THE SORTED EFFECTS METHOD: DISCOVERING HETEROGENEOUS EFFECTS BEYOND THEIR AVERAGES

VICTOR CHERNOZHUKOV, IVAN FERNANDEZ-VAL, YE LUO

Abstract. The partial (ceteris paribus) effects of interest in nonlinear and interactive linear models are heterogeneous as they can vary dramatically with the underlying observed or unobserved covariates. Despite the apparent importance of heterogeneity, a common practice in modern empirical work is to largely *ignore it* by reporting average partial effects (or, at best, average effects for some groups). While average effects provide very convenient scalar summaries of typical effects, by definition they fail to reflect the entire variety of the heterogeneous effects. In order to discover these effects much more fully, we propose to estimate and report *sorted effects* – a collection of estimated partial effects sorted in increasing order and indexed by percentiles. By construction the sorted effect curves *completely represent* and help visualize the range of the heterogeneous effects in one plot. They are as convenient and easy to report in practice as the conventional average partial effects. They also serve as a basis for *classification analysis*, where we divide the observational units into most or least affected groups and summarize their characteristics. We provide a quantification of uncertainty (standard errors and confidence bands) for the estimated sorted effects and related classification analysis, and provide confidence sets for the most and least affected groups. The derived statistical results rely on establishing key, new mathematical results on Hadamard differentiability of a multivariate sorting operator and a related classification operator, which are of independent interest.

We apply the *sorted effects method* and classification analysis to demonstrate several striking patterns in the gender wage gap. We find that this gap is particularly strong for married women, ranging from −60% to 0% between the 2% and 98% percentiles, as a function of observed and unobserved characteristics; while the gap for never married women ranges from −40% to +20%. The most adversely affected women tend to be married, do not have college degrees, work in sales, and have high levels of potential experience.

Keywords: Sorting, Partial Effect, Marginal Effect, Sorted Effect, Classification Analysis, Nonlinear Model, Functional Analysis, Differential Geometry, Gender Wage Gap

1. introduction

In nonlinear and interactive linear models the partial (ceteris paribus) effects of interest often vary with respect to the underlying covariates. For example, consider a binary response model with conditional choice probability $\mathrm{P}(Y = 1 \mid X) = F(X^{\mathsf{T}}\beta)$, where Y is a binary response variable, X is a vector of covariates, F is a distribution function such as the standard normal or logistic, and β is a vector of coefficients. The partial or predictive effect (PE) of a marginal

Date: May 25, 2018.

MIT, BU and University of Florida. First version of October, 2011. We thank the coeditor, three anonymous referees, Alberto Abadie, Daron Acemoglu, Isaiah Andrews, Joshua Angrist, Debopam Bhattacharya, Jaroslav Borovicka, Kasey Buckles, Denis Chetverikov, Jerry Hausman, Tetsuya Kaji, Pat Kline, Roger Koenker, Siyi Luo, Maddie McKelway, Anna Mikusheva, Whitney Newey, Claudia Olivetti, Andres Santos, Camille Terrier, Frank Windmeijer, Abigail Wozniak, and seminar participants at multiple institutions for helpful discussions. We gratefully acknowledge research support from the National Science Foundation.

change in a continuous covariate X_j with coefficient β_j on the conditional choice probability is

$$\Delta(X) = f(X^{\mathsf{T}}\beta)\beta_j, \quad f(v) = \partial F(v)/\partial v,$$

which generally varies in the population of interest with the covariate vector X, as X varies according to some distribution, say μ. A common empirical practice is to report the average partial effect (APE),

$$\mathrm{E}[\Delta(X)] = \int \Delta(x) d\mu(x),$$

as a single summary measure of the PE (e.g., Wooldridge (2010, Chap. 2)), or to report effects for some groups (e.g., Angrist and Pischke (2008)). However, the APE completely disregards the heterogeneity of the PE and may give a very incomplete picture of the impact of the covariates.

In this paper we propose complementing the APE by reporting the entire set of PEs sorted in increasing order and indexed by a ranking with respect to the distribution of the covariates in the population of interest. These sorted effects correspond to percentiles of the PE,

$$\Delta^*_\mu(u) = u^{th}\text{-quantile of } \Delta(X), \quad X \sim \mu,$$

and provide a more complete representation of the heterogeneity of $\Delta(X)$. We shall call these effects as sorted predictive or partial effects (SPE) by default, as most models are predictive.[1] We also show how to use the SPEs to carry out classifications analysis (CA). This analysis consists of classifying the observational units into most or least affected depending on whether their PEs are above or below some tail SPE, and then comparing the moments or distribution of the covariates of the most and least affected groups.

Heterogeneous effects also arise in the most basic linear models with interactions (Oaxaca, 1973; Cox, 1984). Consider a conditional mean model for the Mincer earnings function:

$$Y = P(T,W)^{\mathsf{T}}\beta + \epsilon, \quad \mathrm{E}[\epsilon \mid T, W] = 0, \quad X = (T,W),$$

where Y is log wage, T is an indicator of gender (or race, treatment, or program participation), and W is a vector of labor market characteristics. The vector $P(T,W)$ is a collection of transformations of T and W, involving some interaction between T and W. For example, Oaxaca (1973) used the specification $P(T,W) = (TW, (1-T)W)$. Then, the PE of changing $T = 0$ to $T = 1$ is

$$\Delta(X) = P(1,W)^{\mathsf{T}}\beta - P(0,W)^{\mathsf{T}}\beta,$$

which is a measure of the gender wage gap conditional on worker characteristics. The function $u \mapsto \Delta^*_\mu(u)$ provides again a complete summary of the entire range of PEs. The left panel of Figure 1 illustrates the SPE of the conditional gender wage gap for women. The SPE varies sharply from around -40 to 6.5%, and does not coincide with the average PE of -20%. The PE is especially (negatively) large for women who have any of the following characteristics: married, low educated, high experience, and working on sales occupations – this follows from the classification analysis,

[1] When the underlying model has a structural or causal interpretation, we may use the name sorted structural effects or sorted treatment effects.

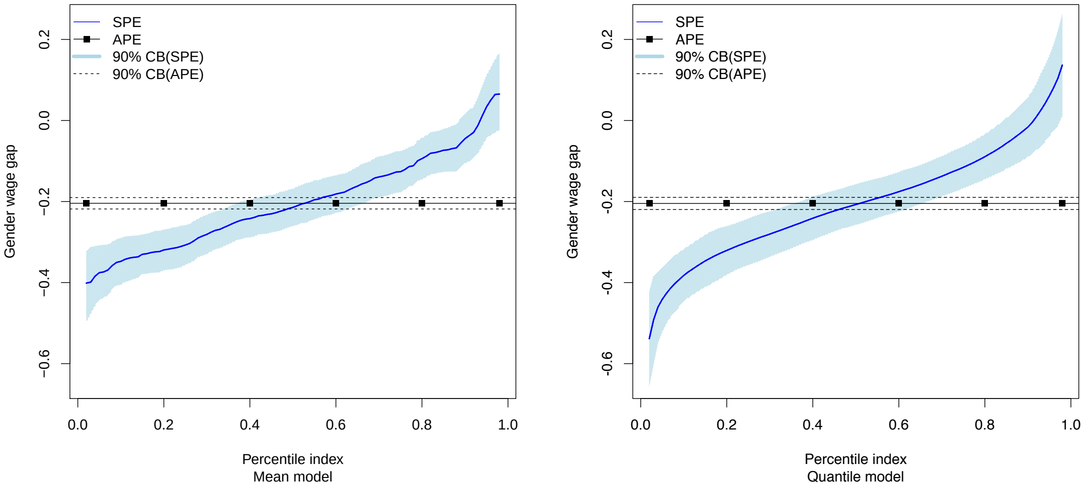

FIGURE 1. APE and SPE (introduced in this paper) of the gender wage gap for women. Estimates and 90% bootstrap uniform confidence bands (derived in this paper) based on a linear model with interactions for the conditional expectation (left) and quantile (right) functions.

where we compare the average characteristics of the subpopulations of women with covariate values X such that $\Delta(X)$ is above the 90% percentile and below the 10% percentile. We refer the reader to Section 3 for a detailed discussion of this example.

The general settings that we deal with in this paper as well as the specific results we obtain are as follows: Let X denote a covariate vector, $\Delta(X)$ denote a generic PE of interest, μ denote the distribution of X in the population of interest, and $\mathcal{X}$ denote the interior of the support of X in this population. The SPE is obtained by sorting the multivariate function $x \mapsto \Delta(x)$ in increasing order with respect to μ. Using tools from differential geometry, we prove that this multivariate sorting operator is Hadamard differentiable with respect to the PE function Δ and the distribution μ at the regular values of $x \mapsto \Delta(x)$ on $\mathcal{X}$. This key and new mathematical result allows us to derive the large sample properties of the empirical SPE, which replace Δ and μ by sample analogs, obtained from parametric or semi-parametric estimators, using the functional delta method. In particular, we derive a functional central limit theorem and a bootstrap functional central limit theorem for the empirical SPE. The main requirement of these theorems is that the empirical Δ and μ also satisfy functional central limit theorems, which hold for many estimators used in empirical economics under general sampling conditions. We use the properties of the empirical SPE to construct confidence sets for the SPE that hold uniformly over quantile indices. We also show under the same conditions that the empirical version of the objects in the classification analysis follow functional central limit theorems and bootstrap functional central limit theorems.

We derive these result by establishing the Hadamard differentiability of a classification operator related to the multivariate sorting operator.

Related technical literature: Previously, Chernozhukov, Fernández-Val, and Galichon (2010) derived the properties of the rearrangement (sorting) operator in the univariate case with known μ (standard uniform distribution). Those results were motivated by a completely different problem – namely, restoration of monotonicity in conditional quantile estimation – rather than the problem of summarizing heterogeneous effects by the SPEs. These prior technical results are not applicable to our case as soon as the dimension of X is greater than one, which is the case in all modern applications where effects are of interest. Moreover, the previous results are not applicable even in the univariate case since the measure μ is not known in all envisioned applications. The properties of the sorting operator are different in the multivariate case and require tools from differential geometry: computation of functional (Hadamard) derivatives of the sorting operator with respect to perturbations of Δ require us to work with integration on $(d_x - 1)$-dimensional manifolds of the type $\{\Delta(x) = \delta\}$, where $d_x = \dim X$. Moreover, we also need to compute functional derivatives with respect to suitable perturbations of the measure μ. In econometrics or statistics, Sasaki (2015) also used differential geometry to characterize the structural properties of derivatives of conditional quantile functions in nonseparable models; and Kim and Pollard (1990) used tools from differential geometry to derive the large sample properties of the maximum score and other cube root consistent estimators. Relative to these papers, we share the use of differential geometry tools as a general proof strategy, but we apply these tools to establish the analytical properties of different functionals – namely, the SPEs. Moreover, our results on the functional differentiability of the sorting and classification operators in the multivariate case constitute new mathematical results, which are of interest in their own right.

Organization of the paper: In Section 2 we discuss the quantities of interest in nonlinear and interactive linear models with examples; introduce the SPE and related CA, along with their empirical counterparts; and outline the main inferential results. In Section 3 we provide an empirical application to the gender wage gap in the U.S. in 2015. We derive the properties of the empirical SPE and CA in large samples and show how to use these properties to make inference in Section 4. Appendix A provides some key mathematical results on the differentiability of the multivariate sorting and classification operators and Appendix B contains the proof of the main results. All other proofs are given in the online appendix with supplementary material (SM), which also contains additional technical material, and results from Monte Carlo simulations and an empirical application to mortgage denials using binary response models (Chernozhukov, Fernandez-Val, and Luo, 2017).

2. Sorted Effects and Classification Analysis

We start by discussing the objects of interest in nonlinear and interactive linear models.

2.1. **Effects of Interest.** We consider a general model characterized by a predictive function $g(X)$, where X is a d_x-vector of covariates that may contain unobserved components, as in quantile regression models. The function g usually arises from a model for a response variable Y, which can be discrete or continuous. We call the function g predictive because the underlying model can be either predictive or causal under additional assumptions, but we do not insist on estimands having a causal interpretation. For example, in a mean regression model, $g(X) = \mathrm{E}[Y \mid X]$ corresponds to the expectation function of Y conditional on X; in a binary response model, $g(X) = \mathrm{P}[Y = 1 \mid X]$ corresponds to the choice probability of $Y = 1$ conditional on X; in a quantile regression model, $g(X) = \mathrm{Q}_Y[\epsilon \mid Z]$, where the covariate $X = (\epsilon, Z)$ consists of the unobservable rank variable ϵ with a uniform distribution, $\epsilon \mid Z \sim U(0,1)$, and the observed covariate vector Z, and where $\mathrm{Q}_Y[\tau \mid Z]$ is the conditional τ^{th}-quantile of Y given Z.

Let $X = (T, W)$, where T is the key covariate or treatment of interest, and W is a vector of control variables. We are interested in the effects of changes in T on the function g holding W constant. These effects are usually called partial effects, marginal effects, or treatment effects. We call them predictive effects (PE) throughout the paper, as such a name most accurately describes the meaning of the estimand (especially when a causal interpretation is not available). If T is discrete, the PE is

$$\Delta(x) = \Delta(t, w) = g(t_1, w) - g(t_0, w), \tag{2.1}$$

where t_1 and t_0 are two values of T that might depend on t (e.g., $t_0 = 0$ and $t_1 = 1$, or $t_0 = t$ and $t_1 = t + 1$). This PE measures the effect of changing T from t_0 to t_1 holding W constant at w. If T is continuous and $t \mapsto g(t, w)$ is differentiable, the PE is

$$\Delta(x) = \Delta(t, w) = \partial_t g(t, w), \tag{2.2}$$

where ∂_t denotes $\partial/\partial t$, the partial derivative with respect to t. This PE measures the effect of a marginal change of T from the level t holding W constant at w.[2]

We consider the following examples in the empirical applications of Section 3 and SM.

