cemmap

centre for microdata methods and practice

Distribution regression with sample selection and UK wage decomposition

Victor Chernozhukov
Iván Fernández-Val
Siyi Luo

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP09/23

Economic and Social Research Council

DISTRIBUTION REGRESSION WITH SAMPLE SELECTION AND UK WAGE DECOMPOSITION

VICTOR CHERNOZHUKOV, IVÁN FERNÁNDEZ-VAL, AND SIYI LUO

ABSTRACT. We develop a distribution regression model under endogenous sample selection. This model is a semi-parametric generalization of the Heckman selection model. It accommodates much richer effects of the covariates on outcome distribution and patterns of heterogeneity in the selection process, and allows for drastic departures from the Gaussian error structure, while maintaining the same level tractability as the classical model. The model applies to continuous, discrete and mixed outcomes. We provide identification, estimation, and inference methods, and apply them to obtain wage decomposition for the UK. Here we decompose the difference between the male and female wage distributions into composition, wage structure, selection structure, and selection sorting effects. After controlling for endogenous employment selection, we still find substantial gender wage gap – ranging from 21% to 40% throughout the (latent) offered wage distribution that is not explained by composition. We also uncover positive sorting for single men and negative sorting for married women that accounts for a substantive fraction of the gender wage gap at the top of the distribution.

Keywords: Sample selection, distribution regression, quantile, heterogeneity, uniform inference, gender wage gap, assortative matching, glass ceiling

---

Date: April 20, 2023.

MIT, BU and BU. First Arxiv version: November 2018. We thank the editor James Heckman, three anonymous referees, Manuel Arellano, Marianne Bitler, Richard Blundell, Stephane Bonhomme, Shuowen Chen, Matt Hong, Sylvia Klosin, Dennis Kristensen, Victor Quintas Martinez, Enrique Sentana and the seminar participants at Banco Central de Chile, Banco de España, Bristol, BU, Cambridge, Cemfi, Chicago, Erasmus, ESWC 2020, LSE, Michigan, Northwestern, NYU, Oxford, PUC, Queen Mary, UCL and Virginia for helpful comments. We are extremely grateful to Richard Blundell and Barra Roantree at the IFS for providing us with the data. We gratefully acknowledge research support from the British Academy's visiting fellowships and National Science Foundation. Fernández-Val and Chernozhukov were visiting UCL CEMMAP over different points in time, while working on this paper; they are both grateful for their hospitality.

1. INTRODUCTION

Sample selection is ubiquitous in empirical economics. For example, it arises naturally in the estimation of wage equations because we do not observe wages of individuals who do not work (Gronau, 1974; Heckman, 1974). Sample selection biases the estimation of causal and predictive effects when the reasons for not observing the data are related to the outcome of interest. In the wage example, there is sample selection bias whenever the employment status and offered wage depend on common unobserved variables such as ability, motivation or skills. The most widely used solution to the sample selection bias is the Heckman selection model (HSM) introduced in Heckman (1974). The classical HSM provides a parsimonious and convenient way to account for sample selection by assuming parametric Gaussian distributions for the outcome and selection processes and imposing strong homogeneity assumptions on the impact of exogenous covariates. In this paper, we present a generalization of the HSM that allows for expressly non-Gaussian structures and eliminates the strong homogeneity restrictions, resulting in a semi-parametric model with some key parameters represented as nonparametric functions. This generalization enables a tractable econometric approach, similar in ease to the classical HSM.

To illustrate the central concepts, we take labor supply as an example. Let Y^* denote the latent offered wage (outcome), and D^* be the net disutility from working (selection), e.g., difference between reservation and offered wage. We observe the employment status indicator equal to one when the offered wage is greater than the reservation wage, $D = 1(D^* \leqslant 0)$, and wage, $Y = Y^*$, only in the case of employment, $D = 1$. Assume that the drivers of the outcome and selection obey the additive structure:[1]

$$Y^* = \mu + U, \quad D^* = \nu(Z) + V, \tag{1.1}$$

where $(\mu, \nu(Z))$ are the means of the outcome and selection, Z is an instrumental variable that shifts the disutility of working but does not affect the offered wage, and (U, V) are centered stochastic shocks independent of Z. The classical HSM results from restricting (U, V) to follow a Gaussian distribution:

$$\mathrm{P}(U \leqslant u, V \leqslant v) = \Phi_2(u/\sigma_U, v/\sigma_V; \rho), \tag{1.2}$$

where $\Phi_2(\cdot, \cdot; \rho)$ denotes the standard bivariate normal distribution function with correlation ρ, and σ_U and σ_V the standard deviations of U and V. This model allows one to identify the distribution of the offered wage from the distribution of observed wage and provides tractable inference. The classical HSM has been challenged for its reliance on parametric Gaussian assumptions to obtain identification, and the data often reject these assumptions (e.g., due to clustering of offered and observed wages at the minimum wage and other levels). These challenges motivate the generalizations we consider in this study.

[1]We can have other exogenous covariates X affecting all model components. We interpret the discussion here as conditional on X having taken a fixed value. We suppress the dependence on X here for notational convenience.

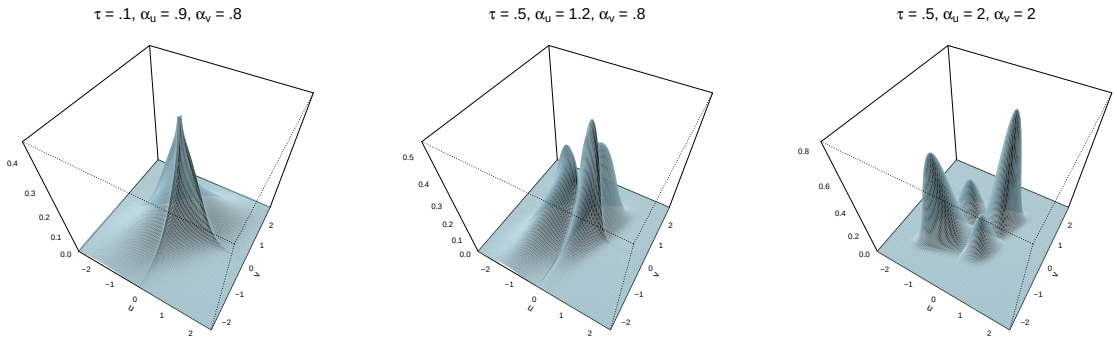

FIGURE 1. Beyond Gaussian: The richness of joint density functions permitted by (1.3).

Notes. In the model (1.3) the marginals F_U and F_V are nonparametric. To generate the panels, F_U and F_V are taken to be transformed power functions $u \mapsto \Phi(a\,(\text{sign}(u)|u|^{\alpha} + \tau\ \max(0,u)^2) + b)$ for different choices of τ and α, with a and b set such that the marginals have zero mean and unit variance; these examples extend those of Liu, Lafferty, and Wasserman (2009). The correlation parameter ρ is .7. The case $\tau = 0$ and $\alpha = 1$ for both F_U and F_V recovers the joint Gaussian density (not shown).

We generalize the classical HSM in several dimensions. We start by relaxing the Gaussian assumption on the marginal distributions of the stochastic shocks in (1.2):

$$\mathrm{P}(U \leqslant u, V \leqslant v) = \Phi_2(\Phi^{-1}(F_U(u)), \Phi^{-1}(F_V(v)); \rho), \tag{1.3}$$

where F_U and F_V are non-parametric marginal distributions of the stochastic shocks U and V, which can be continuous, discrete or mixed, and Φ^{-1} is the quantile function of the standard normal. The distribution of the latent offered wage Y^* is characterized by the parameters (μ, F_U). Despite the still partly Gaussian nature, the model can astutely generate drastically non-Gaussian distributional shapes, as demonstrated by Liu, Lafferty, and Wasserman (2009) and illustrated in Figure 1. Our contribution is to demonstrate that the distribution of the latent offered wage Y^* is still identified from the observed distribution of wages and employment if the instrumental variable Z takes on at least two values, and to provide tractable estimation and inference methods. Thus, by using (1.3) we break away from parametric assumptions and Gaussian behaviors of the classical sample selection model but maintain the same level of tractability.

The generalized HSM. The semi-parametric model (1.3) presents an expressive generalization of the classical model, but for us it merely serves as a starting point to fix ideas. We can get even more flexibility and expressivity while retaining the tractability of the identification and inference. We further relax (1.1)-(1.2) by allowing the effect of Z on D^* to be non-additive and the dependence between U and V to be heterogeneous. The general form of our model specifies the joint distribution of latent outcome

and selection drivers as

$$\mathrm{P}(Y^* \leqslant y, D^* \leqslant d \mid Z) = \Phi_2(\Phi^{-1}\left(F_{Y^*}(y)\right), \Phi^{-1}(F_{D^*}(d \mid Z)); \rho(y,d)), \tag{1.4}$$

where the entire conditional distribution of D^* can now depend on Z; and $\rho(y,d)$ is a local correlation parameter that measures the *strength of selection or sorting*. The model imposes exclusion restrictions – Z affects the marginal distribution of D^*, but does not affect the marginal distribution of Y^* nor the strength of selection. We show that this model admits identification and inference as convenient as the classical model, requiring again only that Z takes on at least two values. Moreover, the model is overidentified (hence testable) when Z takes on more than two values.

Finally, we incorporate exogenous covariates yielding the model:

$$\mathrm{P}(Y^* \leqslant y, D^* \leqslant d \mid Z, X) = \Phi_2(\Phi^{-1}(F_{Y^*}(y \mid X)), \Phi^{-1}(F_{D^*}(d \mid Z, X)), \rho(y, d \mid X)), \tag{1.5}$$

where the covariates X can affect both the marginal conditional distributions of Y^* and D^* and the sorting mechanism. This model expressly allows for heterogenous effects of X's on different parts of distribution. The data on U.K. wages lends strong empirical support for this property, with marital status and other variables heterogenously impacting different parts of the wage distribution. The heterogeneity in X property also contrasts sharply with the classical model (1.2) or its generalization (1.3), which confine the covariates to shift the location of the distribution, but not other shape properties; see Mulligan and Rubinstein (2008) for a use of the Gaussian model where covariates shift both location and scale, but not other properties of the distribution. The data firmly reject the location shift restrictions.

Our econometric development proceeds by specifying a flexible pointwise approximation to $\mathrm{P}(Y^* \leqslant y, D^* \leqslant d \mid X, Z)$ based upon the model's logic, akin to those used in univariate distribution regression (DR) models (Foresi and Peracchi, 1995; Chernozhukov, Fernández-Val, and Melly, 2013).[2] This leads to a tractable Heckman-type approach to estimating the model parameter functions. The first step consists of a probit regression for the selection equation, as in the Heckman two-step method (Heckman, 1979). The second step estimates multiple bivariate probit regression with sample selection correction. We derive functional central limit theorems for all the estimators and their bootstrapped versions and use these results to perform uniform inference on function-valued parameters, for example, the distribution of the offered wages as well as counterfactual distributions induced by various hypotheses. This type of inference provides simultaneous confidence bands and hypotheses tests about the functions of interest. Specifically, we rely on multiplier bootstrap (Giné and Zinn, 1984), applied to the estimated influence functions as in Lewbel (1995), Hansen (1996), Chernozhukov and Hansen (2006), Kline and Santos (2012) and Chernozhukov, Chetverikov, and Kato (2013).

We utilize the proposed methods to analyze data on wages and employment in the UK, covering the period from 1978 to 2013. By estimating the conditional wage distributions for men and women and

[2]Flexible refers to the use of regressors constructed from original raw regressors by taking series transformations and interactions, for example.

conducting wage decompositions that account for the endogeneity of employment selection, we shed light on the relationship between wage and employment in the UK. The results highlight the existence of positive sorting among single men and negative sorting among married women, which is consistent with the assortative matching hypothesis in the marriage market. Furthermore, the findings show that this difference in selection sorting explains a substantial portion of the gender wage gap at the upper end of the distribution, aligning with recent theories centered on glass ceiling. We still find that most of the gender gap in offered and observed wages can be attributed to differences in the wage structure that are often associated with gender discrimination in the labor market. The effect of education on wages is positive and increases with the distribution. In summary, our results demonstrate the significance of considering selection and flexible heterogeneity in empirical analysis and provide support for the use of the generalized Heckman selection model (1.5). Our results also align with those of Maasoumi and Wang (2019) which applied a quantile-regression based methodology of Arellano and Bonhomme (2017) to U.S. data over a similar period.

This paper contributes to the extensive literature on sample selection in economics and statistics. The classical references in this field include works by Gronau (1974), Heckman (1974), Lee (1982), Goldberger (1983), Amemiya (1985, Section 10.7), Maddala (1986, Section 9.4), Manski (1989), Manski (1994), and Vella (1998). A commonly utilized approach to sample selection is the Heckman selection model developed by Heckman (1974; 1976; 1979; 1990). Several extensions to the HSM have been made, including parametric extensions by Lee (1983), Prieger (2002) and Smith (2003) using different distributions, and a bivariate t-distribution extension by Marchenko and Genton (2012) for heavy-tailed data. Semi-parametric versions have been developed by Ahn and Powell (1993), Powell (1994), Andrews and Schafgans (1998), and Newey (1999), while Das, Newey, and Vella (2003) introduced a nonparametric version with additive shocks, all of which focus on location effect models with homogeneous effects. None of these extensions can accommodate all sources of heterogeneity considered in our model. Previous studies on partial identification of unrestricted sample selection models include works by Manski (1990), Balke and Pearl (1994), Manski (1994), Heckman and Vytlacil (2001), and Manski (2003), among others. Our work contributes with new partial identification results to this development, with the main emphasis on achieving point identification. In the online Supplemental material, we consider relaxed forms of the exclusion restriction on the sorting that obtain a nested sequence of bounds, which starts from a single point and ends with the (agnostic) Balke and Pearl (1994) bounds. This sequence provides a form of sensitivity analysis, analogous in spirit to Christensen and Connault (2019)'s sensitivity analysis for structural models.

Arellano and Bonhomme (2017, AB17) proposed another extension of the HSM, which like our model accounts for multiple sources of heterogeneity. Their approach is based on quantile regression, which models the marginal distribution of the latent outcome, and a parametric copula model that links the latent selection and outcome variables. However, compared to our approach, this model requires the latent variables (Y^*, D^*) to be continuous, while our approach accommodates mixed discrete-continuous distributions, which is important given that offered wages are often constrained to be above

a minimum wage level, and observed wages have point masses at minimum wage and other levels due to wage rounding. Additionally, AB17 models the effect of covariates on the conditional quantile of the latent distribution, while we model the effect on the conditional latent distribution. The identification assumptions of the two models are also distinct, as our approach imposes more structure on the dependence between the outcome and selection processes, whereas AB17 requires continuous variation in the instrument Z and real analyticity of the copula function governing the dependence of the shocks in the outcome and selection structures. The analyticity condition implies extrapolability of the conditional distribution of outcome to the case where there is no selection, enabling identification at infinity. In this case, however, identification does not automatically imply consistent estimability, because the continuation of an analytical function is known to be an ill-posed problem. To overcome the latter problem, regularization must be used: AB17 impose parametric assumptions on the copula as a means of such regularization to carry-out estimation.

Finally, there are relevant connections to the work of Liu, Lafferty, and Wasserman (2009) on the estimation of graphical models using the semi-parametric family (1.3). While our approach considers the more general model (1.5), we contribute by offering a DR based approach, which can incorporate covariates (X, Z) in a general yet tractable manner. Furthermore, our focus is on addressing the sample selection problem in the context of these models.

Outline. Section 2 examines the identification problem under sample selection using a new representation of a joint distribution. Section 3 introduces the DR model with selection and associated functionals, estimators of the model parameters and functionals, and a multiplier bootstrap method to perform functional inference. Section 4 reports the results of the empirical application. The proofs of the results of Section 2 are gathered in Appendix A. The online Supplemental Material (SM) contains deferred discussions of Sections 2–4, the asymptotic theory for our estimation and inference methods, additional empirical results, a Monte Carlo simulation calibrated to the empirical application and other technical results.

2. Local Gaussian Representation and Sample Selection

2.1. **Local Gaussian Representation of a Joint Distribution.** Our first result shows that any joint distribution of two random variables has a local Gaussian representation (LGR). This is a useful representation for econometric analysis, because it naturally nests the joint normal distribution and rich semi-parametric models that allow for nonparametric marginal distributions– with much more room to spare. Indeed, we shall use the LGR to provide a new view of the identification problem with sample selection and motivate our modeling choices later.

Let Y^* and D^* be two random variables with joint cumulative distribution function (CDF) F_{Y^*,D^*} and marginal CDFs F_{Y^*} and F_{D^*}. We label these variables with asterisks because they will be latent variables when we introduce sample selection.

Lemma 2.1 (Local Gaussian Representation of Any Joint Distribution)**.** *The joint distribution F_{Y^*,D^*} can be represented via a standard bivariate normal distribution at any point (y,d) as*

$$F_{Y^*,D^*}(y,d) \equiv \Phi_2(\Phi^{-1}(F_{Y^*}(y)), \Phi^{-1}(F_{D^*}(d)); \rho(y,d)), \tag{2.6}$$

where $\Phi_2(\cdot,\cdot;\rho)$ is the joint CDF of a standard bivariate normal random variable with parameter ρ; and $\rho(y,d) \in [-1,1]$ is the implied or local correlation parameter that depends on (y,d), whose value is unique.

Lemma 2.1 establishes that any bivariate CDF admits a unique pointwise representation by standard bivariate normal distributions. This result is related to Sklar's copula representation of the joint distributions but is different, especially in using the localized correlation. Also this result is stronger than the comprehensive property of the Gaussian copula that establishes that this copula includes the two Frechet bounds and independent copula by suitable choice of the correlation parameter, e.g., Smith (2003). We also note that Lemma 2.1 easily extends to CDFs conditional on covariates by making all the parameters dependent on the value of the covariates.

In what follows, it is convenient to define the LGR parameters:

$$\mu(y) := \Phi^{-1}(F_{Y^*}(y)), \;\; \nu(d) := \Phi^{-1}(F_{D^*}(d)),$$

noting that $\mu(y) \in \overline{\mathbb{R}}$ and $\nu(d) \in \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ is the extended real number line. By LGR, the mapping between $F_{Y^*,D^*}(y)$ and its LGR parameters $(\mu(y), \nu(d), \rho(d,y))$ is bijective. Thus LGR carries the same information as the joint CDF.

The implied correlation $\rho(y,d)$ is a key parameter that measures local dependence. Indeed, $\rho(y,d) = 0$ if and only if the distribution F_{Y^*,D^*} factorizes at (y,d) because

$$F_{Y^*,D^*}(y,d) = \Phi_2(\Phi^{-1}(F_{Y^*}(y)), \Phi^{-1}(F_{D^*}(d)); 0) = F_{Y^*}(y) F_{D^*}(d),$$

that is, $\rho(y,d) = 0$ if and only if the events $\{Y^* \leqslant y\}$ and $\{D^* \leqslant d\}$ are independent. Moreover, $\rho(y,d)$ is positive if and only if correlation of $1(Y^* \leqslant y)$ and $1(D^* \leqslant d)$ is positive. Thus, if $\rho(y,d)$ is positive everywhere then Y^* and D^* are positively quadrant dependent (Lehmann, 1966). We prove these assertions and provide additional discussion in Appendix A.1 of the SM.[3]

2.2. **Identification of Sample Selection Model.** We consider now the sample selection problem where we observe two random variables D and Y, which can be defined in terms of the latent variables D^* and Y^* as

$$\begin{aligned} D &= 1(D^* \leqslant 0), \\ Y &= Y^* \text{ if } D = 1, \end{aligned}$$

i.e., D is an indicator for $D^* \leqslant 0$ and Y^* is only observed when $D = 1$. The goal is to identify features of the joint distribution of the latent variables from the joint distribution of the observed variables.

[3]See also Tjøstheim, Otneim, and Støve (2018) for a recent review of copula-based measures of local dependence.

We can write the distribution of the observed variables as

$$\mathrm{P}(D=1)=\Phi(\nu) \text{ and } \mathrm{P}(Y\leqslant y, D=1)=\Phi_2(\mu(y),\nu;\rho(y)),$$

where $\mu(y)$, $\nu := \nu(0)$ and $\rho(y) := \rho(y,0)$ are the parameters of LGR for the latent distribution F_{Y^*,D^*}. The identified set for these parameters determine the identified set for $F_{Y^*}(y)$. Note that in particular,

$$F_{Y^*}(y)=\Phi(\mu(y)).$$

As shown below, $\mu(y)$ and $\rho(y)$ are partially identified. We proceed by characterizing the identified set for these parameters and provide exclusion restrictions to achieve point identification. We also examine the role of relaxed exclusion and other restrictions in reducing the size of the identified set in Appendix A.4 of the SM.

