DISTRIBUTION REGRESSION AND COUNTERFACTUAL ANALYSIS

ABSTRACT. We use binary regressions to model conditional distributions of real outcomes given covariates and develop models for counterfactual analysis. An application to (predictive) gender discrimination effects in the labor market finds that these effects explain most of the observed difference between the distribution of wages for men and women. An application to racial differences in mental ability of young children finds that two thirds of the gap between black and white children is explained by differences in background characteristics, but the remaining unexplained gap is still significant in both economic and statistical terms.

1. DISTRIBUTION REGRESSION

We use the binary regression framework to model and estimate the conditional distribution of an outcome given covariates. The outcome is a real-valued variable and can be either of the following:

- continuous (log wages),
- count (number of patents),
- nonnegative (durations, capital levels),
- discrete (rounded wages) or binary as in the previous lecture.

The key, simple observation is that the conditional distribution of the outcome Y given the set of covariates X can be expressed as

$$F_{Y|X}(y \mid x) = \mathrm{E}[1\{Y \leq y\} \mid X = x].$$

Accordingly, even when the outcome is not binary, we can always construct a collection of binary response variables, which record the events that the outcome falls bellow a set of thresholds:

$$1\{Y \leq y\}, \quad y \in T \subset \mathbb{R},$$

where T is countable subset of $\mathbb{R}$. For estimation and other practical purposes, we take T to be a finite collection of grid points. We can then use binary regressions for a collection of binary response variables to model the conditional distribution of Y given X,

$$F_{Y|X}(y \mid x) = \mathrm{P}(Y \leq y \mid X = x) = F_y(B(x)'\beta(y)),$$

where F_y is a link function, which may vary with the threshold y, $B(X)$ is a dictionary of transformations of X including a 1 as the first entry, and $\beta(y)$ is the parameter vector that may vary with the threshold y. This is the *distribution regression model*.

The distribution regression model is quite flexible and nests a variety of classical models for conditional distribution functions.

Example 1 (Classical Normal Regression Model)**.** In the classical normal regression model, $Y \mid X \sim N(B(X)'\gamma, \sigma^2)$, the conditional distribution of Y given X is

$$F_{Y|X}(y \mid x) = \Phi((y - B(x)'\gamma)/\sigma),$$

where Φ is the standard normal distribution. This conditional distribution is a special case of the distribution regression model with F_y equal to the probit link Φ, and $\beta(y) = (y - \gamma_1, -\gamma'_{-1})'$, for $\gamma = (\gamma_1, \gamma'_{-1})'$. Note that the slopes here don't vary with y. ■

Example 2 (Cox Proportional Hazard Model)**.** The Cox duration regression model is:

$$F_{Y|X}(y \mid x) = 1 - \exp(-\exp(t(y) - B(x)'\gamma)),$$

where $t(\cdot)$ is an unknown monotonic transformation, is a common approach to model conditional distributions in duration and survival analysis. It has also been used to model non-negative outcomes, such as capital in (S, s) models and wages. It corresponds to the following location-shift representation:

$$t(Y) = B(X)'\gamma + V,$$

where V has an extreme value distribution and is independent of X. The model is called proportional hazard model, because the hazard rate,

$$h(y \mid x) = \frac{\partial}{\partial y} d \ln(1 - F_{Y|X}(y \mid x)) = -\frac{\partial t(y)}{\partial y} \exp(t(y)) \exp(B(x)'\gamma),$$

depends proportionally on $\exp(B(x)'\gamma)$. The conditional distribution is a special case of the distribution regression model with F_y equal to the complementary log-log link, $F_y(u) = 1 - \exp(-\exp(u))$, and $\beta(y) = (t(y) - \gamma_1, -\gamma'_{-1})'$, for $\gamma = (\gamma_1, \gamma'_{-1})'$. Note that the slopes here don't vary with y, while distributional regression allows for slopes to be varying with y. ■

Example 3 (Poissson Regression Model)**.** The Poisson distribution is frequently used to model count variables taking values in $0, 1, 2, ...$ In the conditional (regression) version of the Poisson model, the conditional distribution of the count variable Y given X takes the form:

$$F_{Y|X}(y \mid x) = \sum_{k=0}^{y} \frac{\exp(B(x)'\gamma)^k \exp(-\exp(B(x)'\beta))}{k!} = Q\left(y, \exp\left(B(x)'\gamma\right)\right)$$

where Q called the incomplete Gamma function. Thus, the distribution regression model with link function $F_y(u) = Q(y, \exp(u))$ and $\beta(y) = \gamma$ nests the Poisson regression model. The Poisson regression model assumes that the same index governs the whole distribution. This assumption has been often challenged in applications. The zero-inflated Poisson

regression model makes one step in the direction of a more flexible model by allowing the coefficients to be different at 0. The distribution regression model does not restrict the heterogeneity of the coefficients at any level. ■

Given the conditional distribution we can also look at the conditional quantiles:

$$F^{\leftarrow}_{Y|X}(u \mid x), \quad u \in [0,1],$$

where $\leftarrow$ denotes the left-inverse of the map $y \mapsto F_{Y|X}(y \mid x)$ on T. Here we define the left-inverse of a function $G : T \to [0,1]$ on T as

$$G^{\leftarrow}(u) := \inf\{t \in T : G(t) \geq u\} \wedge \sup\{t \in T\}. \tag{1.1}$$

Given a graph of a distribution function $t \mapsto G(t)$ we can obtain the graph of the quantile function $u \mapsto G^{\leftarrow}(u)$ by simply *flipping* the axes and *mirroring* the resulting image.

There are many ways to estimate the distribution regression models. Here we focus on the following method. We can estimate the conditional distribution by:

$$\hat{F}_{Y|X}(y|x) = F_y(B(x)'\hat{\beta}(y)), \quad y \in T,$$

where for each $y \in T$ the estimator $\hat{\beta}(y)$ is the maximum likelihood estimator,

$$\hat{\beta}(y) \in \arg\max_{b(y)\in\mathcal{B}} \mathbb{E}_n[1(Y_i \leq y)\ln F_y(B(X_i)'b(y)) + 1(Y_i > y)\ln(1 - F_y(B(X_i)'b(y)))],$$

where $\mathcal{B}$ is the parameter space for $\beta(y)$. For example, $\hat{\beta}(y)$ is the probit estimator if we use the normal link $F_y = \Phi$ or the logit estimator if we use the logistic link $F_y = \Lambda$. We can handle inference on $y \mapsto F_y(B(x)'\hat{\beta}(y))$ for $y \in T$ by using the delta method in conjunction with the GMM formulation of the problem. Indeed, we can view estimation of

$$\theta_0 = \text{vec}(\beta(y) : y \in T)$$

as a GMM problem with the score:

$$g(Z_i, \theta) = \text{vec}\left(g_y(Z_i, b(y)) : y \in T\right),$$

$$g_y(Z_i, b(y)) = \frac{\partial}{\partial b(y)}\left\{1(Y_i \leq y)\ln F_y(B(X_i)'b(y)) + 1(Y_i > y)\ln(1 - F_y(B(X_i)'b(y)))\right\},$$

which simply stacks the scores of many binary regressions. The joint parameter vector is $\theta = \text{vec}(b(y) : y \in T)$, which simply stacks the parameters of many binary regressions. The map $y \mapsto F_y(B(x)'\hat{\beta}(y))$, $y \in T$, is a smooth transformation of the estimators $\hat{\beta}(y)$, $y \in T$, so the delta method delivers the large sample properties of the estimators $F_y(B(x)'\hat{\beta}(y))$, $y \in T$. This also means that we can use the bootstrap for inference.

In practice, the map $y \mapsto F_y(B(x)'\hat{\beta}(y))$ may be non-monotone, in which case we can rearrange it into monotone function by simply sorting the values of function in a nondecreasing order. This typically improves the finite-sample properties of the estimator. We discuss the rearrangement procedure in Section 3.2.

2. Unconditional and Counterfactual Distributions and Quantiles

Once we get the conditional distribution, we can construct an unconditional distribution of the form:

$$F(y) = \int F_{Y|X}(y|x)dM(x), \quad y \in T,$$

which is simply the conditional distribution integrated against a marginal distribution M for the covariate values. When M is the distribution of X, namely $M = F_X$, we obtain the marginal distribution of Y:

$$F(y) = F_Y(y), \quad y \in T.$$

We can also look at the marginal quantile functions $Q = F^{\leftarrow}$ which is the left-inverse of the distribution function $y \mapsto F(y)$ on T, as defined in (1.1).

> When M is not the distribution of X, then F is a *counterfactual distribution* which corresponds to a sampling experiment where covariates X' are sampled from M, but outcomes are sampled from the conditional distribution $F_{Y|X}(\cdot|X')$.