Example 1 (Binary response model)**.** Let Y be a binary response variable such as an indicator for mortgage denial, and X be a vector of covariates related to Y. The predictive function of the probit or logit model takes the form:

$$g(X) = \mathrm{P}(Y = 1 \mid X) = F(P(X)^{\mathsf{T}}\beta),$$

where $P(X)$ is a vector of known transformations of X, β is a parameter vector, and F is a known distribution function (the standard normal distribution function in the probit model or standard logistic distribution function in the logit model). If T is a binary variable such as an indicator

[2] We can also consider high-order and crossed effects. For example, $\Delta(x) = \partial^2_{t^2} g(t, w)$ gives the second-order PE of the continuous treatment T if $t \mapsto g(t, w)$ is twice differentiable; and, letting $X = (T, S, W)$ where T and S are discrete, $\Delta(x) = g(t_1, s_1, w) - g(t_0, s_1, w) - g(t_1, s_0, w) + g(t_0, s_0, w)$ gives the crossed effect or interaction of T and S.

for black applicant and W is a vector of controls such as the applicant characteristics relevant for the bank decision, the PE,

$$\Delta(x) = F(P(1,w)^{\mathsf{T}}\beta) - F(P(0,w)^{\mathsf{T}}\beta),$$

describes the difference in predicted probability of mortgage denial for a black applicant and a white applicant, conditional on a specific value w of the observable characteristics W. □

Example 2 (Interactive linear model with additive error)**.** Let Y be the logarithm of the wage. Suppose $X = (T, W)$, where T is an indicator for female worker and W are other worker characteristics. We can model the conditional expectation function of log wage using the linear interactive model:

$$Y = g(X) + \epsilon = P(T,W)^{\mathsf{T}}\beta + \epsilon, \quad \mathrm{E}[\epsilon \mid T, W] = 0, \quad X = (T, W),$$

where $P(T,W)$ is a collection of transformations of T and W, involving some interaction between T and W. For example, $P(T,W) = (TW, (1-T)W)$. Then the PE

$$\Delta(x) = P(1,w)^{\mathsf{T}}\beta - P(0,w)^{\mathsf{T}}\beta$$

is the (average) gender wage gap or difference between the expected log wage of a woman and a man, conditional on a specific value w of the characteristics W. □

Example 3 (Linear model with non-additive error, or QR model)**.** Let Y be log wage, T be an indicator for female worker, and W be a vector of worker characteristics as in the previous example. Suppose we model the conditional quantile function of log wage using the linear interactive model:

$$Y = g(X) = P(T,W)^{\mathsf{T}}\beta(\epsilon), \quad \epsilon \mid T, W \sim U(0,1), \quad X = (T, W, \epsilon),$$

where $P(T,W)^{\mathsf{T}}\beta(\tau)$ is the conditional τ^{th}-quantile of Y given T and W. Thus the covariate vector $X = (T, W, \epsilon)$ includes the observed covariates (T, W) as well as the rank variable ϵ, which is an unobserved factor (e.g., "ability rank"). Here $P(T,W)$ is a collection of transformations of T and W, e.g., $P(T,W) = (TW, (1-T)W)$. Then the PE

$$\Delta(x) = P(1,w)^{\mathsf{T}}\beta(\tau) - P(0,w)^{\mathsf{T}}\beta(\tau), \quad x = (t, w, \tau),$$

is the (τ^{th}-quantile) gender wage gap or difference between the conditional τ^{th}-quantile of log-wage of a woman and a man, conditional on a specific value w of the characteristics W.

Note that in this case,

$$X \sim \mu, \quad \mu = F_{T,W} \times F_{\epsilon},$$

where F_{ϵ} is the distribution function of the standard uniform random variable, and $F_{T,W}$ is the distribution of (T, W). For estimation purposes, we will have to exclude the tail quantile indices, so F_{ϵ} will be redefined to have support on a set of the form $[\ell, 1-\ell]$, where $0 < \ell < 0.5$ is a small positive number. □

The set of examples listed above are the most basic, leading cases, arising mostly in predictive analysis and program evaluation. Our theoretical results are rather general and are not limited to these cases. Thus, they allow for both Δ and μ to originate from causal or structural models and to be estimated by structural methods. For example, in treatment effects models with selection on observables (Rosenbaum and Rubin, 1983), the PE is the conditional average treatment effect $\Delta(x) = \mathrm{E}(Y_1 - Y_0 \mid X = x)$, where Y_1 and Y_0 are potential outcomes in the treated and non-treated statuses and X is a vector of covariates. The standard approach is to aggregate the conditional average treatment effects by integration with respect to the distribution of the covariates in the population of interest μ. This yields the average treatment effect if μ is the distribution in the entire population or the average treatment effect on the treated if μ is the distribution in the treated subpopulation. The SPE can be used to complement the analysis by reporting the entire range of conditional average treatment effects, and also to determine the optimal treatment allocation with budget constraints. Thus, Bhattacharya and Dupas (2012) showed that under some conditions this optimal allocation has a cutoff determined by a tail percentile of the conditional average treatment effects, i.e. by a SPE. Another example is the welfare analysis described in Hausman and Newey (2017), where $\Delta(x)$ is the compensating or equivalent variation of a price change conditional on covariates such as income and demographic characteristics, and μ is the distribution of covariates in the population of interest.

In all the previous examples, the PE $\Delta(x)$ is a function of x and therefore can be different for each observational unit. To summarize this effect in a single measure, a common practice in empirical economics is to average the PEs. Averaging, however, masks most of the heterogeneity in the PE allowed by nonlinear or interactive linear models. We propose reporting the entire set of values of the PE sorted in increasing order and indexed by a ranking $u \in [0,1]$ with respect to the population of interest. These sorted effects provide a more complete representation of the heterogeneity in the PE than the average effects.

Definition 2.1 (u-SPE)**.** The u^{th}*-sorted predictive effect* with respect to μ is

$$\Delta^*_\mu(u) := \inf\{\delta \in \mathbb{R} : F_{\Delta,\mu}(\delta) \geqslant u\}, \quad F_{\Delta,\mu}(\delta) := \mathrm{E}_\mu[1\{\Delta(X) \leqslant \delta\}],$$

where E_μ denotes expectation with respect to μ.

The u-SPE is the u^{th}-quantile of $\Delta(X)$ when X is distributed according to μ. As for the average effect, μ can be chosen to select a target subpopulation from the entire population. For example, when T is a treatment indicator:

- If μ is set to the marginal distribution of X in the entire population, then $\Delta^*_\mu(u)$ is the *population u-SPE.*
- If μ is set to the distribution of X conditional on $T = 1$, then $\Delta^*_\mu(u)$ is the *u-SPE on the treated or exposed.*

By considering $\Delta^*_\mu(u)$ at multiple quantile indices, we obtain a one-dimensional representation of the heterogeneity of the PE. Accordingly, our object of interest is the *SPE-function*

$$\{u \mapsto \Delta^*_\mu(u) : u \in \mathcal{U}\}, \quad \mathcal{U} \subseteq [0,1],$$

where $\mathcal{U}$ is the set of quantile indices of interest.

We also show how to use the u-SPE for classification analysis. Let $u \in \mathcal{U}$, with $u < 1/2$, and Z be a d_z-dimensional random vector that includes X and possibly other variables such as Y in Examples 1–3. By abuse of notation, we also denote the distribution of Z over its support $\mathcal{Z}$ as μ.

Definition 2.2 (u-CA)**.** The u^{th}-classification analysis consists of 2 steps: (i) Assign all Z with $\Delta(X) < \Delta^*_\mu(u)$ to the *u-least affected* subpopulation, and all Z with $\Delta(X) > \Delta^*_\mu(1-u)$ to the *u-most affected* subpopulation. (ii) Obtain the moments and distribution of Z in the least and most affected subpopulations. We denote by $\Lambda^{-u}_{\Delta,\mu}(t)$ and $\Lambda^{+u}_{\Delta,\mu}(t)$ generic objects indexed by $t \in \mathbb{R}^{d_z}$ in the least and most affected subpopulations, respectively. For example, $\Lambda^{-u}_{\Delta,\mu}(t) = \mathrm{E}_\mu[Z^t \mid \Delta(X) < \Delta^*_\mu(u)]$ corresponds to the t-moment of Z in the u-least affected subpopulation, and $\Lambda^{-u}_{\Delta,\mu}(t) = \mathrm{E}_\mu[1(Z \leqslant t) \mid \Delta(X) < \Delta^*_\mu(u)]$ to the distribution of Z at t in the u-least affected subpopulation.[3] We define the same quantities in the u-most affected subpopulation replacing $\Delta(X) < \Delta^*_\mu(u)$ by $\Delta(X) > \Delta^*_\mu(1-u)$ in the conditioning set.

2.2. Empirical SPE.

In practice, we replace the PE Δ and the distribution μ by sample analogs to construct plug-in estimators of the SPE. Let $\widehat{\Delta}(x)$ and $\widehat{\mu}(x)$ be estimators of $\Delta(x)$ and $\mu(x)$ obtained from $\{(Y_i, T_i, W_i) : 1 \leqslant i \leqslant n\}$, an independent and identically distributed sample of size n from (Y, T, W).

Definition 2.3 (Empirical u-SPE)**.** The estimator of Δ^*_μ is

$$\widehat{\Delta^*_\mu}(u) := \widehat{\Delta}^*_{\widehat{\mu}}(u) = \inf\{\delta \in \mathbb{R} : F_{\widehat{\Delta},\widehat{\mu}}(\delta) \geqslant u\}, \quad F_{\widehat{\Delta},\widehat{\mu}}(\delta) = \mathrm{E}_{\widehat{\mu}}[1\{\widehat{\Delta}(X) \leqslant \delta\}] =: \widehat{F_{\Delta,\mu}}(\delta).$$

Then the empirical SPE-function is

$$\{u \mapsto \widehat{\Delta^*_\mu}(u) : u \in \mathcal{U}\}, \quad \mathcal{U} \subseteq [0,1],$$

where $\mathcal{U}$ is the set of indices of interest that typically excludes tail indices and satisfies other technical conditions stated in Section 4.

Example 1 (Binary response model, cont.) The estimator of the PE is

$$\widehat{\Delta}(x) = F\left(P(1,w)^\intercal \widehat{\beta}\right) - F\left(P(0,w)^\intercal \widehat{\beta}\right),$$

[3] For a d_z-dimensional random variable $Z = (Z_1, \ldots, Z_{d_z})$ and $t = (t_1, \ldots, t_{d_z}) \in \mathbb{R}^{d_z}$, we denote $Z^t := \prod_{j=1}^{d_z} Z_j^{t_j}$ and $\{Z \leqslant t\} := \{Z_1 \leqslant t_1, \ldots, Z_{d_z} \leqslant t_{d_z}\}$.

where $\widehat{\beta}$ is the maximum likelihood (ML) estimator of β,

$$\widehat{\beta} \in \arg\max_{b\in\mathbb{R}^{d_p}} \sum_{i=1}^{n} [Y_i \log F(P(T_i, W_i)^{\mathsf{T}} b) + (1 - Y_i) \log\{1 - F(P(T_i, W_i)^{\mathsf{T}} b)\}], \quad d_p = \dim P(T, W).$$

□

Example 2 (Interactive linear model with additive error, cont.) The estimator of the PE is

$$\widehat{\Delta}(x) = P(1, w)^{\mathsf{T}} \widehat{\beta} - P(0, w)^{\mathsf{T}} \widehat{\beta},$$

where $\widehat{\beta}$ is the ordinary least squares (OLS) estimator of β,

$$\widehat{\beta} \in \arg\min_{b\in\mathbb{R}^{d_p}} \sum_{i=1}^{n} [Y_i - P(T_i, W_i)^{\mathsf{T}} b]^2, \quad d_p = \dim P(T, W).$$

□

Example 3 (Linear model with non-additive error, cont.) The estimator of the PE is

$$\widehat{\Delta}(x) = P(1, w)^{\mathsf{T}} \widehat{\beta}(\tau) - P(0, w)^{\mathsf{T}} \widehat{\beta}(\tau),$$

where $\widehat{\beta}(\tau)$ is the Koenker and Basset (1978) quantile regression (QR) estimator of $\beta(\tau)$,

$$\widehat{\beta}(\tau) \in \arg\min_{b\in\mathbb{R}^{d_p}} \sum_{i=1}^{n} \rho_\tau(Y_i - P(T_i, W_i)^{\mathsf{T}} b), \quad d_p = \dim P(T, W), \quad \rho_\tau(v) = (\tau - 1\{v < 0\})v.$$

□

Remark 2.1 (Estimation of μ)**.** Let S denote the indicator for an observational unit belonging to the subpopulation of interest. For example, if $S = T$, then $S = 1$ indicates the unit is in the subpopulation of the treated and $S = 0$ indicates the unit is in the subpopulation of the untreated. The indicator S can also incorporate other restrictions, for example $S = 1\{X \in \mathcal{X}\}$ restricts the support of covariate X to the region $\mathcal{X}$. Finally, if S is always 1, then this means that we work with the entire population. Estimation of μ can be done using the empirical distribution:

$$\widehat{\mu}(x) = \sum_{i=1}^{n} S_i 1\{X_i \leqslant x\} / \sum_{i=1}^{n} S_i,$$

provided that $\sum_{i=1}^{n} S_i > 0$. An alternative would be to use the smoothed empirical distribution.

If μ can be decomposed into known and unknown parts, then we only need to estimate the unknown parts. Thus, $\mu = F_{T,W} \times F_\epsilon$ in Example 3, where F_ϵ is known to be the uniform distribution and $F_{T,W}$ is unknown, but can be estimated by the empirical distribution of (T, W) in the part of the population of interest. □

2.3. **Empirical CA.** The empirical version of the u-CA classifies the observations in the sample using the empirical PEs and u-SPE, and computes the moments and distributions in the resulting most and least affected subsamples.

Definition 2.4 (Empirical u-CA)**.** The empirical u^{th}-classification analysis consists of 2 steps: (1) Assign all Z_i with $\widehat{\Delta}(X_i) < \widehat{\Delta^*_\mu}(u)$ to the u-least affected subsample, and all Z_i with $\widehat{\Delta}(X_i) > \widehat{\Delta^*_\mu}(1-u)$ to the u-most affected subsample. (2) Estimate the moments and distribution of Z in the least and most affected subpopulations by the empirical analogs in the least and most affected subsamples, i.e. $\widehat{\Lambda}^{-u}_{\Delta,\mu}(t) = \Lambda^{-u}_{\widehat{\Delta},\widehat{\mu}}(t)$ and $\widehat{\Lambda}^{+u}_{\Delta,\mu}(t) = \Lambda^{+u}_{\widehat{\Delta},\widehat{\mu}}(t)$. For example, $\widehat{\Lambda}^{-u}_{\Delta,\mu}(t) = \mathrm{E}_{\widehat{\mu}}\left[Z^t \mid \widehat{\Delta}(X) < \widehat{\Delta^*_\mu}(u)\right]$ estimates the t-moment of Z in the u-least affected subpopulation and $\widehat{\Lambda}^{-u}_{\Delta,\mu}(t) = \mathrm{E}_{\widehat{\mu}}\left[1(Z \leqslant t) \mid \widehat{\Delta}(X) < \widehat{\Delta^*_\mu}(u)\right]$ the distribution of Z at t in the u-least affected subpopulation. The corresponding estimators in the u-most affected subpopulation are constructed replacing $\widehat{\Delta}(X) < \widehat{\Delta^*_\mu}(u)$ by $\widehat{\Delta}(X) > \widehat{\Delta^*_\mu}(1-u)$ in the conditioning set. Here we use the same notation as in Definition 2.2.