To understand the source of partial identification under sample selection, note that there are two free probabilities, $\mathrm{P}(D=1)$ and $\mathrm{P}(Y\leqslant y, D=1)$, to identify three parameters, $\mu(y)$, ν and $\rho(y)$. The selection probability pins down ν: $\nu=\Phi^{-1}(\mathrm{P}(D=1))$. The parameters $\mu(y)$ and $\rho(y)$ are partially identified by the set of solutions in (μ,ρ) to the equation

$$\mathrm{P}(Y\leqslant y, D=1)=\Phi_2(\mu,\Phi^{-1}(\mathrm{P}(D=1));\rho).$$

These solutions form a one-dimensional manifold in $\mathbb{R}\times(-1,1)$ (Spivak, 1965; Munkres, 1991).[4]

Exclusion Restrictions. To state the exclusion restrictions, let Z be a candidate instrumental variable and $F_{Y^*,D^*|Z}$ be the joint CDF of Y^* and D^* conditional on Z. Then, $F_{Y^*,D^*|Z}$ admits the LGR:

$$F_{Y^*,D^*|Z}(y,d\mid z)=\Phi_2(\mu(y\mid z),\nu(d\mid z);\rho(y,d\mid z)),$$

with parameters $\mu(y\mid z)\in\overline{\mathbb{R}}$, $\nu(d\mid z)\in\overline{\mathbb{R}}$, and $\rho(y,d\mid z)\in[-1,1]$. The exclusion restrictions are the following.

Assumption 1 (Exclusion Restrictions)**.** *There is a binary random variable Z that satisfies:*

(1) *Non-Degeneracy:* $0<\mathrm{P}(D=1)<1$ *and* $0<\mathrm{P}(Z=1\mid D=1)<1$.
(2) *Relevance:* $\mathrm{P}(D=1\mid Z=0)<\mathrm{P}(D=1\mid Z=1)<1$.
(3) *Outcome exclusion:* $\mu(y\mid z)=\mu(y)$ *for all* $y\in\mathbb{R}$ *and* $z\in\{0,1\}$.
(4) *Selection Sorting exclusion:* $\rho(y,0\mid z)=\rho(y,0)$ *for all* $y\in\mathbb{R}$ *and* $z\in\{0,1\}$.

The condition that Z is binary emphasizes that our identification strategy does not rely on large variation of Z. If Z is not binary, we only require that Assumption 1 be satisfied for two values of Z. If it is satisfied for more than two values of Z, then the model is overidentified and the exclusion restrictions become testable.[5] Non-degeneracy states that there is sample selection and that Z has variation in the selected population. Relevance requires that Z affects the probability of selection. The condition $\mathrm{P}(D=1\mid Z=1)<1$ precludes identification at infinity, which we analyze separately in Remark 2.1.

[4]This is because $\partial\Phi_2(\mu,\cdot;\rho)/\partial\mu>0$, $\partial\Phi_2(\cdot,\cdot;\rho)/\partial\rho>0$, and $\partial^2\Phi_2(\cdot,\cdot;\rho)/\partial\mu\partial\rho>0$.

[5]We leave the development of such specification test to future research.

The sign of the first inequality can be reversed by relabelling the values of Z. Outcome exclusion is a standard exclusion restriction, which is not sufficient for point identification in the presence of sample selection (Balke and Pearl, 1994; Manski, 1994; Heckman and Vytlacil, 2001; Manski, 2003). It holds when Y^* is independent of Z.[6] Selection sorting exclusion requires the local correlation function to be independent of Z.[7] AB17 consider an alternative condition to selection sorting exclusion based on real analyticity and continuous variation of Z. We refer to Appendices A.2 and A.3 in the SM for a comparison with the analyticity approach and another alternative approach based on imposing semi-parametric structure on the sorting mechanism when Z takes on more than 2 values.

We can get some intuition about the outcome and selection exclusion restrictions with the parametric and semi-parametric models of labor supply given in (1.1)–(1.3). These models trivially satisfy the conditions stated above, because Z only affects the latent disutility D^* through the location, and the local correlation parameter is constant and does not depend on Z (or y for that matter). Yet the selection sorting exclusion does entail some loss of generality, which motivates forms of relaxed conditions on selection sorting that we consider in Appendix A.4 of the SM.

We now show how the presence of exclusion restrictions helps identify the parameters. Under Assumption 1 the conditional LGR at $d = 0$ simplifies to

$$F_{Y^*,D^*|Z}(y,0 \mid z) = \Phi_2(\mu(y), \nu(z); \rho(y)), \quad z \in \{0,1\}, \tag{2.7}$$

where we simplify the notation $\nu(z) := \nu(0 \mid z)$ and $\rho(y) := \rho(y,0)$. We can relate this representation to the conditional distribution of observed variables as

$$\mathrm{P}(D = 1 \mid Z = z) = \Phi(\nu(z)); \quad \mathrm{P}(Y \leqslant y, D = 1 \mid Z = z) = \Phi_2(\mu(y), \nu(z); \rho(y)), \quad z \in \{0,1\}.$$

As before, $\nu(z)$ is identified from the conditional selection probability:

$$\nu(z) = \Phi^{-1}\left(\mathrm{P}(D = 1 \mid Z = z)\right), \quad z \in \{0,1\}. \tag{2.8}$$

Moreover, $\mu(y)$ and $\rho(y)$ are identified as the solution in (μ, ρ) to

$$\mathrm{P}(Y \leqslant y, D = 1 \mid Z = z) = \Phi_2(\mu, \Phi^{-1}\left(\mathrm{P}(D = 1 \mid Z = z)\right); \rho), \quad z \in \{0,1\}. \tag{2.9}$$

This is a nonlinear system of two equations in two unknowns. The result below states that the solution exists and is unique.

Theorem 2.1 (Identification under Assumption 1)**.** *Suppose that Assumption 1 holds. Suppose that the distribution of the observed variables* (Y, D, Z) *implies* $\rho(y)^2 < 1$*, then* $\mu(y)$ *and* $\rho(y)$ *are point identified as the unique interior solution of* (2.9) *in* (μ, ρ)*.*

The result follows by showing that the Jacobian of the equations in (2.9) is a P-matrix for all $\mu \in \mathbb{R}$ and $\rho \in (-1, 1)$, so uniqueness follows by the global univalence result in Theorem 4 of Gale and Nikaido

[6] Kitagawa (2010) developed a test for the outcome exclusion.

[7] Torgovitsky (2010) previously used this type of restriction to analyze identification of nonseparable models with continuous endogenous explanatory variables.

(1965). We defer to Appendix A.2 the analysis of the case $\rho(y)^2 = 1$ due to its more technical nature. This case deals with boundary situations that can occur at extreme values of y or other non-typical circumstances. In these situations, we can have either point or partial identification. They are easy to detect empirically.

Remark 2.1 (Identification at Infinity)**.** When the instruments are strong enough to set

$$\mathrm{P}(D = 1 \mid Z = 1) = 1$$

and provided that the exclusion conditions holds $\mu(y \mid z) = \mu(y)$, the conditional LGR at $z = 1$ gives

$$\mathrm{P}(Y \leqslant y, D = 1 \mid Z = 1) = \lim_{v \nearrow +\infty} \Phi_2(\mu(y), v; \rho(y \mid 1)) = \Phi(\mu(y)),$$

which identifies $\mu(y)$ by

$$\mu(y) = \Phi^{-1}(\mathrm{P}(Y \leqslant y, D = 1 \mid Z = 1)),$$

without the selection sorting exclusion.[8] This result is analogous to the identification at infinity of Chamberlain (1986) where Z is continuous with unbounded support and $\lim_{z \nearrow +\infty} \mathrm{P}(D = 1 \mid Z = z) = 1$; see also Lewbel (2007) for an alternative identification at infinity strategy using the special regressor approach. Interestingly we note that $\rho(y \mid 1)$ is not point-identified without further restrictions. □

2.3. **Selection Sorting Exclusion.** We provide a further interpretation of the selection exclusion restriction in a formulation of the sample selection problem in terms of the propensity score. The sample selection process is often represented as $D = 1\{V \geqslant v(Z)\}$, where $v(z) := \Phi^{-1}(p(Z))$, with $p(Z) := \mathrm{P}(D = 1 \mid Z = z)$ being the propensity score, and $V \mid Z \sim N(0,1)$ is the unobserved selection normal score. Here we shall maintain Assumption 1(1)–(3). The following theorem shows the condition that selection sorting exclusion imposes on the relationship between Y^* and V. Let

$$F_{Y^*,V|Z}(y, v \mid z) = \Phi_2(\Phi^{-1}(F_{Y^*}(y)), v; \tilde{\rho}(y, v \mid z))$$

be the LGR of the joint CDF of (Y^*, V) conditional on Z, where we use that Y^* is independent of Z and $V \mid Z \sim N(0,1)$.

Theorem 2.2 (Interpretation of Selection Sorting Exclusion)**.** *Suppose Assumption 1(1)–(3) hold. Then Assumption 1(4) holds if and only if $\tilde{\rho}(y, v(z) \mid z) = \tilde{\rho}(y)$ for any z in the support of Z.*

Theorem 2.2 shows that selection exclusion holds whenever the implied correlation $\tilde{\rho}(y, v \mid z)$ between Y^* and V does not depend on v and z. A sufficient condition is that (Y^*, V) are jointly independent of Z and $\tilde{\rho}(y, v) = \tilde{\rho}(y)$ for all v in the support of $v(Z)$, where $\tilde{\rho}(y, v)$ is the local dependence parameter of the joint CDF of (Y^*, V). In the context of labor supply, Y^* is latent wage and V is the ranking in the conditional distribution of utility/net benefit of employment. The classical HSM in (1.2)

[8]Despite its extreme nature, this identification strategy is useful in empirical economics: e.g., Mulligan and Rubinstein (2008) use this reasoning to analyze wage penalty for highly educated women.

and its semi-parametric generalization in (1.3) trivially satisfy this condition, because they have constant local correlation, $\tilde{\rho}(y, v) = \rho$, that is, the strength of selection does not vary with the level of offered wage or the employment ranking. Relative to such traditional models, the condition $\tilde{\rho}(y, v) = \tilde{\rho}(y)$ leaves some room for additionally flexibility – namely, the strength of selection can vary with the level of offered wage. For example, $y \mapsto \tilde{\rho}(y)$ increasing means that workers are more likely to select into the labor force if the offered wage's rank is higher (holding the employment ranking fixed).

The selection exclusion $\tilde{\rho}(y, v) = \tilde{\rho}(y)$ is essentially equivalent to the single index restriction:

$$\mathrm{P}(Y^* \leqslant y \mid V = v) = \Phi\big(a(y) + b(y)\, v\big), \tag{2.10}$$

where $y \mapsto a(y)$ and $y \mapsto b(y)$ are nonparametric functions linked to the LGR of Y^* and V, $F_{Y^*,V}(y, v) = \Phi_2(\mu(y), v; \tilde{\rho}(y))$, via $a(y) = \mu(y)/\sqrt{1 - \tilde{\rho}(y)^2}$ and $b(y) = -\tilde{\rho}(y)/\sqrt{1 - \tilde{\rho}(y)^2}$. This index form enhances interpretability and also connects to the econometric literature that utilizes the simplifying index restrictions; e.g., Ichimura (1993), Klein and Spady (1993), Powell (1994) and Vytlacil (2002). The proof of the equivalence in (2.10) is given in Appendix A.4. Homogeneous models, such as the classical Heckman's labor supply model or the semi-parametric generalization (1.3), impose the much stronger single index condition:

$$\mathrm{P}(Y^* \leqslant y \mid V = v) = \Phi\big(a(y) + bv\big),$$

for all y, for $b = -\rho/\sqrt{(1 - \rho^2)}$ being a rescaled correlation coefficient; note that the Gaussian restriction on Y^* of the HSM corresponds to the further linearity restriction: $a(y) = ay$. This connection highlights the generalization our model brings: in the general model $a(y) \neq ay$ allows for expressive departures from Gaussianity, and $b(y) \neq b$ allows for much richer patterns of dependence across the conditional distribution.

We now elaborate on the value that the variation $\rho(y)$ with respect to y brings to the table. What does this property mean in terms of the flexibility of dependence patterns between net offered wage Y^* and employment ranking V? The standard models that have homogeneous $\rho(y) = \rho$ or *equivalently* homogeneous $b(y) = b$, are only able to generate limited forms of dependence. The left panel of Figure 2 presents an example of the contour of the joint pdf of (Y^*, V) when $\rho(y) = .14$. In contrast, our model is able to generate richer forms of dependence between (Y^*, V) as illustrated in the middle and right panels. The middle panel has the same correlation between Y^* and V as the left panel, but we see that the dependence increases from small values as we move towards the upper-right tail corner. The right panel is another example, where the correlation between Y^* and V is zero, but (Y^*, V) are positively dependent when $Y^* > 0$ and negative dependent when $Y^* < 0$. In all examples, the marginal distributions are fixed to be normal, but the joint distributions are not normal, except for the left panel.

The above discussions aim to support the idea of using sorting exclusion, but one may and should challenge it in applications. To this end, we emphasize three points: First, as we mentioned in the discussion of Assumption 1, selection exclusion is testable when Z takes on more than 2 values. Second, one can work with the relaxed forms of exclusion restrictions, namely the sign restrictions and r-relaxed exclusion restrictions and compute the bounds provided by Theorem A.1 in the SM. Third,

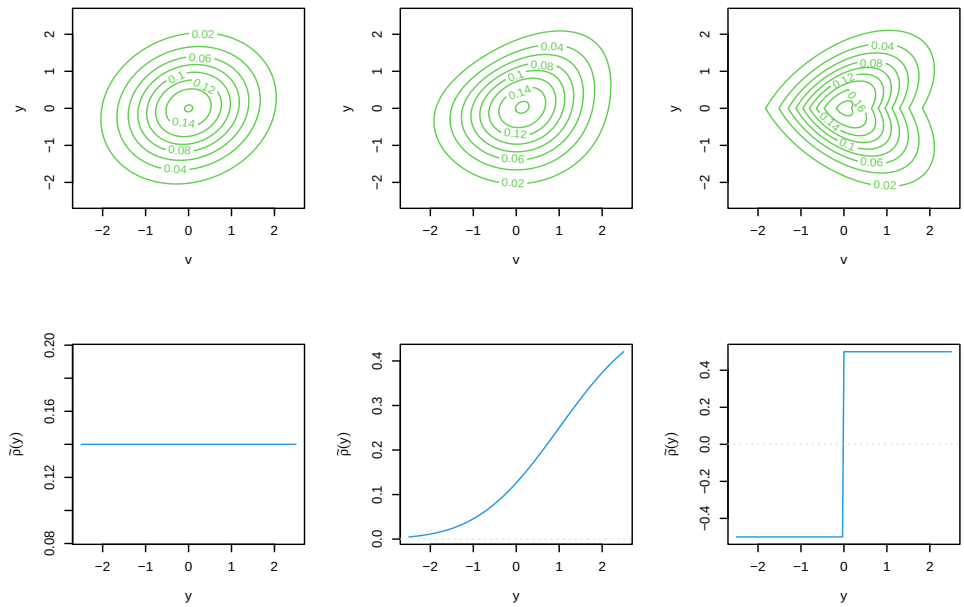

FIGURE 2. Joint CDF of (Y^*, V) for different local correlation parameters $\tilde{\rho}(y)$; marginal CDFs of (Y^*, V) are standard normal.

Notes. Left panel: $\tilde{\rho}(y) = .14$; middle panel: $\tilde{\rho}(y) = \Phi(2(y-1)/3)/3$; right panel: $\tilde{\rho}(y) = 1(y \geqslant 0)/2 - 1(y < 0)/2$. The marginal distribution of Y^* and V are chosen as standard normal for simplicity.

if richer instruments are available, one can still obtain results for identification under analyticity or other regularity assumptions on the local correlation function without selection sorting exclusion; see Appendix A.3 in the SM.

3. ECONOMETRICS OF DISTRIBUTION REGRESSION MODEL WITH SAMPLE SELECTION

3.1. **The Model.** We consider a semi-parametric version of the LGR with covariates:

$$F_{Y^*,D^*}(y, 0 \mid Z = z) = \Phi_2(-x'\beta(y), -z'\pi; \rho(x'\delta(y))), \tag{3.11}$$

where Y^* is the latent outcome of interest, which can be either continuous, discrete or mixed; D^* is a latent variable that determines sample selection; X is a vector of covariates; $Z = (Z_1, X)$; and Z_1

are excluded covariates, i.e., observed covariates that satisfy the exclusion restrictions.[9] The excluded covariates avoid reliance on functional form assumptions to achieve identification.

The model (3.11) is a semi-parametric version of (1.5), where we replace the components $\Phi^{-1}(F_{Y^*}(y \mid X))$, $\Phi^{-1}(F_{D^*}(d \mid Z, X))$, and $\rho(y, d \mid X)$ by three indexes.[10] We shall refer to $-x'\beta(y)$ as the outcome equation, to $-z'\pi$ as the selection equation, and to $\rho(x'\delta(y))$ as the selection sorting equation. We observe the selection indicator $D = 1(D^* > 0)$ and the outcome $Y = Y^*$ when $D = 1$.[11] In the empirical application that we consider below, Y^* is offered wage, D^* is the net utility from working, e.g., the difference between offered wage and reservation wage, D is an employment indicator, Y is the observed wage, X includes labor market characteristics such as education, age, number of children and marital status, and Z_1 includes measures of out-of-work income. We shall discuss the validity of these measures as excluded covariates in Section 4.

The model (3.11) is semi-parametric because $y \mapsto \beta(y)$ and $y \mapsto \delta(y)$ are unknown functions, i.e. infinite dimensional parameters in general. This flexibility allows the effect of X on the outcome and selection sorting equations to vary across the distribution. For example, it allows the return to education to vary across the distribution, the selection sorting to be different for high and low educated individuals, or to have positive selection sorting at the upper tail and negative at the bottom tail or vice versa.[12] The function $u \mapsto \rho(u)$ is a known link with range $[-1, 1]$, e.g. the Fisher transformation (Fisher, 1915), $\rho(u) = \tanh(u)$. The corresponding distribution of Y^* conditional on Z is

$$F_{Y^*}(y \mid Z = z) = \lim_{v \nearrow +\infty} F_{Y^*, D^*}(y, v \mid Z = z) = \Phi(-x'\beta(y)), \;\; z = (x, z_1).$$

The selection bias arises because this distribution is different from the distribution of the observed outcome Y, i.e. $F_{Y^*}(y \mid Z = z) \neq F_Y(y \mid Z = z, D = 1)$

Example 3.1 (Semi-Parametric version of Heckman Selection Model)**.** Here we revisit the semi-parametric special case given in the Introduction, where

$$D^* = \nu(Z) + V, \quad Y^* = \mu(X) + U,$$

[9] It is understood that the models are flexible in the following sense. Given x and z, we can generate constructed regressors $t(x)$ and $b(z)$ as technical transformations of x and z, for example, by taking powers or splines of the components and their interactions. We then can reassign notation $x \leftarrow t(x)$ and $z \leftarrow b(z)$. This convention entails no loss of generality provided that $t(x)$ and $b(z)$ contain the same information as x and z. In this paper, we do not formally consider the case where the dimensions of t and b grow with the sample size, but this is possible along the lines of the series literature, with the inference being operationally equivalent to the case with the fixed number of series terms, provided that the number of series terms is small compared to the sample size and that the approximation errors are negligible; e.g., Newey (1997) and Chen (2007).

[10] Here we include the covariates X in Z to lighten the notation and denote the covariates that satisfy the exclusion restriction by Z_1 instead of Z.

[11] The minus signs in (3.11) are included to take into account that the selection is defined by $D^* > 0$ instead of $D^* \leqslant 0$. We use this definition to facilitate the interpretation of the parameters and the comparison with the classical HSM; see Example 3.1.

[12] The parametric copula model of AB17 imposes that the sign of the sorting is the same across the latent wage distribution.

where (U, V) is independent of Z (Z contains X here) such that

$$F_{U,V}(u, v \mid Z = z) = \Phi_2\left(\Phi^{-1}(F_U(u)), \Phi^{-1}(F_V(v)); \rho\right).$$

Therefore,

$$F_{Y^*,D^*}(y, 0 \mid Z = z) = \Phi_2\left(\Phi^{-1}(F_U(y - \mu(x))), \Phi^{-1}(F_V(-\nu(z))); \rho\right).$$

As noted in the Introduction, this (strictly) nests the classical Gaussian selection model where $F_U(u) = \Phi(u/\sigma_U)$ and $F_V(v) = \Phi(v/\sigma_V)$, where σ_U and σ_V are the standard deviations of U and V. For the econometric specification, we use parametrization:

$$F_U(y - \mu(x)) = \Phi\left(-x'\beta(y)\right); \quad F_V(-\nu(z)) = \Phi(-z'\pi).$$

□

The parameters $\beta(y)$ and π have the same interpretation as in probit models. Their signs are informative about the signs of the partial effects of the corresponding covariates on the conditional distribution of the outcome or probability of selection, and ratios of their components yield ratios of partial effects. Indeed, if X is continuous,

$$\frac{\partial F_{Y^*}(y \mid Z = z)}{\partial x} = -\beta(y)\phi(-x'\beta(y)),$$

where ϕ is the standard normal PDF. The parameter $\delta(y)$ determines the sign of the effect of the covariates on the sorting because

$$\frac{\partial \rho(x'\delta(y))}{\partial x} = \delta(y)\dot{\rho}(x'\delta(y)),$$

and $\dot{\rho}(u) = \partial\rho(u)/\partial u = 1 - \tanh(u)^2 > 0$. Appendix B in the SM provides additional discussion on the interpretation of the model parameters.