This construction is very useful for counterfactual analysis. For example, let F_{X_k} denote the distribution of job-relevant characteristics (education, experience, etc.) for men when $k = m$, and women when $k = w$. Let $F_{Y_j|X_j}$ denote the conditional distributions of wages given job-relevant characteristics for group $j \in \{w, m\}$. This conditional distribution describes the stochastic wage schedule that a given group faces. Using these distributions we can construct $F_{\langle j|k\rangle}$, the distribution of wages for group k facing group j wage schedule as

$$F_{\langle j|k\rangle}(y) = \int F_{Y_j|X_j}(y|x)dF_{X_k}(x), \quad y \in T.$$

For example, $F_{\langle m|m\rangle}$ is the distribution of wages for men who face men's wage schedule, and $F_{\langle w|w\rangle}$ is the distribution of wages for women who face women's wage schedule. These are observed distributions. We can also look at $F_{\langle m|w\rangle}$, the counterfactual distribution of wage for women if they would face the men's wage schedule. We can interpret $F_{\langle m|w\rangle}$ as the distribution of wages for women in the absence of "gender discrimination", although we do have to be careful and say that this is just a predictive distribution, which does not have a causal interpretation without further (strong) assumptions.

We can use the counterfactual distributions to decompose the differences in the observed wage distributions:

$$F_{\langle m|m\rangle} - F_{\langle w|w\rangle} = \underbrace{(F_{\langle m|m\rangle} - F_{\langle m|w\rangle})}_{\text{composition effect}} + \underbrace{(F_{\langle m|w\rangle} - F_{\langle w|w\rangle})}_{\text{discrimination effect}},$$

where the first term on the right is the composition effect, which results from the populations of men and women having different distributions of job market characteristics, and

the second is the *discrimination effect* or *price effect*, which results from women facing different wage schedules than men. Analogous decomposition could be made for the observed quantile functions:

$$Q_{\langle m|m\rangle} - Q_{\langle w|w\rangle} = \underbrace{(Q_{\langle m|m\rangle} - Q_{\langle m|w\rangle})}_{\text{composition effect}} + \underbrace{(Q_{\langle m|w\rangle} - Q_{\langle w|w\rangle})}_{\text{discrimination effect}}. \tag{2.1}$$

What we see above are the distribution and quantile versions of the Oaxaca-Blinder decomposition.

In the empirical example based on U.S. data, which we describe below, we find that the distributions of wages for men and women are different, with the distribution of wages for men being shifted to the right by 20% or 30% along the horizontal axis, or the quantile function for men being shifted up by 20% or 30%. Almost all of the difference can be attributed to the discrimination/price effect. The composition effect is close to zero and is slightly negative due to the fact that men's characteristics distribution is slightly worse than women's characteristic distribution.

The estimation of the counterfactual distributions can be done using the plug-in principle:

$$\hat{F}_{\langle j|k\rangle}(y) = \int \hat{F}_{Y_j|X_j}(y|x)d\hat{F}_{X_k}(x), \quad y \in T, \tag{2.2}$$

where the conditional distribution estimator $\hat{F}_{Y_j|X_j}$ is the distribution regression estimator applied to observations of (Y, X) for group j, and the covariate distribution estimator $\hat{F}_{X_k}$ is the empirical distribution of observations of X for group k. Just like other estimation problems we have seen, this estimator can be treated as the GMM estimator by simply stacking together various scores corresponding to different steps of the procedure. This also means that we can use the bootstrap as a practical inference tool.

In practice, the map $y \mapsto \hat{F}_{\langle j|k\rangle}(y)$ may be non-monotone if $y \mapsto \hat{F}_{Y_j|X_j}(y|x)$ is non-monotone, but we can rearrange it into monotone function by simply sorting the values of function in a nondecreasing order. This typically improves the finite-sample properties of the estimator as we show in Section 3.2.

3. Generic Inference Method for Distributions and Quantiles

3.1. **The Method.** Let $\mathbb{D}$ denote the set of weakly increasing functions, mapping T to $[0, 1]$. We will call elements of this set "distribution functions", albeit some of them need not be proper distribution functions. Let $y \mapsto F(y)$ in $\mathbb{D}$ denote some target distribution function. This target could be a conditional distribution function, an unconditional distribution function, or a counterfactual distribution function. Another target will be the left-inverse

$a \mapsto F^{\leftarrow}(a)$ of F on T, which we can call the "quantile function" of F, where the left-inverse was defined in (1.1). We construct confidence sets for the distribution and quantile functions.

The idea of the approach given here is as follows:

1. Construct confidence bands $I = [L, U]$ for F based upon bootstrapping an estimator $\hat{F}$. We can impose logical shape restrictions on these estimator and bands if necessary.
2. Convert these confidence bands into confidence bands for the quantile function defined as $F^{\leftarrow}$ by taking the inverse $I^{\leftarrow} = [U^{\leftarrow}, L^{\leftarrow}]$ of the bands $I = [L, U]$.

Consider a confidence band $I = [L, U]$ for F, with lower and upper bands L and U. Specifically, given two functions $y \mapsto U(y)$ and $y \mapsto L(y)$ in the set $\mathbb{D}$ such that $L \leq U$, pointwise, we define a band $I = [L, U]$ as the collection of intervals

$$I(y) = [L(y), U(y)], \quad y \in T.$$

We say that I covers F if $F \in I$ pointwise namely $F(y) \in I(y)$ for all $y \in T$. If U and L are some data-dependent bands, we say that $I = [L, U]$ is a confidence band for F of level p, if I covers F with probability at least p. Below we provide a bootstrap algorithm for computing a confidence band. Such method works if $\hat{F}$, the estimator of F, is sufficiently regular. Examples include estimators discussed in the previous sections.

Our ultimate goal is to construct a confidence set for $F^{\leftarrow}$ from a generic confidence set $[L, U]$ for F. The following result provides a confidence set $I^{\leftarrow}$. Here, we say that a collection of sets $I^{\leftarrow} = \{I^{\leftarrow}(a), a \in [0, 1]\}$ covers $F^{\leftarrow}$ if $F^{\leftarrow}(a) \in I^{\leftarrow}(a)$ for each $a \in [0, 1]$.

Theorem 1 (Generic Bands for Quantile Functions)**.** *Consider a distribution function F and band functions L and U in the class $\mathbb{D}$.*

(1) *If F is covered by the band $I := [L, U]$, then the quantile function $F^{\leftarrow}$ is covered by the band $I^{\leftarrow}$ defined by*

$$I^{\leftarrow}(a) := [U^{\leftarrow}(a), L^{\leftarrow}(a)].$$

(2) *Thus if the distribution function F is covered by I with probability at least p, then the quantile function $F^{\leftarrow}$ is covered by $I^{\leftarrow}$ with probability at least p.*

Proof. The result is immediate from the definition of the left inverse: For any $a \in \mathcal{A}$, since $L(y) \leq F(y)$ for each $y \in T$, and F and L are in $\mathbb{D}$,

$$\begin{aligned} F^{\leftarrow}(a) &= \inf\{y \in T : F(y) \geq a\} \wedge \sup T \\ &\leq \inf\{y \in T : L(y) \geq a\} \wedge \sup T = L^{\leftarrow}(a). \end{aligned}$$

Analogously, conclude that $F^{\leftarrow}(a) \geq U^{\leftarrow}(a)$. ∎

The band $I^{\leftarrow}$ can be narrowed without affecting coverage by exploiting the support restrictions. Suppose that T is the support of the distribution function F (and not merely a set of grid points at which we measure F). This is relevant, for example, when outcomes are discrete or counts. Then it makes sense to exploit the support restriction that $F^{\leftarrow}(a) \in T$ by intersecting the confidence sets for $F^{\leftarrow}(a)$ with T. Clearly this won't affect the coverage properties of the sets.

Corollary 1 (Imposing Support Restrictions)**.** *Consider the set $\tilde{I}^{\leftarrow}$ defined by pointwise intersection of $I^{\leftarrow}$ with T, namely $\tilde{I}^{\leftarrow}(a) := I^{\leftarrow}(a) \cap T$. Then, $\tilde{I}^{\leftarrow} \subseteq I^{\leftarrow}$ pointwise, and if $I^{\leftarrow}$ covers $F^{\leftarrow}$ then so does $\tilde{I}^{\leftarrow}$.*

The corollary is immediate because pointwise intersection of $I^{\leftarrow}$ with the set T does not change the coverage property, since $F^{\leftarrow}$ only takes values in T.

Figure 1 illustrates the construction of bands using Theorem 1. It shows an $F : [0, 10] \mapsto [0, 1]$ covered by a band $I = [L, U]$. It also shows that the inverse map $F^{\leftarrow} : [0, 1] \mapsto [0, 10]$ is covered by the inverted band $I^{\leftarrow} = [U^{\leftarrow}, L^{\leftarrow}]$. The band $I^{\leftarrow}$ is easy to obtain but does not exploit the fact that the support of the distribution F in this example is the set $T = \{0, 1, \ldots, 10\}$. By intersecting $I^{\leftarrow}$ with T for each $a \in \mathcal{A} = [0, 1]$ we obtain the band $\tilde{I}^{\leftarrow}$ which reflects such support restrictions.