2.4. **Inference on SPE.** The main inferential result for the SPE can be previewed as follows. Assume that the PE function $x \mapsto \Delta(X)$ is not locally flat in the sense that the norm of its gradient does not vanish anywhere over the support, and other regularity conditions stated in Section 4. Then, the empirical SPE-process is $\sqrt{n}$-consistent and converges in distribution to a centered Gaussian process, namely

$$\sqrt{n}(\widehat{\Delta}^*_{\widehat{\mu}}(u) - \Delta^*_\mu(u)) \rightsquigarrow Z_\infty(u) \text{ in } \ell^\infty(\mathcal{U}),$$

the metric space of bounded functions on $\mathcal{U}$, as a stochastic process indexed by $u \in \mathcal{U}$, where $\mathcal{U}$ is a compact subset of $(0,1)$. Moreover, the exchangeable bootstrap algorithm specified in Algorithm 2.1 estimates consistently the law of $Z_\infty(u)$.

The next corollary to Theorem 4.1 in Section 4 provides uniform bands that cover the SPE-function simultaneously over a region of values of u with prespecified probability in large samples. It does cover pointwise confidence bands for the SPE-function at a specific quantile index u as a special case by simply taking $\mathcal{U}$ to be the singleton set $\{u\}$.

Corollary 2.1 (Inference on SPE-function using Limit Theory and Bootstrap)**.** *Under the assumptions of Theorem 4.1, for any* $0 < \alpha < 1$,

$$\mathrm{P}\left\{\Delta^*_\mu(u) \in \left[\widehat{\Delta^*_\mu}(u) - \widehat{t}_{1-\alpha}(\mathcal{U})\widehat{\Sigma}(u)^{1/2}/\sqrt{n}, \widehat{\Delta^*_\mu}(u) + \widehat{t}_{1-\alpha}(\mathcal{U})\widehat{\Sigma}(u)^{1/2}/\sqrt{n}\right] : u \in \mathcal{U}\right\} \to 1-\alpha,$$

where $\widehat{t}_{1-\alpha}(\mathcal{U})$ *is any consistent estimator of* $t_{1-\alpha}(\mathcal{U})$*, the* $(1-\alpha)$*-quantile of*

$$t(\mathcal{U}) := \sup_{u \in \mathcal{U}} |Z_\infty(u)| \Sigma(u)^{-1/2},$$

and $u \mapsto \widehat{\Sigma}(u)$ *is a uniformly consistent estimator of* $u \mapsto \Sigma(u)$*, the variance function of* $u \mapsto Z_\infty(u)$*. We provide consistent estimators of* $t_{1-\alpha}(\mathcal{U})$ *and* $u \mapsto \Sigma(u)$ *in Algorithm 2.1.*

We now describe a practical bootstrap algorithm to estimate the quantiles of $t(\mathcal{U})$. Let $(\omega_1, \dots, \omega_n)$ denote the bootstrap weights, which are nonnegative random variables independent of the data obeying the conditions stated in van der Vaart and Wellner (1996). For example, $(\omega_1, \dots, \omega_n)$ is a multinomial vector with dimension n and probabilities $(1/n, \dots, 1/n)$ in the empirical bootstrap. In what follows B is the number of bootstrap draws, such that $B \to \infty$. In our experience, setting $B \geqslant 500$ suffices for good accuracy.

Algorithm 2.1 (Bootstrap law of $t(\mathcal{U})$ and its quantiles)**.** ***1)*** *Draw a realization of the bootstrap weights* $(\omega_1, \dots, \omega_n)$. ***2)*** *For each* $u \in \mathcal{U}$, *compute* $\widetilde{\Delta^*_\mu}(u) = \widetilde{\Delta}^*_{\widetilde{\mu}}(u)$, *a bootstrap draw of* $\widehat{\Delta^*_\mu}(u) = \widehat{\Delta}^*_{\widehat{\mu}}(u)$, *where* $\widetilde{\Delta}$ *and* $\widetilde{\mu}$ *are the bootstrap versions of* $\widehat{\Delta}$ *and* $\widehat{\mu}$ *that use* $(\omega_1, \dots, \omega_n)$ *as sampling weights in the computation of the estimators. Construct a bootstrap draw of* $Z_\infty(u)$ *as* $\widetilde{Z}_\infty(u) = \sqrt{n}(\widetilde{\Delta^*_\mu}(u) - \widehat{\Delta^*_\mu}(u))$. ***3)*** *Repeat steps (1)-(2) B times.* ***4)*** *For each* $u \in \mathcal{U}$, *compute a bootstrap estimator of* $\Sigma(u)^{1/2}$ *such as the bootstrap interquartile range rescaled with the normal distribution,* $\widehat{\Sigma}(u)^{1/2} = (q_{0.75}(u) - q_{0.25}(u))/(z_{0.75} - z_{0.25})$, *where* $q_p(u)$ *is the pth sample quantile of* $\widetilde{Z}_\infty(u)$ *in the B draws and* z_p *is the pth quantile of* $N(0,1)$. ***5)*** *Use the empirical distribution of* $\widetilde{t}(\mathcal{U}) = \sup_{u \in \mathcal{U}} |\widetilde{Z}_\infty(u)| \widehat{\Sigma}(u)^{-1/2}$ *across the B draws to approximate the distribution of* $t(\mathcal{U}) = \sup_{u \in \mathcal{U}} |Z_\infty(u)| \Sigma(u)^{-1/2}$. *In particular, construct* $\widehat{t}_{1-\alpha}(\mathcal{U})$, *an estimator of* $t_{1-\alpha}(\mathcal{U})$, *as the* $(1-\alpha)$*-quantile of the B draws of* $\widetilde{t}(\mathcal{U})$.

Remark 2.2 (Monotonization of the bands)**.** While the SPE-function $u \mapsto \Delta^*_\mu(u)$ is increasing by definition, the end functions of the confidence band $u \mapsto \widehat{\Delta^*_\mu}(u) \pm \widehat{t}_{1-\alpha}(\mathcal{U}) \widehat{\Sigma}(u)^{1/2}/\sqrt{n}$ might not be increasing. Chernozhukov, Fernández-Val, and Galichon (2009) showed that monotonizing the end functions via rearrangement reduces the width of the band in uniform norm, while increases coverage in finite-samples. We use this refinement in the empirical examples.[4] □

Remark 2.3 (Finite-Sample Bias Corrections)**.** The empirical u-SPE might be biased in small samples, specially at the the tails. Bootstrap is also useful to improve the estimator and confidence bands. Thus, a corrected estimator can be formed as $2\widehat{\Delta^*_\mu} - \overline{\Delta^*_\mu}$, and a corrected $(1-\alpha)$-confidence band as $\left[2\widehat{\Delta^*_\mu} - \overline{\Delta^*_\mu} \pm \widehat{t}_{1-\alpha}(\mathcal{U}) \widehat{\Sigma}(u)^{1/2}/\sqrt{n}\right]$, where $\overline{\Delta^*_\mu}$ is the mean of the bootstrap draw of the estimator. In Appendix H of the SM, we show that this correction reduces the bias of the estimator and increases the coverage of the confidence bands in a simulation calibrated to the gender wage gap application. □

2.5. **Inference on CA.** Let $\Lambda^u_{\Delta,\mu}(t) := [\Lambda^{-u}_{\Delta,\mu}(t), \Lambda^{+u}_{\Delta,\mu}(t)]$ and $\widehat{\Lambda}^u_{\Delta,\mu}(t) := [\widehat{\Lambda}^{-u}_{\Delta,\mu}(t), \widehat{\Lambda}^{+u}_{\Delta,\mu}(t)]$. The main inferential result for CA can be previewed as follows: the empirical CA-process converges in distribution to a centered bivariate Gaussian process, namely

$$\sqrt{n}(\widehat{\Lambda}^u_{\Delta,\mu}(t) - \Lambda^u_{\Delta,\mu}(t)) \rightsquigarrow Z^u_\infty(t) \;\text{ in }\; \ell^\infty(\mathbb{R}^{d_z})^2, \tag{2.3}$$

[4] In practice, the rearrangement simply consists in sorting the two vectors containing the discretized version of the end-functions in increasing order; see Chernozhukov, Fernández-Val, and Galichon (2009) for more details.

as a stochastic process indexed by $t \in \mathbb{R}^{d_z}$. Moreover, exchangeable bootstrap estimates consistently the law of $Z^u_\infty(t)$.

The next corollary to Theorem 4.2 in Section 4 provides uniform bands that cover L linear combinations of the 2-dimensional vector $\Lambda^u_{\Delta,\mu}(t)$ with coefficients $c_1, \ldots, c_L$ simultaneously over $t \in \mathcal{T}$ with prespecified probability in large samples. It covers pointwise confidence intervals for the mean of the k^{th} component of Z for least affected as a special case with $L = 1$ linear combination, $c_1 = (1, 0)'$, and $\mathcal{T} = \{e_k\}$, where e_k is a unit vector with a one in the k^{th} position. Joint confidence intervals for s differences of means of the k_1^{th}, …, k_s^{th} components of Z between most and least affected are a special case with $L = 1$ linear combination, $c_1 = (-1, 1)'$, and $\mathcal{T} = \{e_{k_1}, \ldots, e_{k_s}\}$. Joint uniform bands for the distribution of the k^{th} component of Z for most and least affected are also a special case with $L = 2$ linear combinations, $c_1 = (1, 0)$, $c_2 = (0, 1)$, and $\mathcal{T} = \{t \in \mathbb{R}^{d_z} : t_j = \bar{T}, j \neq k\}$, where $\bar{T}$ is an arbitrarily large number. By appropriate choice of the linear combinations and the index set T, we can therefore conduct multiple tests while preserving the significance level from simultaneous inference problems (Romano, Shaikh, and Wolf, 2010a; Romano, Shaikh, and Wolf, 2010b; List, Shaikh, and Xu, 2016). We show examples in the empirical application of Section 3.

Corollary 2.2 (Inference on CA-function using Limit Theory and Bootstrap)**.** *Under the assumptions of Theorem 4.2, for any* $0 < \alpha < 1$,

$$\mathrm{P}\left\{c_\ell' \Lambda^u_{\Delta,\mu}(t) \in c_\ell' \widehat{\Lambda}^u_{\Delta,\mu}(t) \pm \widehat{t}^u_{1-\alpha}(\mathcal{T}, L)[c_\ell' \widehat{\Sigma}^u(t) c_\ell]^{1/2}/\sqrt{n} : t \in \mathcal{T}, \ell = 1, \ldots, L\right\} \to 1 - \alpha,$$

where $\widehat{t}^u_{1-\alpha}(\mathcal{T}, L)$ *is any consistent estimator of* $t^u_{1-\alpha}(\mathcal{T}, L)$*, the* $(1-\alpha)$*-quantile of*

$$t^u(\mathcal{T}, L) := \sup_{t \in \mathcal{T}, \ell = 1, \ldots, L} |c_\ell' Z^u_\infty(t)| [c_\ell' \Sigma^u(t) c_\ell]^{-1/2},$$

and $t \mapsto \widehat{\Sigma}^u(t)$ *is a uniformly consistent estimator of* $t \mapsto \Sigma^u(t)$*, the variance function of* $t \mapsto Z^u_\infty(t)$*. A p-value of the null hypothesis* $c_\ell' \Lambda^u_{\Delta,\mu}(t) = r_\ell(t)$ *for all* $t \in \mathcal{T}$ *and* $\ell = 1, \ldots, L$ *of the realization of the statistic* $\sup_{t \in \mathcal{T}, \ell = 1, \ldots, L} |c_\ell' \widehat{\Lambda}^u_{\Delta,\mu}(t) - r_\ell(t)| [c_\ell' \widehat{\Sigma}^u(t) c_\ell]^{-1/2} = s$ *is*

$$S_{t^u(\mathcal{T},L)}(s) = \mathrm{P}\left(t^u_{1-\alpha}(\mathcal{T}, L) > s\right).$$

We provide consistent estimators of $t^u_{1-\alpha}(\mathcal{T}, L)$*,* $u \mapsto \Sigma^u(t)$ *and* $S_{t^u(\mathcal{T},L)}(t)$ *in Algorithm 2.2.*

Algorithm 2.2 (Bootstrap law of $t(\mathcal{T}, L)$, quantiles and p-values)**.** ***1)*** *Draw a realization of the bootstrap weights* $(\omega_1, \ldots, \omega_n)$*.* ***2)*** *For each* $t \in \mathcal{T}$*, compute* $\widetilde{\Lambda}^{-u}_{\Delta,\mu}(t) = \Lambda^{-u}_{\widetilde{\Delta},\widetilde{\mu}}(t)$ *and* $\widetilde{\Lambda}^{+u}_{\Delta,\mu}(t) = \Lambda^{+u}_{\widetilde{\Delta},\widetilde{\mu}}(t)$*, a bootstrap draw of* $\widehat{\Lambda}^{-u}_{\Delta,\mu}(t) = \Lambda^{-u}_{\widehat{\Delta},\widehat{\mu}}(t)$ *and* $\widehat{\Lambda}^{+u}_{\Delta,\mu}(t) = \Lambda^{+u}_{\widehat{\Delta},\widehat{\mu}}(t)$*, where* $\widetilde{\Delta}$ *and* $\widetilde{\mu}$ *are the bootstrap versions of* $\widehat{\Delta}$ *and* $\widehat{\mu}$ *that use* $(\omega_1, \ldots, \omega_n)$ *as sampling weights in the computation of the estimators. Construct a bootstrap draw of* $Z^u_\infty(t)$ *as* $\widetilde{Z}^u_\infty(t) = \sqrt{n}(\widetilde{\Lambda}^u_{\Delta,\mu}(t) - \widehat{\Lambda}^u_{\Delta,\mu}(t))$*, where* $\widetilde{\Lambda}^u_{\Delta,\mu}(t) = [\widetilde{\Lambda}^{-u}_{\Delta,\mu}(t), \widetilde{\Lambda}^{+u}_{\Delta,\mu}(t)]$*.* ***3)*** *Repeat steps (1)-(2) B times.* ***4)*** *For each* $t \in \mathcal{T}$ *and* $\ell = 1, \ldots, L$*, compute a bootstrap estimator of* $[c_\ell' \Sigma^u(t) c_\ell]^{1/2}$ *such as the bootstrap interquartile range rescaled with the normal distribution* $[c_\ell' \widehat{\Sigma}^u(t) c_\ell]^{1/2} = (q^u_{0.75}(t, \ell) - q^u_{0.25}(t, \ell))/(z_{0.75} - z_{0.25})$*, where* $q^u_p(t, \ell)$ *is the pth sample quantile of* $c_\ell' \widetilde{Z}^u_\infty(t)$ *in the B draws and* z_p *is the pth quantile of* $N(0, 1)$*.* ***5)*** *Use* *the empirical distribution of $\widetilde{t}(\mathcal{T}, L) = \sup_{t\in\mathcal{T},\ell=1,\ldots,L} |c_\ell' \widetilde{Z}_\infty^u(t)| [c_\ell' \widehat{\Sigma}^u(t) c_\ell]^{-1/2}$ across the B draws to approximate the distribution of $t(\mathcal{T}, L) = \sup_{t\in\mathcal{T},\ell=1,\ldots,L} |c_\ell' Z_\infty^u(t)| [c_\ell' \Sigma^u(t) c_\ell]^{-1/2}$. In particular, construct $\widehat{t}_{1-\alpha}(\mathcal{T}, L)$, an estimator of $t_{1-\alpha}(\mathcal{T}, L)$, as the $(1-\alpha)$-quantile of the B draws of $\widetilde{t}(\mathcal{T}, L)$, and an estimation of the p-value $S_{t^u(\mathcal{T},L)}(s)$ as the proportion of the B draws of $\widetilde{t}(\mathcal{T}, L)$ that are greater than s.*