Example 3.2 (Data Generating Process)**.** The model (3.11) has multiple data generating process representations as nonseparable systems. One example is

$$\begin{aligned} D^* &= Z'\pi + V, \;\; V \mid Z \sim \mathcal{N}(0,1), \\ X'\beta(Y^*) &= \rho(X'\delta(Y^*))V + \sqrt{1 - \rho(X'\delta(Y^*))^2}U, \;\; U \mid Z \sim \mathcal{N}(0,1), \end{aligned}$$

where U and V are independent. For example, in the wage application V can be interpreted as unobserved utility from working (unobserved benefit of working for money-metric utility), net of what Z already captures, and U as unobserved skills or innate ability net of what V and X already capture. This representation is similar to the semi-parametric HSM in Example 3.1 with the difference that the equation for Y^* is nonseparable. □

3.2. **Functionals: Factual, Counterfactual** & **Decomposition.** Several key functionals of the model's parameters (3.11) can be of interest. One is the marginal distribution of the latent outcome Y^*

$$F_{Y^*}(y) = F_{Y^*}(y;\beta,F_X) := \int F_{Y^*}(y \mid Z=z)dF_Z(z) = \int \Phi(-x'\beta(y))dF_X(x),$$

where F_Z and F_X are the marginal distributions of Z and X, respectively. In the case of the wage application, F_{Y^*} corresponds to the distribution of the offered wage, which is a potential or latent outcome free of selection. Using the formula above, we can also construct *counterfactual* distributions by replacing $\beta(y)$ and F_X by coefficients and distributions from different populations or groups, $\bar{\beta}(y)$ and $\bar{F}_X$. These distributions are useful to decompose the distribution of offered wages between females and males or between blacks and whites, which can be the basis to uncover discrimination in the labor market. Another functional is the probability of selection

$$\mathrm{P}(D=1) = \mathrm{P}(D=1;\pi,F_Z) := \int \mathrm{P}(D=1 \mid Z=z)dF_Z(z) = \int \Phi(z'\pi)dF_Z(z),$$

which we can also use to define counterfactuals and employ them to decompose differences in employment rates between employment structure effects, π, and composition effects, F_Z.

We can also use the model to construct distributions for the observed outcome using that

$$\begin{aligned} F_{Y|D}(y \mid 1) = F_Y(y;\beta,\pi,\delta,F_Z) &:= \int \frac{\Phi_2\left(-x'\beta(y), z'\pi; -\rho(x'\delta(y))\right)}{\Phi(z'\pi)} dF_Z(z \mid D=1) \\ &= \frac{\int \Phi_2\left(-x'\beta(y), z'\pi; -\rho(x'\delta(y))\right) dF_Z(z)}{\int \Phi(z'\pi)dF_Z(z)}, \end{aligned}$$

where the second equality follows from the Bayes rule. We can again construct counterfactual distributions by changing $\beta(y)$, π, $\delta(y)$ and F_Z. In the wage application, we will decompose the differences in the wage distribution between genders or across time into changes in the worker composition F_Z, wage structure $\beta(y)$, selection structure π, and selection sorting $\delta(y)$. Both selection effects are new to this model.

Quantiles and other functionals of the distributions of latent and observed outcomes can be constructed by applying the appropriate operator. For example, the τ-quantile of the latent outcome is $Q_{Y^*}(\tau) = \mathbf{Q}_\tau(F_{Y^*})$, where $\mathbf{Q}_\tau(F) := \inf\{y \in \mathbb{R} : F(y) \geqslant \tau\}$ is the quantile or left-inverse operator.

3.3. **Estimation.** To estimate the model parameters and functionals of interest, we assume that we have a random sample of size n from (D, DY, Z), $\{(D_i, D_iY_i, Z_i)\}_{i=1}^n$, where we use DY to indicate that we only observe Y when $D=1$.

Before describing the estimators, it is convenient to introduce some notation. Let $\mathscr{Y}$ be the region of interest of Y, and denote $\theta_y := (\beta(y), \delta(y))$, where we replace the arguments in y by subscripts to lighten the notation.[13]

[13]If the support of Y is finite, $\mathscr{Y}$ can be the entire support, otherwise $\mathscr{Y}$ should be a subset of the support excluding low density areas such as the tails.

The estimation relies on the relationship between conditional distributions and binary regressions. Thus, the CDF of Y at a point y conditional on X is the expectation that an indicator that Y is less than y conditional on X,

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

To implement this idea, we construct the set of indicators for the selected observations

$$I_{yi} = 1(Y_i \leqslant y) \text{ if } D_i = 1,$$

for each $y \in \mathscr{Y}$. In the presence of sample selection, we cannot just run a probit binary regression of I_{yi} on X_i to estimate the parameter $\beta(y)$ as in Foresi and Peracchi (1995) and Chernozhukov, Fernández-Val, and Melly (2013). The problem is similar to running least squares in the HSM. Instead, we use that

$$\begin{aligned}\ell_i(\pi, \theta_y) = \left[1 - \Phi(Z_i'\pi)\right]^{1-D_i} &\times \Phi_2(-X_i'\beta(y), Z_i'\pi; -\rho(X_i'\delta(y)))^{D_i I_{yi}} \\ &\times \Phi_2(X_i'\beta(y), Z_i'\pi; \rho(X_i'\delta(y)))^{D_i(1-I_{yi})}\end{aligned}$$

is the likelihood of (D_i, I_{yi}) conditional on Z_i. This likelihood is the same as the likelihood of a bivariate probit model or more precisely a probit model with sample selection (Zellner and Lee, 1965; Poirier, 1980; Van de Ven and Van Praag, 1981).

We estimate the model parameters using a computationally attractive two-step method to maximize the average log-likelihood, similar to the Heckman two-step method. The first step is a probit regression for the probability of selection to estimate π, which is identical to the first step in the Heckman two-step method. The second step consists of multiple distribution regressions (DRs) with sample selection corrections to estimate $\beta(y)$ and $\delta(y)$ for each value of $y \in \mathscr{Y}$. These steps are summarized in the following algorithm:

Algorithm 3.1 (Two-Step DR Method)**.** *(1) Run a probit for the selection equation to estimate π: (2) Run multiple DRs with sample selection correction to estimate θ_y. That is,*

$$\begin{aligned}&(1) \quad \widehat{\pi} = \underset{c \in \mathbb{R}^{d_\pi}}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \left[D_i \log\Phi(Z_i'c) + (1 - D_i)\log\Phi(-Z_i'c)\right], \\ &(2) \quad \widehat{\theta}_y = \underset{t=(b,d)\in\Theta}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} D_i \left[\, I_{yi} \log\Phi_2\left(-X_i'b, Z_i'\widehat{\pi}; -\rho(X_i'd)\right)\right. \\ &\qquad\qquad \left. + (1 - I_{yi})\log\Phi_2\left(X_i'b, Z_i'\widehat{\pi}; \rho(X_i'd)\right)\right], \quad y \in \mathscr{Y}.\end{aligned}$$

where $I_{yi} = 1(Y_i \leqslant y)$ and $\Theta \in \mathbb{R}^{d_\theta}$ is a compact parameter set, and

$$d_\pi := \dim \pi, \; d_\theta := \dim \theta_u, \;\; \rho(u) := \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \in [-1, 1].$$

In practice we replace the set $\mathscr{Y}$ by a finite grid $\bar{\mathscr{Y}}$ if $\mathscr{Y}$ contains many values.

The estimators of the functionals of interest are constructed from the estimators of the parameters using the plug-in method. For example, the estimator of the distribution of the latent outcome and the

estimator of the probability of selection are

$$\widehat{F}_{Y^*}(y) := F_{Y^*}(y;\widehat{\beta},\widehat{F}_X) = \frac{1}{n}\sum_{i=1}^{n}\Phi(-X_i'\widehat{\beta}(y)), \quad \widehat{\mathrm{P}}(D=1) := \mathrm{P}(D=1;\widehat{\pi},\widehat{F}_Z) = \frac{1}{n}\sum_{i=1}^{n}\Phi(z'\widehat{\pi}), \tag{3.12}$$

and the estimators of the counterfactual distributions of the observed outcome are constructed from

$$\widehat{F}_{Y|D}(y \mid 1) := F_Y(y;\widehat{\beta},\widehat{\pi},\widehat{\delta},\widehat{F}_Z) = \frac{\sum_{i=1}^{n}\Phi_2(-X_i'\widehat{\beta}(y), Z_i'\widehat{\pi}; -\rho(X_i'\widehat{\delta}(y)))}{\sum_{i=1}^{n}\Phi(Z_i'\widehat{\pi})}, \tag{3.13}$$

by choosing the estimators of $\widehat{\beta}(y)$, $\widehat{\pi}$, and $\widehat{\delta}(y)$ and the sample values of Z appropriately. Estimators of quantiles and other functionals of these distributions are obtained by applying the operators that define the functionals to the estimator of the distribution. See Section D of SM for details.

3.4. **Inference on Functional Parameters.** The model parameters and functionals of interest are generally function-valued. We show how to construct confidence bands for them that can be used to test functional hypotheses such as the entire function being zero, non-negative or constant. To explain the construction, consider the case where the functional of interest is a linear combination of the model parameter θ_y, that is the function $y \mapsto c'\theta_y$, $y \in \mathscr{Y}$, where $c \in \mathbb{R}^{d_\theta}$. The set $CB_p(c'\theta_y)$ is an asymptotic p-confidence band for $c'\theta_y$ if it satisfies

$$\mathrm{P}\left[c'\theta_y \in CB_p(c'\theta_y), \text{ for all } y \in \mathscr{Y}\right] \to p.$$

We form $CB_p(c'\theta_y)$ as $CB_p(c'\theta_y) := c'\widehat{\theta}_y \pm c\nu(p)SE(c'\widehat{\theta}_y)$, where $\widehat{\theta}_y$ is the estimator of θ_y defined in Algorithm 3.1, $SE(c'\widehat{\theta}_y)$ is the standard error of $c'\widehat{\theta}_y$, and $c\nu(p)$ is a critical value, i.e. a consistent estimator of the p-quantile of the statistic

$$t_{\mathscr{Y}} = \sup_{y\in\mathscr{Y}} \frac{|c'\widehat{\theta}_y - c'\theta_y|}{SE(c'\widehat{\theta}_y)}.$$

We obtain the standard error and critical value from the limit distribution of the stochastic process $y \mapsto \widehat{\theta}_y$ derived in Section D of the SM. In practice, it is convenient to estimate the critical value using resampling methods. Multiplier bootstrap is computationally attractive in our setting because it does not require parameter re-estimation and therefore avoids the nonlinear optimization in both steps of Algorithm 3.1. The multiplier bootstrap is implemented using the following algorithm:

Algorithm 3.2 (Multiplier Bootstrap)**.** *(i) For $b \in 1,\dots,B$ and the finite grid $\tilde{\mathscr{Y}} \subseteq \mathscr{Y}$, repeat the steps: (1) Draw the bootstrap multipliers $\{\omega_i^b : 1 \leqslant i \leqslant n\}$ independently from the data and normalized them to have zero mean, $\omega_i^b = \tilde{\omega}_i^b - \sum_{i=1}^n \tilde{\omega}_i^b/n$, $\tilde{\omega}_i^b \sim$ i.i.d. $\mathscr{N}(0,1)$. (2) Obtain the bootstrap estimator of the model parameter*

$$\widehat{\theta}_y^b = \widehat{\theta}_y + n^{-1}\sum_{i=1}^{n}\omega_i^b\,\widehat{\psi}_i(\widehat{\theta}_y,\widehat{\pi}),$$

where $\widehat{\psi}_i(\widehat{\theta}_y,\widehat{\pi})$ is an estimator of the influence function of $\widehat{\theta}_y$ given in equation (D.24) *of the SM. (3) Construct bootstrap realization of maximal t-statistic $t_{\mathscr{Y}}$ for the functional of interest,*

$$t_{\mathscr{Y}}^b = \max_{y\in\tilde{\mathscr{Y}}} \frac{|c'\widehat{\theta}_y^b - c'\widehat{\theta}_y|}{SE(c'\widehat{\theta}_y)}, \quad SE(c'\widehat{\theta}_y) = \sqrt{c'\widehat{\Sigma}_{\theta_y\theta_y}c},$$

where $\widehat{\Sigma}_{\theta_y\theta_y}$ *is an estimator of the asymptotic variance-covariance matrix of* $\widehat{\theta}_y$ *given in equation* (D.23) *of the SM. (ii) Compute the critical value* $cv(p)$ *as the simulation* $p-$*quantile of* $t^b_{\widetilde{\mathscr{Y}}}$, $cv(p) = p-$*quantile of* $\{t^b_{\widetilde{\mathscr{Y}}} : 1 \leqslant b \leqslant B\}$

The centering of the multipliers in step (i1) of the algorithm is a finite sample adjustment. Confidence bands for other functionals of the model parameter can be constructed using a similar bootstrap method.

4. Wage Decompositions in the UK

We apply the DR model with sample selection to carry out wage decompositions accounting for endogenous employment participation using data from the United Kingdom.

4.1. **Data.** The data come from the U.K. Family Expenditure Survey (FES) for the years 1978 to 2001, Expenditure and Food Survey (EFS) for the years 2002 to 2007, and Living Costs and Food Survey (LCFS) for the years 2008 to 2013. Despite the differences in the name, these surveys contain comparable information. Indeed, the FES was combined to the National Food Survey to form the EFS, which was renamed LCFS when it became a module of the Integrated Household Survey. The data from the FES has been previously used by Gosling, Machin, and Meghir (2000), Blundell, Reed, and Stoker (2003), Blundell, Gosling, Ichimura, and Meghir (2007) and Arellano and Bonhomme (2017) to study wage equations in the U.K. labor market. We are not aware of any previous use of the data from the EFS and LCFS for this purpose.[14] The three surveys contain repeated cross-sectional observations for women and men. The selection of the sample is similar to the previous work that used the FES. Thus, we keep individuals with ages between 23 to 59 years, and drop full-time students, self-employed workers, those married with spouse absent, and those with missing education or employees whose wages are missing. This leaves a sample of 258,900 observations, 139,504 of them correspond to women and 119,396 to men. The sample size per survey year and gender ranges from 2,197 to 4,545.

The outcome of interest, Y, is the logarithm of real hourly wage rate. We construct this variable as the ratio of the weekly usual gross main nominal earning to the weekly usual working hours, deflated by the U.K. quarterly retail price index. The selection variable, D, is an indicator for being employed.[15] The covariates, X, include 5 indicators for age when ceasing school ($\leqslant$15, 16, 17–18, 19–20, 21–22 and $\geqslant$ 23), a quartic polynomial in age, an indicator of being married or cohabiting, 6 variables with the number of kids by age categories (1, 2, 3–4, 5–10, 11–16, and 17-18), 36 survey year indicators, and 11 region indicators (Northern 5.48%, Yorkshire 9.56%, North Western 10.20%, East Midlands 7.36%, West Midlands 9.13%, East Anglia 5.31%, Greater London 10.06%, South Eastern 16.82%, South Western

[14] See Roantree and Vira (2018) for another recent application of the data to the analysis of female labor force participation.

[15] For data before 1990, $D = 0$ if the individual is in one of the following status: seeking work, sick but seeking work, sick but not seeking work, retired and unoccupied. For those in and after 1990, $D = 0$ if the individual is seeking work and available, waiting to start work, sick or injured, retired or unoccupied.

7.94%, Wales 4.99%, Scotland 8.92%, and Northern Ireland 4.23%).[16] We provide descriptive statistics of the variables and some background on the U.K. labor market using our data in Section G of the SM.

The excluded covariate, Z_1, is a potential out-of-work income benefit interacted with the marital status indicator used before in Blundell, Reed, and Stoker (2003) and Blundell, Gosling, Ichimura, and Meghir (2007). This benefit is constructed with the Institute for Fiscal Studies (IFS) tax and welfare-benefit model (TAXBEN). TAXBEN is a static tax and benefit micro-simulation model of taxes on personal incomes, local taxes, expenditure taxes, and entitlement to benefits and tax credits that operates on large-scale, representative household surveys (Brewer, 2009). It is designed to calculate the income of a tax unit if the individual was considered out-of-work.[17] It is composed of eligible unemployment and housing benefits, which are determined by the demographic composition of the tax unit and the housing costs that the tax unit faces. These costs vary by region and over time due to numerous policy changes that have occurred over time. There is no consensus in the literature about the validity of this variable as an excluded covariate. In our case, the outcome and selection sorting exclusions imply that, conditional on the observed covariates, the offered wage and dependence between offered wage and net reservation wage do not depend on the level of the benefit. We shall assume that the exclusion restrictions are satisfied and refer to Blundell, Reed, and Stoker (2003) and Blundell, Gosling, Ichimura, and Meghir (2007) for a discussion on the plausibility of the outcome restriction. In the Introduction, we stated a rich semi-parametric generalization of Heckman's labor supply model that trivially satisfies the selection sorting exclusion, motivating their use in our analysis.

4.2. **Empirical Specifications.** We estimate the DR model for different samples and carry out several wage decompositions where we compare the distributions of men and women, or the distributions over time within genders. The specifications of the selection and outcome equations include all the covariates described above except for the excluded covariates in the outcome equation. The parameter of the selection sorting function is notoriously more difficult to estimate than the parameters of the selection and outcome equations. We consider four simplified specifications of the sorting function where the covariates included in the index $X'\delta(y)$ are:

- Specification 1: a constant.
- Specification 2: a constant and the marital status indicator.
- Specification 3: a constant and a linear trend on the year of the survey.
- Specification 4: a constant and a linear trend on the year of the survey interacted with the marital status indicator.

[16] In the rest of the paper we shall refer to an individual being married or cohabiting as married.

[17] Our definition of the out-of-work benefit income is slightly different from the definition of Blundell, Reed, and Stoker (2003) and Blundell, Gosling, Ichimura, and Meghir (2007). They calculated it as the income of a tax unit if all the individuals within the tax unit were out of work. In our view our definition might better reflect the opportunity cost or outside value option of working that the individual faces.

We also experimented with other specifications that include the education indicators, indicators of survey year, or age. We do not report these results because they do not show any clear pattern mainly due to imprecision in the estimation of the parameter $\delta(y)$.[18]

4.3. **Selection Sorting.** We report point estimates and 95% confidence bands for the local correlation function $y \mapsto \rho(x'\delta(y))$ in the selection sorting equation. Estimates and 95% confidence bands for the coefficients of the selection, outcome and selection sorting equations are given in Section G of the SM. The estimates are obtained with Algorithm 3.1 replacing $\mathscr{Y}$ by a finite grid containing the sample quantiles of log real hourly wage with indexes $\{0.10, 0.11, \ldots, 0.90\}$ in the pooled sample of men and women. We report all the estimates as a function of the quantile index. The confidence bands are constructed by Algorithm 3.2 with $B = 500$ bootstrap repetitions and the same finite grid as for the estimates. We also report estimates from the HSM of Example 3.1 with dashed lines as a benchmark of comparison.

Figures 3–5 display the estimates of the sorting functions for specifications 1–3, respectively. Figure 3 shows positive selection sorting for men and negative selection sorting for women. In both cases we cannot reject that the sorting is constant across the distribution. This finding is refined in Figure 4, where we uncover that the positive male sorting comes mainly from bachelors, whereas the negative female sorting comes from married women. This pattern is consistent with a marriage market where there is assortative matching in offered wages given observable characteristics, where women with high potential wages are married to highly paid working men and decide not to work (Neal, 2004). Figure 5 shows that the sorting homogeneity found in the pooled sample hides some heterogeneity across time. Thus, we find that the male sorting is heterogeneous in the early years, negative at the bottom and positive at the top of the distribution, and progressively becomes homogenous. The female sorting is more homogenous over time, but also displays a positive trend, especially at the bottom of the distribution. Figure 16 in the SM shows that the trends in sorting are driven by married individuals at the bottom of the distribution and single individuals at the top of the distribution.[19]

4.4. **Distributions of Offered and Observed Wages, and Wage Decompositions.** Figure 6 shows point estimates of the quantiles of offered and observed wages for men and women based on specification 4. Estimates for the other specifications and confidence bands for all the specifications are given in the SM. The offered wage is a latent variable defined for all the individuals that is free of sample selection. As we showed in Section 3, the distributions of both types of wages can be expressed as functionals of the model parameters, and estimated using the plug-in estimators (3.12) and (3.13).[20] We find different sample selection biases for men and women. Thus, the quantiles of the observed wages are below

[18]The main results on the wage decompositions presented below are not sensitive to the specification of the sorting equation.

[19]We do not report confidence bands for specifications 3 and 4 to avoid cluttering. The confidence bands for the coefficients of the selection sorting function $\delta(y)$ in the SM show that the results on the trends are statistically significant.

[20]The model-based estimator of the observed distribution in (3.13) produces almost identical estimates to the empirical distribution of the observed wages.