3.2. **Imposing Logical Shape Constraints.** In many applications the point estimates $\hat{F}$ and interval estimates $[L', U']$ for the target distribution F do not satisfy the logical monotonicity or range restrictions, namely they don't take values in the set $\mathbb{D}$. Given such an ordered triple $L' \leq \hat{F} \leq U'$, we can always transform it into another ordered triple $L \leq \check{F} \leq U$ that obey the logical monotonicity and shape restrictions. For example, we can set

$$\check{F} = \mathcal{S}(\hat{F}), \quad L = \mathcal{S}(L'), \quad U = \mathcal{S}(U'), \tag{3.1}$$

where $\mathcal{S}$ is the shaping operator that given a function f yields a mapping $t \mapsto \mathcal{S}(f)(t) \in \mathbb{D}$ with

$$\mathcal{S}(f) = \mathcal{M}(0 \vee f \wedge 1),$$

where the maximum and minimum are taken pointwise, and $\mathcal{M}$ is the rearrangement operator that given a function $f : T \mapsto [0, 1]$ yields a map $t \mapsto \mathcal{M}(f)(t) \in \mathbb{D}$.[1]

The *rearrangement operator* is defined as follows. Let T be a countable set (in practice, a finite set). Given $f : T \mapsto [0, 1]$, we first consider $\mathcal{M}f$ as a vector of sorted values of the set $\{f(t) : t \in T\}$, where the sorting is done in a non-decreasing order. Since T is an ordered

[1] Other monotonization operators, such as the projection on the set of weakly increasing functions, can also be used, as we remark further below.

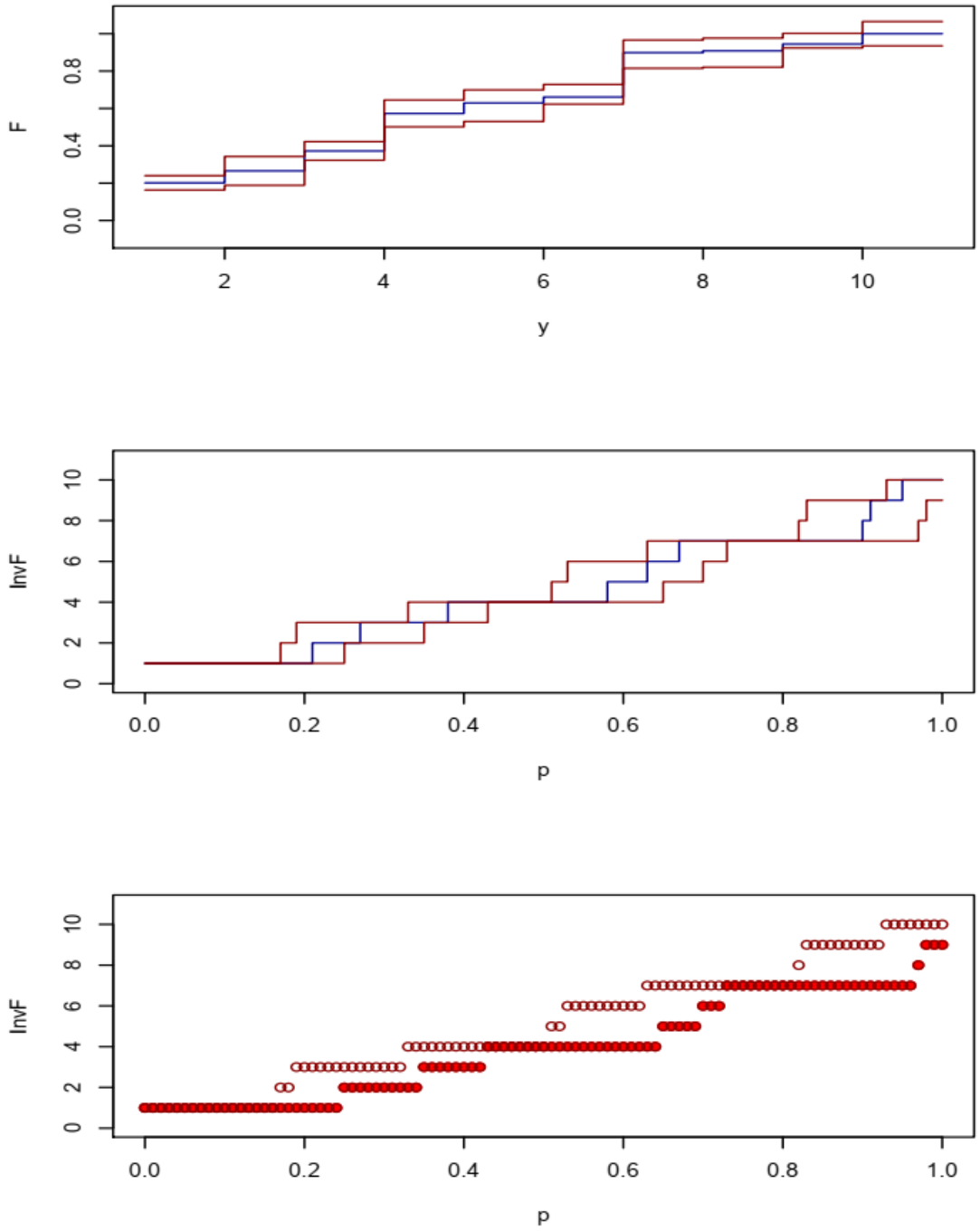

FIGURE 1. Top: the distribution function F defined over $T = \{0, 1, ..., 10\}$ and a confidence band $I = [L, U]$. Both the distribution function and the confidence bands are interpolated by piecewise constant outside T. Middle: the quantile function $F^{\leftarrow}$ and the confidence band $I^{\leftarrow} = [U^{\leftarrow}, L^{\leftarrow}]$. Bottom: The support-restricted confidence band $\tilde{I}^{\leftarrow} = I^{\leftarrow} \cap T$ shown by circles and the inverse $F^{\leftarrow}$ shown by red balls; the red balls are encircled because $F^{\leftarrow} \in \tilde{I}^{\leftarrow}$.

set of the same cardinality as $\mathcal{M}f$, we can assign the elements of $\mathcal{M}f$ to T in one-to-one manner: to the k-th smallest element of T we assign the k-th smallest element of $\mathcal{M}f$. The resulting mapping $t \mapsto \mathcal{M}f(t)$ is the rearrangement map.

The following lemma shows that shape restrictions *improve* the finite-sample properties of the estimators and confidence bands.

Lemma 1 (Shaping Improves Point and Interval Estimates)**.** *The shaping operator $\mathcal{S}$*

(a) *is weakly contractive under the max distance:*

$$\|\mathcal{S}(A) - \mathcal{S}(B)\|_\infty \leq \|A - B\|_\infty, \quad \text{for any } A, B\text{: } T \to [0,1],$$

(b) *is shape-neutral,*

$$\mathcal{S}(F) = F \text{ for any } F \in \mathbb{D},$$

(c) *preserves the partial order:*

$$A \leq B \implies \mathcal{S}(A) \leq \mathcal{S}(B), \quad \text{for any } A, B\text{: } T \to [0,1].$$

Consequently,

(1) *the re-shaped point estimate constructed via (3.1) is weakly better than the initial estimate under the max distance:*

$$\|\check{F} - F\|_\infty \leq \|\hat{F} - F\|_\infty,$$

(2) *the re-shaped confidence band constructed via (3.1) has weakly better coverage than the initial confidence band:*

$$\mathrm{P}(L' \leq F \leq U') \leq \mathrm{P}(L \leq F \leq U),$$

(3) *and re-shaped confidence band constructed via (3.1) is weakly shorter than the original confidence band under the max norm,*

$$\|U - L\|_\infty \leq \|U' - L'\|_\infty.$$

The band $[L, U]$ is weakly better than the original band $[L', U']$, in the sense that coverage is preserved while the width of the confidence band is weakly shorter.

Remark 1 (Isotonization is Another Option)**.** An alternative to rearrangement is to use the isotonization, which projects a given function on the set of weakly increasing functions that map T to $[0, 1]$. This also has the improving properties stated in Lemma 1. In fact any convex combination between isotonization and rearrangement has the improving properties stated in Lemma 1. ■

Remark 2 (Shape Restrictions on Confidence Bands by Intersection)**.** An alternative way for imposing shape restrictions on the confidence band, is to intersect the initial band

$[L', U']$ with the set of weakly increasing functions WI that map T to $[0, 1]$. That is, we simply set

$$[L^I, U^I] = \text{WI} \cap [L', U'] = \{w \in \text{WI} : L'(y) \leq w(y) \leq U'(y), \quad \forall y \in T\}.$$

Thus, U^I is the greatest weakly increasing minorant of $0 \vee U' \wedge 1$ and L^I is the smallest majorant of $0 \vee L' \wedge 1$. This approach gives the tightest confidence bands, in particular

$$[L^I, U^I] \subseteq [L, U].$$

However, this construction might mis-behave under misspefication, where rearrangement continues to deliver meaningful confidence sets. Imagine that we have constructed an estimator $\hat{F}$ that is consistent for the target F, which is not monotone, i.e. $F \notin \mathbb{D}$. Then if the confidence bands $[L', U']$ are sufficiently tight, then we can end up with empty intersection bands, $[L^I, U^I] = \emptyset$. By contrast $[L, U]$ is non-empty and covers the probability limit $F^* = \mathcal{S}(F)$ of $\check{F}$, where the limit target F^* does belongs to $\mathbb{D}$. ■

3.3. **The bootstrap algorithm.** The following algorithm implements the ideas presented above.