2.6. **Inference on Most and Least Affected Subpopulations.** In addition to moments and distributions, we can conduct inference on the subpopulations of most and least affected.[5] Let

$$\mathcal{M}^{-u} := \{(x,y) \in \mathcal{Z} : \Delta(x) \leqslant \Delta_\mu^*(u)\}, \quad \mathcal{M}^{+u} := \{(x,y) \in \mathcal{Z} : \Delta(x) \geqslant \Delta_\mu^*(1-u)\},$$

be the sets representing the u-least and u-most affected subpopulation, respectively. Here we assume that $\mathcal{Z}$ is compact or that the support of (X, Y) has been intersected with a compact set to form $\mathcal{Z}$. We can construct an outer $(1-\alpha)$-confidence set for $\mathcal{M}^{-u}$ as[6]

$$\mathcal{CM}^{-u}(1-\alpha) = \{(x,y) \in \mathcal{Z} : \widehat{\Sigma}^{-1/2}(x,u)\sqrt{n}[\widehat{\Delta}(x) - \widehat{\Delta}_\mu^*(u)] \leqslant \widehat{c}(1-\alpha)\},$$

where $\widehat{c}(1-\alpha)$ is a consistent estimator of $c(1-\alpha)$, the $(1-\alpha)$-quantile of the random variable

$$V_\infty = \sup_{\{x\in\mathcal{X}:\Delta(x)=\Delta_\mu^*(u)\}} \Sigma^{-1/2}(x,u)[G_\infty(x) - Z_\infty(u)],$$

and $x \mapsto \widehat{\Sigma}(x,u)$ is a uniformly consistent estimator of $x \mapsto \Sigma(x,u)$, the variance function of the process $G_\infty(x) - Z_\infty(u)$ defined in Section 4. The estimator $\widehat{c}(1-\alpha)$ can be obtained as the $(1-\alpha)$-quantile of the bootstrap version of V_∞,

$$\widetilde{V}_\infty^* = \sup_{\{x\in\mathcal{X}:\widehat{\Delta}(x)=\widehat{\Delta}_\mu^*(u)\}} \widehat{\Sigma}^{-1/2}(x,u)\sqrt{n}\Big([\widetilde{\Delta}(x) - \widetilde{\Delta_\mu^*}(u)] - [\widehat{\Delta}(x) - \widehat{\Delta}_\mu^*(u)]\Big),$$

where $\widetilde{\Delta}(x)$ and $\widetilde{\Delta_\mu^*}(u)$ are defined as in Algorithm 2.1. A similar $(1-\alpha)$-confidence set, $\mathcal{CM}^{+u}(1-\alpha)$, can be constructed for $\mathcal{M}^{+u}$. These sets can be visualized by plotting all 2 or 3 dimensional projections of their elements. We provide an example of such plots in Section 3. An immediate consequence of the set inference results in Chernozhukov, Kocatulum, and Menzel (2015) and the results of this paper is the following corollary:

Corollary 2.3 (Inference on Most and Least Affected Subpopulations)**.** *The sets $\mathcal{CM}^{-u}(1-\alpha)$ and $\mathcal{CM}^{+u}(1-\alpha)$ cover $\mathcal{M}^{-u}$ and $\mathcal{M}^{+u}$ with probability approaching $1-\alpha$, and $\mathcal{CM}^{-u}(1-\alpha)$ and $\mathcal{CM}^{+u}(1-\alpha)$ are consistent in the sense that they approach to $\mathcal{M}^{-u}$ and $\mathcal{M}^{+u}$ at a $\sqrt{n}$-rate with respect to the Hausdorff distance.*

[5] Here we follow the set inference approach described in Chernozhukov, Kocatulum, and Menzel (2015), which builds on Chernozhukov, Hong, and Tamer (2007). In addition our results justify the use of subsapling-based methods as in Chernozhukov, Hong, and Tamer (2007) and Romano and Shaikh (2010).

[6] Note that we can also similarly construct an inner confidence region, which is the complement of the outer confidence region of $\mathcal{X} \setminus \mathcal{M}^{-u}$, see Chernozhukov, Kocatulum, and Menzel (2015) for relevant discussion.

3. Empirical Analysis of the Gender Wage Gap

We report the main results of the application to the gender wage gap using data from the U.S. March Supplement of the Current Population Survey (CPS) in 2015. In Appendix H of the SM, we complement the analysis with supporting results from a simulation calibrated to this application. There, we find that our estimation and inference methods perform well in finite samples that closely mimic the characteristics of the CPS data. This exercise serves to indirectly verify the plausibility of the main regularity conditions mentioned in Section 2 and formally stated in Section 4.

Our sample consists of white, non-hispanic individuals who are aged 25 to 64 years and work 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; individuals with allocated or missing information in any of the variables used in the analysis; and individuals with hourly wage rate below \$3. The resulting sample contains $32,523$ workers including $18,137$ men and $14,382$ of women.

We estimate interactive linear models with additive and non-additive errors, using mean and quantile regressions, respectively. The outcome variable 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. The key covariate T is an indicator for female worker, and the control variables W include 5 marital status indicators (widowed, divorced, separated, never married, and married); 5 educational attainment indicators (less than high school graduate, high school graduate, some college, college graduate, and advanced degree); 4 region indicators (midwest, south, west, and northeast); a quartic in potential experience constructed as the maximum of age minus years of schooling minus 7 and zero, i.e., $experience = \max(age - education - 7, 0)$; 5 occupation indicators (management, professional and related; service; sales and office; natural resources, construction and maintenance; and production, transportation and material moving); 12 industry indicators (mining, quarrying, and oil and gas extraction; construction; manufacturing; wholesale and retail trade; transportation and utilities; information; financial services; professional and business services; education and health services; leisure and hospitality; other services; and public administration); and all the two-way interactions between the education, experience, occupation and industry variables except for the occupation-industry interactions.[7] All calculations use the CPS sampling weights to account for nonrandom sampling in the March CPS.

[7] The sample selection criteria and the variable construction follow Mulligan and Rubinstein (2008). The occupation and industry categories follow the 2010 Census Occupational Classification and 2012 Census Industry Classification, respectively.

Table 1. Descriptive Statistics of Workers

	All	Women	Men		All	Women	Men
Log wage	3.15	3.02	3.25	O.manager	0.48	0.55	0.43
Female	0.44	1.00	0.00	O.service	0.10	0.10	0.09
MS.married	0.65	0.61	0.68	O.sales	0.23	0.31	0.16
MS.widowed	0.01	0.02	0.01	O.construction	0.09	0.01	0.15
MS.separated	0.02	0.02	0.02	O.production	0.11	0.04	0.17
MS.divorced	0.13	0.16	0.10	I.minery	0.03	0.01	0.04
MS.Nevermarried	0.19	0.18	0.20	I.construction	0.06	0.01	0.09
E.lhs	0.02	0.02	0.03	I.manufacture	0.14	0.08	0.18
E.hsg	0.25	0.21	0.28	I.retail	0.13	0.11	0.14
E.sc	0.28	0.29	0.27	I.transport	0.04	0.02	0.06
E.cg	0.28	0.30	0.27	I.information	0.02	0.02	0.03
E.ad	0.16	0.18	0.15	I.finance	0.08	0.10	0.07
R.northeast	0.19	0.19	0.19	I.professional	0.11	0.10	0.13
R.midwest	0.27	0.28	0.27	I.education	0.24	0.40	0.11
R.south	0.35	0.35	0.35	I.leisure	0.05	0.05	0.04
R.west	0.18	0.18	0.19	I.services	0.03	0.03	0.04
Experience	21.68	21.72	21.65	I.public	0.07	0.06	0.07

Source: March Supplement CPS 2015.

Table 1 reports sample means of the variables used in the analysis. Working women are more highly educated than working men, have about the same potential experience, and are less likely to be married and more likely to be divorced. They work relatively more often in managerial and sales occupations and in the industries providing education and health services. Working men are relatively more likely to work in construction and production occupations within non-service industries. The unconditional gender wage gap is 23%.

Figure 1 of Section 1 plots estimates and 90% confidence bands for the APE and SPE-function on the treated (women) of the conditional gender wage gap using additive and non-additive error models. The PEs are obtained as described in Examples 2 and 3 with $P(T,W) = (TW, (1-T)W)$. In this case $\dim P(T,W) = 332$, which makes it very difficult to identify any pattern about the gender wage gap just by looking at the regression coefficients. The distribution $F_{T,W}$ is estimated by the empirical distribution of (T,W) for women, and F_ϵ is approximated by a uniform distribution over the grid $\{.02, .03, \dots, .98\}$. The confidence bands are constructed using Algorithm 2.1 with standard exponential weights (weighted bootstrap) and $B = 500$, and are uniform for the SPE-function over the grid $\mathcal{U} = \{.01, .02, \dots, .98\}$. We monotonize the bands using the rearrangement method described in Remark 2.2, and implement the finite sample corrections described in Remark 2.3. After controlling for worker characteristics, the gender wage gap for women remains on average around 20%. More importantly, we uncover a striking amount of heterogeneity, with the PE ranging between -6.5 and 40% in the additive error model and between -14 and 54% in the non-additive error model.[8]

Table 2. Classification Table – Average Characteristics of the 10% Least and Most Affected Women by Gender Wage Gap

	10% Least		10% Most			10% Least		10% Most	
	Est.	S.E.	Est.	S.E.		Est.	S.E.	Est.	S.E
Log wage	3.08	0.03	2.97	0.03	O.manager	0.67	0.04	0.38	0.04
M.married	0.28	0.03	0.87	0.02	O.service	0.08	0.02	0.10	0.02
M.widowed	0.02	0.01	0.01	0.01	O.sales	0.19	0.03	0.42	0.04
M.separated	0.02	0.01	0.01	0.00	O.construction	0.02	0.01	0.01	0.00
M.divorced	0.15	0.02	0.07	0.02	O.production	0.03	0.01	0.09	0.02
M.nevermarried	0.52	0.03	0.04	0.01	I.minery	0.01	0.01	0.02	0.01
E.lhs	0.01	0.01	0.06	0.01	I.construction	0.02	0.01	0.02	0.01
E.hsg	0.08	0.02	0.30	0.04	I.manufacture	0.02	0.01	0.11	0.02
E.sc	0.15	0.03	0.23	0.04	I.retail	0.06	0.02	0.19	0.03
E.cg	0.37	0.04	0.17	0.03	I.transport	0.01	0.01	0.04	0.01
E.ad	0.39	0.04	0.24	0.03	I.information	0.04	0.01	0.05	0.01
R.ne	0.24	0.02	0.18	0.02	I.finance	0.03	0.02	0.09	0.03
R.mw	0.26	0.02	0.28	0.02	I.professional	0.06	0.02	0.04	0.02
R.so	0.31	0.02	0.39	0.03	I.education	0.46	0.04	0.33	0.04
R.we	0.19	0.02	0.15	0.02	I.leisure	0.10	0.03	0.01	0.01
Experience	13.05	1.03	26.32	0.75	I.services	0.09	0.02	0.01	0.01
					I.public	0.09	0.02	0.08	0.02

PE estimated from a linear conditional quantile model with interactions.
Standard Errors obtained by weighted bootstrap with 500 repetitions.

Table 2 shows the results of a classification analysis, exhibiting characteristics of women that are most and least affected by the gender wage gap together with standard errors obtained by weighted bootstrap. We focus here on the non-additive model, but the results from the additive model are similar. Since the PE are predominantly negative, we define the most affected as $\Delta(X) < \Delta^*_\mu(u)$ and the lest affected as $\Delta(X) > \Delta^*_\mu(1-u)$ to facilitate the interpretation. According to this model the 10% of the women *most affected* by the gender wage gap on average earn lower wages, are much more likely to be married, much less likely to be never married, have lower education, live in the South, possess much more potential experience, are more likely to have sales and non managerial occupations, and work more often in manufacture and retail and less often in education industries than the 10% least affected women.

Table 3 tests if the differences found in table 2 are statistically significant. It reports p-values for the test of equality of means for most and least affected women. The first p-value accounts for simultaneous inference on all variables within a given category. For example, it accounts that

[8] In the 2016 version of the paper we found similar patterns of heterogeneity using CPS 2012 data with a specification that did not include occupation and industry indicators.