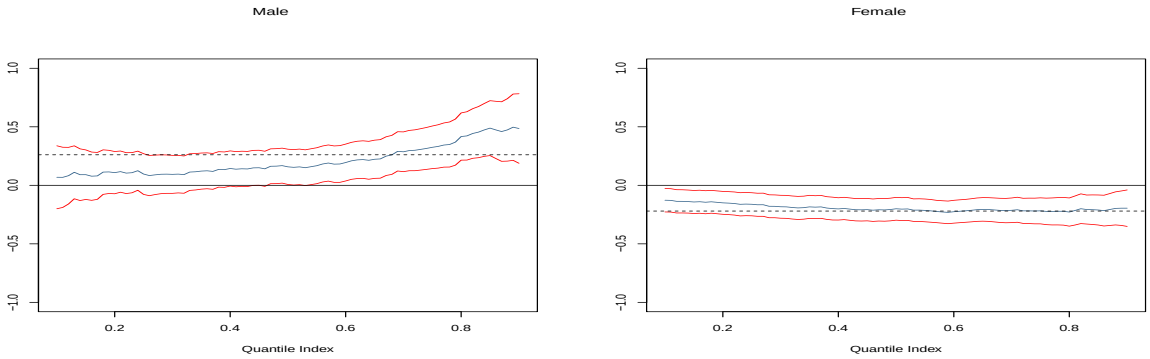

FIGURE 3. Estimates and 95% confidence bands for the selection sorting function: specification 1

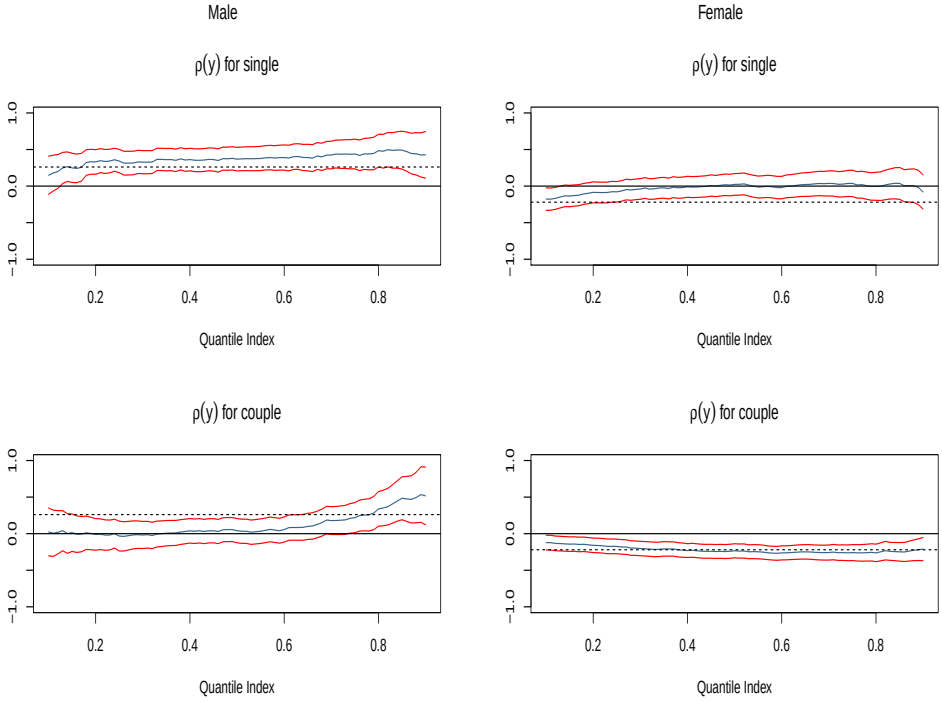

FIGURE 4. Estimates and 95% confidence bands for the selection sorting function: specification 2

the quantiles of latent wages for men, consistent with the positive selection in the sorting equation, whereas they are similar for women. In this case, the bias coming from the negative sorting is almost exactly offset by the difference in the composition between working and non-working women.

Figure 7 compares the quantile function of offered wages between men and women and carries out a gender wage gap analysis based on specification 4. The estimates and 95% confidence bands for the

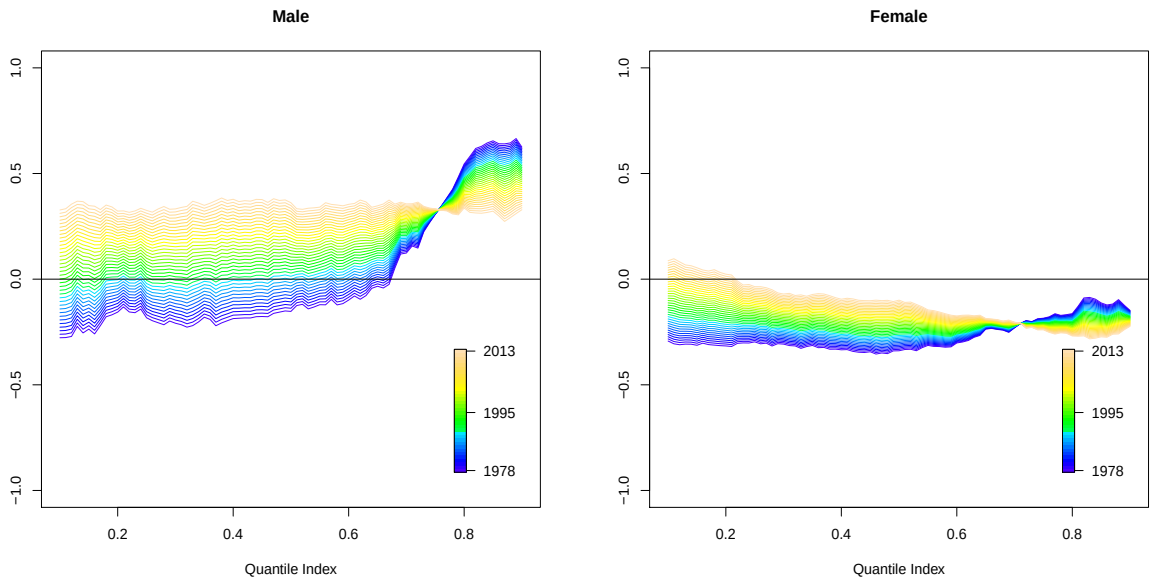

FIGURE 5. Estimates of the selection sorting function: specification 3

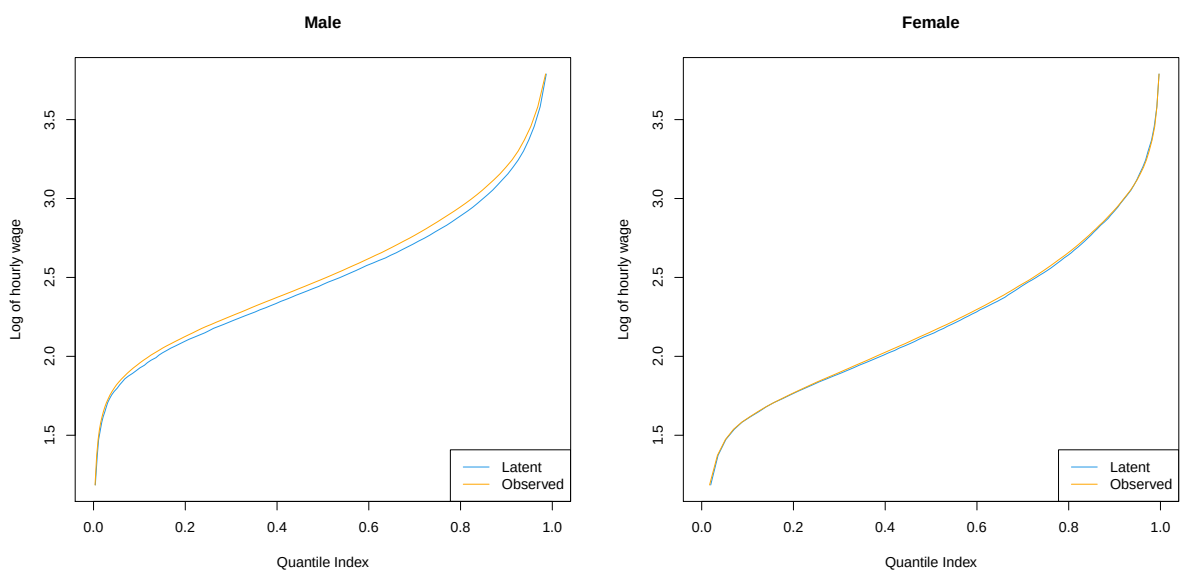

FIGURE 6. Estimates of the quantiles of observed and offered (latent) wages: specification 4

other specifications are reported in the SM. The gender wage gap analysis is based on the counterfactual distributions

$$F_{Y^*\langle j,k\rangle}(y) = F_{Y^*}(y;\beta_j, F_{X_k}) = \int \Phi(-x'\beta_j(y))dF_{X_k}(x),$$

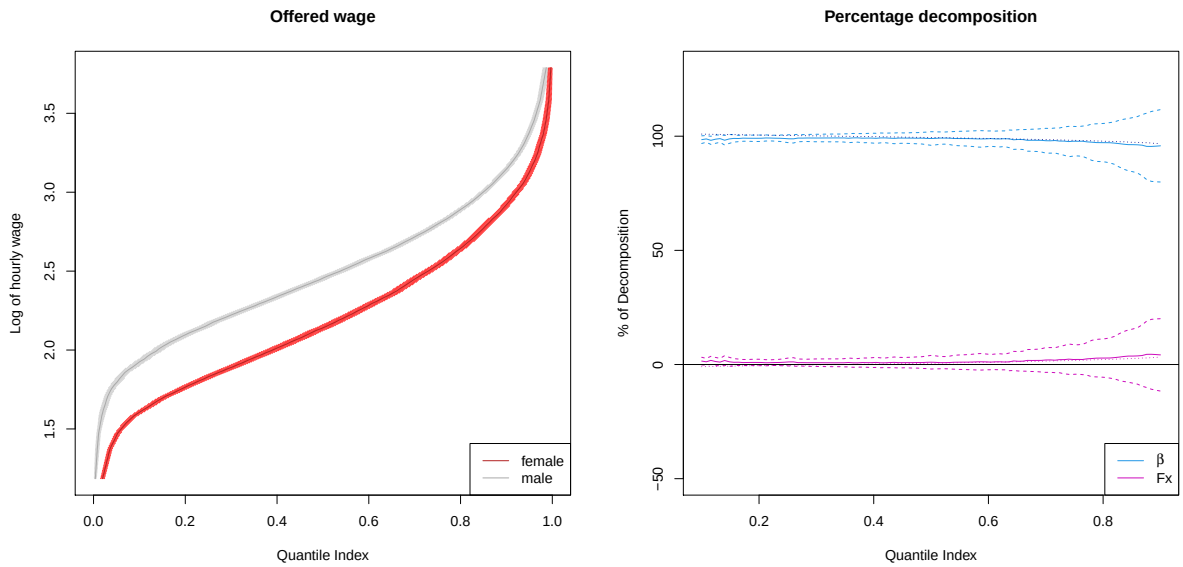

FIGURE 7. Estimates and 95% confidence bands for the quantiles of offered (latent) wages and decomposition between women and men: specification 4

where $\beta_j(y)$ is the coefficient of the wage equation in group j, F_{X_k} is the distribution of the characteristics in group k, and j and k are group indices for women and men. $F_{Y^*\langle j,k\rangle}$ corresponds to the distribution of offered wages that we would observed when the wage structure is as in group j and the distribution of characteristics is as in group k. We decompose the difference in the quantile functions of the latent wages between women (group 1) and men (group 0) using the counterfactual distributions as

$$F_{Y^*\langle 1,1\rangle} - F_{Y^*\langle 0,0\rangle} = [F_{Y^*\langle 1,1\rangle} - F_{Y^*\langle 0,1\rangle}] + [F_{Y^*\langle 0,1\rangle} - F_{Y^*\langle 0,0\rangle}],$$

where the first term is the wage structure or discrimination effect and the second term is the composition effect. We obtain estimates of the counterfactual distributions and quantile functions using the plug-in estimator in (3.12) and the quantile and increasing rearrangement operators, see Section D of the SM. We find that the wages offered to women are between 21 and 40% lower than the wages offered to men at the same quantile index. The majority of this difference is explained by differences in the wage structure, $\beta(y)$, whereas differences in composition, F_X, have very little explanatory power. This result can be interpreted as evidence of gender discrimination in the labor market.

We next use the DR model to decompose changes in the distribution of the observed wage between women and men, and between the first and second halves of the sample period for each gender. We extract four components that correspond to different inputs of the DR model: (1) selection (employment) sorting: $\delta(y)$; (2) selection (employment) structure: π; (3) outcome (wage) structure: $\beta(y)$; and (4) composition: F_Z. To define the effects of these components, let $F_{Y\langle t,s,r,k\rangle}$ be the counterfactual distribution of wages when the sorting is as in group t, the employment structure is as in group s, the wage

structure is as in group r, and the composition of the population is as in group k. The actual distribution in group t therefore corresponds to $F_{Y\langle t,t,t,t\rangle}$. We assume that there are two groups indexed by 0 and 1 that correspond to demographic populations such as men and women, or time periods such as the first and second halves of the sample years. Then, we can decompose the distribution of observed wage between group 1 and group 0 as:

$$F_{Y\langle 1,1,1,1\rangle} - F_{Y\langle 0,0,0,0\rangle} = [F_{Y\langle 1,1,1,1\rangle} - F_{Y\langle 0,1,1,1\rangle}] + [F_{Y\langle 0,1,1,1\rangle} - F_{Y\langle 0,0,1,1\rangle}] \\ + [F_{Y\langle 0,0,1,1\rangle} - F_{Y\langle 0,0,0,1\rangle}] + [F_{Y\langle 0,0,0,1\rangle} - F_{Y\langle 0,0,0,0\rangle}],$$

where the first term in square brackets of the right hand side is a sorting effect, the second an employment structure effect, the third a wage structure effect, and the forth a composition effect. This is a distributional version of the classical Oaxaca-Blinder decomposition that accounts for sample selection (Kitagawa, 1955; Oaxaca, 1973; Blinder, 1973). It is well-known that the order of extraction of the components in this type of decompositions might matter. As a robustness check, we estimate the decomposition changing the ordering of the components. In results not reported, we find that the main findings are not sensitive to the change of ordering.

In terms of the DR model, the counterfactual distribution can be expressed as the functional

$$F_{Y\langle t,s,r,k\rangle}(y) = F_Y(y;\beta_r,\pi_s,\delta_t,F_{Z_k}) = \frac{\int \Phi_2\left(-x'\beta_r(y), z'\pi_s; -\rho(x'\delta_t(y))\right) dF_{Z_k}(z)}{\int \Phi(z'\pi_s) dF_{Z_k}(z)},$$

where δ_t is the coefficient of the sorting function in group t, π_s is the coefficient of the employment equation in group s, β_r is the coefficient of the wage equation in group r, and F_{Z_k} is the distribution of characteristics in group k. Given random samples for groups 0 and 1, we construct a plug-in estimator of $F_{Y\langle t,s,r,k\rangle}$ by suitably combining the estimators of the model parameters and distribution of covariates from the two groups.

Remark 4.1 (Selection Effects)**.** To interpret the selection effects, it is useful to consider a simplified version of the model without covariates where $F_Y(y;\pi,\rho) = \Phi_2\left(-\beta,\pi;-\rho\right)/\Phi(\pi)$. Here we drop the dependence of β and ρ on y to lighten the notation, and make explicit the dependence of F_Y on the selection parameters π and ρ to carry out comparative statics with respect to them. Then, by the properties of the normal distribution

$$\frac{\partial F_Y(y;\pi,\rho)}{\partial \rho} = -\frac{\phi_2(-\beta,\pi;-\rho)}{\Phi(\pi)} < 0,$$

and

$$\frac{\partial F_Y(y;\pi,\rho)}{\partial \pi} \propto \Phi\left(\frac{-\beta+\rho\pi}{\sqrt{1-\rho^2}}\right)\Phi(\pi) - \int_{-\infty}^{\pi} \Phi\left(\frac{-\beta+\rho x}{\sqrt{1-\rho^2}}\right)\phi(x)dx \begin{cases} <0 \text{ if } \rho<0, \\ =0 \text{ if } \rho=0, \\ >0 \text{ if } \rho>0, \end{cases}$$

where Φ and ϕ are the standard normal CDF and PDF, and $\phi_2(\cdot,\cdot;\rho)$ be the joint PDF of a standard bivariate normal random variable with parameter ρ.[21] Increasing ρ, therefore, shifts the distribution to the right (increases quantiles) because it makes selection sorting more positive while the size of the

[21] To obtain the derivative we use that $\Phi_2\left(-\beta,\pi;-\rho\right) = \int_{-\infty}^{\pi} \Phi\left(\frac{-\beta+\rho x}{\sqrt{1-\rho^2}}\right)\phi(x)dx$.

selected population is fixed. The effect of increasing π is more nuanced and depends on the sign of ρ. Intuitively, π affects the size of the selected population and the relative importance of observables and unobservables in the selection. For example, when selection sorting is negative, increasing the size of the selected population by increasing π shifts the distribution of the right (increases quantiles) because the newly selected individuals have smaller (more negative) selection unobservables that correspond to larger (more positive) outcome unobservables. In other words, the newly selected individuals are relatively less adversely selected. The sign of the selection effects might be different in the presence of covariates if the parameter variation changes the composition of the selected population. We provide an example of this sign reversal in Appendix C of the SM. □

Figure 8 reports estimates of the quantile functions of observed wages for men and women, together with the relative contributions of each component to the decomposition between men (group 0) and women (group 1) based on specification 4. The bands for the contributions are joint for all the components and rely on the delta method; see Remark D.3 in the SM. Estimates of the components of the decomposition and the analysis based on specifications 1–3 are given in the SM. The distribution for men first order stochastically dominates the distribution for women. Most of this gender wage gap is explained by differences in the wage structure, i.e. differences in the returns to observed characteristics. However, differences in sorting and employment structure also account for an important percentage of the gap, especially at the top of the distribution. Thus, we uncover that the negative female sorting explains about 30–40% of the gap at the top of the distribution. A possible explanation is that women with very high potential wages decide not to work because there are no high-paid jobs available to them due to glass ceiling (Albrecht, Bjorklund, and Vroman, 2003). The negative contribution of the employment structure can be explained by the order of the decomposition where we are applying the male employment structure to the female distribution with positive male sorting.[22] In this case we are increasing the proportion of employed women, where the added women come from a pool with lower positive selection, and this negative effect is not reversed by a change in the composition of the working women; see Remark 4.1 for more details. The aggregate selection effect, defined as the sum of the selection sorting and selection structure effects, is positive and statistically significant at the top of the distribution; see Figure 23 in the SM. Differences in the composition of the characteristics contribute very little to explain the gender gap. Finally, the estimates from the HSM in dotted lines pick up the average contributions of the components, but miss all the heterogeneity across the distribution.

Figures 9 and 10 report estimates of the quantile functions of observed wages for the first and second halves of the sample period, together with the relative contributions of each component to the decomposition between second half (group 0) and first half (group 1) based on specification 2 for women and men, respectively. Estimates of the components of the decompositions are given in the SM. The distribution for the second half first order stochastically dominates the distribution for the first half in both cases. For women, the most important components are the wage structure and composition effects

[22]While the sign of the employment structure contribution changes with the order of the decomposition, neither its importance nor the significance of the contributions of the other components are sensitive to this order.

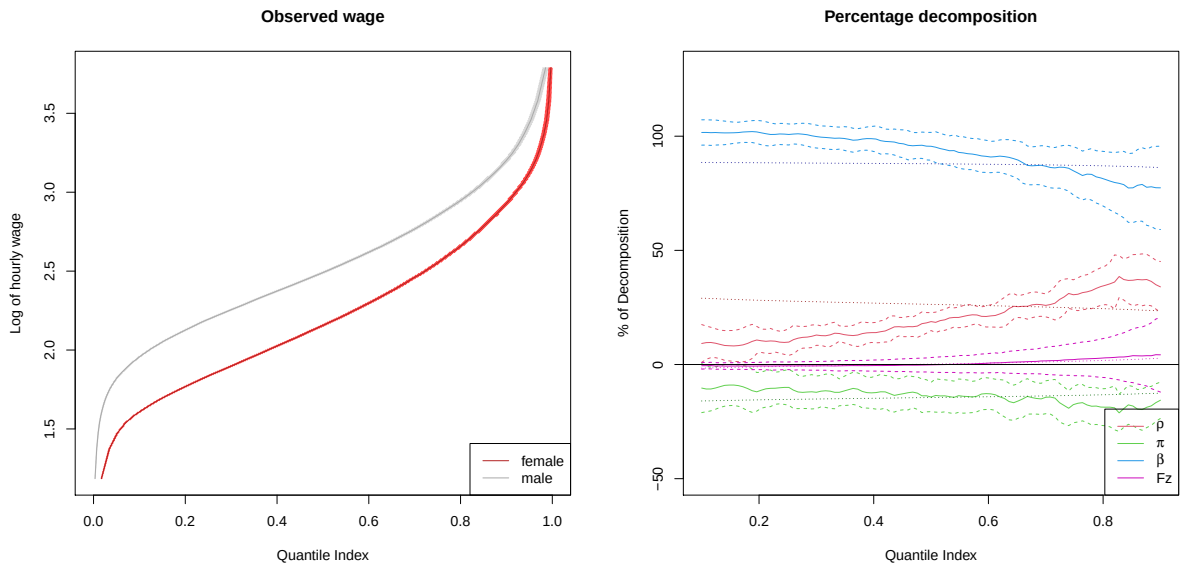

FIGURE 8. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women in specification 4

in this order. The importance of the wage structure is decreasing along the distribution, whereas the importance of the composition is increasing. Composition and wage structure are also the most important components for men. The small contributions of the selection sorting component to the change in the distribution of wages between the two time period for both genders seem to contradict the linear time trends that we found in the coefficient of the sorting selection function. This might be explained by the inability of a coarse partition of the sample into two halves to capture the gradual increase in selection sorting, together with the changes in the composition.

