\}$

The previous RIF transformation for the Gini index is based on the entire distribution F_Y. A different RIF transformation needs to be computed for a component of the Gini index; for example, to measure the disparities due to the bottom 25% of the distribution, the RIF transformation needs to account for the condition of being part of the lower quartile of the distribution. In some sense, modeling the conditional quantile function is an intermedia step that allows me to split the analysis by quartiles.

F.4 Empirical Implementation

I implemented the FFL decomposition using the hourly wage series I constructed from the CPS ORG data for years 1986 ($t = 0$) and 2015 ($t = 1$), spanning three decades of rising wage disparity. For the empirical analysis, I model two statistics, ν_t: a quantile τ and the Gini index, computed from the entire distribution of wages at times $t = 0$ and $t = 1$. I denote their corresponding RIF transformations as $RIF(y_t; q_{t,\tau}, F_t)$ for quantiles and $RIF(y_t; G_t, F_t)$ for the Gini index. I assume linearity of the RIF as

$$RIF(y_{t,isj}; \nu_t, F_t) = x'_{isj}\beta_t^{\nu} + \eta_{t,s}^{\nu} + \gamma_{t,j}^{\nu} + \varepsilon_{isj}, \qquad t = 0,1.$$

Here, s indexes states, and j indexes industries. To model the counterfactual distribution of $Y_0|\Psi = 1$, I use the appropriate reweighting functions and assume linearity of $RIF(y_{1,isj}; v_C, F_C)$ employing the same specification.

The vector x_{isj} includes job-related attributes such as unionization status, public sector employment, manufacturing employment, and part-time work status. Demographic characteristics encompass indicators for nonwhite, female, and marital status. Additional controls include a quadratic in potential experience, urban residency indicators, education categories, decade-based experience indicators, and interactions between education and experience. Finally, $\eta_{t,s}^{\nu}$ and γ_j^{ν} represent state and industry fixed effects, respectively.

Before estimating the RIF regressions, it is important to examine the wage density for irregularities that could affect the estimation of the RIF at key quantiles or the wage model specification. Figure F.1 shows kernel density estimates of hourly wages for 1986 and 2015, with bandwidths of 0.06 and 0.08, respectively. It also includes the 1986 density reweighted to match the 2015 distribution of characteristics. Notable features include cliffs at the lower end due to minimum wage effects and mid-distribution peaks from wage heaping, where workers round their wages to the nearest dollar. The impact of minimum wages is evident in Figure F.1, with vertical lines marking the binding federal minimum wage in 1986 and the binding state minimum wage in 2015. Since minimum wages are not explicitly modeled, the 1986 and reweighted densities overlap in those ranges, indicating that the included covariates may not fully capture wage-setting mechanisms. Thus, any effects observed at the lower end of the distribution should be interpreted with caution.

Figure F.2 presents coefficient estimates from RIF regressions for various quantiles, calculated at 95 points along the (0,1) interval. The dashed lines represent estimates for 1986, while the solid lines correspond to 2015. These regressions control for individual, job-related, demographic, and macroeconomic factors, as described before. The general trends in the coefficient estimates obtained using UQR largely mirror those observed in Figure 1 of the main text, which presents results from conditional quantile regression, with the notable exception of education.

Union membership exhibited a positive wage premium across all quantiles in 1986, but this premium diminished by 2015. For public sector employees, a wage premium is observed at the

lower end of the wage distribution, while a wage penalty appears at the upper end. These patterns remain similar in magnitude between 1986 and 2015. In contrast, manufacturing jobs were associated with lower wages in 1986 but became a positive determinant of wages across all quantiles by 2015, according to UQR estimates. Part-time, nonwhite, and female workers experienced wage penalties across the wage distribution, with the gender wage gap appearing more pronounced. Married workers consistently benefited from a wage premium, reinforcing the well-documented marriage wage effect. Additionally, an urban wage premium is present in both years, though it has declined over time, suggesting a narrowing of urban-rural wage disparities.

The RIF regression estimates for education highlight varying effects across the wage distribution. For high school, some college, and associate degrees, the impact is positive and increasing up to the 30th percentile, then declines toward the 80th percentile, turning negative at the top of the distribution. This hill-shaped pattern suggests these education levels boost wages at lower quantiles but reduce them at the upper end.

For college degrees, the pattern is also hill-shaped but consistently positive. The effect peaks around the 30th percentile, indicating strong returns for lower-wage earners, and then diminishes at higher quantiles, though it remains positive across the distribution. These findings show that education plays a key role in shaping wages but does so in a non-uniform way, with its effects on inequality depending on both the level of education and position in the wage distribution.