Table 3. Classification Table – Difference in the Average Characteristics of the 10% Most and Least Affected Women by Gender Wage Gap

	Est.	S.E.	P-val.[1]	JP-val.[2]		Est.	S.E.	P-val.[1]	JP-val.[2]
Log wage	-0.10	0.04	0.03	0.70	O.manager	-0.29	0.06	0.00	0.00
M.married	0.59	0.04	0.00	0.00	O.service	0.02	0.03	0.99	1.00
M.widowed	-0.02	0.02	0.93	1.00	O.sales	0.22	0.06	0.00	0.03
M.separated	-0.01	0.01	0.89	1.00	O.construction	-0.01	0.01	0.67	1.00
M.divorced	-0.08	0.04	0.46	0.86	O.production	0.06	0.02	0.08	0.42
M.nevermarried	-0.48	0.04	0.00	0.00	I.minery	0.01	0.01	0.97	1.00
E.lhs	0.05	0.01	0.01	0.17	I.construction	-0.00	0.01	1.00	1.00
E.hsg	0.22	0.05	0.00	0.00	I.manufacture	0.08	0.02	0.02	0.15
E.sc	0.08	0.06	0.70	1.00	I.retail	0.12	0.04	0.06	0.32
E.cg	-0.19	0.06	0.01	0.16	I.transport	0.04	0.01	0.08	0.39
E.ad	-0.15	0.06	0.07	0.46	I.information	0.01	0.02	1.00	1.00
R.ne	-0.06	0.04	0.35	0.99	I.finance	0.06	0.04	0.78	0.99
R.mw	0.02	0.04	0.95	1.00	I.professional	-0.01	0.03	1.00	1.00
R.so	0.08	0.04	0.23	0.97	I.education	-0.13	0.06	0.29	0.74
R.we	-0.04	0.03	0.69	1.00	I.leisure	-0.09	0.03	0.04	0.22
Experience	13.27	1.54	0.00	0.00	I.services	-0.07	0.02	0.02	0.15
					I.public	-0.01	0.03	1.00	1.00

PE estimated from a linear conditional quantile model with interactions.

Standard Errors and p-values obtained by weighted bootstrap with 500 repetitions.

[1] These p-values are adjusted for multiplicity to account for joint testing of zero coefficients on for all variables within a category: M E, R, O, or I.

[2] These p-values are adjusted for multiplicity to account for joint testing of zero coefficients on all the variables in the table.

we are conducting five tests corresponding to the five categories of marital status. For the non categorical variables log wage and experience the p-values are for one test. The second p-value accounts for simultaneous inference of all the differences displayed in the table.[9] These p-values are obtained by Algorithm 2.2 with the appropriate choice of vectors of linear combinations and set $\mathcal{T}$, and 500 weighted bootstrap repetitions. The p-values show that most of the differences from table 2 are statistically significant at conventional significant levels after controlling for simultaneous inference. In particular, the most affected women are significantly more likely to be married, high-school graduates, more experienced, and in sales occupations, and less likely to be never married and in managerial occupations under the most strict simultaneous inference correction. Blau and Kahn (2017) have recently documented the importance of differences in occupation and industry to explain the gender wage gap using data from the Panel Study of Income Dynamics (PSID) 1980-2010 and a different methodology based on wage decompositions. Consistent with our findings,

[9] We employ the so called "single-step" methods for controlling the family-wise error rate. To generate a (somewhat) higher power, we recommend to employ the p-values generated via "step-down" methods, such as those reported in Romano and Wolf (2016) and List, Shaikh, and Xu (2016).

they argue that this importance might be due to compensating differentials. Unlike Blau and Kahn (2017) and previous studies in the literature, our analysis uncovers significant heterogeneity in the extent of the gender wage gap and relates this heterogeneity to human capital, occupation, industry and other characteristics.

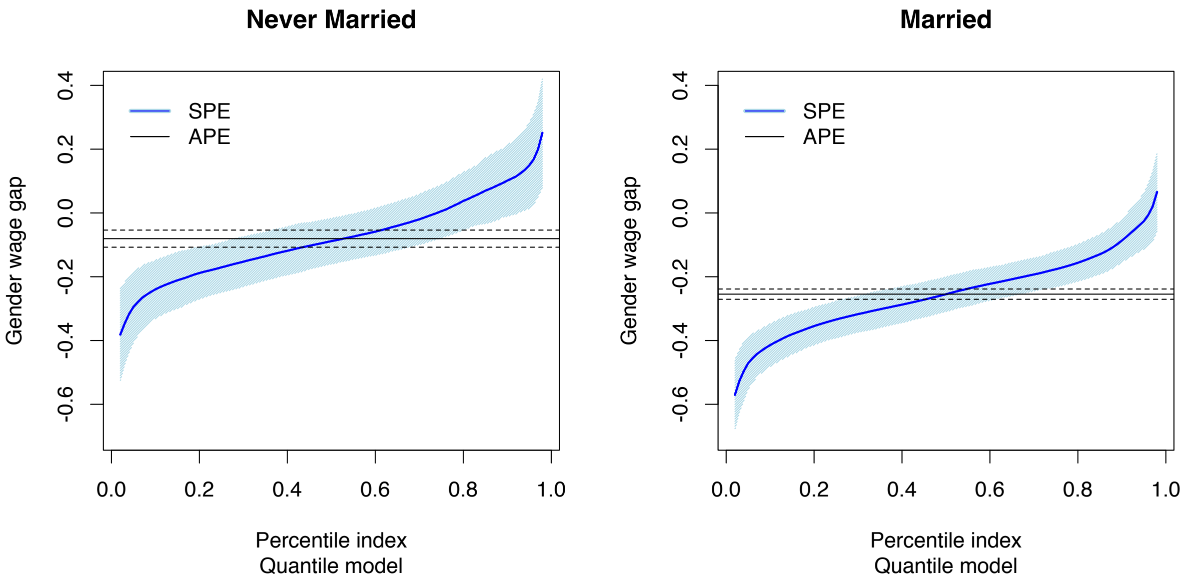

Figure 2. APE and SPE of the gender wage gap for women by marital status. Estimates and 90% bootstrap uniform confidence bands for the conditional quantile function.

We further explore these findings by analyzing the APE and SPE on the treated *conditional* on marital status and unobserved rank in the non-additive error model. Figures 2 and 3 show estimates and 90% confidence bands of the APE and SPE-function of the gender wage gap for 2 subpopulations defined by marital status (married and never married) and 3 subpopulations defined by unobserved rank (first decile, median and ninth decile, where the unobserved rank is .1, .5 and .9, respectively). The confidence bands are constructed as in fig. 1. We find significant heterogeneity in the gender gap within each subpopulation, and also between subpopulations defined by marital status and unobserved rank. The SPE-function is more negative for married women and at the tails of the conditional distribution. Married women at the top decile suffer from the highest gender wage gaps. This pattern is consistent with "glass-ceiling" effects behind the gender wage gap (Albrecht, Bjorklund, and Vroman, 2003).

Figure 4 plots simultaneous 90% confidence bands for the distribution of experience and log wage for the most and least affected women. They are obtained by Algorithm 2.2 with 500 weighted bootstrap replications. The estimated distribution of experience for the most affected first-order stochastically dominates the same estimated distribution for the least affected women.

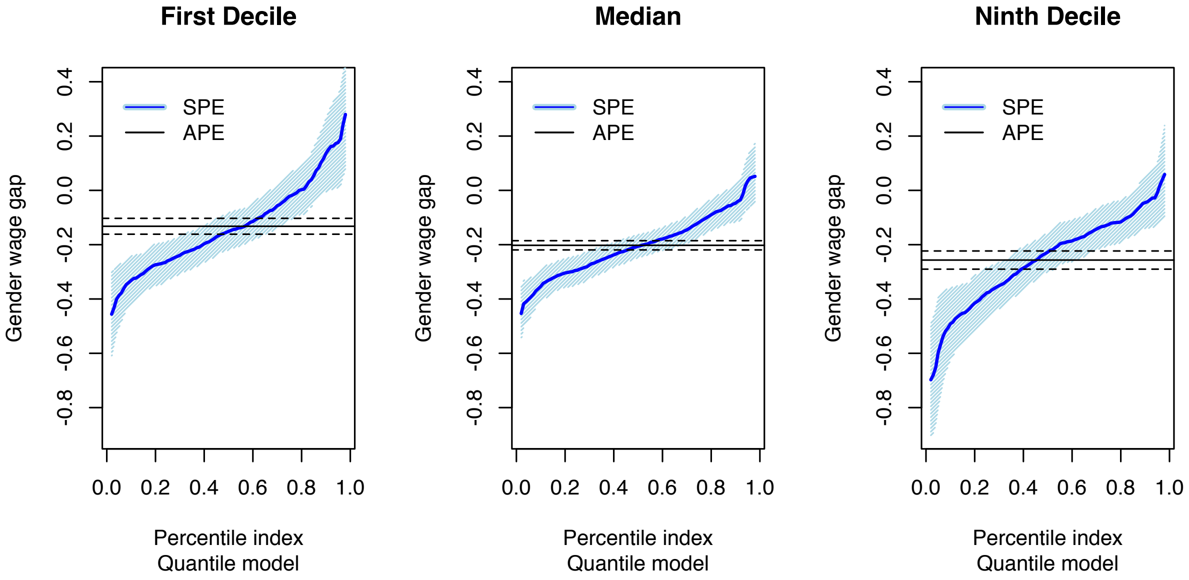

FIGURE 3. APE and SPE of the gender wage gap for women by unobserved ranking in the conditional distribution. Estimates and 90% bootstrap uniform confidence bands for the conditional quantile function.

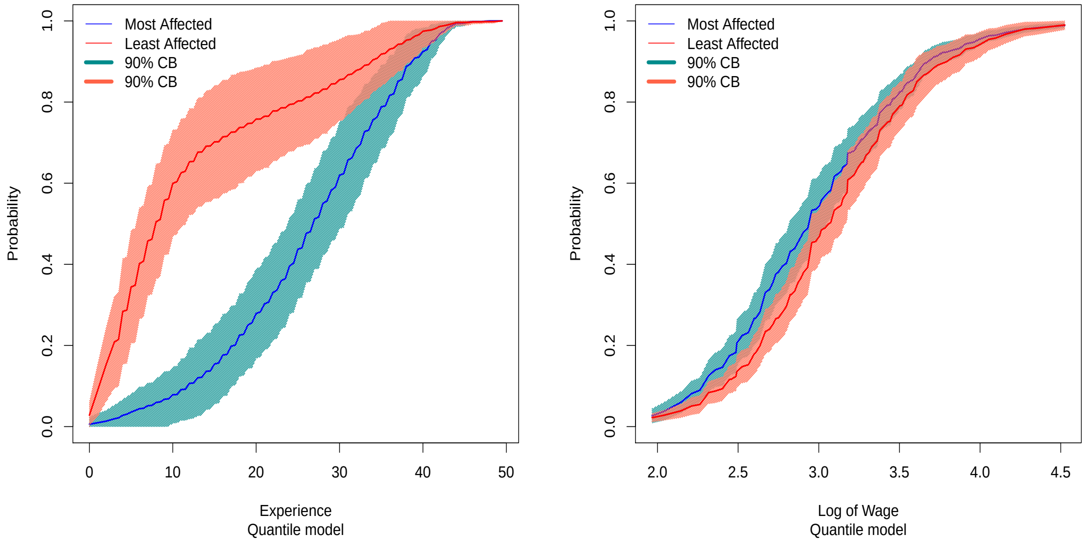

FIGURE 4. Estimates and 90% weighted bootstrap joint uniform confidence bands for the distributions of experience and log wages of the 10% most and least affected women by gender wage gap.

Moreover, the uniform bands confirm that this dominance is statistically significant at the 90% confidence level for the underlying distributions. The estimated (marginal) distribution of log wage for the least affected first-order dominates the same estimated distribution for most affected, but we cannot reject that the underlying distributions are equal at the 10% significance level. The results of the classification analysis are consistent with preferences that make never married highly educated young women working on managerial occupations be more career-oriented.[10]

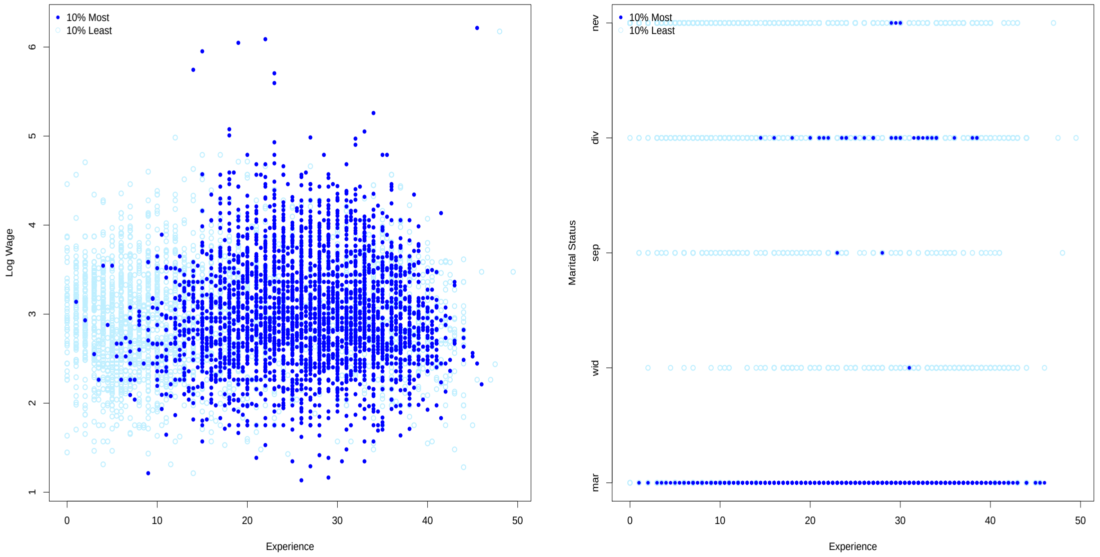

FIGURE 5. Estimates and 90% weighted bootstrap confidence bands for projections of the confidence sets for characteristics of 10% most and least affected women by gender wage gap.

Finally, Figure 5 plots two dimensional projections of experience-log wage and experience-marital status of the confidence sets for the 10% most and least affected subpopulations. We show the results from the additive error model for the conditional expectation. Here we use a simplified specification that excludes the two-way interactions from W to get more precise estimates of all the PEs. We obtain 90% confidence sets for the most and least affected subpopulations by weighted bootstrap with standard exponential weights and 500 repetitions. The sets $\mathcal{CM}^{-0.1}(0.90)$ and $\mathcal{CM}^{+0.1}(0.90)$ include 23% and 19% of the women in the sample, respectively.[11] The projections show that there are relatively more least affected women with low experience at all wage levels, more high affected women with high wages with between 15 and 25 years of experience, and more least affected women which are not married at all experience levels.

[10] We find similar results using the additive error model. We do not report these results for the sake of brevity.

[11] Recall that in this application the set $\mathcal{CM}^{-0.1}(0.90)$ corresponds to most affected women and $\mathcal{CM}^{+0.1}(0.90)$ to least affected women. We drop one woman that is included in both sets.