4.5. **Discussion.** The main findings can be summarized as: (1) positive sorting for men and negative sorting for women driven by single men and married women, which is consistent with assortative matching in the marriage market; (2) heterogeneity in selection sorting decreases gradually over time; (3) differences in returns to characteristics in the wage equation, which might be associated with gender discrimination in the labor market, account for most of the gender wage gap; (4) selection sorting on unobservables explains up to 39% of the gender wage gap at the top of the distribution, which can be taken as evidence of glass ceiling; and (5) changes in the structure of the wage equation and composition of the characteristics account for most of the differences in the wage distribution between the two halves of the sample period within each gender.

We compare and contrast these findings with previous results from the literature that studied similar issues. These results were obtained from different data and/or using different methodology. Blundell, Gosling, Ichimura, and Meghir (2007) applied a bound approach that does not require of exclusion restrictions to study the evolution of wage inequality using the FES data for the period 1978–2000. They

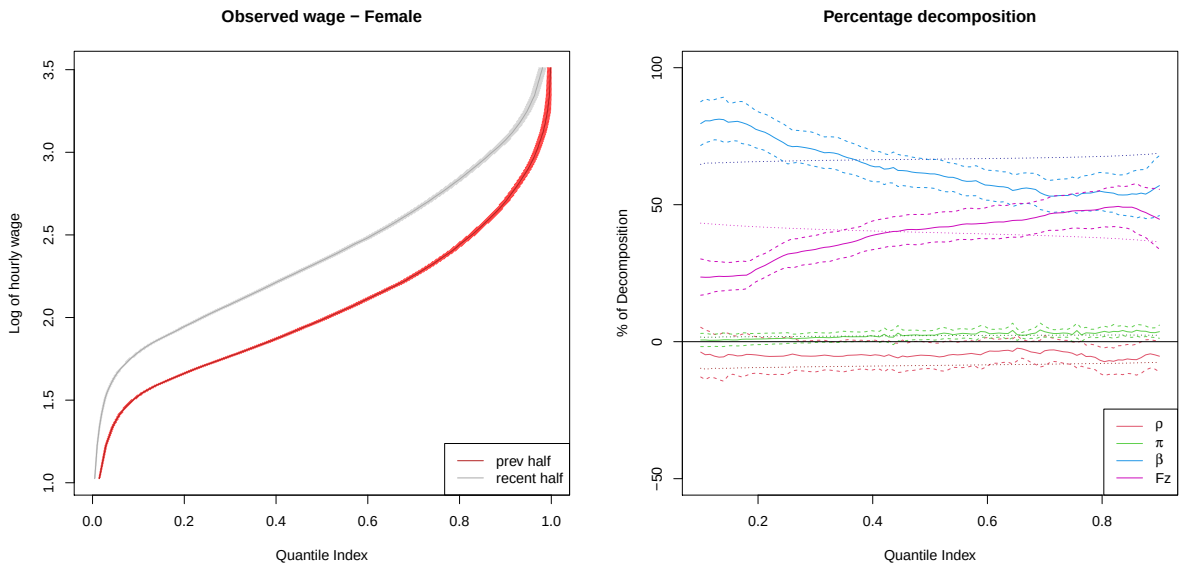

FIGURE 9. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between first and second half of the sample period for women in specification 2

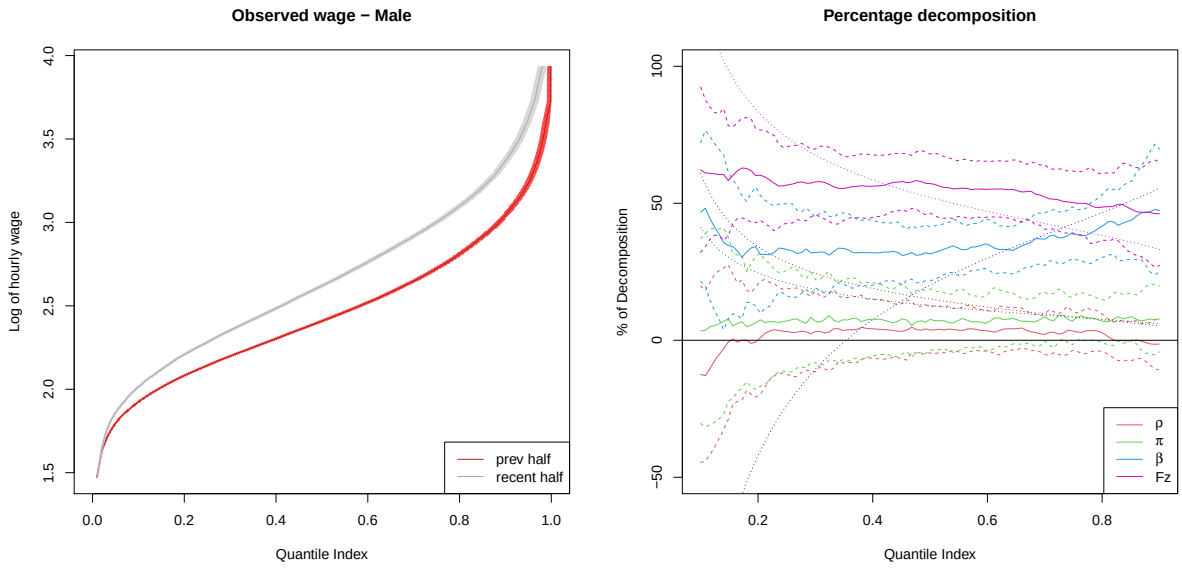

FIGURE 10. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between first and second half of the sample period for men in specification 2

assumed positive sorting for men and women in some of their estimates to make the bounds more informative. Interestingly, they mentioned the possibility that the assumption is violated for married women due to assortative matching in the marriage market.[23] They also found evidence against the validity of out-of-work benefit income as a valid excluded covariate for men. AB17 using the same data from the FES, also found positive sorting for men, stronger for single than for married men, but they employed an alternative methodology that combines quantile regression for the marginal distributions with a parametric model for the copula. Contrary to our findings, they also found positive selection for women, which is statistically significant only for married women. Mulligan and Rubinstein (2008) estimated a HSM using data from the US-CPS for the periods 1975-1979 and 1995-1999. They found that the selection sorting for women shifted from negative to positive between the two periods. We also find for the UK that the sorting for most women has a positive trend over time, but remains negative even in 2013 for most of the distribution. Maasoumi and Wang (2019) applied the methodology of AB17 to data from the US-CPS for the period 1976–2014. They also found negative sorting for women at the beginning of the sample period that became positive during the 90s, and positive sorting for men throughout the entire period. Bertrand (2017) pointed out multiple possible explanations for the glass ceiling based on the field of education, psychological attributes or preferences for job flexibility that are compatible with our finding on the importance of sorting on unobservables at the top of the distribution. None of the previous papers distinguished between the selection sorting and selection structure effects.

One limitation of our dataset is that it does not contain a direct measure of work experience. As a final robustness check, we find that the results are not sensitive to the exclusion of college graduates from the sample by redoing the analysis excluding all the individuals who cease school after age 18. This is a relevant exclusion because work experience is a more relevant determinant of wage for highly educated workers.[24]

5. Conclusion

We develop a distribution regression model with sample selection that accommodates rich patterns of heterogeneity in the effects of covariates on outcomes and selection. The model is semi-parametric in nature, as it has function-valued parameters, and is able to considerably generalize the classical selection model of Heckman (1974). Furthermore, the model accounts for richer covariate effects than the previous semi-parametric generalizations which allowed only the location effects for covariates. We propose to estimate the model by a process of bivariate probit regressions, indexed by threshold-dependent parameters. We show that the resulting estimators of the function-valued parameters are approximately Gaussian and concentrate in a $1/\sqrt{n}$ neigborhood of the true values. We present an

[23] In results not reported, we find that the negative sorting for married women is robust to the definition of the out-of-work benefit income variable. Thus, we find similar estimates using the income of a tax unit if all the individuals within the tax unit were out of work as the excluded covariate, as in Blundell, Gosling, Ichimura, and Meghir (2007).

[24] These results are available from the authors upon request.

extensive wage decomposition analysis for the U.K. using new data, generating both new findings and demonstrating the power of the method. Our identification approach is especially designed for the sample selection problem but can be applied to other settings. In work in progress, we show how the selection exclusion can be used to identify causal effects in treatment effects models with endogeneity in the presence of an instrumental variable that can be binary.

REFERENCES

AHN, H., AND J. L. POWELL (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58(1-2), 3–29.

ALBRECHT, J., A. BJORKLUND, AND S. VROMAN (2003): "Is There a Glass Ceiling in Sweden?," *Journal of Labor Economics*, 21(1), 145–177.

AMEMIYA, T. (1985): *Advanced Econometrics*. Harvard University Press.

ANDREWS, D. W. K., AND M. M. A. SCHAFGANS (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *The Review of Economic Studies*, 65(3), 497–517.

ARELLANO, M., AND S. BONHOMME (2017): "Quantile Selection Models With an Application to Understanding Changes in Wage Inequality," *Econometrica*, 85(1), 1–28.

BALKE, A., AND J. PEARL (1994): "Counterfactual probabilities: Computational methods, bounds and applications," in *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier.

BERTRAND, M. (2017): "The Glass Ceiling," Discussion paper, Becker Friedman Institute for Research in Economics, Working Paper No. 2018-38.

BLINDER, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *The Journal of Human Resources*, 8(4), 436–455.

BLUNDELL, R., A. GOSLING, H. ICHIMURA, AND C. MEGHIR (2007): "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds," *Econometrica*, 75(2), 323–363.

BLUNDELL, R., H. REED, AND T. M. STOKER (2003): "Interpreting Aggregate Wage Growth: The Role of Labor Market Participation," *American Economic Review*, 93(4), 1114–1131.

BREWER, M. (2009): "TAXBEN: the IFS' static tax and benefit microsimulation model," Discussion paper, Institute for Fiscal Studies, Presentation given at the BSPS Conference, University of Sussex, September 2009.

CHAMBERLAIN, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), 189–218.

CHEN, X. (2007): "Large Sample Sieve Estimatin of Semi-Nonparametric Models," *Handbook of Econometrics*, 6, 5559–5632.

CHERNOZHUKOV, V., D. CHETVERIKOV, AND K. KATO (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *Ann. Statist.*, 41(6), 2786–2819.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81(6), 2205–2268.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, B. MELLY, AND K. WÜTHRICH (2016): "Generic Inference on Quantile and Quantile Effect Functions for Discrete Outcomes," *ArXiv e-prints*.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND A. GALICHON (2010): "Quantile and probability curves without crossing," *Econometrica*, 78(3), 1093–1125.

CHERNOZHUKOV, V., AND C. HANSEN (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132(2), 491–525.

CHRISTENSEN, T., AND B. CONNAULT (2019): "Counterfactual sensitivity and robustness," *Econometrica (to appear); arXiv preprint arXiv:1904.00989*.

DAS, M., W. K. NEWEY, AND F. VELLA (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70(1), 33–58.

DETTE, H., M. HALLIN, T. KLEY, AND S. VOLGUSHEV (2015): "Of copulas, quantiles, ranks and spectra: An L_1-approach to spectral analysis," *Bernoulli*, 21(2), 781–831.

FISHER, R. A. (1915): "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, 10(4), 507–521.

FORESI, S., AND F. PERACCHI (1995): "The Conditional Distribution of Excess Returns: An Empirical Analysis," *Journal of the American Statistical Association*, 90(430), 451–466.

GALE, D., AND H. NIKAIDO (1965): "The Jacobian Matrix and Global Univalence of Mappings.," *Mathematische Annalen*, 159, 81–93.

GINÉ, E., AND J. ZINN (1984): "Some limit theorems for empirical processes," *Ann. Probab.*, 12(4), 929–998, With discussion.

GOLDBERGER, A. S. (1983): "Abnormal selection bias," in *Studies in econometrics, time series, and multivariate statistics*, pp. 67–84. Elsevier.

GOSLING, A., S. MACHIN, AND C. MEGHIR (2000): "The Changing Distribution of Male Wages in the U.K.," *The Review of Economic Studies*, 67(4), 635–666.

GRONAU, R. (1974): "Wage Comparisons-A Selectivity Bias," *Journal of Political Economy*, 82(6), 1119–43.

HANSEN, B. E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica: Journal of the econometric society*, pp. 413–430.

HECKMAN, J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42(4), 679–94.

——— (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," in *Annals of Economic and Social Measurement, Volume 5, number 4*, pp. 475–492. National Bureau of Economic Research, Inc.

——— (1990): "Varieties of Selection Bias," *American Economic Review*, 80(2), 313–18.

HECKMAN, J., AND B. SINGER (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2), 271–320.

HECKMAN, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

HECKMAN, J. J., AND E. J. VYTLACIL (2001): "Instrumental variables, selection models, and tight bounds on the average treatment effect," in *Econometric Evaluation of Labour Market Policies*, pp. 1–15. Springer.

HÖFFDING, W. (1940): "Masstabinvariante korrelationstheorie," *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin*, 5, 181–233.

ICHIMURA, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of econometrics*, 58(1-2), 71–120.

KITAGAWA, E. M. (1955): "Components of a Difference Between Two Rates," *Journal of the American Statistical Association*, 50(272), 1168–1194.

KITAGAWA, T. (2010): "Testing for Instrument Independence in the Selection Model," .

KLEIN, R. W., AND R. H. SPADY (1993): "An efficient semiparametric estimator for binary response models," *Econometrica: Journal of the Econometric Society*, pp. 387–421.

KLINE, P., AND A. SANTOS (2012): "A Score Based Approach to Wild Bootstrap Inference," *Journal of Econometric Methods*, 1.1, 23–41.

KOLEV, N., U. D. ANJOS, AND B. V. D. M. MENDES (2006): "Copulas: A Review and Recent Developments," *Stochastic Models*, 22(4), 617–660.

KRANTZ, S. G., AND H. R. PARKS (2002): *A primer of real analytic functions*. Springer Science & Business Media.

LEE, L.-F. (1982): "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 49(3), 355–372.

LEE, L.-F. (1983): "Generalized econometric models with selectivity," *Econometrica*, pp. 507–512.

LEHMANN, E. L. (1966): "Some Concepts of Dependence," *Ann. Math. Statist.*, 37(5), 1137–1153.

LEWBEL, A. (1995): "Consistent nonparametric hypothesis tests with an application to Slutsky symmetry," *Journal of Econometrics*, 67(2), 379–401.

——— (2007): "Endogenous selection or treatment model estimation," *Journal of Econometrics*, 141(2), 777–806.

LIU, H., J. LAFFERTY, AND L. WASSERMAN (2009): "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs.," *Journal of Machine Learning Research*, 10(10).

MAASOUMI, E., AND L. WANG (2019): "The gender gap between earnings distributions," *Journal of Political Economy*, 127(5), 2438–2504.

MADDALA, G. S. (1986): *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

MANSKI, C. (1989): "Anatomy of the Selection Problem," *Journal of Human Resources*, 24(3), 343–360.

MANSKI, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80(2), 319–323.

MANSKI, C. F. (1994): "The selection problem," in *Advances in econometrics : sixth World Congress*, ed. by C. A. Sims, chap. 4, pp. 143–170. Cambridge University Press Cambridge [England] ; New York, NY.

——— (2003): *Partial identification of probability distributions*. Springer Science & Business Media.

MARCHENKO, Y. V., AND M. G. GENTON (2012): "A Heckman Selection-t Model," *Journal of the American Statistical Association*, 107(497), 304–317.

MAS-COLELL, A. (1979): "Homeomorphisms of compact, convex sets and the Jacobian matrix," *SIAM Journal on Mathematical Analysis*, 10(6), 1105–1109.

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.

MUNKRES, J. R. (1991): *Analysis on manifolds*. Addison-Wesley Publishing Company, Advanced Book Program, Redwood City, CA.

NEAL, D. (2004): "The Measured Black-White Wage Gap among Women Is Too Small," *Journal of Political Economy*, 112(S1), 1–28.

NEWEY, W., AND D. MCFADDEN (1986): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, vol. 4, chap. 36, pp. 2111–2245. Elsevier, 1 edn.

NEWEY, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

——— (1999): "Consistency of two-step sample selection estimators despite misspecification of distribution," *Economics Letters*, 63(2), 129–132.

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

PEARSON, K. (1900): "Mathematical Contributions to the Theory of Evolution. VII. On the Correlation of Characters not Quantitatively Measurable," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 195, 1–405.

POIRIER, D. J. (1980): "Partial observability in bivariate probit models," *Journal of Econometrics*, 12(2), 209–217.

POWELL, J. L. (1994): "Estimation of semiparametric models," *Handbook of econometrics*, 4, 2443–2521.

PRIEGER, J. E. (2002): "A flexible parametric selection model for non-normal data with application to health care usage," *Journal of Applied Econometrics*, 17(4), 367–392.

ROANTREE, B., AND K. VIRA (2018): "The rise and rise of women's employment in the UK," Discussion paper, Institute for Fiscal Studies, IFS Briefing Note BN234.

ROTHENBERG, T. J. (1971): "Identification in parametric models," *Econometrica: Journal of the Econometric Society*, pp. 577–591.

SIBUYA, M. (1959): "Bivariate extreme statistics, I," *Annals of the Institute of Statistical Mathematics*, 11(2), 195–210.

SMITH, M. D. (2003): "Modelling sample selection using Archimedean copulas," *The Econometrics Journal*, 6(1), 99–123.

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

SUNGUR, E. A. (1990): "Dependence Information in Parameterized Copulas," *Communications in Statistics - Simulation and Computation*, 19(4), 1339–1360.

TJØSTHEIM, D., H. OTNEIM, AND B. STØVE (2018): "Statistical dependence: Beyond Pearson's ρ," *ArXiv e-prints*.

TORGOVITSKY, A. (2010): "Identification and estimation of nonparametric quantile regressions with endogeneity," Discussion paper, Yale University.

TREFETHEN, L. N. (2020): "Quantifying the ill-conditioning of analytic continuation," *BIT Numerical Mathematics*, 60(4), 901–915.

VAN DE VEN, W. P., AND B. M. VAN PRAAG (1981): "The demand for deductibles in private health insurance: A probit model with sample selection," *Journal of Econometrics*, 17(2), 229 – 252.

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.

VELLA, F. (1998): "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources*, 33(1), 127–169.

VYTLACIL, E. J. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.

ZELLNER, A., AND T. H. LEE (1965): "Joint Estimation of Relationships Involving Discrete Random Variables," *Econometrica*, 33(2), 382–394.

APPENDIX A. PROOFS OF RESULTS IN SECTION 2

A.1. **Proof of Lemma 2.1.** Let $\phi_2(\cdot,\cdot;\rho)$ be the joint probability density function (PDF) of a standard bivariate normal random variable with parameter ρ. The representation (2.6) exists and is unique by the following properties of the standard bivariate normal distribution:

(1) $\rho \mapsto \Phi_2(\cdot,\cdot;\rho)$ is continuously differentiable and $\partial\Phi_2(\cdot,\cdot;\rho)/\partial\rho = \phi_2(\cdot,\cdot;\rho) > 0$ (Sibuya, 1959; Sungur, 1990);

(2) $\lim_{\rho\nearrow 1}\Phi_2(x,y;\rho) = \min[\Phi(x),\Phi(y)]$;

(3) $\lim_{\rho\searrow -1}\Phi_2(x,y;\rho) = \max[\Phi(x)+\Phi(y)-1,0]$;

together with the Frechet-Hoeffding bounds

$$\max[\Phi(\mu(y))+\Phi(\nu(d))-1,0] \leqslant F_{Y^*,D^*}(y,d) \leqslant \min[\Phi(\mu(y)),\Phi(\nu(d))].$$

□

A.2. **Full Statment and Proof of Theorem 2.1.** Let

$$F_{Y,D|Z}(y,1\mid z) := \mathrm{P}(D=0\mid Z=z)+\mathrm{P}(Y\leqslant y, D=1\mid Z=z).$$

The full statement of the theorem is:

(1) If $F_{Y,D|Z}(y,1\mid z)=1$, $z\in\{0,1\}$, then,

$$\rho(y)=1,\ \ \mu(y)\in[\Phi^{-1}(\mathrm{P}(D=1\mid Z=1)),+\infty).$$

(2) If $F_{Y,D|Z}(y,1\mid z)=\mathrm{P}(D=0\mid Z=z)$, $z\in\{0,1\}$, then,

$$\rho(y)=-1,\ \mu(y)\in(-\infty,\Phi^{-1}(\mathrm{P}(D=0\mid Z=1))].$$

(3) If $F_{Y,D|Z}(y,1\mid 1)-\mathrm{P}(D=0\mid Z=1)=F_{Y,D|Z}(y,1\mid 0)-\mathrm{P}(D=0\mid Z=0)>0$, then,

$$\rho(y)=1,\ \ \mu(y)=\Phi^{-1}\left(F_{Y,D|Z}(y,1\mid 1)-\mathrm{P}(D=0\mid Z=1)\right).$$

(4) If $F_{Y,D|Z}(y,1\mid 1)<1$ and $F_{Y,D|Z}(y,1\mid 0)=1$, then,

$$\rho(y)=1,\ \ \mu(y)=\Phi^{-1}\left(F_{Y,D|Z}(y,1\mid 1)-\mathrm{P}(D=0\mid Z=1)\right).$$

(5) If $F_{Y,D|Z}(y,1\mid 1)>\mathrm{P}(D=0\mid Z=1)$ and $F_{Y,D|Z}(y,1\mid 0)=\mathrm{P}(D=0\mid Z=0)$, then,

$$\rho(y)=-1,\ \mu(y)=\Phi^{-1}\left(F_{Y,D|Z}(y,1\mid 1)\right).$$

(6) If $F_{Y,D|Z}(y,1\mid 1)=F_{Y,D|Z}(y,1\mid 0)<1$, then,

$$\rho(y)=-1,\ \mu(y)=\Phi^{-1}\left(F_{Y,D|Z}(y,1\mid 1)\right).$$

(7) Otherwise, $\mu(y)$ and $\rho(y)$ are point identified as the solution in (μ,ρ) to (2.9). This solution exists and is unique.