Unlike conditional quantile regression, which measures the effect of covariates on the conditional distribution of wages (i.e., wages given a specific set of characteristics), the UQR approach captures the total effect of changing the distribution of covariates on the overall wage distribution. While UQR can provide insights into how changes in education levels affect wage inequality across the entire labor market, the conditional quantile approach offers a complementary perspective by focusing on within-group dynamics. This focus on conditional distributions may be more aligned with understanding inequality within specific demographic or job-related groups, which is often central to discussions of wage disparities and targeted policy interventions.

F.4.1 Decomposition Results

Figure F.3, panel A, illustrates the overall change in (real log) wages at each percentile and decomposes this change into composition and wage structure effects using the reweighting

procedure. The overall change in wages exhibits a U-shaped pattern, with increasing wage dispersion at the top end of the distribution and declining wages at the lower end. The composition effects play a significant role in driving inequality, contributing to much of the observed increase in wage dispersion. Once composition effects are accounted for, the remaining wage structure effects—estimated using reweighting—exhibit a cleaner U-shaped pattern. Wage declines are concentrated in the middle of the distribution (20th to 80th percentile), while wage gains at both the top and bottom ends become comparable. Notably, composition effects alone cannot explain the U-shaped nature of wage changes, suggesting that structural factors are key drivers of this pattern.

Panel B of Figure F.3 moves to the next step of the decomposition, using RIF regressions to attribute the composition effect to specific sets of covariates. This panel compares the overall composition effect obtained through reweighting (shown in panel A) with the composition effect derived from RIF regressions. The difference between these two curves represents the specification error. While the error is relatively small for many quantiles, it is more pronounced at specific points, particularly in the upper 80th percentile and some middle and upper-middle quantiles (40th, 55th, 60th, and 65th percentiles). This suggests that the RIF regression may struggle to capture non-linear relationships, as changes in the distribution of covariates can alter the wage distribution, which in turn affects the RIF values used in the regression.

From an implementation perspective, relying on the conditional distribution through conditional quantile regression may offer certain advantages. The conditional quantile regression focuses on within-group changes in the wage distribution, which can provide more precise insights into how covariates affect wage inequality within specific demographic or job-related groups. This can be particularly useful for policy applications aimed at targeted interventions, where understanding within-group dynamics is crucial. In contrast, the unconditional approach provides a broader view but may be less suited to contexts where detailed group-specific effects are important.

Finally, I examine the impact of each factor on overall wage inequality, focusing on the 90–10 log wage differential, as well as the 50–10 and 90–50 differentials, which represent changes in the lower and upper parts of the distribution, respectively. Tables F.1 and F.2 present the decomposition of these differentials alongside the Gini of log wages. Table F.1 provides a simple

Oaxaca-Blinder-type decomposition using RIF regressions without reweighting, while Table F.2 includes results from the full reweighting procedure.

Table F.1 highlights a significant increase in inequality in the middle of the distribution, as measured by the total 90–10 log wage differential. This increase is driven by a rise in total composition effects and a decline in total structure effects. However, for the lower tail (50–10 gap), the upper tail (90–50 gap), and the entire distribution (Gini of log wages), the FFL method does not detect significant changes.