4. Detailed Large Sample Theory

4.1. **Detailed Large Sample Theory for SPE.** For an open set $\mathcal{K}$, let the class $\mathcal{C}^1$ on $\mathcal{K}$ denote the set of continuously differentiable real valued functions on $\mathcal{K}$. We make the following technical assumptions about the PE function $\Delta : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ and the distribution of the covariates:

S.1. The part of the domain of the PE function $x \mapsto \Delta(x)$ of interest, $\mathcal{X}$, is open and its closure $\overline{\mathcal{X}}$ is compact. The distribution μ is absolutely continuous with respect to the Lebesgue measure with density μ'. There exists an open set $B(\mathcal{X})$ containing $\overline{\mathcal{X}}$ such that $x \mapsto \Delta(x)$ is $\mathcal{C}^1$ on $B(\mathcal{X})$, and $x \mapsto \mu'(x)$ is continuous on $B(\mathcal{X})$ and is zero outside the domain of interest, i.e. $\mu'(x) = 0$ for any $x \in B(\mathcal{X}) \setminus \mathcal{X}$.

S.2. Let $\mathcal{M}_\Delta(\delta) := \{x \in \mathcal{X} : \Delta(x) = \delta\}$. For any regular value δ of Δ on $\overline{\mathcal{X}}$, we assume that the closure of $\mathcal{M}_\Delta(\delta)$ has a finite number of connected branches.

The following property of the set $\mathcal{M}_\Delta(\delta)$ is a useful implication of Assumptions S.1 and S.2 that we will exploit in the analysis.

Remark 4.1 (Properties of $\mathcal{M}_\Delta(\delta)$). By Theorem 5-1 in Spivak (1965, p. 111), S.1 and S.2 imply that $\mathcal{M}_\Delta(\delta)$ is a $(d_x - 1)$-manifold without boundary in $\mathbb{R}^{d_x}$ of class $\mathcal{C}^1$ for any δ that is a regular value of $x \mapsto \Delta(x)$ on $\overline{\mathcal{X}}$.

Assumption S.1 imposes mild smoothness conditions on the PE function $x \mapsto \Delta(x)$. It also requires that all the components of the covariate X are continuous random variables. We defer the treatment of the case where X has both continuous and discrete components to the SM. As a matter of generalization, our theoretical analysis allows us to replace that $x \mapsto \mu'(x)$ vanishes on $\partial\mathcal{X}$, by the weaker condition that the intersection of $\mathcal{M}_\Delta(\delta)$ and the boundary of $\mathcal{X}$ have zero volume with respect to μ, namely

$$\int_{\mathcal{M}_\Delta(\delta)} 1\{x \in \partial\mathcal{X}\} \frac{\mu'(x)}{\|\partial\Delta(x)\|} d\text{Vol} = 0, \tag{4.4}$$

where $\partial\mathcal{X}$ denotes the boundary of $\overline{\mathcal{X}}$, $\partial\Delta(x)$ is the gradient of $x \mapsto \Delta(x)$, and $\int_{\mathcal{M}} f(x) d\text{Vol}$ denotes the integral of the function f on the manifold $\mathcal{M}$ with respect to volume; see Appendix D in the SM for a brief review on Differential Geometry. This relaxation is relevant to cover the case where X includes an uniformly distributed component such as the unobserved rank in Example 3.[12]

Assumption S.2 imposes shape restrictions on $x \mapsto \Delta(x)$ that rule out cases such as infinite cyclical oscillations or flat areas. A simple sufficient condition for S.2 is that the map $x \mapsto \Delta(x)$ does not have critical points on $\overline{\mathcal{X}}$. This means that $x \mapsto \Delta(x)$ is not locally flat anywhere on $\overline{\mathcal{X}}$, which we define to mean that the norm of the gradient, $\|\partial\Delta(x)\|$, does not vanish on $x \in \overline{\mathcal{X}}$.

[12] In the numerical examples of Section H in the SM, the first two designs only satisfy this relaxed condition.

In this case, any δ in the image of $\overline{\mathcal{X}}$ under Δ is regular. This condition is probably the most relevant for practice and can be verified in applications, at least informally.

Remark 4.2 (Verification of Regularity Conditions in Practice)**.** The main regularity condition is that PE function $x \mapsto \Delta(x)$ be smooth and not locally flat, namely $\|\partial\Delta(x)\|$ does not vanish. Our inferential results are developed under this assumption, and they do not apply otherwise. To verify if these results apply in practice, we strongly recommend to conduct a Monte Carlo experiment using a data generating process that mimics the application at hand.[13] Indeed, failure of the inference method in the simulation experiment implies failure of the regularity conditions. We provide an application of this supporting analysis to the gender wage gap example in Appendix H. Looking forward, it would be useful to develop further an inference method with good robustness properties with respect to the regularity conditions, i.e. that remains uniformly valid when the PE function is (close to being) locally flat. We delegate this line of research to future work.[14]

We make the following assumptions about the estimator of the PE. Let $\ell^\infty(\mathcal{T})$ denote the set of bounded and measurable functions $g : \mathcal{T} \to \mathbb{R}$ and $\mathcal{F}$ a fixed subset of continuous functions on $B(\mathcal{X})$. Let $\ell^\infty(B(\mathcal{X}))$ be the set of bounded and measurable functions on $B(\mathcal{X})$ and $\rightsquigarrow$ denote weak convergence (convergence in distribution).

S.3. $\widehat{\Delta}$, the estimator of Δ, belongs to $\mathcal{F}$ with probability approaching 1 and obeys a functional central limit theorem, namely,

$$a_n(\widehat{\Delta} - \Delta) \rightsquigarrow G_\infty \text{ in } \ell^\infty(B(\mathcal{X})),$$

where a_n is a sequence such that $a_n \to \infty$ as $n \to \infty$, and $x \mapsto G_\infty(x)$ is a tight process that has almost surely uniformly continuous sample paths on $B(\mathcal{X})$.

In the parametric and semiparametric models of Examples 1–3, S.3 holds under weak conditions that guarantee asymptotic normality of the ML, OLS and QR estimators. For the QR estimator in Example 3 where the unobserved rank is one of the covariates, these conditions include that the density of Y conditional on X be bounded away from zero (Koenker, 2005), which is facilitated by excluding tail quantile indexes.

Let $\widehat{\mu}$ be the estimator of the distribution μ. It is convenient to identify μ and $\widehat{\mu}$ with the operators:

$$g \mapsto \mu(g) = \int g(x) d\mu(x), \quad g \mapsto \widehat{\mu}(g) = \int g(x) d\widehat{\mu}(x),$$

mapping from the set $\mathcal{G} := \{x \mapsto 1(f(x) \leqslant \delta) : f \in \mathcal{F}, \delta \in \mathcal{V}\}$ to $\mathbb{R}$, where $\mathcal{F}$ is the fixed subset of continuous functions on $B(\mathcal{X})$ containing Δ, and $\mathcal{V}$ is any compact set of $\mathbb{R}$. We require $\mathcal{G}$ to be

[13] In fact, we recommend doing this for every econometric method.

[14] For instance, it is of interest to determine whether the use of subsampling instead of bootstrap can deliver a more robust inference method when the PE function is close to being flat; see, e.g. Romano and Shaikh (2012).

totally bounded under the $L^2(\mu)$ norm. Define $\mathbb{H}$ as the set of all bounded linear operators H on $\mathcal{G}$ of the form

$$g \mapsto H(g),$$

which are uniformly continuous on $g \in \mathcal{G}$ under the $L^2(\mu)$ norm. We define the boundedness of these operators with respect to the norm:

$$\|H\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |H(g)|,$$

and define the corresponding distance between two operators H and $\widetilde{H}$ in $\mathbb{H}$ as $\|H - \widetilde{H}\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} |H(g) - \widetilde{H}(g)|$. Clearly, $\mu \in \mathbb{H}$.

We make the following assumption about $\widehat{\mu}$.

S.4. The function $x \mapsto \widehat{\mu}(x)$ is a distribution over $B(\mathcal{X})$ obeying in $\mathbb{H}$,

$$b_n(\widehat{\mu} - \mu) \rightsquigarrow H_\infty, \tag{4.5}$$

where $g \mapsto H_\infty(g)$ is a.s. an element of $\mathbb{H}$ (i.e. it has almost surely uniformly continuous sample paths on $\mathcal{G}$ with respect to the $L^2(\mu)$ metric) and b_n is a sequence such that $b_n \to \infty$ as $n \to \infty$.

When $\widehat{\mu}$ is the empirical distribution based on a random sample from the population with distribution μ, then $b_n = \sqrt{n}$ and $H_\infty = B_\mu$, where B_μ is a μ-Brownian Bridge, i.e. a Gaussian process with zero mean and covariance function $(g_1, g_2) \mapsto \mu(g_1 g_2) - \mu(g_1)\mu(g_2)$. In this case condition S.4 imposes that the function class

$$\mathcal{G} = \{x \mapsto 1(f(x) \leqslant \delta) : f \in \mathcal{F}, \delta \in \mathcal{V}\}$$

is μ-Donsker. Note that $\mathcal{F}$ is the parameter space that contains $\Delta(x)$ as well as $\widehat{\Delta}(x)$ in S.3. In parametric models for the PE where $\mathcal{F} = \{f(x,\theta) : \theta \in \Theta\}$, f is known, $\theta \subseteq \mathbb{R}^{d_\theta}$ with $d_\theta < \infty$, and $x \mapsto f(x,\theta)$ is $\mathcal{C}^1$ on $\mathcal{X}$ for all $\theta \in \Theta$, the class $\mathcal{G}$ is μ-Donsker under mild conditions specified for example in van der Vaart (1998, Chap. 19). Examples 1 and 2 specify the PE parametrically. Lemma F.1 in the SM gives other sufficient conditions for the Donsker property.

The following result is derived as a consequence of the new mathematical results on the Hadamard differentiability of the sorting operator, stated in Lemma A.2 in the Appendix (proof given in SM due to space constraints), in conjunction with the functional delta method. It shows that the empirical SPE-function follows a FCLT over sets of quantiles corresponding to Δ^*_μ pre-images of compact sets of $\mathbb{R}$.

Define $\mathcal{D}$ as a compact set consisting of regular values of $x \mapsto \Delta(x)$ on $\overline{\mathcal{X}}$, and $\mathcal{U} := \{\widetilde{u} \in [0,1] : \Delta^*_\mu(\widetilde{u}) \in \mathcal{D}, f_{\Delta,\mu}(\Delta^*_\mu(\widetilde{u})) > \varepsilon\}$, for a fixed $\varepsilon > 0$, where $f_{\Delta,\mu}(\Delta^*_\mu(\widetilde{u}))$ is the density of $\Delta(X)$ defined in Lemma A.1(a). Let $r_n := a_n \wedge b_n$, the slowest of the rates of convergence of $\widehat{\Delta}$ and $\widehat{\mu}$. Assume $r_n/a_n \to s_\Delta \in [0,1]$ and $r_n/b_n \to s_\mu \in [0,1]$, where $s_\Delta = 0$ when $b_n = o(a_n)$ and $s_\mu = 0$ when $a_n = o(b_n)$. For example, $s_\mu = 0$ if μ is treated as known.

Theorem 4.1 (FCLT for $F_{\widehat{\Delta},\widehat{\mu}}$ and $\widehat{\Delta}^*_{\widehat{\mu}}$)**.** *Suppose that* S.1-S.4 *hold, and the convergence in* S.3 *and* S.4 *holds jointly. Then, as $n \to \infty$,*

(a) The estimator of the distribution of PE obeys a functional central limit theorem, namely, in $\ell^\infty(\mathcal{D})$,

$$r_n(F_{\widehat{\Delta},\widehat{\mu}}(\delta) - F_{\Delta,\mu}(\delta)) \rightsquigarrow s_\Delta T_\infty(\delta) + s_\mu H_\infty(g_{\Delta,\delta}),$$

as a stochastic process indexed by $\delta \in \mathcal{D}$, where

$$T_\infty(\delta) := -\int_{\mathcal{M}_\Delta(\delta)} \frac{G_\infty(x)\mu'(x)}{\|\partial\Delta(x)\|} d\mathrm{Vol}.$$

(b) The empirical SPE-process obeys a functional central limit theorem, namely in $\ell^\infty(\mathcal{U})$,

$$r_n(\widehat{\Delta}^*_{\widehat{\mu}}(u) - \Delta^*_\mu(u)) \rightsquigarrow -\frac{s_\Delta T_\infty(\Delta^*_\mu(u)) + s_\mu H_\infty(g_{\Delta,\Delta^*_\mu(u)})}{\int \frac{\mu'(x)}{\|\partial\Delta(x)\|} d\mathrm{Vol}} \quad =: \quad Z_\infty(u), \tag{4.6}$$

as a stochastic process indexed by $u \in \mathcal{U}$.

Remark 4.3 (Critical values)**.** (a) Theorem 4.1 shows that $\delta \mapsto F_{\widehat{\Delta},\widehat{\mu}}(\delta)$ ($u \mapsto \widehat{\Delta}^*_{\widehat{\mu}}(u)$) follows a FCLT over any compact set $\mathcal{D}$ (the Δ^*_μ pre-image of $\mathcal{D}$), where $\mathcal{D}$ excludes the critical values of $x \mapsto \Delta(x)$ on $\overline{\mathcal{X}}$. Thus, we can set $\mathcal{D} = \Delta(\overline{\mathcal{X}}) := \{\Delta(x) : x \in \overline{\mathcal{X}}\}$ when the map $x \mapsto \Delta(x)$ does not have critical points on $\overline{\mathcal{X}}$. This case is nice because it allows us not to worry about critical values when performing inference, and practically relevant as it occurs very naturally in many applications. For instance, it arises whenever $\Delta(x)$ is strictly locally monotonic in some direction. (b) In numerical examples reported in the SM, we find that the bootstrap inference method proposed performs well even in models where $x \mapsto \Delta(x)$ has critical points, without excluding the corresponding critical values from $\mathcal{D}$. This evidence suggests that the exclusion of critical values might not be necessary for inference. □

4.2. **Detailed Large Sample Theory for CA.** It is convenient to modify the notation for the u-CA separating the dependence on $\Delta^*_\mu(u)$ from Δ and μ and specifying the characteristic of interest as φ_t. Moreover, when $Z = (X, Y)$ we remove the dependence on Y by taking expectations conditional on X. Let $\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t) := \widehat{\Lambda}^u_{\Delta,\mu}(t)$ and $\Lambda_{\Delta,\mu,\Delta^*_\mu(u)}(\varphi_t) := \Lambda^u_{\Delta,\mu}(t)$, where $\varphi_t \in \mathcal{F}_M \cup \mathcal{F}_I$, $t = (t_1, \ldots, t_{d_z}) \in \mathbb{R}^{d_z}$, $u \in \mathcal{U}$, $\mathcal{F}_M := \{\int z_1^{t_1} \cdots z_{d_z}^{t_{d_z}} d\mu(y \mid x) : t_1, ..., t_{d_z} \in \{0, 1, 2, \ldots\}, \int |z_1^{t_1} \cdots z_{d_z}^{t_{d_z}}| d\mu(z) < \infty, t_1 + \ldots + t_{d_z} \leqslant M\}$, M is some fixed integer, $\mu(y \mid x)$ is the distribution of Y at y conditional on $X = x$, and $\mathcal{F}_I := \{\int 1(z_1 \leqslant t_1, ..., z_{d_z} \leqslant t_{d_z}) d\mu(y \mid x) : t_1, ..., t_{d_z} \in \mathbb{R}\}$. For example, $\varphi_t(x) = x^{t_x} \mathrm{E}[Y^{t_y} \mid X = x]$ or $\varphi_t(x) = 1(x \leqslant t_x)\mu(t_y \mid x)$ for $t = (t_x, t_y)$. To derive the properties of $\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t)$, we use that the class of functions $\widetilde{\mathcal{G}} = \{1(f \leqslant \delta)\varphi : \varphi \in \mathcal{F}_M \cup \mathcal{F}_I, \delta \in \mathcal{V}, f \in \mathcal{F}\}$ is μ-Donsker. When $x \mapsto \mu(y \mid x)$ is continuous, this property holds by assumption S.4 when $\widehat{\mu}$ is the empirical distribution.[15]

[15] Lemma F.2 in the SM gives other sufficient conditions for the Donsker property.