Proof. The identification of $\nu(z)$ follows from equalizing the marginals of $F_{Y^*,D^*|Z}$ with respect to D^* and the conditional LGR at $D^*=0$. Since $\nu(z)$ is identified, we shall use

$$\Phi(\nu(z))=\mathrm{P}(D=1\mid Z=z)\text{ and }\bar{\Phi}(\nu(z))=\mathrm{P}(D=0\mid Z=z),z\in\{0,1\},$$

to lighten the notation. We also recall the Frechet-Hoeffding bounds in our setting, as they will be extensively used in the proof:

$$\max[\Phi(\mu(y))-\bar{\Phi}(\nu(z)),0]\leqslant\Phi_2(\mu(y),\nu(z);\rho(y))\leqslant\min[\Phi(\mu(y)),\Phi(\nu(z))],z\in\{0,1\},\tag{FHB}$$

where the upper bound is only attained at $\rho(y)=1$ and the lower bound at $\rho(y)=-1$.

Cases (1)–(2). We consider first the partially identified cases. In case (1),

$$\Phi_2(\mu(y),\nu(z);\rho(y))=F_{Y,D|Z}(y,1\mid z)-\bar{\Phi}(\nu(z))=\Phi(\nu(z)),z\in\{0,1\},$$

implies that $\rho(y)=1$ by the upper FHB. The identified set for $\mu(y)$ is obtained from

$$\Phi(\mu(y))\geqslant\max[\Phi(\nu(0)),\Phi(\nu(1))]=\Phi(\nu(1))$$

by the upper FHB and Assumption 1(2). In case (2),

$$\Phi_2(\mu(y),\nu(z);\rho(y))=F_{Y,D|Z}(y,1\mid z)-\bar{\Phi}(\nu(z))=0,z\in\{0,1\},$$

implies that $\rho(y)=-1$ by the lower FHB. The identified set for $\mu(y)$ is obtained from

$$\Phi(\mu(y))\leqslant\min[\bar{\Phi}(\nu(0)),\bar{\Phi}(\nu(1))]=\bar{\Phi}(\nu(1))$$

by the lower FHB and Assumption 1(2).

In the rest of the proof we can assume that the conditions that define cases (1)–(2) do not hold.

Cases (3)–(4). These boundary cases correspond to $\rho(y) = 1$. They are identifiable because $\rho(y) = 1$ if and only if $F_{Y,D|Z}(y, 1 \mid 1) - \bar{\Phi}(v(1)) = F_{Y,D|Z}(y, 1 \mid 0) - \bar{\Phi}(v(0))$ or $F_{Y,D|Z}(y, 1 \mid 0) = 1$. The only if part follows from the upper FHB and Assumption 1(2). Indeed, there are 2 cases depending on the values of $\Phi(v(0))$ and $\Phi(\mu(y))$. If $\Phi(\mu(y)) < \Phi(v(0))$, then $\min[\Phi(v(z)), \Phi(\mu(y))] = \Phi(\mu(y))$ for $z \in \{0, 1\}$ so that $F_{Y,D|Z}(y, 1 \mid 1) - \bar{\Phi}(v(1)) = F_{Y,D|Z}(y, 1 \mid 0) - \bar{\Phi}(v(0))$. If $\Phi(v(0)) \leqslant \Phi(\mu(y))$, then $\Phi_2(\mu(y), v(0); 1) = \Phi(v(0))$ so that $F_{Y,D|Z}(y, 1 \mid 0) = 1$. For the case $F_{Y,D|Z}(y, 1 \mid 1) - \bar{\Phi}(v(1)) = F_{Y,D|Z}(y, 1 \mid 0) - \bar{\Phi}(v(0))$, the if part follows because $v \mapsto \Phi_2(\cdot, v; \rho)$ is strictly monotonic when $\rho \in (-1, 1)$ and $v(1) > v(0)$ by Assumption 1(2) so that $\Phi_2(\mu(y), v(0); \rho(y)) \neq \Phi_2(\mu(y), v(1); \rho(y))$. This shows that $\rho(y) \notin (-1, 1)$. Moreover, this case is ruled out when $\rho(y) = -1$ by the lower FHB, $F_{Y,D|Z}(y, 1 \mid 1) > \bar{\Phi}(v(1))$, and Assumption 1(2).[25] The case $F_{Y,D|Z}(y, 1 \mid 0) = 1$ implies that $\Phi_2(\mu(y), v(0); \rho(y)) = \Phi(v(0))$, which is only possible when $\rho(y) = 1$ by the upper FHB.

Now, we can analyze the identification of $\mu(y)$ using the upper FHB. Case (3) corresponds to $F_{Y,D|Z}(y, 1 \mid z) - \bar{\Phi}(v(z)) = \Phi(\mu(y))$, $z \in \{0, 1\}$, which identify $\mu(y)$. Case (4) corresponds to $F_{Y,D|Z}(y, 1 \mid 0) = 1$ and $F_{Y,D|Z}(y, 1 \mid 1) - \bar{\Phi}(v(1)) = \Phi(\mu(y))$. The second equation identifies $\mu(y)$.

Cases (5)–(6). These boundary cases correspond to $\rho(y) = -1$. They are identifiable because $\rho(y) = -1$ if and only if $F_{Y,D|Z}(y, 1 \mid 0) = F_{Y,D|Z}(y, 1 \mid 1)$ or $F_{Y,D|Z}(y, 1 \mid 0) = \bar{\Phi}(v(0))$. Symmetrically to $\rho(y) = 1$, the only if part follows from the lower FHB and Assumption 1(2), whereas the if part for the case $F_{Y,D|Z}(y, 1 \mid 0) = F_{Y,D|Z}(y, 1 \mid 1)$ follows from the upper FHB, strict monotonic of $v \mapsto \Phi_2(\cdot, v; \rho)$ when $\rho \in (-1, 1)$, and Assumption 1(2).[26] The if part for $F_{Y,D|Z}(y, 1 \mid 0) = \bar{\Phi}(v(0))$ follows because this case implies that $\Phi_2(\mu(y), v(0); \rho(y)) = 0$, which is only possible when $\rho(y) = -1$ by the lower FHB.

Now, we can analyze the identification of $\mu(y)$ using the lower FHB. Case (5) corresponds to $F_{Y,D|Z}(y, 1 \mid 1) = \Phi(\mu(y))$, which identifies $\mu(y)$. Case (6) corresponds to $F_{Y,D|z}(y, 1 \mid z) = \Phi(\mu(y))$, $z \in \{0, 1\}$. Both of these equations have the same solution that identifies $\mu(y)$.

Case (7). Finally, consider now the non-boundary case where $\rho(y) \in (-1, 1)$. The parameters $\mu(y)$ and $\rho(y)$ are identified as the solution in (μ, ρ) to (2.9). This nonlinear system of 2 equations has unique solution under Assumption 1(2). This result follows from Theorem 4 of Gale and Nikaido (1965), after showing that the Jacobian of the system (2.9) is a P-matrix when $\rho(y) \in (-1, 1)$, which is what we demonstrate in the remainder of the proof.

Let $\partial_\mu \Phi_2(\mu, v; \rho) = \partial \Phi_2(\mu, v; \rho)/\partial \mu$ and $\partial_\rho \Phi_2(\mu, v; \rho) = \partial \Phi_2(\mu, v; \rho)/\partial \rho$. The Jacobian matrix of the system,

$$J(\mu(y), \rho(y)) = \begin{pmatrix} \partial_\mu \Phi_2(\mu(y), v(1); \rho(y)) & \partial_\rho \Phi_2(\mu(y), v(1); \rho(y)) \\ \partial_\mu \Phi_2(\mu(y), v(0); \rho(y)) & \partial_\rho \Phi_2(\mu(y), v(0); \rho(y)) \end{pmatrix},$$

[25] Note that in case (4), $F_{Y,D|Z}(y, 1 \mid 1) - \bar{\Phi}(v(1)) = \max[\Phi(\mu(y)) - \bar{\Phi}(v(1)), 0] \geqslant \max[\Phi(\mu(y)) - \bar{\Phi}(v(0)), 0] = \Phi(v(0)) > 0$ when $\rho(y) = -1$.

[26] Indeed, this result follows from a similar argument to the case $F_{Y,D|Z}(y, 1 \mid 1) - \bar{\Phi}(v(1)) = F_{Y,D|Z}(y, 1 \mid 0) - \bar{\Phi}(v(0))$ using that $\Phi_2(\mu(y), v(z); \rho(y)) + \bar{\Phi}(v(0)) = \Phi_2(-\mu(y), v(z); -\rho(y))$.

is a P-matrix for all $\mu(y) \in \mathbb{R}$ and $\rho(y) \in (-1,1)$ because by the properties of the bivariate normal CDF:

$$\partial_\mu \Phi_2(\mu(y), \nu(1); \rho(y)) = \Phi\left(\frac{\nu(1) - \rho(y)\mu(y)}{\sqrt{1-\rho(y)^2}}\right)\phi(\mu(y)) > 0,$$
$$\partial_\rho \Phi_2(\mu(y), \nu(0); \rho(y)) = \phi_2(\mu(y), \nu(0); \rho(y)) > 0,$$

and

$$\det(J(\mu(y), \rho(y))) = \phi(\mu(y))^2 \left[\Phi\left(\tilde{\nu}(1,y)\right)\phi\left(\tilde{\nu}(0,y)\right) - \Phi\left(\tilde{\nu}(0,y)\right)\phi\left(\tilde{\nu}(1,y)\right)\right] > 0,$$

where $\tilde{\nu}(0,y) = [\nu(0) - \rho(y)\mu(y)]/\sqrt{1-\rho(y)^2}$ and $\tilde{\nu}(1,y) = [\nu(1) - \rho(y)\mu(y)]/\sqrt{1-\rho(y)^2}$. In the last result we use that, by the properties of the normal distribution,

$$\phi_2(\mu, \nu; \rho) = \phi\left([\nu - \rho\mu]/\sqrt{1-\rho^2}\right)\phi(\mu)$$

and the inverse Mills ratio $\nu \mapsto \lambda(\nu) := \phi(\nu)/\Phi(\nu)$ is strictly decreasing in $\mathbb{R}$, so that

$$\Phi\left(\tilde{\nu}(1,y)\right)\phi\left(\tilde{\nu}(0,y)\right) - \Phi\left(\tilde{\nu}(0,y)\right)\phi\left(\tilde{\nu}(1,y)\right) > 0,$$

since $\tilde{\nu}(0,y) < \tilde{\nu}(1,y)$. □

A.3. **Proof of Theorem 2.2.** Let $D^* = V - \nu(Z)$ and $V \mid Z \sim N(0,1)$ such that $p(z) = F_{D^*|Z}(0 \mid z) = \Phi(\nu(z))$.[27] Note that $\Phi^{-1}(F_{Y^*}(y)) = \mu(y)$, $\Phi^{-1}(F_{D^*|Z}(0 \mid z)) = \nu(z)$, and

$$\begin{aligned}\Phi_2(\mu(y), \nu(z); \tilde{\rho}(y, \nu(z) \mid z)) &= F_{Y^*,V|Z}(y, \nu(z) \mid z) = \mathrm{P}(Y^* \leqslant y, V \leqslant \nu(z) \mid Z = z) \\ &= \mathrm{P}(Y^* \leqslant y, D^* \leqslant 0 \mid Z = z) = F_{Y^*,D^*|Z}(y, 0 \mid z) = \Phi_2(\mu(y), \nu(z); \rho(y \mid z)),\end{aligned}$$

by definition of the LGR, such that $\tilde{\rho}(y, \nu(z) \mid z) = \rho(y \mid z)$ The result holds because $\tilde{\rho}(y, \nu(z) \mid z) = \rho(y)$ if $\rho(y \mid z) = \rho(y)$ and $\rho(y \mid z) = \tilde{\rho}(y)$ if $\tilde{\rho}(y, \nu(z) \mid z) = \tilde{\rho}(y)$. □

A.4. **Proof of Equivalence** (2.10)**.** Assume joint independence, $(Y^*, V) \perp\!\!\!\perp Z$. Then, $\tilde{\rho}(y, v) = \tilde{\rho}(y)$ for all $v \in \mathcal{V}_0$, where $\mathcal{V}_0$ is an open interval containing the support of $\nu(Z)$, implies the single index property because

$$\mathrm{P}(Y^* \leqslant y \mid V = v) = \frac{(\partial/\partial v)F_{Y^*,V}(y, v)}{(\partial/\partial v)F_V(v)} = \frac{(\partial/\partial v)\Phi_2(\mu(y), v; \tilde{\rho}(y))}{(\partial/\partial v)\Phi(v)} = \Phi\left(a(y) + b(y)v\right),$$

for $a(y) = \mu(y)/\sqrt{1-\tilde{\rho}(y)^2}$ and $b(y) = -\tilde{\rho}(y)/\sqrt{1-\tilde{\rho}(y)^2}$. Conversely, assuming that the index restriction holds for all $v \in \mathcal{V}_0$, the partial differential equation

$$(\partial/\partial v)F_{Y^*,V}(y, v) = \Phi\left(a(y) + b(y)v\right)(\partial/\partial v)\Phi(v)$$

has a solution at the LGR, $F_{Y^*,V}(y, v) = \Phi_2\left(\mu(y), v; \tilde{\rho}(y)\right)$, for all $v \in \mathcal{V}_0$, where $\mu(y) = a(y)/\sqrt{1+b(y)^2}$ and $\tilde{\rho}(y) = -b(y)/\sqrt{1+b(y)^2}$. In both cases we use that $V \sim N(0,1)$.

[27] Note that any definition of D^* that yields the same probability of $D = 1$ conditional on Z is observationally equivalent. In particular, we can choose without loss of generality D^* with continuous CDF.

Supplementary Material to "Distribution Regression with Sample Selection, with an Application to Wage Decompositions in the UK "

Victor Chernozhukov, Iván Fernández-Val, and Siyi Luo

ABSTRACT. The supplementary material includes seven appendices. Appendices A–C contain deferred discussions of Sections 2–4. They include properties of the LGR, comparison with AB17, identification approaches with rich instruments, bound analysis with an example, models with homogeneous parameters and an example of sign reversal in the selection effects. Appendices D and E contain the asymptotic theory for the estimation and inference methods, and the corresponding proofs. Appendix F reports the results of a Monte Carlo simulation calibrated to the empirical application. Appendix G shows additional results of the empirical application. They include descriptive statistics and background on the U.K. labor market, a model of offered and reservation wages, estimates of the coefficients of the employment (selection) equation, estimates and 95% confidence intervals for the coefficients of the wage equation, estimates and 95% confidence intervals for the sorting equation not reported in the main text, estimates and 95% confidence intervals for the decomposition of the employment rate, estimates and 95% confidence bands for the components of the wage decomposition of observed wages in the specification 4, estimates and 95% confidence bands for the offered and observed wages and their decompositions for the specifications 1–3, and estimates and 95% confidence bands for the wage decomposition between first and second half of the sample period for men and women in specification 2.

APPENDIX A. DEFERRED DISCUSSIONS OF SECTION 2

A.1. **Discussion of the LGR and Local Correlation.** In the LGR, the marginal CDFs of Y^* and D^* are represented by local Gaussian links

$$F_{Y^*}(y) = \Phi(\mu(y)), \quad F_{D^*}(d) = \Phi(\nu(d)),$$

and the copula of Y^* and D^* is represented by a local Gaussian copula

$$\begin{aligned} &C_{Y^*,D^*}(u,v) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \rho(y_u, d_v)), \\ &\forall (u,v) \in [0,1]^2 : \exists y_u \in \mathbb{R} : F_{Y^*}(y_u) = u, \quad \exists d_v \in \mathbb{R} : F_{D^*}(d_v) = v. \end{aligned} \tag{A.14}$$

Kolev, Anjos, and Mendes (2006) developed a closely related result to (A.14) for the copula. They established that the copula of any bivariate distribution can be represented by the bivariate Gaussian copula with a local correlation parameter. Like the copula, the LGR is convenient because it separates $\mu(y)$ and $\nu(d)$ as two parameters determining the marginals of Y^* and D^* from $\rho(y,d)$ as a parameter determining the dependence between Y^* and D^*. Unlike the copula, the arguments of the LGR are the same as the arguments of the joint CDF and the domain of the LGR is therefore $\mathbb{R}^2$.

Here we present additional discussion of the local correlation $\rho(y,d)$.[1] In general, the sign of $\rho(y,d)$ determines the sign of the local dependence between Y^* and D^* at (y,d) as measured by the covariance between $1\{Y^* \leqslant y\}$ and $1\{D^* \leqslant d\}$.

Lemma A.1 (Correlation interpretation for $\rho(y,d)$)**.** *For (Y^*, D^*) having LGR with local correlation parameter $\rho(y,d)$, the map $\rho(y,d) \mapsto \mathrm{Cov}(1\{Y^* \leqslant y\}, 1\{D^* \leqslant d\})$, mapping $[-1,1]$ to the real line, is a strictly increasing (hence bijective) function. Moreover, $\mathrm{Cov}(1\{Y^* \leqslant y\}, 1\{D^* \leqslant d\})$ and $\rho(y,d)$ have the same sign (positive, negative, or 0).*

It follows that $\rho(y,d)$ is positive (resp. negative) everywhere if and only if Y^* and D^* are positively (resp. negatively) quadrant dependent in the sense defined by (Lehmann, 1966). The local dependence measure $\mathrm{Cov}(1\{Y^* \leqslant y\}, 1\{D^* \leqslant d\})$ is called Laplace covariance kernel by Dette, Hallin, Kley, and Volgushev (2015).

When $\rho(y,d) = \rho(y)$ we can further characterize the local dependence in terms of the normal scores $Y^s := \Phi^{-1}(F_{Y^*}(Y^*))$ and $D^s := \Phi^{-1}(F_{D^*}(D^*))$. Let $F_{Y^s,D^s}(y,d) = \Phi_2(\mu_s(y), \nu_s(d); \rho_s(y,d))$ be the LGR of the distribution of (Y^s, D^s).

Lemma A.2 (Other Properties of $\rho(y,d)$)**.** *If $\rho(y,d) = \rho(y)$ and $d \mapsto F_{D^*}(d)$ is continuous on $\mathbb{R}$,*

$$\rho_s(y) = -\frac{\mathrm{Cov}(1\{Y^s \leqslant \mu_s(y)\}, D^s)}{\phi(\mu_s(y))},$$

where $\mathrm{Cov}(Y^s, D^s) = \mathrm{E}[\rho_s(Y^s)]$, provided that $\mathrm{Cov}(Y^s, D^s)$ exists, where $\mu_s(y) = \sup\{y^* \in \mathbb{R} : F_{Y^*}(y^*) \leqslant \Phi(y)\}$, and $\rho_s(y) = \rho(y^*)$ for y^* such that $F_{Y^*}(y^*) = \Phi(\mu_s(y))$.

[1] The parameter $\rho(y,d)$ is the tetrachoric correlation coefficient between $\{Y^* \leqslant y\}$ and $\{D^* \leqslant d\}$ (Pearson, 1900).

Note that for the points y in the support of Y^s, $\rho_s(y) = -\mathrm{Cov}(1\{Y^s \leqslant y\}, D^s)/\phi(y)$. Lemma A.2 therefore characterizes $\rho_s(y)$, and therefore also $\rho(y)$, as a measure of local dependence between Y^s and D^s. It also shows that the covariance between these normal scores, a measure of global dependence, can be expressed as the average of the local dependence parameter. The case considered with $\rho(y,d) = \rho(y)$ and continuous D^* is relevant for the discussion of the sample selection problem in Section 2.3.