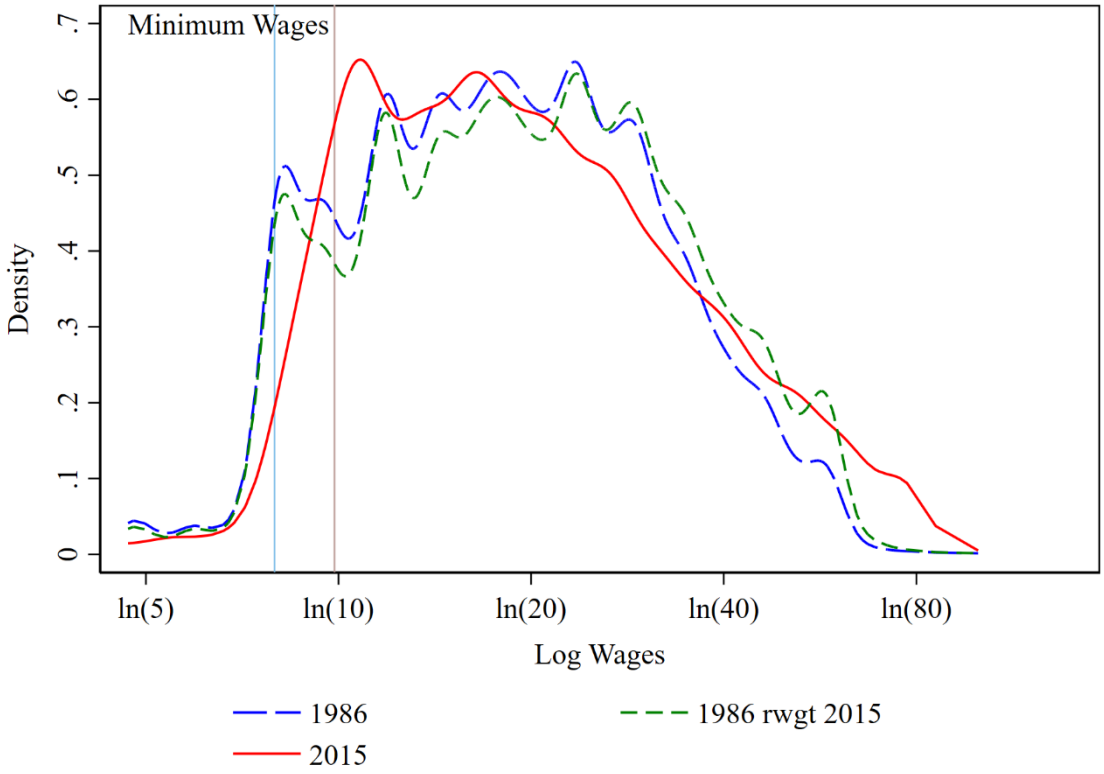
In contrast, my proposed method (as shown in Table 3b of the main text) also identifies a positive and significant increase in inequality in the middle of the distribution (Q2–Q3), driven by changes in composition (changes in returns) and mitigated by a reduction in structure effects (changes in covariates). Additionally, similar to the FFL procedure, my method finds a small and statistically insignificant change in the Gini coefficient for the entire distribution of log wages. However, unlike the FFL method, my approach disentangles a significant reduction in inequality in the lower quartile of log wages and a significant increase in the upper quartile.

Table F.1 also reveals notable changes in composition and wage structure effects in the middle of the distribution, primarily due to unionization, education, occupation, and industry. However, most factors are statistically insignificant for the lower tail (50–10 gap) and the upper tail (90–50 gap), except for changes in composition effects grouped by industry. Moreover, none of the factors show statistically significant effects on the Gini index for the entire distribution of log wages. In contrast, my proposed methodology identifies significant effects of covariates on the aggregate Gini index and individual quartiles of the distribution.

Table F.2 shows the results using the reweighting procedure. The reweighting procedure creates a counterfactual distribution of the wages that would have prevailed under the worker's characteristics in the year 1986, but with the distributions of observed and unobserved characteristics in year 2015. The magnitude, statistical significance, using the reweighting procedure are very similar to those a from the simple Oaxaca-Blinder decomposition using RIF regressions. In addition, the table shows the specification error for the composition effects as well as the reweighting error.