The following result is derived as a consequence of the new mathematical results on the Hadamard differentiability of the classification operator, stated in Lemma A.3 in the Appendix (proof given in SM due to space constraints), in conjunction with the functional delta method.

Theorem 4.2 (FCLT for $\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t)$)**.** *Suppose that* S.1-S.4 *hold, the convergence in* S.3 *and* S.4 *holds jointly, and $u \in \mathcal{U}$. If $Z = (X, Y)$, then assume that $\mathcal{Y}$ is compact and $x \mapsto \mu(y \mid x)$ is continuous on $B(\mathcal{X})$ for all $y \in \mathcal{Y}$. Then, as $n \to \infty$, (a) $\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t)$ obeys a FCLT with respect to $t \mapsto \varphi_t \in \mathcal{F}_M$, namely, in $\ell^\infty(\mathbb{R}^{d_z})^2$,*

$$r_n\left(\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t) - \Lambda_{\Delta,\mu,\Delta^*_{\mu}(u)}(\varphi_t)\right) \rightsquigarrow \int_{\mathcal{M}_\Delta(\delta)} \widetilde{\varphi}_t(x)\frac{Z_\infty(u) - s_\Delta G_\infty(x)}{\|\partial\Delta(x)\|}\mu'(x)d\mathrm{Vol} + s_\mu H_\infty(h_{\Delta,\delta,\varphi_t}) =: Z^u_\infty(t),$$

as a stochastic process indexed by $t \in \mathbb{R}^{d_z}$, where $\widetilde{\varphi}_t(x) = [\varphi_t(x) - \Lambda_{\Delta,\mu,\delta}(\varphi_t)]/F_{\Delta,\mu}(\delta)$, $\widetilde{h}_{\Delta,\delta,\varphi_t} := \widetilde{\varphi}_t(x)1\{\Delta(x) \leqslant \delta\}$, and $Z_\infty(u)$ is the limit process of Theorem 4.1; and (b) if in addition Assumption AS.1 *holds, then $\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t)$ obeys the same FCLT with respect to $t \mapsto \varphi_t \in \mathcal{F}_I$.*

Assumption AS.1 is a technical condition stated in Appendix E of the SM to deal with the discontinuity of the indicator functions when $\varphi_t \in \mathcal{F}_I$. A sufficient condition for AS.1 is that

$$\int_{\mathcal{M}_\Delta(\delta)\cap\{x: x_k = t_k\}} d\mathrm{Vol} = 0$$

holds uniformly over all $\delta \in \mathcal{V}$, $t_k \in \mathbb{R}$ and $k = 1, 2, ..., d_x$. In other words, the manifold $\mathcal{M}_\Delta(\delta)$ and the set of points $\{x : x_k = t_k\}$ can not have an intersection with positive volume of $(d_x - 1)$-dimension.

4.3. **Bootstrap Inference for SPE and CA.** Corollaries 2.1 and 2.2 use critical values of statistics related to the limit processes Z_∞ and Z^u_∞ to construct confidence bands and p-values. These critical values can be hard to obtain in practice. In principle one can use simulation, but it might be difficult to numerically locate and parametrize the manifold $\mathcal{M}_\Delta(\delta)$, and to evaluate the integrals on $\mathcal{M}_\Delta(\delta)$ needed to compute the realizations of $Z_\infty(u)$ and $Z^u_\infty(t)$. This creates a real challenge to implement our inference methods. To deal with this challenge we employ (exchangeable) bootstrap to compute critical values (Præstgaard and Wellner, 1993; van der Vaart and Wellner, 1996) instead of simulation. We show that the bootstrap law is consistent to approximate the distribution of the limit processes of Theorems 4.1 and 4.2.

To state the bootstrap validity result formally, we follow the notation and definitions in van der Vaart and Wellner (1996). Let D_n denote the data vector and let $\mathrm{B}_n = (\omega_1, \dots, \omega_n)$ be the vector of bootstrap weights. Consider a random element $\widetilde{Z}_n = Z_n(\mathrm{D}_n, \mathrm{B}_n)$ in a normed space $\mathbb{D}$. We say that the bootstrap law of $\widetilde{Z}_n$ consistently estimates the law of some tight random element Z_∞

and write $\widetilde{Z}_n \rightsquigarrow_{\mathrm{P}} Z_\infty$ if

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D})} |\mathrm{E}_{\mathrm{B}_n} h(\widetilde{Z}_n) - \mathrm{E}_{\mathrm{P}} h(Z_\infty)| \to_{\mathrm{P}} 0,$$

where $\mathrm{BL}_1(\mathbb{D})$ denotes the space of functions with Lipschitz norm at most 1; $\mathrm{E}_{\mathrm{B}_n}$ denotes the conditional expectation with respect to B_n given the data D_n; E_{P} denotes the expectation with respect to P, the distribution of the data D_n; and $\to_{\mathrm{P}}$ denotes convergence in (outer) probability.

The next result is a consequence of the functional delta method for the exchangeable bootstrap. Let $\Lambda_{\widetilde{\Delta},\widetilde{\mu},\widetilde{\Delta}^*_{\widetilde{\mu}}(u)}(\varphi_t) := \widetilde{\Lambda}^u_{\Delta,\mu}(t)$, the bootstrap draw of $\widehat{\Lambda}^u_{\Delta,\mu}(t)$ defined in Algorithm 2.2.

Theorem 4.3 (Bootstrap FCLT for $\widehat{\Delta^*_\mu}$ and $\Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\widehat{\mu}}(u)}(\varphi_t)$)**.** *Suppose that the bootstrap is consistent for the law of the estimator of the PE, namely $a_n(\widetilde{\Delta} - \widehat{\Delta}) \rightsquigarrow_{\mathrm{P}} G_\infty$ in $\ell^\infty(B(\mathcal{X}))$, and for the law of the estimated measure, namely $b_n(\widetilde{\mu} - \widehat{\mu}) \rightsquigarrow_{\mathrm{P}} H_\infty$ in $\mathbb{H}$. Then, (1) under the assumptions of Theorem 4.1, the bootstrap is consistent for the law of the empirical SPE-process, namely*

$$r_n(\widetilde{\Delta^*_\mu}(u) - \widehat{\Delta^*_\mu}(u)) \rightsquigarrow_{\mathrm{P}} Z_\infty(u) \text{ in } \ell^\infty(\mathcal{U});$$

and (2) under the assumptions of Theorem 4.2, the bootstrap is consistent for the law of the empirical CA-process, namely

$$r_n\left(\Lambda_{\widetilde{\Delta},\widetilde{\mu},\widetilde{\Delta}^*_{\widetilde{\mu}}(u)}(\varphi_t) - \Lambda_{\widehat{\Delta},\widehat{\mu},\widehat{\Delta}^*_{\mu}(u)}(\varphi_t)\right) \rightsquigarrow Z^u_\infty(t) \text{ in } \ell^\infty(\mathbb{R}^{d_z})^2.$$

Theorem 4.3 employs the high-level condition that the bootstrap can approximate consistently the laws of $\widehat{\Delta}$ and $\widehat{\mu}$, after suitable rescaling. In Examples 1-3 when $\widehat{\mu}$ is the empirical measure based on the random sample of size n, the exchangeable bootstrap method entails randomly reweighing the sample using the weights $(\omega_1, \ldots, \omega_n)$, which include empirical boostrap and i.i.d. exponential weights, for example. In this case the high level condition holds if the weights satisfy the conditions stated in equation (3.6.8) of van der Vaart and Wellner (1996). We refer to van der Vaart and Wellner (1996) and Chernozhukov, Fernández-Val, and Melly (2013) for bootstrap FCLT for parametric and semi parametric estimators of Δ including least squares, quantile regression, and distribution regression, as well as nonparametric estimators of μ including the empirical distribution function.

Appendix A. Key New Mathematical Results: Hadamard Differentiability of Sorting and Classification Operators

A.1. **Notation.** We denote the PE as $\Delta(x)$, the empirical PE as $\widehat{\Delta}(x)$, and $\partial\Delta(x) := \partial\Delta(x)/\partial x$, the gradient of $x \mapsto \Delta(x)$. For a vector $v = (v_1, \ldots, v_{d_v}) \in \mathbb{R}^{d_v}$, $\|v\|$ denotes the Euclidian norm of v, that is $\|v\| = \sqrt{v^{\mathsf{T}} v}$, where the superscript $^{\mathsf{T}}$ denotes transpose.

A.2. **Basic Analytical Properties of Sorted Functions.** The following lemma establishes the properties of the distribution function $\delta \mapsto F_{\Delta,\mu}(\delta)$ and the SPE-function $u \mapsto \Delta^*_\mu(u)$.

Define $\mathcal{D}$ as a compact set consisting of regular values of $x \mapsto \Delta(x)$ on $\overline{\mathcal{X}}$.

Lemma A.1 (Basic Properties of $F_{\Delta,\mu}$ and Δ^*_μ)**.** *Under conditions* S.1 *and* S.2*:*

1. For any $\delta \in \mathcal{D}$, the derivative of $F_{\Delta,\mu}(\delta)$ with respect to δ is:

$$f_{\Delta,\mu}(\delta) := \partial_\delta F_{\Delta,\mu}(\delta) = \int_{\mathcal{M}_\Delta(\delta)} \frac{\mu'(x)}{\|\partial\Delta(x)\|} d\text{Vol}. \tag{A.7}$$

This integral is well-defined because the gradient $x \mapsto \partial\Delta(x)$ is finite, continuous, and bounded away from 0 on $\mathcal{M}_\Delta(\delta) \subseteq \overline{\mathcal{X}}$. The map $\delta \mapsto f_{\Delta,\mu}(\delta)$ is uniformly continuous on $\mathcal{D}$.

2. Fix $\varepsilon > 0$, then for any $u \in \mathcal{U} := \{\widetilde{u} \in [0,1] : \Delta^*_\mu(\widetilde{u}) \in \mathcal{D}, f_{\Delta,\mu}(\Delta^*_\mu(\widetilde{u})) > \varepsilon\}$, the derivative of $\Delta^*_\mu(u)$ respect to u is:

$$\partial_u \Delta^*_\mu(u) = \frac{1}{f_{\Delta,\mu}(\Delta^*_\mu(u))}. \tag{A.8}$$

Moreover, the derivative map $u \mapsto \partial_u \Delta^*_\mu(u)$ is uniformly continuous on $\mathcal{U}$.

A.3. **Functional Derivatives of Sorting-Related Operators.** We consider the properties of the distribution function and the SPE-function as functional operators $(\Delta,\mu) \mapsto F_{\Delta,\mu}$ and $(\Delta,\mu) \mapsto \Delta^*_\mu$. We show that these operators are Hadamard differentiable with respect to (Δ,μ). These results are critical ingredients to deriving the large sample distributions of the empirical versions of $F_{\Delta,\mu}$ and Δ^*_μ in Section 4.

We now recall the definition of uniform Hadamard differentiability from van der Vaart and Wellner (1996).

Definition A.1 (Hadamard Derivative Uniformly in an Index)**.** Suppose the linear spaces $\mathbb{D}$ and $\mathbb{E}$ are equipped with the norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$, and Θ is a compact subset of a metric space. A map $\phi_\theta : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ is called Hadamard-differentiable uniformly in $\theta \in \Theta$ at $f \in \mathbb{D}_\phi$ tangentially to a subspace $\mathbb{D}_0 \subseteq \mathbb{D}$ if there is a continuous linear map $\partial_f \phi_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that uniformly in $\theta \in \Theta$:

$$\frac{\phi_\theta(f + t_n h_n) - \phi_\theta(f)}{t_n} - \partial_f \phi_\theta[h] \to 0, \quad n \to \infty, \tag{A.9}$$

for all converging real sequences $t_n \to 0$ and $\|h_n - h\|_{\mathbb{D}} \to 0$ such that $f + t_n h_n \in \mathbb{D}_\phi$ for every n, and $h \in \mathbb{D}_0$; moreover, the map $(\theta, h) \mapsto \partial_f \phi_\theta[h]$ is continuous on $\Theta \times \mathbb{D}_0$.

In what follows, we let $\mathbb{F}$ denote the space of continuous functions on $B(\mathcal{X})$ equipped with the sup-norm, and $\mathbb{F}_0$ denote a subset of $\mathbb{F}$ that contains uniformly continuous functions.