Algorithm 1 (Bootstrap Algorithm for Confidence Bands for F)**.** Given the ingredients above, we provide an explicit algorithm for the bootstrap construction of the joint confidence band $I = [L, U]$ for $(F(y))_{y \in T}$ based on the bootstrappable estimators $(\hat{F}(y))_{y \in T}$:

(1) Obtain many bootstrap draws of the estimator $(\hat{F}(y))_{y \in T}$,

$$(\hat{F}^{*(j)}(y))_{y \in T}, \quad j = 1, \ldots, B$$

where the index j enumerates the bootstrap draws.

(2) For each y in T compute the bootstrap variance estimate

$$\hat{s}^2(y) = B^{-1} \sum_{j=1}^{B} (\hat{F}(y)^{*(j)} - \hat{F}(y))^2,$$

(or use the estimate based on interquartile range).

(3) Compute the critical value

$$c(1-\alpha) = \ (1-\alpha)\text{-quantile of } \left\{ \max_{y \in T} |\hat{F}(y)^{*(j)} - \hat{F}(y)| / \hat{s}(y) \right\}_{j=1}^{B}.$$

(4) Construct a preliminary $(1-\alpha)$-level confidence band for $(F(y))_{y \in T}$ as

$$[L'(y), U'(y)] = [\hat{F}(y) \pm c(1-\alpha)\hat{s}(y)], \quad y \in T.$$

(5) Impose the shape restrictions on $\hat{F}$ and L' and U' obtaining $\check{F}$, and L and U. Report $I = [L, U]$ as a $(1-\alpha)$-level confidence band for $(F(y))_{y \in T}$.

4. Generic Confidence Bands for Quantile Effects

Our next goal is to construct a confidence band for the quantile effect function $a \mapsto \Delta(a)$ defined by

$$\Delta(a) := F_1^{\leftarrow}(a) - F_0^{\leftarrow}(a),$$

corresponding to the difference of quantile functions of the two distribution functions F_0 and F_1 with support sets T.

The basic idea is as follows:

1. Construct joint confidence bands for F_0 and F_1.
2. Convert these into confidence bands for $F_0^{\leftarrow}$ and $F_1^{\leftarrow}$ by inversion.
3. Contruct the confidence band for the quantile effect Δ by taking the pointwise Minkowski difference between the confidence bands for $F_1^{\leftarrow}$ and $F_0^{\leftarrow}$.

Specifically, suppose we have the confidence bands $I_0^{\leftarrow}$ for $F_0^{\leftarrow}$ and $I_1^{\leftarrow}$ for $F_1^{\leftarrow}$, which jointly cover $F_0^{\leftarrow}$ and $F_1^{\leftarrow}$ with probability p. For example, we can construct these sets using Theorem 1 in conjunction with the Bonferroni inequality.[2] Algorithm 2 provides a construction of the confidence bands that has joint coverage property and is less conservative than using Bonferroni. Then we can convert these bands to confidence bands for Δ by taking the pointwise Minkowski difference $\ominus$ of the pairs of intervals at each a treated as sets of points. Recall that the Minkowski difference between two subsets V and U of a vector space is $V \ominus U := \{v - u : v \in V, u \in U\}$. Note that if V and U are intervals $[v_1, v_2]$ and $[u_1, u_2]$, then

$$V \ominus U = [v_1, v_2] \ominus [u_1, u_2] = [v_1 - u_2, v_2 - u_1].$$

Theorem 2 (Generic Bands for Quantile Effect Functions). *Consider the distribution functions F_0 and F_1 and the band functions L_0, U_0, L_1 and U_1 in the class $\mathbb{D}$.*

(1) *If F_k is covered by $I_k := [L_k, U_k]$ for $k = 0$ and $k = 1$, then the quantile effect function $\Delta = F_1^{\leftarrow} - F_0^{\leftarrow}$ is covered by the band $I_\Delta^{\leftarrow} = [U_1^{\leftarrow}, L_1^{\leftarrow}] - [U_0^{\leftarrow}, L_0^{\leftarrow}]$, where the minus operator is defined by a pointwise Minkowski difference:*

$$I_\Delta^{\leftarrow}(a) \quad := \quad [U_1^{\leftarrow}(a), L_1^{\leftarrow}(a)] \ominus [U_0^{\leftarrow}(a), L_0^{\leftarrow}(a)].$$

[2] The joint coverage of two confidence sets with marginal coverage probabilities $\tilde{p}$ is at least $p = 2\tilde{p} - 1$ by Bonferroni inequality.

(2) *If the distribution functions F_0 and F_1 are jointly covered by I_0 and I_1 with probability at least p, then the quantile effect function $\Delta = F_1^{\leftarrow} - F_0^{\leftarrow}$ is covered by $I_\Delta^{\leftarrow}$ with probability at least p.*

Proof. The results is immediate from the definition of the Minkowski sum. ∎

As in Theorem 1, we can narrow the band $I_\Delta^{\leftarrow}$ without affecting coverage by imposing support restrictions on the bands for the quantile functions.

Corollary 2 (Imposing Support Restrictions)**.** *Consider the band $\tilde{I}_\Delta^{\leftarrow} = \tilde{I}_1^{\leftarrow} - \tilde{I}_0^{\leftarrow}$ defined by:*

$$\tilde{I}_\Delta^{\leftarrow}(a) := \tilde{I}_1^{\leftarrow}(a) \ominus \tilde{I}_0^{\leftarrow}(a), \quad \tilde{I}_k^{\leftarrow}(a) := \{[U_k^{\leftarrow}(a), L_k^{\leftarrow}(a)] \cap T\}, \quad k \in \{0,1\}.$$

Then $\tilde{I}_\Delta^{\leftarrow} \subseteq I_\Delta^{\leftarrow}$, and if $I_\Delta^{\leftarrow}$ covers Δ then so does $\tilde{I}_\Delta^{\leftarrow}$.

Figure 2 shows bands for a pair of quantile functions together with bands for the quantile effect function constructed using Theorem 2. The top plot shows the bands $I_0^{\leftarrow}$ and $I_1^{\leftarrow}$ for the quantile functions $F_0^{\leftarrow}$ and $F_1^{\leftarrow}$. The middle plot shows the band I_Δ for the quantile effect function $\Delta = F_1^{\leftarrow} - F_0^{\leftarrow}$, obtained by taking the Minkowski difference of $I_1^{\leftarrow}$ and $I_0^{\leftarrow}$. The bottom plot shows the confidence band $\tilde{I}_\Delta$ for the quantile effect function Δ resulting from imposing the support constraints. As the Theorem 2 predicts, the quantile function Δ is covered by the band I_Δ.

In what follows we write down an explicit algorithm that implements the proposal above.

Algorithm 2 (Bootstrap Algorithm for Confidence Bands for Quantile Effects)**.**

(1) Obtain many bootstrap draws of the estimator $\{(\hat{F}_k(y))_{y\in T}\}_{k\in\{0,1\}}$,

$$\{(\hat{F}_k^{*(j)}(y))_{y\in T}\}_{k\in\{0,1\}}, \quad j = 1, \ldots, B$$

where the index j enumerates the bootstrap draws.

(2) For each y in T and $k \in \{0,1\}$ compute the bootstrap variance estimate

$$\hat{s}_k^2(y) = B^{-1} \sum_{j=1}^{B} (\hat{F}(y)_k^{*(j)} - \hat{F}_k(y))^2,$$

(or use the estimate based on the interquartile range).