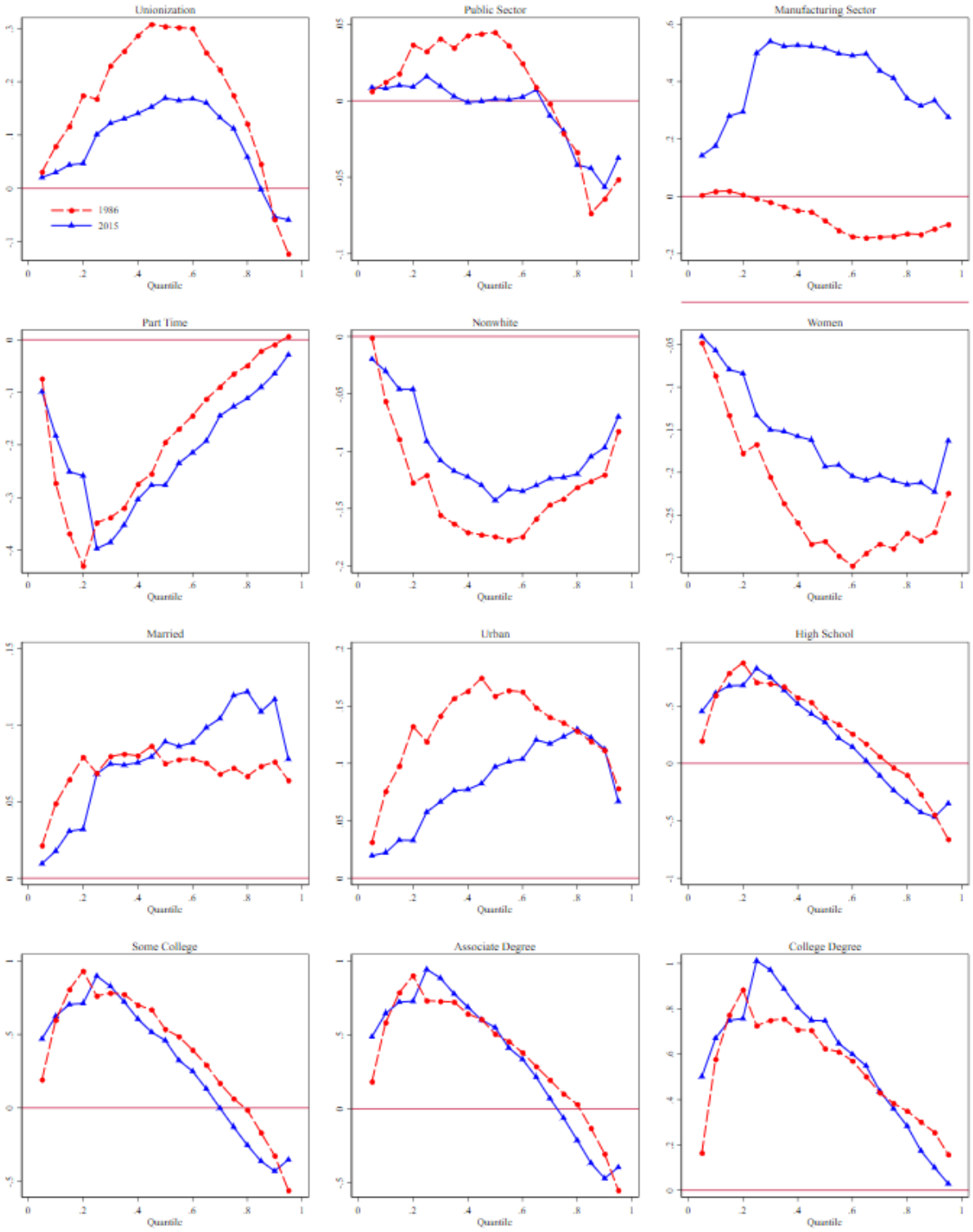
A key finding in Table F.2 is that the specification error is statistically significant, suggesting that the RIF regression may struggle to capture non-linear relationships. Changes in the distribution of covariates can alter the overall wage distribution, which in turn affects the RIF values used in the regression. This introduces a potential limitation in decomposing sources of inequality using the FFL method, as the results may be sensitive to shifts in the distributional shape, making it more challenging to precisely identify composition and structure effects. These issues highlight the importance of considering alternative approaches, such as conditional quantile regression, which directly models within-group wage variation and is less affected by distributional changes.

Figure F.1: Kernel density estimates of (log) real wages 1986 - 2015 ($2020) and counterfactual