A.2. **Comparison with AB17.** We compare our identification conditions with AB17. We start by restating AB17 assumptions in terms of the LGR of the joint distribution of (Y^*, V). In addition to Assumption 1(1)–(3), AB17 assumed that (i) (Y^*, V) are jointly independent of Z, (ii) $v \mapsto \tilde{\rho}(y, v)$ is real analytic on the unit interval, (iii) the support of $p(Z)$ contains an open interval, and (iv) $y \mapsto F_{Y*}(y)$ is continuous and strictly increasing.[2] Assumption 1 neither implies nor is implied by conditions (i)–(iv). Thus, condition (iii) requires Z to have continuous variation and is therefore more restrictive than our assumption that Z can be binary. Unlike (iv), we do not require the CDF of Y^* to be continuous. Our identification result applies to discrete and mixed outcomes. However, under condition (i), Theorem 2.2 shows that selection exclusion requires that $v \mapsto \tilde{\rho}(y, v)$ is constant, which is stronger than (ii). In other words, we impose stronger restrictions in the dependence between the latent outcome and unobserved selection ranking, but require less variation in the excluded covariate Z and outcome Y^*.

More specifically, let $p(z) = \mathrm{P}(D = 1 \mid Z = z)$ and $V = F_{D^*|Z}(D^* \mid Z)$ such that $V \mid Z \sim U(0,1)$.[3] AB17 assumed that (i) (Y^*, V) are jointly independent of Z, (ii) $v \mapsto C_{Y^*,V}(\cdot, v)$ is real analytic on the unit interval, where $C_{Y^*,V}$ is the copula of (Y^*, V), and (iii) the support of $p(Z)$ contains an open interval. We now show that our selection sorting exclusion neither implies nor is implied by conditions (i) and (ii). Selection sorting exclusion implies that for any $u \in [0,1]$ that satisfies $F_{Y^*}(y_u) = u$ for some y_u,

$$C_{Y^*,V|Z}(u, p(z) \mid z) = C_{Y^*,D^*|Z}(u, p(z) \mid z) = \Phi_2(\Phi^{-1}(u), \Phi^{-1}(p(z)); \rho(y_u, 0)) = C_{Y^*,V}(u, p(z)),$$

since $p(z) = F_{D^*|Z}(0 \mid z)$. This implication is weaker than condition (i) but it suffices for the identification argument in AB17. However, it only guarantees that $v \mapsto C_{Y^*,V}(\cdot, v)$ is real analytic on the support of $p(Z)$.[4] Therefore, we conclude that selection exclusion implies conditions (i) and (ii) only if the support of $p(Z)$ is the unit interval. To verify that the converse is also not true, note that the LGR of (Y^*, V) conditional on Z under condition (i) is

$$F_{Y^*,V|Z}(y, v \mid z) = \Phi_2(\tilde{\mu}(y), \tilde{\nu}(v); \tilde{\rho}(y, v)).$$

This, together with $\tilde{\mu}(y) = \mu(y)$ and $\tilde{\nu}(p(z)) = \nu(z)$, imply that

$$F_{Y^*,D^*|Z}(y, 0 \mid z) = \Phi_2(\tilde{\mu}(y), \tilde{\nu}(p(z)); \tilde{\rho}(y, p(z))) = \Phi_2(\mu(y), \nu(z); \tilde{\rho}(y, p(z))),$$

[2] Heckman and Singer (1984) also exploited a real analyticity assumption to solve a censoring problem in duration analysis.

[3] We assume that D^* is absolutely continuous with strictly increasing distribution. This assumption is without loss of generality because the distribution of D^* is only identified at $D^* = 0$.

[4] Note that $v \mapsto \Phi_2(\cdot, \Phi^{-1}(v); \rho(\cdot, 0))$ is a real analytic function.

which satisfies the selection exclusion only if $\tilde{\rho}(y,v) = \tilde{\rho}(y)$ for all v in the support of $p(Z)$, i.e. the local dependence between Y^* and V does not vary with the value of V in this region. We finally note that condition (i) together with $\tilde{\rho}(y,v) = \tilde{\rho}(y)$ for all v in the unit interval imply condition (ii) because

$$C_{Y^*,V}(\cdot,v) = \Phi_2(\cdot,\Phi^{-1}(v);\tilde{\rho}(\cdot))$$

is a real analytic function with respect to v in the unit interval. Alternatively, condition (ii) is equivalent to $v \mapsto \tilde{\rho}(\cdot,v)$ being real analytic, which is weaker than $\tilde{\rho}(y,v) = \tilde{\rho}(y)$.

A.3. **Further Identification Results with Richer Instruments.** Suppose that (Y^*,V) are jointly independent of Z:

$$(Y^*,V) \perp\!\!\!\perp Z. \tag{A.15}$$

This assumption implies that the sorting mechanism depends only on the instrumental values through the propensity score

$$\begin{aligned} \mathrm{P}(Y^* \leqslant y, D^* \leqslant 0 \mid Z = z) &= \Phi_2(\mu(y), \nu(z); \rho(y,\nu(z)) \\ &= \mathrm{P}(Y^* \leqslant y, D^* \leqslant 0 \mid \nu(Z) = \nu(z)), \quad z \in \mathcal{Z}, \end{aligned} \tag{A.16}$$

where $\nu(z)$ is the transformed propensity score $\Phi^{-1}(p(z))$, and $\mathcal{Z}$ is the support of Z.

Using (A.16), the system of equations for the identification analysis can be written as:

$$\mathsf{R}(y,v) := \mathrm{P}(Y^* \leqslant y, D^* \leqslant 0 \mid \nu(Z) = v) = \Phi_2(\mu(y), v; \rho(y,v)) \quad v \in \mathcal{V}, \tag{A.17}$$

where $\mathcal{V}$ is the support of $\nu(Z)$. Intuitively, if the propensity scores have enough variation, the equations above should contain identifying information. Still, considerable care is required to turn this information into point identification and estimability.

Real Analyticity Approach. This approach imposes real-analyticity assumptions on $v \mapsto \rho(y,v)$ over $\mathbb{R}$, but avoids making parametric assumptions. Specifically, analyticity means that over each open neighborhood the mapping can be represented by a convergent Taylor expansion around a point in that neighborhood (Krantz and Parks, 2002, Def. 1.1.5). For example, a continuously infinite-differentiable function with derivatives bounded by a common constant has this property. The key property of real analytical functions is the continuation: knowing a function over an open region, implies we know the function over entire $\mathbb{R}$ (Krantz and Parks, 2002, Cor. 1.2.6). Thus, intuitively, such assumption rules out "wilder" forms of $\rho(y,v)$ that cause identification failure and attain the agnostic Balke-Pearl bounds.[5]

Lemma A.3 (Identification Under Real Analyticity)**.** *Fix a value of y in what follows. Suppose that $v \mapsto \rho(y,v)$ is real analytic on $\mathbb{R}$, mapping into $[-1,+1]$. Also assume that $v \mapsto \rho(y,v)$ can be continuously*

[5]The Balke-Pearl bounds are attained by having $\rho(y,v)$ vary between +1 or -1 over each small neigborhood of a given point v_0, to trace the smallest root and the maximal root as bounds for $\mu(y)$, then taking the smallest upper bound and taking the largest lower bound across instrumental values v. Such oscillatory (and unreasonable) behavior of the sorting function ρ can be ruled out by a variety of means, which underlie but don't exhaust all of the point identification results in this paper.

extended to $\overline{\mathbb{R}}$*, and the propensity score variable* $p(Z)$ *is continuous on an open subinterval of* $(0,1)$*. Then the pair* $(\mu(y), \rho(y,v))$ *are identified:* $\mu(y)$ *is identified at infinity as*

$$\mu(y) = \lim_{v\to\infty} \Phi^{-1}(\mathsf{R}(y,v))$$

and $\rho(v)$ *is identified as the unique solution to* $\mathsf{R}(y,v) = \Phi_2(\mu(y), v, \rho(y,v))$ *for any* $v \in \mathcal{V}$*.*

The lemma above extends the argument in AB17 to the present LGR-based approach. The drawback of the analyticity approach is its restrictive nature; for example, functions generated from finite spline sieves are not real analytical. Moreover, identification does not immediately translate into estimability due to the ill-conditioned nature of the analytical continuation, especially at "infinity" – see Trefethen (2020) for further discussion. Imposing parametric functional forms on $v \mapsto \rho(y,v)$ provides necessary regularization, as used, for example, in AB17 for estimation purposes.[6] But once we make the parametric assumptions, we can also get identification and estimability without requiring continuous instruments and without relying on the identification at infinity arguments.

Semi-Parametric Approach. We assume that $\rho(y,v) = \rho(y,v;\theta_0(y))$ for $\theta_0(y) \in \Theta \subset \mathbb{R}^d$, where $\rho(y,v;\theta)$ is a known function up to the parameter θ and Θ is a compact region with non-empty open interior in $\mathbb{R}^d$; and that the true value of $\rho(y,v)$ is induced by some value $\theta_0(y)$ in the interior of Θ. Let

$$\mathsf{L}_y(m,v;\theta) := \Phi_2(m,v;\rho(y,v;\theta)).$$

Moreover, we assume that $\theta \mapsto \rho(y,v;\theta)$ is continuously differentiable at each v and y of interest. Define the map

$$(m,\theta) \mapsto \mathsf{L}_y(m,\theta) := \left\{\mathsf{L}_y(m,v;\theta), v \in \mathcal{V}\right\},$$

where $\mathcal{V}$ is the support of $v(Z)$, which we assume is a finite set. Define the Jacobian of this map as $\mathsf{JL}_y(m,\theta)$. The first row of $\mathsf{JL}_y(m,\theta)$ is $\{(\partial/\partial m)\mathsf{L}_y(m,\theta)\}'$ and the $(k+1)$-st row is $\{(\partial/\partial\theta_k)\mathsf{L}_y(m,\theta)\}'$, $k = 1,\ldots,d$, where θ_k is the kth component of θ.

We have the system of equations

$$\mathsf{R}(y,v) = \mathsf{L}_y(m,v;\theta), \quad v \in \mathcal{V}. \tag{A.18}$$

We shall apply univalence results to deduce global and local identifiability.

To proceed, we assume that $v \mapsto \rho(y,v;\theta)$ is non-redundantly parameterized, namely, each set of values of this function over the set $\mathcal{V}$ can only correspond to one value of θ. Define $\mathcal{M}$ as a compact interval in $\mathbb{R}$ containing $\mu(y)$ in its interior.

Lemma A.4 (Identification with Discrete Instruments)**.** *Given the set-up in this subsection, assume that* $\mathcal{V}$*, the support of* $v(Z)$*, is a finite set with cardinality* $|\mathcal{V}|$ *at least* $d+1$*. Then the system of equations (A.18) has the unique local solution at* $(m,\theta) = (\mu(y),\theta_0(y))$ *if the Jacobian* $\mathsf{JL}_y(\mu(y),\theta_0(y))$ *has rank* $d+1$*. Moreover, if* Θ *is rectangular and* $\mathsf{JL}_y(m,\theta)$ *is full rank and* $R\mathsf{JL}(m,\theta)$ *is a P-matrix over* $\mathcal{M}\times\Theta$*, for some* $(d+1)\times|\mathcal{V}|$ *matrix* R*, then* $(\mu(y),\theta_0(y))$ *are globally identified. More generally if* Θ *is a convex*

[6] See Trefethen (2020) for making this point explicit through the use of polynomials.

polyhedron or a set with C^1 boundary, with non-empty interior, and if the generalized P-matrix or positive quasi-definite matrix conditions of Mas-Colell (1979) in Theorem 1 and Theorem 2 are satisfied for the projection matrix $R\mathsf{JL}_y(m,\theta)$ for some $(d+1)\times|\mathcal{V}|$ matrix R, then $(\mu(y),\theta_0(y))$ are globally identified.

Unlike Lemma A.3, Lemma A.4 requires only that the instrument and the induced propensity score take on $d+1$ values, because it does not rely on extrapolation or identification at infinity (the parametric function needs to be only defined over $\mathcal{V}$). The conditions for identification above for $d=1$ have particularly interpretable form, leading to Theorem 2.1 in the main text.

Intuition. A simple geometrical insight that helps interpret the local identification conditions is that in order for $\mu(y)$ to be a non-unique root locally, we need to be able to move the path of roots for m, m_t, along a time axis with constant non-zero velocity $c\neq 0$, that is $\dot{m}_t = c$, starting from $m_0=\mu(y)$. Since we can rescale time, without loss of generality, set $|c|=1$. Then, this move is feasible if we can find a corresponding move $\dot{\theta}_t=\delta_t$ that can implement $m_t = m_0 + ct$ as a root, starting at θ_0. In summary, a path of local alternative solutions can be potentially generated by solving the following differential equation system: for all $t\in(0,\epsilon)$

$$m_t(y) = m_0 + ct,\ \ \theta_t = \theta_0 + \int_0^t \delta_s ds$$

such that

$$c = \dot{m}_t(y) = \frac{(\partial/\partial\theta_t')\mathsf{L}_y(m_t,v;\theta_t)}{(\partial/\partial m_t)\mathsf{L}_y(m_t,v;\theta_t)}\delta_t,\quad \forall v\in\mathcal{V},$$

and subject to the boundary conditions $\rho(y,v;\theta_t)\in[-1,1]$. If this path exists, then we have non-identification of $\mu(y)$. Otherwise, we have local identification of $\mu(y)$.

Since $\epsilon>0$ is arbitrary and the problem is smooth, it suffices to check local identification by just examining the failure of local identification at the origin $t=0$. If the instrument is not rich enough, namely, the cardinality of $\mathcal{V}$ is smaller than $d+1$, a local path direction δ_0 always exists. More generally, δ_0 exists if the matrix M with rows

$$\left.\frac{(\partial/\partial\theta_t')\mathsf{L}_y(m_t,v;\theta_t)}{(\partial/\partial m_t)\mathsf{L}_y(m_t,v;\theta_t)}\right|_{t=0},\quad v\in\mathcal{V},$$

linearly spans the vector $c\mathbf{1}$, where $\mathbf{1}$ is a vector of ones. If the spanning condition does not hold, δ_0 does not exist, leading to local identification. For example, if $d=1$, and the matrix M has two scalar entries $M_1\neq M_2$ in its rows, then solving $M_1\delta_0=c, M_2\delta_0=c$ for the same δ_0 is impossible, yielding local identification. This case corresponds to Theorem 2.1. This view of local identification is equivalent to local identification based on the Jacobian $\mathsf{JL}_y(\mu(y),\theta_0)$ having full rank, but it is perhaps more illuminating.

A.4. **Relaxing Selection Sorting Exclusion.** We next consider replacing the selection sorting exclusion with weaker conditions. While the key motivating semi-parametric example (1.3) easily meets the exclusion restriction, we analyze how deviations from such structures can impact the identification results. This also helps us connect the results to the Balke-Pearl bounds that only rely on the outcome exclusion. We provide additional theoretical discussions for motivating this investigation in Section 2.3

Assumption 2 (Relaxed Selection Restrictions)**.** *Suppose Assumption 1 (1)-(3) hold, namely non-degeneracy, relevance, and outcome exclusion, and in addition either:*

(5) *r-Relaxed Selection Exclusion:* $|\rho(y \mid 0) - \rho(y \mid 1)| \leqslant r$ *for some* $0 < r \leqslant 2$*; or*
(6) *Positive Selection:* $\rho(y \mid z) \geqslant 0$ *for each* $z \in \{0,1\}$.

The first condition allows the strength of selection to respond to Z, but in a limited fashion governed by the parameter r. As r varies from 0 to 2, we can trace a nested sequence of identification regions, starting with a point and ending with the Balke-Pearl bounds resulting from outcome exclusion.

The second condition restricts the sign of the implied correlation to be positive (the sign in this condition can be reversed by changing the sign of outcome). This also leads to tighter bounds relative to the case with just outcome exclusion. It is useful to note that the implied correlation is positive, $\rho(y) > 0$, if and only if $1(Y^* > y)$ and $1(D^* > 0)$ are positively correlated for all y. In the context of labor supply, this condition means that there is positive sorting into employment throughout the offered wage distribution. Therefore, the positivity condition has a rather natural interpretation in terms of intrinsic economic quantities.

With this relaxed condition, we work with the two equations:

$$\mathrm{P}(Y \leqslant y, D = 1 \mid Z = z) = \Phi_2(\mu, \Phi^{-1}(\mathrm{P}(D = 1 \mid Z = z)); \rho_z), \quad z \in \{0,1\}, \tag{A.19}$$

that share one unknown μ, and each has its own unknown ρ_z. The following statement is straightforward, but we record it formally nonetheless. In the statement we let the parameter (μ, ρ_0, ρ_1) range over the parameter set $\overline{\mathbb{R}} \times [-1,1] \times [-1,1]$.

Theorem A.1 (Identification under Assumption 2)**.** *Let* $S_z = \{(\mu_z, \rho_z)\}$ *denote all solutions to equation (A.19) for* $z = 0$ *and* $z = 1$*. Each* S_z *is not empty, and, restricted to the non-degenerate case* $\rho_z^2 < 1$*, is a one-dimensional manifold. Under outcome exclusion, the identification region for* $(\mu(y), \rho(y \mid 0), \rho(y \mid 1))$ *is* $R_0 = \{(\mu, \rho_0, \rho_1) : (\mu, \rho_z) \in S_z,$ *for each* $z \in \{0,1\}\}$*. Under outcome exclusion and the relaxed selection, the identification region is* $R_0(r) = R_0 \cap \{(\mu, \rho_0, \rho_1) : |\rho_0 - \rho_1| \leqslant r\}$*. Under outcome exclusion and positive selection, the region is* $R_0^+ = R_0 \cap \{(\mu, \rho_0, \rho_1) : \min(\rho_0, \rho_1) \geqslant 0\}$*. The identification regions for* $\mu(y)$ *are given by applying the projection operator* π_1 *taken with respect to the first coordinate: this gives regions* $\pi_1 R_0$*,* $\pi_1 R_1(r)$*, and* $\pi_1 R_0^+$*, respectively.*

We observe that that the identification region $\pi_1 R_0$ is (equivalent to) the Balke-Pearl bound. The regions $R_1(r)$ and R_0^+ tighten the bounds: by construction, these sets are weakly smaller than R_0, and

the example below demonstrates they can be much smaller. In fact, $R_1(r)$ converges to a singleton as $r \searrow 0$. We can also intersect the two sets if both forms of relaxation are assumed to hold.

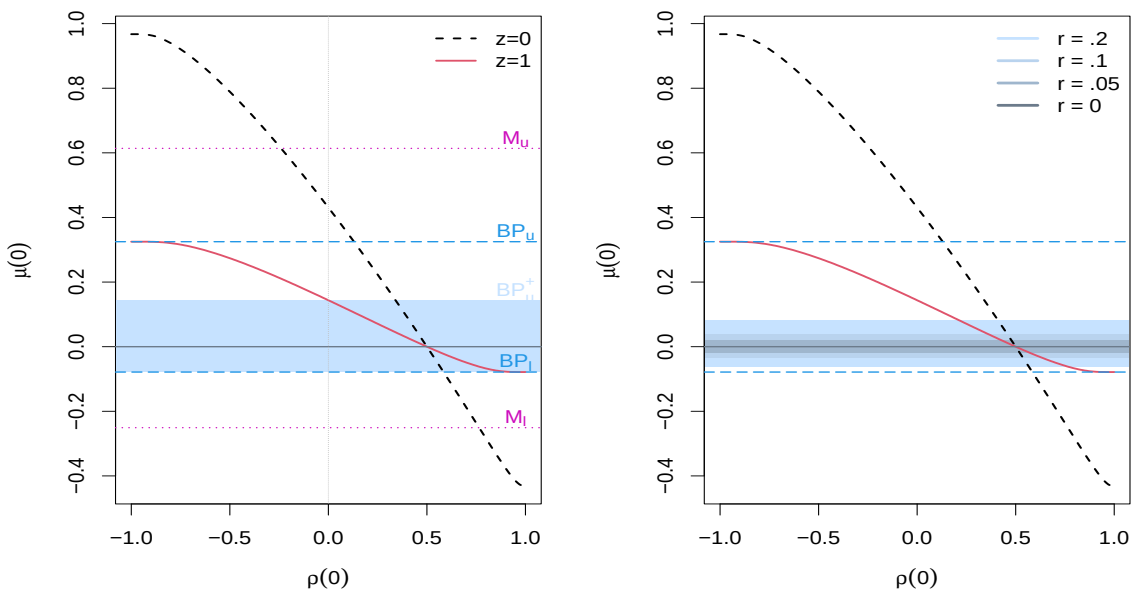

FIGURE 1. Illustration of identification analysis via exclusion restrictions.

Notes. The curves represent solutions in (μ, ρ) to $\Phi_2(0, z; 0.5) = \Phi_2(\mu, z; \rho)$ for $z \in \{0, 1\}$. The point identification $\mu(0) = 0$ occurs at the intersection of the curves. The left panel also demonstrates the construction of identified sets for $\mu(0)$ when the selection exclusion is removed ($[BP_l, BP_u]$) or replaced by positive sorting ($[BP_l, BP^+]$). The right panel demonstrates the construction of identified sets when the selection takes the relaxed form of Assumption 2(5).