(3) Compute the critical value

$$c(1-\alpha) = \ (1-\alpha)\text{-quantile of } \left\{ \max_{y\in T, k\in\{0,1\}} |\hat{F}_k(y)^{*(j)} - \hat{F}_k(y)|/\hat{s}_k(y) \right\}_{j=1}^{B}.$$

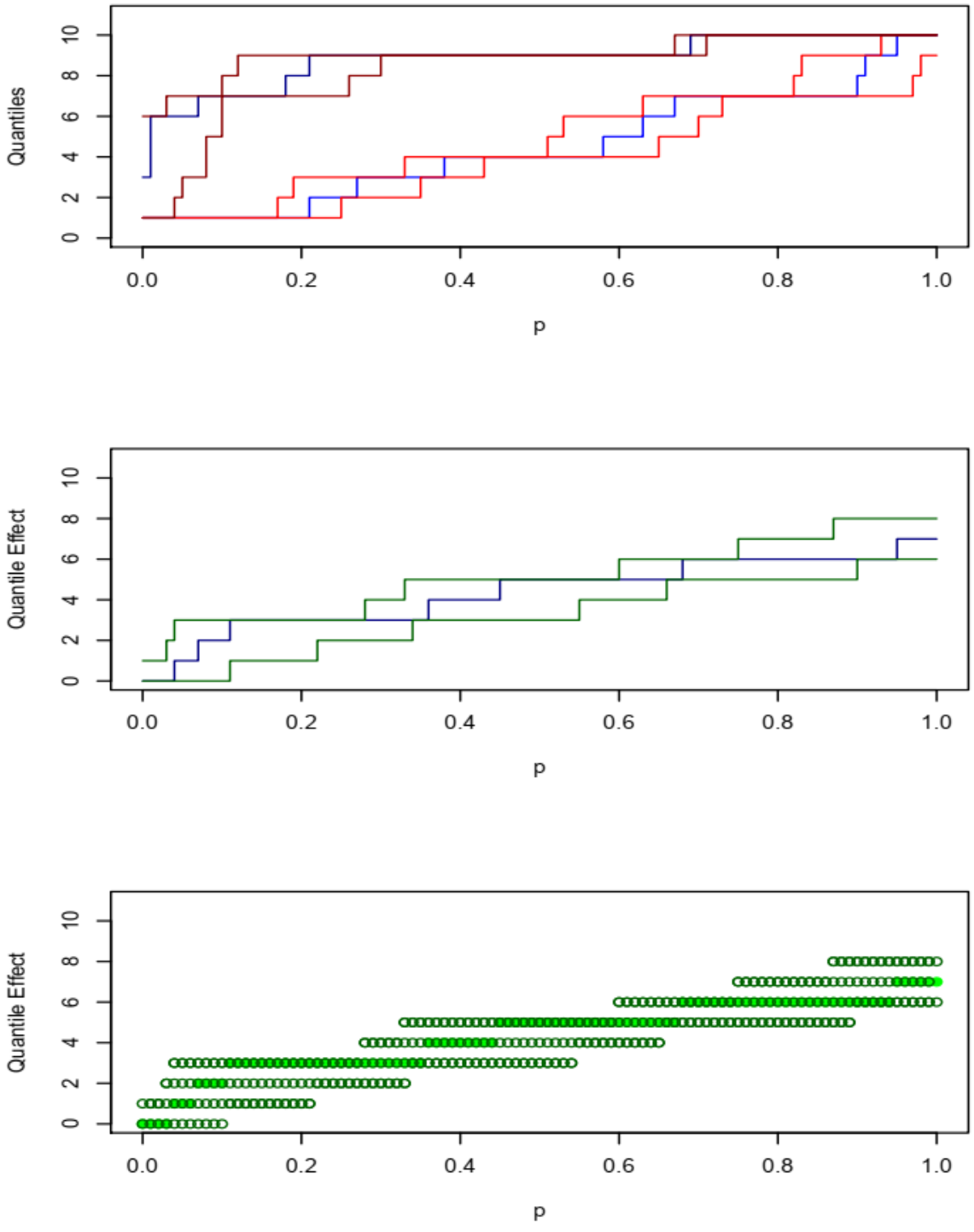

FIGURE 2. Top: the two quantiles functions $F_0^{\leftarrow}$ and $F_1^{\leftarrow}$ and confidence bands $I_0^{\leftarrow}$ and $I_1^{\leftarrow}$. Middle: the quantile effect $F_1^{\leftarrow} - F_0^{\leftarrow}$ and the confidence band $I_\Delta^{\leftarrow}$. Bottom: the support restricted confidence band $\tilde{I}_\Delta^{\leftarrow}$.

(4) Construct a preliminary $(1-\alpha)$-level joint confidence region for $\{(F_k(y))_{y\in T}\}_{k\in\{0,1\}}$ as

$$[L'_k(y), U'_k(y)] = [\hat{F}_k(y) \pm c(1-\alpha)\hat{s}_k(y)], \quad y \in T, \quad k \in \{0,1\}.$$

(5) Impose the shape restrictions on $\hat{F}_k$ and L'_k and U'_k by setting:

$$\check{F}_k = \mathcal{S}(\hat{F}_k), \quad [L_k, U_k] = [\mathcal{S}(L'_k), \mathcal{S}(U'_k)].$$

(6) Report $I_k = [L_k, U_k]$ for $k \in \{0,1\}$ as a $(1-\alpha)$-level joint confidence band for $(F_k(y))_{y\in T}$ for $k \in \{0,1\}$. Report $I_k^{\leftarrow} = [U_k^{\leftarrow}, L_k^{\leftarrow}]$ for $k \in \{0,1\}$ as a $(1-\alpha)$-level joint confidence band for $(F_k^{\leftarrow}(y))_{y\in T}$ for $k \in \{0,1\}$.

(7) Report

$$I_\Delta^{\leftarrow} = [U_1^{\leftarrow}, L_1^{\leftarrow}] - [U_0^{\leftarrow}, L_0^{\leftarrow}]$$

as a $(1-\alpha)$-level confidence band for quantile effect.

(8) If necessary, impose support restrictions in the previous two steps, following Corollaries 1 and 2.

5. Gender Wage Gap in 2012

To illustrate the use of the bands in practice we consider an application to gender wage gap using data from the U.S. March Supplement of the Current Population Survey (CPU) in 2012. We select white non-hispanic individuals, aged 25 to 64 years, and working more than 35 hours per week during at least 50 weeks of the year. We exclude self-employed workers; individuals living in group quarters; individuals in the military, agricultural or private household sectors; individuals with inconsistent reports on earnings and employment status; and individuals with allocated or missing information in any of the variables used in the analysis. The resulting sample consists of $29,217$ workers including $16,690$ men and $12,527$ of women. The variable of interest Y is the logarithm of the hourly wage rate constructed as the ratio of the annual earnings to the total number of hours worked, which is constructed in turn as the product of number of weeks worked and the usual number of hours worked per week.[3]

In this application F_0 and F_1 correspond to observed and counterfactual distributions of wages for women and men. Following Section 2, we denote these distributions by $F_{\langle j|k\rangle}$, where

$$F_{\langle j|k\rangle}(y) = \int F_{Y_j|X_j}(y|x)dF_{X_k}(x), \quad y \in T,$$

for $j,k \in \{m,w\}$, where m and w refer to men and women, $F_{Y_j|X_j}$ is the wage structure in group j, and F_{X_k} is the distribution of worker characteristics in group k. The worker

[3]This sample selection criteria and the variable construction follow [9].

characteristics X include 5 marital status indicators (widowed, divorced, separated, never married, and married); 6 educational attainment indicators (0-8 years of schooling completed, high school dropouts, high school graduates, some college, college graduate, and advanced degree); 4 region indicators (midwest, south, west, and northeast); and a quartic in potential experience constructed as the maximum of age minus years of schooling minus 7 and zero, i.e., $experience = \max(age - education - 7, 0)$, interacted with the educational attainment indicators.

We estimate $F_{\langle w|w\rangle}$ and $F_{\langle m|m\rangle}$ using the empirical distributions of Y for women and men. We estimate $F_{\langle m|w\rangle}$ by (2.2), where $\hat{F}_{Y_m|X_m}$ is the distribution regression estimator with a logit link and a linear index in X in the sample of men. We use the empirical distribution of X for women to estimate F_{X_w}. All the estimators use sampling weights to account for nonrandom sampling in the March CPS.

TABLE 1. Descriptive Statistics

	All	Men	Women
log wage	2.79	2.90	2.65
female	0.43	0.00	1.00
married	0.66	0.69	0.63
widowed	0.01	0.00	0.02
divorced	0.12	0.10	0.15
separated	0.02	0.02	0.02
never married	0.19	0.19	0.18
0-8 years completed	0.00	0.01	0.00
high school dropout	0.02	0.03	0.02
high school graduated	0.25	0.27	0.23
some college	0.28	0.27	0.30
college graduated	0.28	0.28	0.29
advanced degree	0.15	0.14	0.17
northeast	0.20	0.20	0.19
midwest	0.27	0.27	0.28
south	0.35	0.35	0.35
west	0.18	0.19	0.18
potential experience	18.96	19.01	18.90

Source: March Supplement CPS 2012

Table 1 reports descriptive statistics for the variables used in the analysis. Working women are more highly educated than working men, but have less experience. The unconditional average gender wage gap is 25%. Figure 3 shows that the log hourly wage variable

Y has multiple mass points due to rounding in the reporting of the earnings and labor supply variables. In particular, we observe $2,633$ different values in the $16,690$ observations for men and $2,091$ different values in the $12,527$ observations for women. We note here that the inference methods of Section 3 are fully robust to the process that generates mass points in the variable Y.