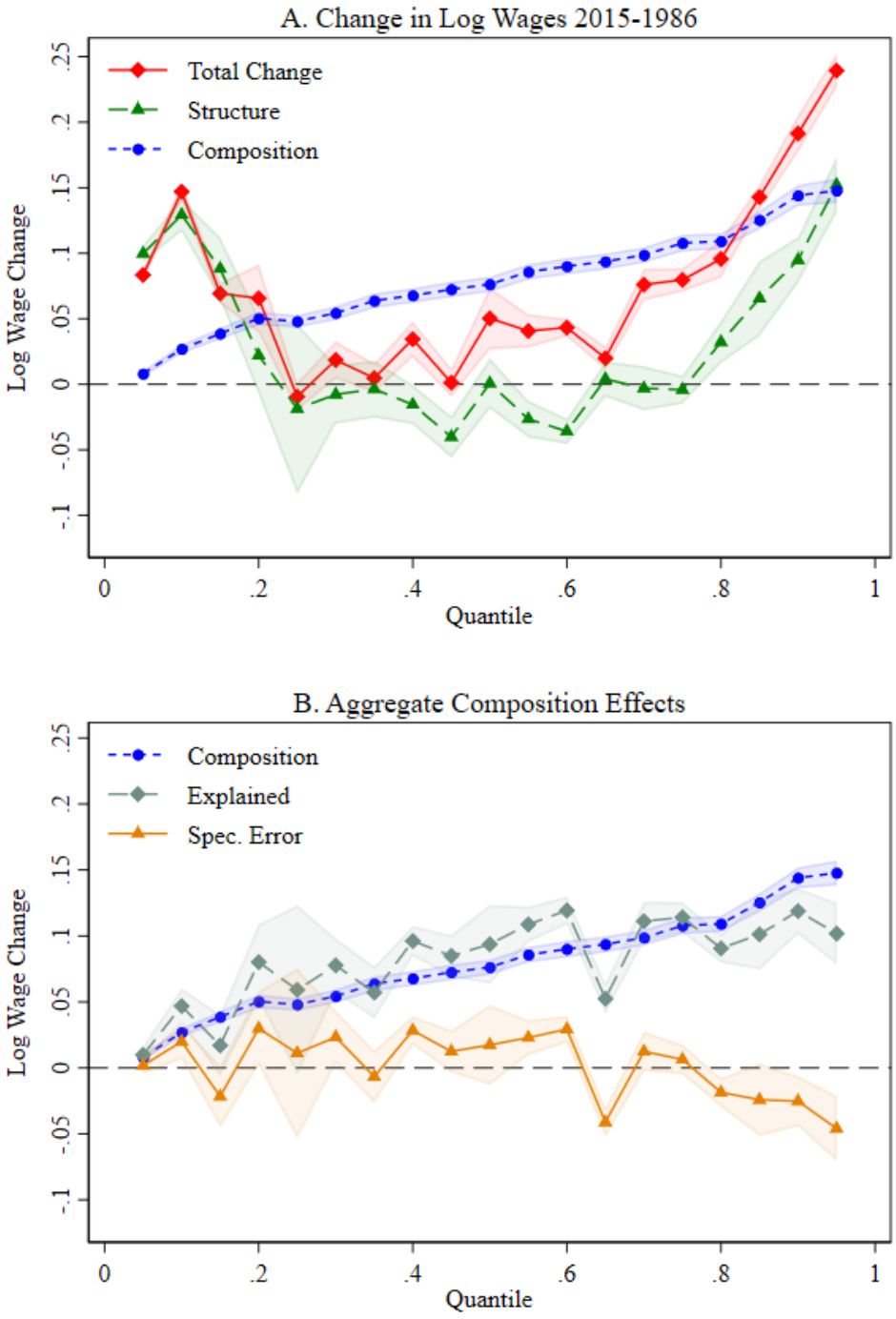
Note: This figure presents kernel density estimates of hourly (log) wages for workers in 1986, 2015, and the counterfactual reweighted distribution. Vertical lines mark the binding federal minimum wage in 1986 and the highest state minimum wage in 2015, which exceeded the federal minimum. All wages are converted to 2020 USD using the CPI series CUSR0000SA0. The data are drawn from the CPS ORG, utilizing the CEPR Uniform Extracts to construct a consistent hourly wage series. The sample is restricted to workers aged 16 to 65, with hourly wages between $1 and $100 in 1979 dollars.

Figure F.2: Selected Coefficients Estimates From the Unconditional Quantile Regression



Note: This figure compares 1986 and 2015 coefficient estimates from RIF regressions for quantiles, calculated for 95 points in the (0,1) interval. It features dashed lines for 1986 and solid lines for 2015. The regression, utilizing CPS ORG hourly wage data, includes individual, job-related, demographic, and macroeconomic factors, plus state and industry fixed effects.

Figure F.3: Decomposition of Total Change and Composition Effects - 2015 vs. 1986



Note: Panel A shows the total change in real log wages (red line) from 1986 to 2015, decomposed into composition (blue dashed line) and wage structure effects (green dashed line). Panel B compares the composition effect from reweighting (blue dashed line) with the explained portion from RIF regressions (gray dashed line) and the specification error (orange solid line). Shaded areas indicate 95% confidence intervals computed using 1,000 bootstraps.