Lemma A.2 (Hadamard differentiability of $(\Delta,\mu) \mapsto F_{\Delta,\mu}$ and $(\Delta,\mu) \mapsto \Delta^*_\mu$)**.** *Let $\mathbb{D} := \mathbb{F} \times \mathbb{H}$ and $\mathbb{D}_0 := \mathbb{F}_0 \times \mathbb{H}$. Assume that* S.1-S.2 *hold. Then,* *(a) The map $(\Delta, \mu) \mapsto F_{\Delta,\mu}(\delta)$, mapping $\mathbb{D} \to \mathbb{R}$, is Hadamard differentiable uniformly in $\delta \in \mathcal{D}$ at (Δ, μ) tangentially to $\mathbb{D}_0$ with the derivative map $\partial_{\Delta,\mu} F_{\Delta,\mu}(\delta) : \mathbb{D}_0 \to \mathbb{R}$ defined by*

$$(G, H) \mapsto \partial_{\Delta,\mu} F_{\Delta,\mu}(\delta)[G, H] := -\int_{\mathcal{M}_\Delta(\delta)} \frac{G(x)\mu'(x)}{\|\partial\Delta(x)\|} d\mathrm{Vol} + H(g_{\Delta,\delta}).$$

(b) The map $(\Delta, \mu) \mapsto \Delta^*_\mu(u)$, mapping $\mathbb{D} \to \mathbb{R}$ is Hadamard differentiable uniformly in $u \in \mathcal{U}$ at (Δ, μ) tangentially to $\mathbb{D}_0$ with the derivative map, $\partial_{\Delta,\mu}\Delta^*_\mu(u) : \mathbb{D}_0 \to \mathbb{R}$, defined by

$$(G, H) \mapsto \partial_{\Delta,\mu}\Delta^*_\mu(u)[G, H] := -\frac{\partial_{\Delta,\mu} F_{\Delta,\mu}(\Delta^*_\mu(u))[G, H]}{f_{\Delta,\mu}(\Delta^*_\mu(u))}.$$

A.4. **Functional Derivatives of Classification Operators.** Let $\widetilde{\mathbb{D}} := \mathbb{F} \times \widetilde{\mathbb{H}} \times \mathbb{R}$ and $\widetilde{\mathbb{D}}_0 := \mathbb{F}_0 \times \widetilde{\mathbb{H}} \times \mathbb{R}$, where $\mathbb{F}$ and $\mathbb{F}_0$ are defined as before; $\widetilde{\mathbb{H}}$ is the set of bounded linear operators mapping from the set $\widetilde{\mathcal{G}} := \{\varphi 1(\Delta \leqslant \delta) : f \in \mathcal{F}, \varphi \in \mathcal{F}_I \cup \mathcal{F}_M, \delta \in \mathcal{V}\}$ to $\mathbb{R}$, with norm

$$\|H\|_{\widetilde{\mathcal{G}}} = \sup_{g \in \widetilde{\mathcal{G}}} |H(g)|,$$

where the map $g \mapsto H(g)$ is uniformly continuous on $g \in \widetilde{\mathcal{G}}$ under the $L^2(\mu)$ norm. We derive the properties of the least affected classification operator $\Lambda^-_{\Delta,\mu,\delta} : \widetilde{\mathbb{D}} \to \mathbb{R}$ defined by

$$\Lambda^-_{\Delta,\mu,\delta}(\varphi_t) := \int \varphi_t(x) 1\{\Delta(x) \leqslant \delta\} d\mu(x) / \int 1\{\Delta(x) \leqslant \delta\} d\mu(x),$$

where $\varphi_t \in \mathcal{F}_M$ for moments and $\varphi_t \in \mathcal{F}_I$ for distributions of the components of Z, and $\delta = \Delta^*_\mu(u)$ for some $u \in \mathcal{U}$. The properties of the most affected operator $\Lambda^+_{\Delta,\mu,\delta} : \widetilde{\mathbb{D}} \to \mathbb{R}$ can be derived using similar arguments, which are omitted for brevity.

Lemma A.3 (Hadamard differentiability of $(\Delta, \mu, \delta) \mapsto \Lambda^-_{\Delta,\mu,\delta}$)**.** *Assume that Assumptions S.1 and S.2 hold, $\delta \in \mathcal{D}$, and $F_{\Delta,\mu}(\delta) > 0$. Then,*

(a) The map $\Lambda^-_{\Delta,\mu,\delta}(\varphi_t) : \widetilde{\mathbb{D}} \to \mathbb{R}$ is Hadamard-differentiable uniformly in $\varphi_t \in \mathcal{F}_M$ at (Δ, μ, δ) tangentially to $\widetilde{\mathbb{D}}_0$.

(b) If in addition Assumption AS.1 *stated in Appendix E of the SM holds, the map $\Lambda^-_{\Delta,\mu,\delta}(\varphi_t) : \widetilde{\mathbb{D}} \to \mathbb{R}$ is Hadamard-differentiable uniformly in $\varphi_t \in \mathcal{F}_I$ at (Δ, μ, δ) tangentially to $\widetilde{\mathbb{D}}_0$.*

(c) The derivative map $\partial_{\Delta,\mu,\delta}\Lambda^-_{\Delta,\mu,\delta}(\varphi_t) : \widetilde{\mathbb{D}} \to \mathbb{R}$ is defined by:

$$(G, H, K) \mapsto \partial_{\Delta,\mu,\delta}\Lambda^-_{\Delta,\mu,\delta}(\varphi_t)[G, H, K] := \int_{\mathcal{M}_\Delta(\delta)} \widetilde{\varphi}_t(x) \frac{K - G(x)}{\|\partial\Delta(x)\|} d\mathrm{Vol} + H(\widetilde{h}_{\Delta,\delta,\varphi_t}),$$

where $\widetilde{\varphi}_t(x) = [\varphi_t(x) - \Lambda^-_{\Delta,\mu,\delta}(\varphi_t)] / \int 1(\Delta(x) \leqslant \delta) d\mu(x)$ and $\widetilde{h}_{\Delta,\delta,\varphi_t} := \widetilde{\varphi}_t(x) 1\{\Delta(x) \leqslant \delta\}$.

Appendix B. Proofs of Section 4

We first recall Theorem 3.9.4 of van der Vaart and Wellner (1996).

Lemma B.1 (Delta-method)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be metrizable topological vector spaces, and Θ is a compact subset of a metric space. Let $\phi_\theta : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ be a Hadamard differentiable mapping uniformly in $\theta \in \Theta$ at $f \in \mathbb{D}$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$, with derivative $\partial_f \phi_\theta$. Let $\widehat{f}_n : \Omega_n \to \mathbb{D}_\phi$ be stochastic maps taking values in $\mathbb{D}_\phi$ such that $r_n(\widehat{f}_n - f) \rightsquigarrow J_\infty$ for some sequence of constants $r_n \to \infty$, where J_∞ is separable and takes values in $\mathbb{D}_0$. Then $r_n(\phi_\theta(\widehat{f}_n) - \phi_\theta(f)) \rightsquigarrow \partial_f \phi_\theta[J_\infty]$, as a stochastic process indexed by $\theta \in \Theta$.*

Proof of Theorem 4.1. The statements follow directly from Lemma A.2, and Lemma B.1, by setting $\phi_\theta = F_{\Delta,\mu}(\delta)$ with $\theta = \delta$ or $\phi_\theta = \Delta^*_\mu(u)$ with $\theta = u$, $\mathbb{D}_\phi = \mathbb{D} = \mathbb{F} \times \mathbb{H}$, $\mathbb{E} = \mathbb{R}$, $\mathbb{D}_0 = \mathbb{F}_0 \times \mathbb{H}$, $f = (\Delta, \mu)$, $\widehat{f}_n = (\widehat{\Delta}, \widehat{\mu})$, and $J_\infty = (s_\Delta G_\infty, s_\mu H_\infty)$. The expression of $\partial_f \phi_\theta$ for each statement is the Hadamard derivative in the corresponding statement of Lemma A.2. □

Proof of Theorem 4.2. The statements follow directly from Lemma A.4, and Lemma B.1, by setting $\phi_\theta = \Lambda^-_{\Delta,\mu,\delta}$, $\theta = t$, $\mathbb{D}_\phi = \mathbb{D} = \mathbb{F} \times \widetilde{\mathbb{H}} \times \mathbb{R}$, $\mathbb{E} = \mathbb{R}$, $\mathbb{D}_0 = \mathbb{F}_0 \times \widetilde{\mathbb{H}} \times \mathbb{R}$, $f = (\Delta, \mu, \Delta^*_\mu(u))$, $\widehat{f}_n = (\widehat{\Delta}, \widehat{\mu}, \widehat{\Delta^*_\mu}(u))$, and $J_\infty = (s_\Delta G_\infty, s_\mu H_\infty, Z_\infty)$. The expression of $\partial_f \phi_\theta$ for each statement is the Hadamard derivative in the corresponding statement of Lemma A.4. □

To prove Theorem 4.3, we recall Theorem 3.9.11 of van der Vaart (1998). Here we use the notation for bootstrap convergence $\rightsquigarrow_{\mathrm{P}}$ defined in Section 4.3.

Lemma B.2 (Delta-method for bootstrap in probability)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be metrizable topological vector spaces, and Θ is a compact subset of a metric space. Let $\phi_\theta : \mathbb{D}_\phi \subseteq \mathbb{D} \mapsto \mathbb{E}$ be a Hadamard-differentiable mapping uniformly in $\theta \in \Theta$ at f tangentially to $\mathbb{D}_0$ with derivative $\partial_f \phi_\theta$. Let $\widehat{f}_n$ be a random element such that $r_n(\widehat{f}_n - f) \rightsquigarrow J_\infty$. Let $\widetilde{f}_n$ be a stochastic map in $\mathbb{D}$, produced by a bootstrap method, such that $r_n(\widetilde{f}_n - \widehat{f}_n) \rightsquigarrow_{\mathrm{P}} J_\infty$. Then, $r_n(\phi_\theta(\widetilde{f}_n) - \phi_\theta(\widehat{f}_n)) \rightsquigarrow_{\mathrm{P}} \partial_f \phi_\theta[J_\infty]$, as a stochastic process indexed by $\theta \in \Theta$.*

Proof of Theorem 4.3. The statement (1) follows directly from Lemma A.2, and Lemma B.2, by setting $\phi_\theta = \Delta^*_\mu(u)$, $\theta = u$, $\mathbb{D}_\phi = \mathbb{D} = \mathbb{F} \times \mathbb{H}$, $\mathbb{E} = \mathbb{R}$, $\mathbb{D}_0 = \mathbb{F}_0 \times \mathbb{H}$, $f = (\Delta, \mu)$, $\widehat{f}_n = (\widehat{\Delta}, \widehat{\mu})$, and $J_\infty = (s_\Delta G_\infty, s_\mu H_\infty)$. The expression of $\partial_f \phi_\theta$ is the Hadamard derivative in statement (b) of Lemma A.2. The statement (2) follows directly from Lemma A.3, and Lemma B.2, by setting $\phi_\theta = \Lambda^-_{\Delta,\mu,\delta}$, $\theta = t$, $\mathbb{D}_\phi = \mathbb{D} = \mathbb{F} \times \widetilde{\mathbb{H}} \times \mathbb{R}$, $\mathbb{E} = \mathbb{R}$, $\mathbb{D}_0 = \mathbb{F}_0 \times \widetilde{\mathbb{H}} \times \mathbb{R}$, $f = (\Delta, \mu, \Delta^*_\mu(u))$, $\widehat{f}_n = (\widehat{\Delta}, \widehat{\mu}, \widehat{\Delta^*_\mu}(u))$, and $J_\infty = (s_\Delta G_\infty, s_\mu H_\infty, Z_\infty)$. The expression of $\partial_f \phi_\theta$ is the Hadamard derivative in statement (c) of Lemma A.3. □

References

Albrecht, J., A. Bjorklund, and S. Vroman (2003): “Is There a Glass Ceiling in Sweden?,” *Journal of Labor Economics*, 21(1), 145–177.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press.

Bhattacharya, D., and P. Dupas (2012): “Inferring welfare maximizing treatment assignment under budget constraints,” *J. Econometrics*, 167(1), 168–196.

Blau, F. D., and L. M. Kahn (2017): “The Gender Wage Gap: Extent, Trends, and Explanations,” *Journal of Economic Literature*, 55(3), 789–865.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009): “Improving point and interval estimators of monotone functions by rearrangement,” *Biometrika*, 96(3), 559–575.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): “Quantile and probability curves without crossing,” *Econometrica*, 78(3), 1093–1125.

Chernozhukov, V., I. Fernandez-Val, and Y. Luo (2017): “Supplement to “The Sorted Effects Method: Discovering Heterogenous Effects Beyond Their Averages”,” Discussion paper, ArXiv.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): “Inference on counterfactual distributions,” *Econometrica*, 81(6), 2205–2268.

Chernozhukov, V., H. Hong, and E. Tamer (2007): “Estimation and Inference on Identified Parameter Sets in Econometric Models,” Econometrica.

Chernozhukov, V., E. Kocatulum, and K. Menzel (2015): “Inference on sets in finance,” *Quant. Econ.*, 6(2), 309–358.

Cox, D. R. (1984): “Interaction,” *Internat. Statist. Rev.*, 52(1), 1–31, With discussion and a reply by the author.

Hausman, J. A., and W. K. Newey (2017): “Nonparametric Welfare Analysis,” *Annual Review of Economics*, 9(1), 521–546.

Kim, J., and D. Pollard (1990): “Cube Root Asymptotics,” *Ann. Statist.*, 18(1), 191–219.

Koenker, R. (2005): *Quantile regression*. Cambridge university press.

Koenker, R., and G. Basset (1978): “Regression Quantiles,” *Econometrica*, 46(1), 33–50.

List, J. A., A. M. Shaikh, and Y. Xu (2016): “Multiple hypothesis testing in experimental economics,” Discussion paper, National Bureau of Economic Research.

Mulligan, C. B., and Y. Rubinstein (2008): “Selection, Investment, and Women’s Relative Wages Over Time,” *The Quarterly Journal of Economics*, 123(3), 1061–1110.

Oaxaca, R. (1973): “Male-Female Wage Differentials in Urban Labor Markets,” *International Economic Review*, 14(3), 693–709.

Præstgaard, J., and J. A. Wellner (1993): “Exchangeably weighted bootstraps of the general empirical process,” *Ann. Probab.*, 21(4), 2053–2086.

Romano, J. P., and A. M. Shaikh (2010): “Inference for the identified set in partially identified econometric models,” *Econometrica*, 78(1), 169–211.

——— (2012): “On the uniform asymptotic validity of subsampling and the bootstrap,” *The Annals of Statistics*, 40(6), 2798–2822.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010a): “Hypothesis Testing in Econometrics,” *Annual Review of Economics*, 2(1), 75–104.

——— (2010b): “multiple testing,” in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan, Basingstoke.

Romano, J. P., and M. Wolf (2016): “Efficient computation of adjusted p-values for resampling-based stepdown multiple testing,” *Statistics & Probability Letters*, 113, 38–40.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.

SASAKI, Y. (2015): "What do quantile regressions identify for general structural functions?," *Econometric Theory*, 31(5), 1102–1116.

SPIVAK, M. (1965): *Calculus on manifolds. A modern approach to classical theorems of advanced calculus*. W. A. Benjamin, Inc., New York-Amsterdam.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer Series in Statistics.

WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press, second edn.