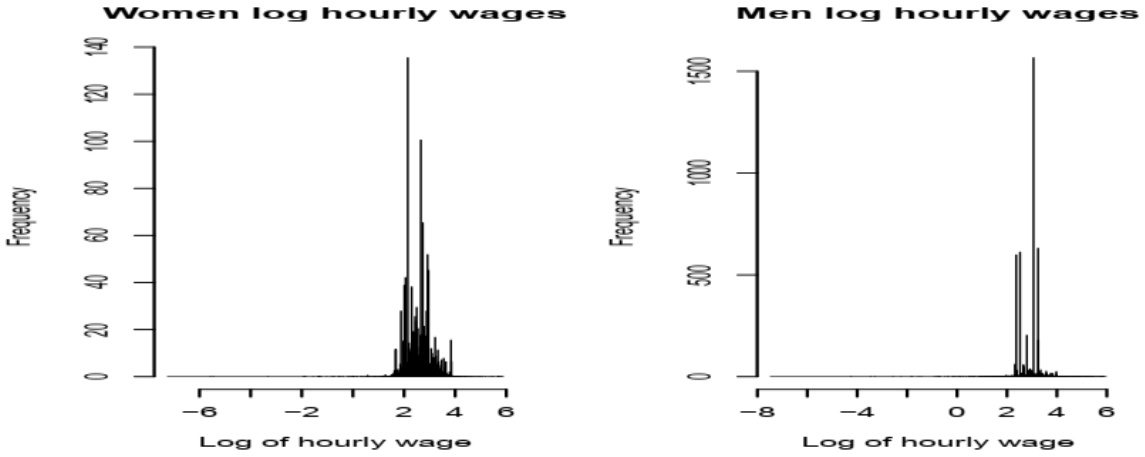

FIGURE 3. Histograms of log hourly wages for women and men.

Figures 4 and 5 show point estimates and 90% joint confidence bands for the observed and counterfactual distribution and quantile functions of wages. We construct the confidence bands for the distributions using Algorithm 2 by empirical bootstrap with 200 repetitions. Following Theorem 1, we apply left-inverse to these bands to obtain the confidence bands for the quantile functions. We do not report bands that impose the support constrains because the variable Y takes on many values. Here we find that the wage distribution for men first order stochastically dominates the wage distribution for women. The counterfactual quantile function $F^{\leftarrow}_{\langle m|w\rangle}$ is very similar to the observed quantile function for men, suggesting that most of the wage gap is due to gender discrimination. This observation is confirmed in Figure 6, which plots point estimates and 90% simultaneous confidence bands for the decomposition of the quantile gender wage gap into composition and discrimination effects of (2.1). We follow Theorem 2 to construct the bands by the Minkowski difference of the bands for the quantiles, as described in Algorithm 2. The gender wage gap ranges from 20% to 30% and is increasing with the quantile index. Most of the gap is explained by gender wage discrimination with the composition having a small

negative effect non statistically significant for most quantiles. This low contribution of the composition effect is explained by the similarity in the observable characteristics of man and women in Table 1.

6. Racial differences in mental ability of young children

As a second empirical application, we analyze the racial IQ test score gap. We use data from the US Collaborative Perinatal Project (CPP) obtained from [7]. These data contain information on children from 30,002 women who gave birth in 12 medical centers between 1959 and 1965. Our main outcome of interest, Y, is the standardized test scores at the age of seven years (both Stanford-Binet and Wechsler Intelligence Test). In addition to the test score measure, the dataset contains a rich set of background characteristics for the children, X, including information on age, gender, region, socioeconomic status, home environment, prenatal conditions, and interviewer fixed effects. [7] provide a comprehensive description of the dataset and extensive descriptive statistics.

A key feature of the test score variable is the discrete nature of its distribution. We observe only 128 different values for the standardized test score. Figure 7 presents the corresponding histogram. Note that each bar corresponds to exactly one value. For instance, more than 4% of the observations have exactly the same score. This is a common feature of test scores, which are necessarily discrete because they are based on a finite number of questions.

In this application F_0 and F_1 correspond to observed and counterfactual distributions of test scores for black and white children. Following Section 2, we denote these distributions by $F_{\langle j|k\rangle}$, where

$$F_{\langle j|k\rangle}(y) = \int F_{Y_j|X_j}(y|x)dF_{X_k}(x), \quad y \in T, \tag{6.1}$$

for $j, k \in \{w, b\}$, where w and b refer to white and black, $F_{Y_j|X_j}$ is the conditional distribution of test scores in group j, and F_{X_k} is the distribution of background characteristics in group k. With these counterfactual test score distributions it is possible to decompose the observed black-white test score gap into

$$F^{\leftarrow}_{\langle w|w\rangle} - F^{\leftarrow}_{\langle b|b\rangle} = [F^{\leftarrow}_{\langle w|w\rangle} - F^{\leftarrow}_{\langle w|b\rangle}] + [F^{\leftarrow}_{\langle w|b\rangle} - F^{\leftarrow}_{\langle b|b\rangle}],$$

where the first term in brackets corresponds is the composition effect due to differences in observable background characteristics and the second term is the unexplained difference.

We estimate $F_{\langle w|w\rangle}$ and $F_{\langle b|b\rangle}$ by the empirical test score distributions for white and black children, respectively. We estimate the counterfactual distribution $F_{\langle w|b\rangle}$ by the sample analog of (6.1) replacing $F_{Y_w|X_w}$ by the DR estimator for white children, and F_{X_b} by the empirical distribution of X for black children. We use the logistic link function for the DR, but the results using the linear link function or the normal link function are similar.

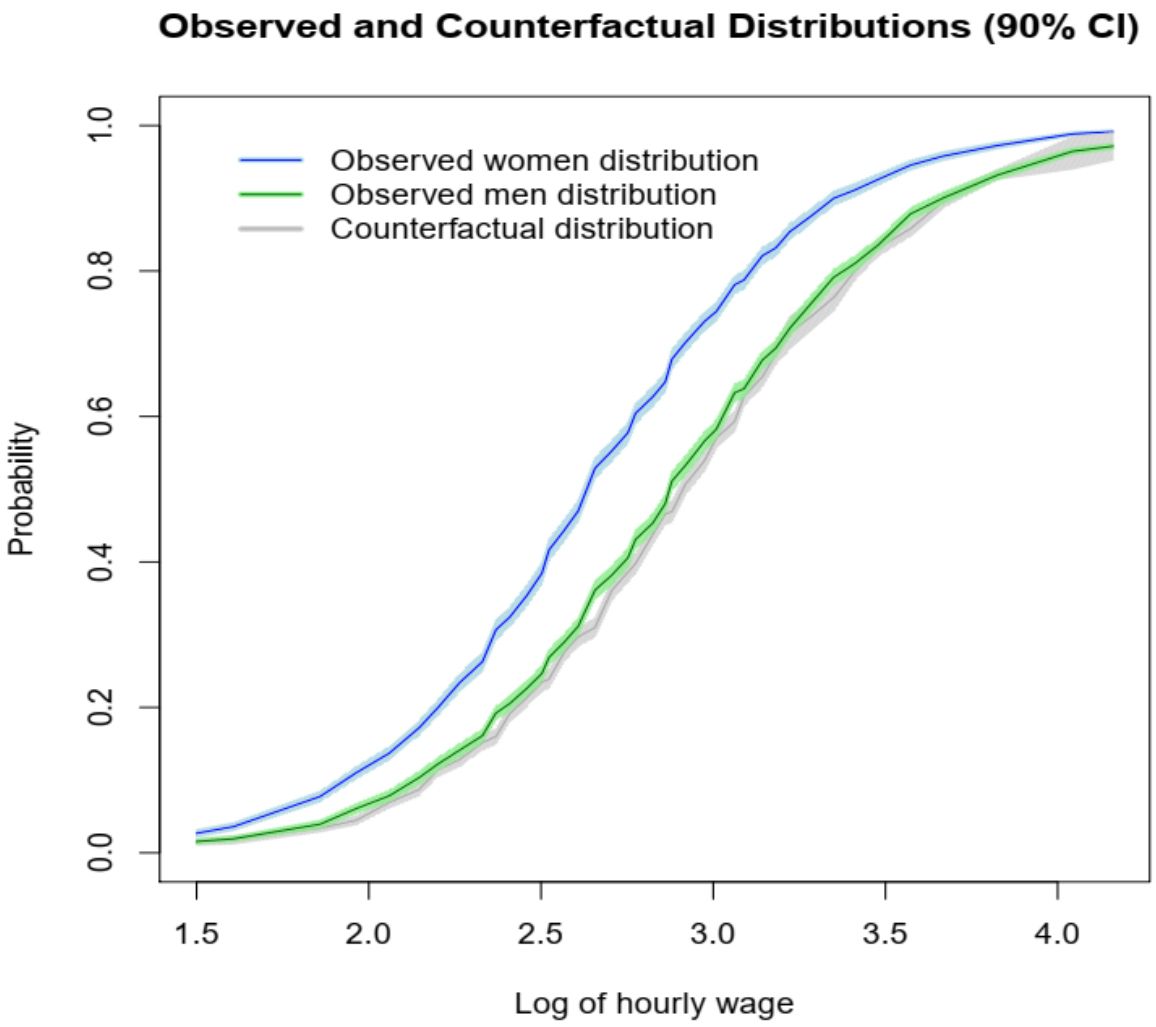

FIGURE 4. The distribution functions of observed wages for women and men, and the distribution function of counterfactual wages for women under men's wage structure, with 90% joint confidence bands. Confidence bands obtained by empirical bootstrap with 200 repetitions.