Table F.1: Oaxaca-Blinder Decomposition of Wage Inequality Using RIF Regressions

Inequality Measure	90-10	50-10	90-50	GiniLog x 100
	(1)	(2)	(3)	(4)
Total Change	0.031***	0.056	0.056	0.056
	(0.008)	(0.057)	(0.049)	(0.077)
Composition	0.093***	-0.01	-0.01	-0.01
	(0.004)	(0.018)	(0.028)	(0.095)
Wage Structure	-0.062***	0.065	0.065***	0.065
	(0.01)	(0.074)	(0.024)	(0.041)
Composition Effects:				
Union	0.01***	-0.001	-0.001	-0.001
	(0.001)	(0.006)	(0.012)	(0.047)
Other	0.009***	-0.002	-0.002	-0.002
	(0.002)	(0.005)	(0.003)	(0.028)
Education	0.067***	0.011	0.011	0.011
	(0.003)	(0.007)	(0.008)	(0.121)
Ocupation	0.033***	-0.002	-0.002	-0.002
	(0.004)	(0.012)	(0.006)	(0.026)
Industry	-0.026***	-0.016***	-0.016***	-0.016
	(0.002)	(0.002)	(0.003)	(0.087)
Wage Structure Effects:				
Union	0.006***	-0.015***	-0.015	-0.015
	(0.002)	(0.003)	(0.014)	(0.049)
Other	0.011	0.023	0.023	0.023
	(0.011)	(0.029)	(0.027)	(0.065)
Education	0.098**	-0.138	-0.138	-0.138
	(0.047)	(0.085)	(0.163)	(0.477)
Ocupation	-0.078**	-0.032	-0.032	-0.032
	(0.034)	(0.037)	(0.048)	(0.253)
Industry	0.105***	-0.05	-0.05	-0.05
	(0.035)	(0.079)	(0.077)	(1.021)

Note: This table presents the Oaxaca-Blinder decomposition of wage inequality measures using RIF regressions without reweighting. The decomposition is applied to the 90–10, 50–10, and 90–50 log wage differentials, as well as the Gini index of log wages. "Other" includes non-white, non-married, and five experience categories. Statistical significance levels are indicated as *** for 1%, ** for 5%, and * for 10%. Bootstrapped standard errors (1,000 replications) were used to compute the p-values, with standard errors presented in parentheses..

Table F.2: Reweighted Decomposition of Wage Inequality Using RIF Regressions

Inequality Measure	90-10	50-10	90-50	GiniLog x 100
	(1)	(2)	(3)	(4)
Total Change	0.031***	0.056	0.056	0.056
	(0.008)	(0.057)	(0.049)	(0.077)
Composition	0.11***	-0.01	-0.01	-0.01
	(0.004)	(0.023)	(0.03)	(0.138)
Wage Structure	-0.065***	-0.065	-0.065	-0.065
	(0.012)	(0.04)	(0.056)	(0.092)
Composition Effects:				
Union	0.008***	-0.001	-0.001	-0.001
	(0.001)	(0.005)	(0.009)	(0.038)
Other	0.005***	-0.002	-0.002	-0.002
	(0.002)	(0.004)	(0.002)	(0.016)
Education	0.083***	0.011	0.011	0.011
	(0.004)	(0.012)	(0.01)	(0.172)
Ocupation	0.035***	-0.002	-0.002	-0.002
	(0.004)	(0.013)	(0.006)	(0.037)
Industry	-0.021***	-0.016***	-0.016***	-0.016
	(0.002)	(0.002)	(0.006)	(0.059)
Specification Error	-0.023*	0.065***	0.065*	0.065
	(0.013)	(0.025)	(0.035)	(0.104)
Wage Structure Effects:				
Union	0.015***	0.015	0.015***	0.015
	(0.003)	(0.01)	(0.005)	(0.059)
Other	-0.023	-0.023	-0.023	-0.023
	(0.091)	(0.065)	(0.161)	(1.513)
Education	0.138**	0.138	0.138*	0.138
	(0.063)	(0.153)	(0.074)	(1.092)
Ocupation	0.032	0.032	0.032	0.032
	(0.052)	(0.082)	(0.078)	(0.466)
Industry	0.05	0.05	0.05	0.05
	(0.103)	(0.079)	(0.166)	(1.759)
Reweighting Error	-0.278***	-0.278**	-0.278**	-0.278
	(0.097)	(0.121)	(0.128)	(2.471)

Note: This table presents the FFL reweighting decomposition of wage inequality measures using RIF regressions. The decomposition is applied to the 90–10, 50–10, and 90–50 log wage differentials, as well as the Gini index of log wages. "Other" includes non-white, non-married, and five experience categories. Statistical significance levels are indicated as *** for 1%, ** for 5%, and * for 10%. Bootstrapped standard errors (1,000 replications) were used to compute the p-values, with standard errors presented in parentheses.