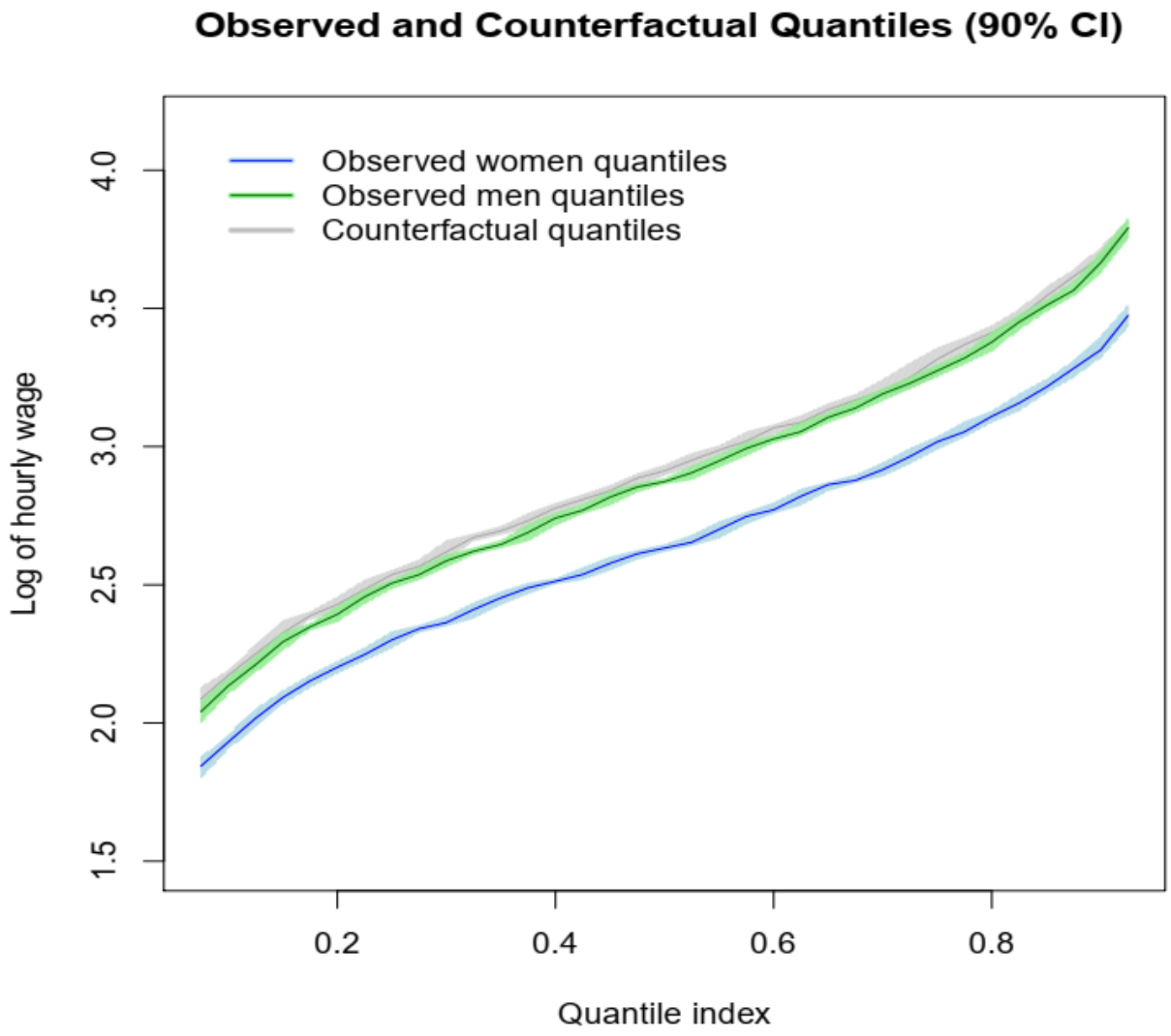

FIGURE 5. The quantile functions of observed wages for women and men, and the quantile function of counterfactual wages for women under men's wage structure, with 90% joint confidence bands.

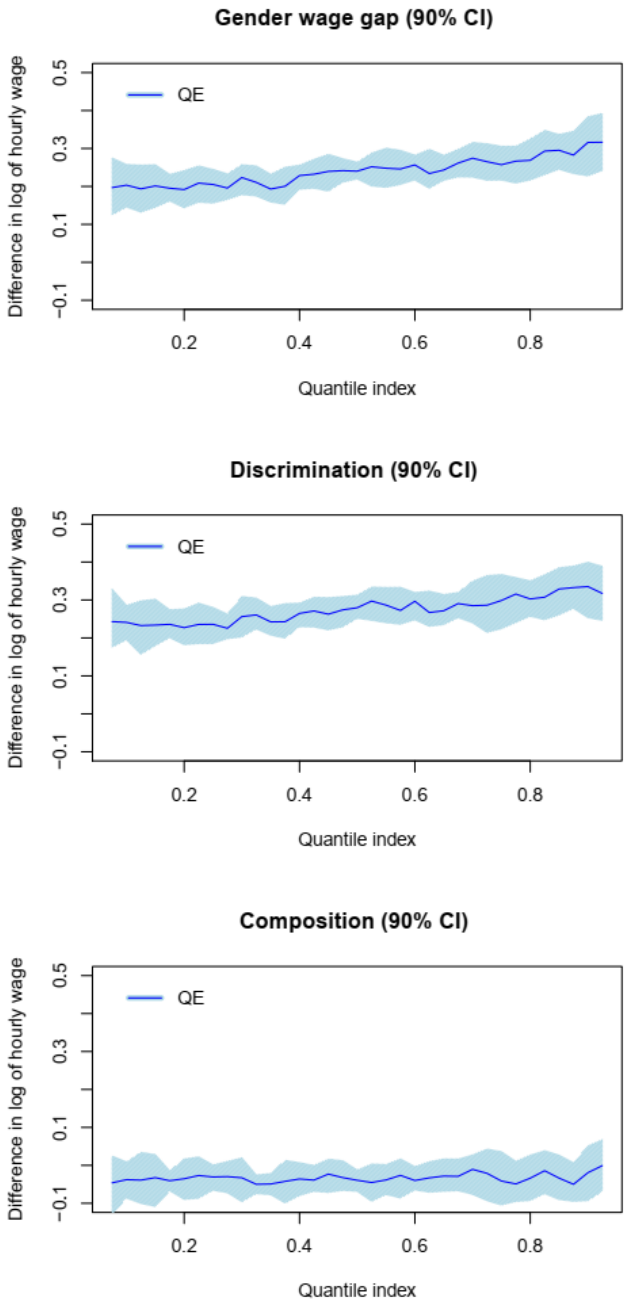

FIGURE 6. Decomposition of observed gender wage gap into composition and discrimination effects, with 90% confidence bands. The composition effect is the difference between the quantile functions of observed wages for men and counterfactual wages for women under the men's wage structure. The discrimination effect is the difference between the quantile functions of counterfactual wages for women under the men's wage structure and observed wages for women.

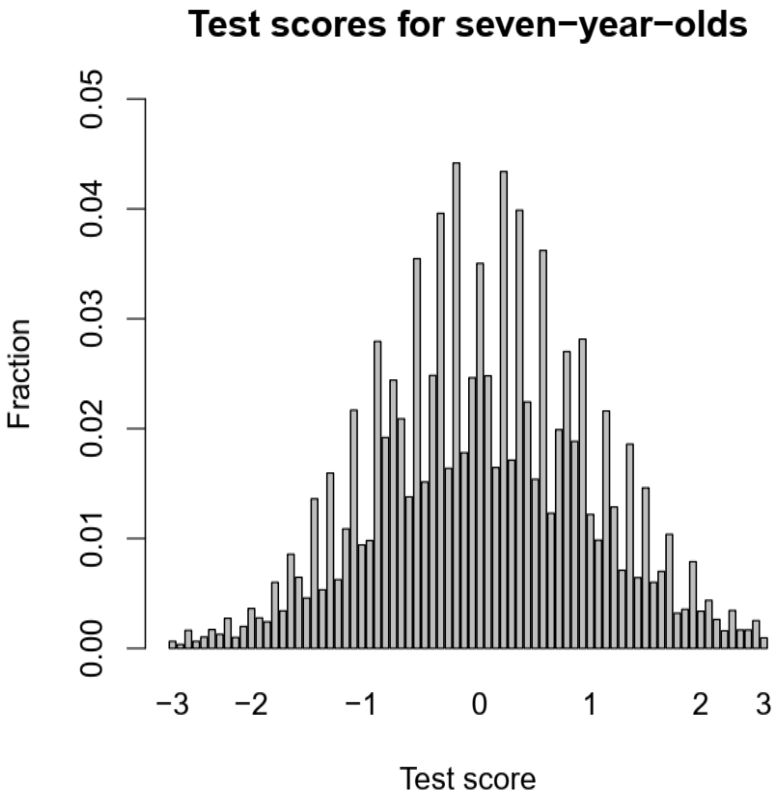

FIGURE 7. Histogram of test scores of seven year old children. Each bind corresponds to a unique value of the test score.

Figure 8 reports the results of the decomposition. The first panel shows the observed and counterfactual quantile functions, $F^{\leftarrow}_{\langle w|w\rangle}$, $F^{\leftarrow}_{\langle b|b\rangle}$ and $F^{\leftarrow}_{\langle w|b\rangle}$. The second panel shows the difference between the observed quantile functions, $F^{\leftarrow}_{\langle w|w\rangle} - F^{\leftarrow}_{\langle b|b\rangle}$. The third and fourth panels decompose these observed differences into the composition effect ($F^{\leftarrow}_{\langle w|w\rangle} - F^{\leftarrow}_{\langle w|b\rangle}$) and the unexplained component ($F^{\leftarrow}_{\langle w|b\rangle} - F^{\leftarrow}_{\langle b|b\rangle}$). The point estimates are shown with their respective 95% simultaneous confidence bands constructed by Algorithm 2 using weighted bootstrap with standard exponential weights and $B = 1,000$ replications. The bands impose the restrictions that the supports of the test scores correspond to the observed values in the sample.

We find a large and statistically significant positive raw black-white gap. A formal test based on the uniform bands rejects the null hypothesis of a zero or a negative racial test score gap at all quantiles. The estimated QE function is increasing in the quantile index ranging from below 0.6 standard deviation units at the lower tail up to over one standard deviation unit at the upper tail of the distribution. The quantile differences at the tails

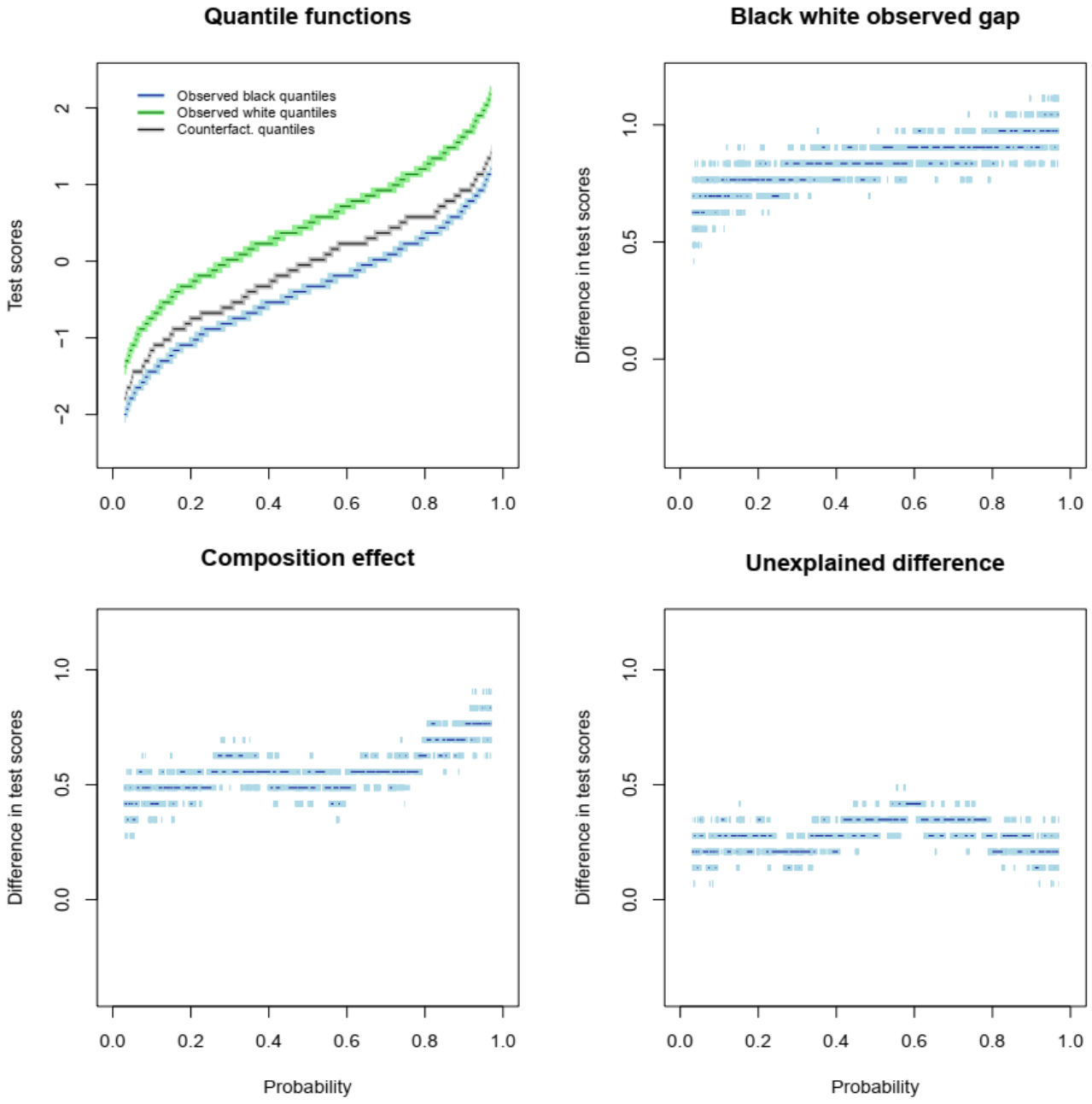

FIGURE 8. Decomposition of observed racial differences in mental ability of seven year old children. Quantile functions, raw difference, composition effect, and unexplained difference including support restricted 95% confidence bands.

substantially differ from the mean difference of 0.85 standard deviation units reported in [7]. In fact, we can formally reject the null hypothesis of a constant raw test score gap across the distribution because we can not draw a horizontal line at any value of the difference

of test scores, which is covered by the confidence band of the QE function at all quantile indexes.

Our decomposition analysis shows that about two third of this gap can be explained by differences in the distribution of observable characteristics. Nevertheless, the remaining unexplained difference is significant, both in economic and in statistical terms. Looking at the quantile effect function, we can see that there is substantial effect heterogeneneity along the distribution. Interestingly, the increase in the test score gap at the upper quantiles can be fully explained by differences in background characteristics between black and white children. The resulting unexplained difference is maximized in the center of the distribution. Finally, our simultaneous confidence bands allow for testing several interesting hypothesis' about the whole quantile effect function. For instance, we can reject the null hypothesis that the composition effect and the unexplained difference are zero, negative, or constant at all quantiles but we cannot reject that they are positive everywhere.

Notes

The distribution regression model was developed in [11], [5], and [3]. Ronald Oaxaca [10] and Alan Blinder [1] pioneered the use of least squares methods to carry out decompositions of mean wages. The decomposition was extended to distributions in [4], [8], [6], and [3], among others. The generic inference method of Section 3 was developed in [2].

Appendix A. Problems

(1) Demonstrate that estimation of the counterfactual distribution can be formulated as a GMM framework.

(2) Replicate the results of the empirical section. Provide a brief explanation for what you are doing. You can use the R code provided, but you need to write comments in the code explaining what each block of the code is doing.

References

[1] Alan S. Blinder. Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, 1973.

[2] V. Chernozhukov, I. Fernández-Val, B. Melly, and K. Wüthrich. Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes. *ArXiv e-prints*, August 2016.

[3] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[4] John DiNardo, Nicole M. Fortin, and Thomas Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044, 1996.

[5] Silverio Foresi and Franco Peracchi. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90(430):451–466, 1995.

[6] Nicole Fortin, Thomas Lemieux, and Sergio Firpo. *Decomposition Methods in Economics*, volume 4 of *Handbook of Labor Economics*, chapter 1, pages 1–102. Elsevier, 2011.

[7] Jr. Fryer, Roland G. and Steven D. Levitt. Testing for racial differences in the mental ability of young children. *American Economic Review*, 103(2):981–1005, April 2013.

[8] José A. F. Machado and José Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *J. Appl. Econometrics*, 20(4):445–465, 2005.

[9] Casey B. Mulligan and Yona Rubinstein. Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, 123(3):1061–1110, 2008.

[10] Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, 1973.

[11] O Dale Williams and James E Grizzle. Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association*, 67(337):55–63, 1972.