A.5. **The role of selection exclusions and their relaxed forms in identification.** We illustrate the role of exclusion restrictions or their relaxed versions, we consider a simple example.[7] Assume that Y^* is binary with $Y^* = 1\{U > 0\}$, $D = 1\{V \leqslant Z\}$, (U, V) are standard bivariate normal with parameter $\rho = 0.5$, $\mathrm{P}(D = 1 \mid Z = z) = \Phi(z)$ and $\mathrm{P}(Z = 1) = 1/2$. The distribution of Y^* is fully characterized by $F_{Y^*}(0) = \mathrm{P}(Y^* = 0)$ or equivalently by $\mu(0) = \Phi^{-1}(F_{Y^*}(0))$. In what follows, none of the bounds to be constructed make use of the parametric distribution of Y^* and D, other than to compute $\mathrm{P}(Y^* \leqslant 0, D = 1 \mid Z = z)$.

Figure 1 plots the solutions in (μ, ρ) to (2.9) at $y = 0$ for each value of z without imposing any exclusion restriction, together with bounds for $\mu(0)$ after sequentially imposing exclusion restrictions. In the right panel, the solid and dashed curves are the solutions in (μ, ρ) to $\Phi_2(0, z; 0.5) = \Phi_2(\mu, z; \rho)$ for

[7]This example is inspired by an example graciously provided to us by one of the referees.

$z \in \{0,1\}$, respectively. The horizontal lines indicate the construction of bounds on $\mu(0)$ by projection. Without imposing any exclusion restriction the identified set is $[M_l, M_u] = [-.25, .61]$. This set is obtained as the mean of the maximum/minimum of the curves (Manski, 1990). When we impose the outcome exclusion we obtain the identified set $\pi_1 R_0 = [BP_l, BP_u] = [-.08, .32]$. This set is obtained as the minimum/maximum of the maximum/minimum of the curves (Balke and Pearl, 1994). When we impose both outcome and exclusion restrictions, we obtain $\pi_1 R_1(0) = \{\mu(0)\} = \{0\}$ as the identified set; that is, we point-identify the parameter value $\mu(0)$. Graphically, we obtain this set by intersecting the two curves and projecting the intersection point on the vertical axis.

In the left panel, we also show how the positive selection restriction in Assumption 2(6) helps tighten the bound. Indeed, the resulting identified set $\pi_1 R_0^+$ is $[BP_l, BP^+] = [-.08, .14]$, which is tighter than the Balke-Pearl set $\pi_1 R_0$. Finally, the right panel shows bounds imposing the relaxed selection restriction in Assumption 2(5) for $r \in \{0, .05, .10, .20\}$. The resulting identified sets are $\pi_1 R_1(0) = \{0\}$, $\pi_1 R_1(.05) = [-.02, .02]$, $\pi_1 R_1(.10) = [-.03, .04]$, and $\pi_1 R_1(.20) = [-.06, .08]$. The set $\pi_1 R_0$ can be obtained by setting r sufficiently large (in particular, $r = 1.13$ is sufficient).

A.6. **Proof of Lemma A.1.** We have that

$$\begin{aligned}\mathrm{Cov}(1\{Y^* \leqslant y\}, 1\{D^* \leqslant d\}) &= \Phi_2(\mu(y), \nu(d); \rho(y,d)) - \Phi(\mu(y))\Phi(\nu(d)) \\ &= \Phi_2(\mu(y), \nu(d); \rho(y,d)) - \Phi_2(\mu(y), \nu(d); 0),\end{aligned}$$

and $\rho \mapsto \Phi_2(\cdot, \cdot; \rho)$ is strictly increasing. The conclusion about signs follows. □

A.7. **Proof of Lemma A.2.** We start by characterizing the marginal and joint CDFs of $Y^s = \Phi^{-1}(F_{Y^*}(Y^*))$ and $D^s = \Phi^{-1}(F_{D^*}(D^*))$. Since $d \mapsto F_{D^*}(d)$ is continuous, $F_{D^*}(D^*) \sim U(0,1)$ by the probability integral transform. Hence,

$$F_{D^s}(d) = \mathrm{P}(F_{D^*}(D^*) \leqslant \Phi(d)) = \Phi(d).$$

Let $\mathscr{Y}^s = \{y \in \overline{\mathbb{R}} : \Phi(y) = F_{Y^*}(y^*) \text{ for some } y^* \in \mathbb{R}\}$ be the support of Y^s. For any $y \in \mathscr{Y}^s$,

$$F_{Y^s}(y) = \mathrm{P}(\Phi^{-1}\left(F_{Y^*}(Y^*)\right) \leqslant \Phi^{-1}\left(F_{Y^*}(y^*)\right)) = F_{Y^*}(y^*) = \Phi(y).$$

The CDF extends to all $y \in \overline{\mathbb{R}}$ by

$$F_{Y^s}(y) = F_{Y^s}(\mu_s(y)) = \Phi(\mu_s(y)), \;\; \mu_s(y) = \sup\{y^* \in \overline{\mathbb{R}} : F_{Y^*}(y^*) \leqslant \Phi(y)\}.$$

Hence, the LGR of the distribution of (Y^s, D^s) is

$$F_{Y^s, D^s}(y,d) = \Phi_2(\mu_s(y), d; \rho_s(y,d)),$$

where $\rho_s(y,d) = \rho_s(\mu_s(y), d)$ because $F_{Y^s,D^s}(y,d) = F_{Y^s,D^s}(\mu_s(y), d)$.

Next, we show that for any integrable random variable X and random variable Z,

$$\mathrm{Cov}(1\{Z \leqslant z\}, X) = -\int_{-\infty}^{\infty} [F_{X,Z}(x,z) - F_X(x)F_Z(z)]\mathrm{d}x \tag{A.20}$$

and

$$\mathrm{Cov}(Z, X) = -\int_{-\infty}^{\infty} \mathrm{Cov}(1\{Z \leqslant z\}, X)\mathrm{d}z. \tag{A.21}$$

Given (A.20), (A.21) follows from Höffding (1940) and Lehmann (1966). To show (A.20), let (X_1, Z_1) and (X_2, Z_2) be two independent copies of (X, Z). Then,

$$\mathrm{Cov}(1\{Z \leqslant z\}, X) = \frac{1}{2}\mathrm{E}[(1\{Z_1 \leqslant z\} - 1\{Z_2 \leqslant z\})(X_1 - X_2)].$$

The result follows from $X_1 - X_2 = \int_{-\infty}^{\infty}[1\{X_2 \leqslant x\} - 1\{X_1 \leqslant x\}]\mathrm{d}x$ and interchanging the expectation and integral.

Finally, the results follow from applying (A.20) and (A.21) to $X = D^s$ and $Z = Y^s$, together with $\rho_s(y, d) = \rho_s(y)$ because $\rho(y, d) = \rho(y)$ by assumption and

$$\begin{aligned}\Phi_2(\mu_s(y), d; \rho_s(y, d)) = F_{Y^s,D^s}(y, d) = F_{Y^s,D^s}(\mu_s(y), d) = F_{Y^*,D^*}(y^*, d^*)\\ = \Phi_2(\mu(y^*), \nu(d^*); \rho(y^*, d^*)) = \Phi_2(\mu_s(y), d; \rho(y^*, d^*)),\end{aligned}$$

where y^* and d^* are the solution to $F_{Y^*}(y^*) = \Phi(\mu_s(y))$ and $F_{D^*}(d^*) = \Phi(d)$. Indeed, by (A.20),

$$\begin{aligned}\mathrm{Cov}(1\{Y^s \leqslant y\}, D^s) &= -\int_{-\infty}^{\infty}[\Phi_2(\mu_s(y), d; \rho_s(\mu_s(y))) - \Phi_2(\mu_s(y), d; 0)]\mathrm{d}d\\ &= -\int_{-\infty}^{\infty}\int_0^{\rho_s(\mu_s(y))} \phi_2(\mu_s(y), d; r)\mathrm{d}r\mathrm{d}d = -\rho_s(\mu_s(y))\phi(\mu_s(y)),\end{aligned}$$

by $\Phi_2(\mu_s(y), d; 0) = \Phi(\mu_s(y))\Phi(d)$, $\phi_2(\cdot, \cdot; r) = \partial\Phi_2(\cdot, \cdot; r)/\partial r$ (Sibuya, 1959; Sungur, 1990), and

$$\int_{-\infty}^{\infty} \phi_2(\mu_s(y), d; r)\mathrm{d}d = \phi(\mu_s(y))$$

after interchanging the order of the integrals. Then, by (A.21),

$$\mathrm{Cov}(Y^s, D^s) = \int_{-\infty}^{\infty} \rho_s(\mu_s(y))\phi(\mu_s(y))\mathrm{d}y = \int_{-\infty}^{\infty} \rho_s(y)\mathrm{d}F_{Y^s}(y) = \mathrm{E}[\rho_s(Y^s)].$$

□

A.8. **Proof of Lemma A.3.** The result is immediate from $y \mapsto \mathsf{R}(y, \nu)$ being real analytical on $\mathbb{R}$, which pins down its value on the entire $\mathbb{R}$ via analytical continuation (Krantz and Parks, 2002, cor. 1.2.6), including the limit value which exists by assumption. This gives the identification of $\mu(y)$ and the identification of $\rho(y, \nu)$ follows using the same argument as in the proof of Lemma 2.1. □

A.9. **Proof of Lemma A.4.** The local identification result is an immediate consequence of Rothenmberg (Rothenberg, 1971), and the global result follows from Mas-Colell (1979), Theorem 1 or Theorem 2. □

A.10. **Proof of Theorem A.1.** The sets S_0 and S_1 are not empty by Lemma 2.1 applied to the conditional CDFs $F_{Y^*,D^*|Z}(y, 0 \mid z)$, $z \in \{0, 1\}$. They are one-dimensional manifolds by the implicit function theorem, because $(\mu, \rho) \mapsto \Phi_2(\mu, \Phi^{-1}(\mathrm{P}(D = 1 \mid Z = z)); \rho) - \mathrm{P}(Y \leqslant y, D = 1 \mid Z = z)$ is continuously differentiable and surjective for $z \in \{0, 1\}$. The rest of the claims follow by definition. □

APPENDIX B. DEFERRED DISCUSSIONS OF SECTION 3

Remark B.1 (Homogeneous parameters)**.** One important case arises when some of the components of $\beta(y)$ do not vary with y. This case is related with homogeneity on quantile effects with respect to the covariates and quantile index. Assume that $y \mapsto \Phi(-x'\beta(y))$ is one-to-one with inverse (quantile) function $u \mapsto Q_{Y^*}(u \mid x)$. Then, $\Phi(-x'\beta(y)) = \Phi(-x_1'\beta_1(y) - x_2'\beta_2)$ for all $y \in \mathbb{R}$ if and only if $\partial Q_{\widetilde{Y}^*}(u \mid x)/\partial x_2 = \beta_2$ for all $u \in [0,1]$, where $Q_{\widetilde{Y}^*}(u \mid x)$ is the u-quantile of $\widetilde{Y}^* = -X_1'\beta_1(Y)$ conditional on $X = x$. The only if part follows from $\mathrm{P}(\widetilde{Y}^* \leqslant \widetilde{y} \mid X = x) = \Phi(\widetilde{y} - x_2'\beta_2)$ as $\Phi(-X_1'\beta_1(Y) - X_2'\beta_2) \mid X \sim U(0,1)$, so that $Q_{\widetilde{Y}^*}(u \mid x) = \Phi^{-1}(u) + x_2'\beta_2$. The if part follows from $Q_{\widetilde{Y}^*}(u \mid x) = x_2'\beta_2 + \Phi^{-1}(u)$, so that $\mathrm{P}(Y \leqslant y \mid X = x) = \mathrm{P}(\widetilde{Y}^* \leqslant -X_1'\beta_1(y) \mid X = x) = \Phi(-x_1'\beta_1(y) - x_2'\beta_2)$. For example, when X_1 only includes a constant, then the DR model corresponds to the transformation model $-\beta_1(Y) = X_2'\beta_2 + \Phi^{-1}(U)$, $U \mid X \sim U(0,1)$, where the covariates X_2 have homogeneous effects on the quantiles of $-\beta_1(Y)$. The HSM is a special case of the transformation model with $y \mapsto -\beta_1(y)$ being linear. □

APPENDIX C. DEFERRED DISCUSSIONS OF SECTIONS 4

Example C.1 (Sign Reversal in Selection Effects)**.** The sign of the selection effects might be different in the presence of covariates if the parameter variation changes the composition of the selected population. Consider the following simple example with only one covariate based on the wage application. Let the covariate be an indicator for high skills. Assume that high-skilled workers are relatively more likely to participate than low-skilled workers, there is no selection sorting on unobservables, which corresponds to $\rho(x'\delta(y)) = 0$ in the model, and the distribution of offered wages for high-skilled workers first-order stochastically dominates the same distribution for low-skilled workers. In this case increasing the probability of participation for high-skilled workers, which corresponds to increasing the component of π associated with the high-skill indicator in the model, both increases the overall probability of participation and shifts the distribution of observed wages to the right (increases quantiles), despite the lack of selection sorting. Intuitively, the distribution of observed wages is a mixture of the distribution of wages for employed high-skilled and low-skilled workers, and we are increasing the relative proportion of employed high-skilled workers. The opposite holds if the distribution of offered wages for high-skilled workers is first-order stochastically dominated by the same distribution for low-skilled workers. □

APPENDIX D. ASYMPTOTIC THEORY

We derive asymptotic theory for the estimators of the model parameters and functionals of interest.

D.1. **Limit distributions.** We first introduce some notation that is useful to state the assumptions that we make to derive the limit distribution of the estimators. Let $\widetilde{S}_1 := \partial_\pi L_1(\pi)$ and $\widetilde{S}_{2y} := \partial_{\theta_y} L_2(\theta_y, \pi)$ be the scores of the first and second steps in Algorithm 3.1 evaluated at the true parameter values, and

$H_1 := \mathrm{E}[\partial_{\pi\pi'} L_1(\pi)]$ and $H_{2y} := \mathrm{E}\left[\partial_{\theta_y\theta_y} L_2(\theta_y, \pi)\right]$ be the corresponding expected Hessians. Let

$$\Sigma_{\theta_y\theta_{\tilde{y}}} := H_{2y}^{-1}\left\{n\mathrm{E}\left[\widetilde{S}_{2y}\widetilde{S}_{2\tilde{y}}'\right] - J_{21y}H_1^{-1}J_{21\tilde{y}}'\right\}H_{2\tilde{y}}^{-1}, \tag{D.22}$$

where $J_{21y} := \mathrm{E}\left[\partial_{\theta_y\pi'} L_2(\theta_y, \pi)\right]$, $d_\pi := \dim \pi$, and $d_\theta := \dim \theta_y$.

Assumption 3 (DR Estimator with Sample Selection)**.** *(1) Random sampling: $\{(D_i^*, Y_i^*, Z_i)\}_{i=1}^n$ is a sequence of independent and identically distributed copies of (D^*, Y^*, Z). We observe $D = 1(D^* > 0)$ and $Y = Y^*$ if $D = 1$. (2) Model: the distribution of (D^*, Y^*) conditional on Z follows the DR model* (3.11)*. (3) The support of Z, $\mathscr{Z}$, is a compact set. (4) The set $\mathscr{Y}$ is either finite or a bounded interval. In the second case, the density function of Y conditional on X and $D = 1$, $f_{Y|X,D}(y \mid x, 1)$, exists, is uniformly bounded above, and is uniformly continuous in (y, x) on $\mathscr{Y} \times \mathscr{X}_1$, where $\mathscr{X}_1$ is the support of X conditional on $D = 1$. (5) Identification and non-degeneracy: the equations $\mathrm{E}[\partial_\pi L_1(\tilde{\pi})] = 0$ and $\mathrm{E}[\partial_{\theta_y} L_2(\tilde{\theta}_y, \tilde{\pi})] = 0$ posses a unique solution at $(\tilde{\pi}, \tilde{\theta}_y) = (\pi, \theta_y)$ that lies in the interior of a compact set $\Pi \times \Theta \subset \mathbb{R}^{d_\pi + d_\theta}$ for all $y \in \mathscr{Y}$; and the matrices H_1, H_{2y} and $\Sigma_{\theta_y\theta_y}$ are nonsingular for each $y \in \mathscr{Y}$.*

Part (1) is a standard condition about the sampling and selection process, which is designed for cross sectional data. Part (2) imposes the semi-parametric DR model on the LGR of the conditional distribution of (D^*, Y^*) at $d = 0$. Part (3) imposes some compactness conditions, which can be generalized at the cost of more complicated proofs. Part (4) covers continuous, discrete and mixed continuous-discrete outcomes. Part (5) imposes directly identification and that the variance-covariance matrix of the first-step estimator and the covariance function of the second-step estimator are well-behaved. Note that H_1, H_{2y} and J_{21y} are finite by Part (3). More primitive conditions for part (5) can be found in the conditional maximum likelihood literature, e.g., Newey and McFadden (1986).

The main result of this section is a functional central limit theorem for $\widehat{\theta}_y$. Let $\ell^\infty(\mathscr{Y})$ be the set of bounded functions on $\mathscr{Y}$, and $\rightsquigarrow$ denote weak convergence (in distribution).

Theorem D.1 (FCLT for $\widehat{\theta}_y$)**.** *Under Assumption 3,*

$$\sqrt{n}(\widehat{\pi} - \pi) = -H_1^{-1}\widetilde{S}_1 + o_P(1) \rightsquigarrow Z_\pi \sim \mathscr{N}(0, -H_1^{-1}), \text{ in } \mathbb{R}^{d_\pi}$$

and

$$\sqrt{n}(\widehat{\theta}_y - \theta_y) = -H_{2y}^{-1}\sqrt{n}\left(\widetilde{S}_{2y} - J_{21y}H_1^{-1}\widetilde{S}_1\right) + o_P(1) \rightsquigarrow Z_{\theta_y} \text{ in } \ell^\infty(\mathscr{Y})^{d_\theta},$$

where $y \mapsto Z_{\theta_y}$ is a zero-mean Gaussian process with uniformly continuous sample paths and covariance function $\Sigma_{\theta_y\theta_{\tilde{y}}}$, $y, \tilde{y} \in \mathscr{Y}$, defined in (D.22)*.*

Remark D.1 (Comparison with Chernozhukov, Fernández-Val, and Melly (2013))**.** The asymptotic distribution in Theorem D.1 for $\widehat{\theta}(y)$ does not follow from the theory of Chernozhukov, Fernández-Val, and Melly (2013) for DR-estimators without sample selection. This theory does not cover two-step M-estimators with an objective function that is not concave. We rely on compactness of the parameter space instead of concavity. □

The first order term in the limit of $\sqrt{n}(\widehat{\theta}_y - \theta_y)$ is the sample average of the influence function of $\widehat{\theta}_y$. We construct an estimator of the covariance function $\Sigma_{\theta_y\theta_{\tilde{y}}}$ based on this function. Thus, we form

$$\widehat{\Sigma}_{\theta_y\theta_{\tilde{y}}} = n^{-2}\sum_{i=1}^{n}\widehat{\psi}_i(\widehat{\theta}_y,\widehat{\pi})\widehat{\psi}_i(\widehat{\theta}_{\tilde{y}},\widehat{\pi})'. \tag{D.23}$$

Here, $\widehat{\psi}_i$ is an estimator of the influence function of $\widehat{\theta}_y$,

$$\widehat{\psi}_i(t,c) = -\widehat{H}_{2y}(t,c)^{-1}\left(S_{2yi}(t,c) - \widehat{J}_{21y}(t,c)\widehat{H}_1(c)^{-1}S_{1i}(c)\right), \tag{D.24}$$

where $S_{1i}(c)$ and $S_{2iy}(t,c)$ are the individual scores of the first and second steps of Algorithm 3.1,

$$\begin{aligned}
S_{1i}(c) &:= \partial_c L_{1i}(c),\ L_{1i}(c) := D_i\log\Phi(Z_i'c) + (1-D_i)\log\Phi(-Z_i'c),\\
S_{2yi}(t,c) &:= \partial_t L_{2yi}(t,c),\ t=(b,d)\\
L_{2yi}(t,c) &:= D_i\left[I_{yi}\log\Phi_2\left(-X_i'b, Z_i'c; -\rho(x'd)\right) + (1-I_{yi})\log\Phi_2\left(X_i'b, Z_i'c; \rho(x'd)\right)\right],
\end{aligned}$$

and

$$\widehat{H}_1(c) := \partial_{cc'}L_1(c),\ \widehat{H}_{2y}(t,c) := \partial_{tt'}L_2(t,c),\ \widehat{J}_{21y}(t,c) := \partial_{tc'}L_2(t,c),$$

are estimators of H_1, H_{2y}, and J_{21y} when evaluated at $c=\widehat{\pi}$ and $t=\widehat{\theta}_y$.

We now establish a functional central limit theorem for the estimators of functionals of the model parameters. This result is based on expressing the functional as a suitable operator of the model parameters and using the functional delta method (van der Vaart and Wellner, 1996, Chapter 3.9). To present the result in a concise manner, we consider a generic functional

$$u \mapsto \Delta_u = \varphi_u(\pi,\theta_\cdot,F_Z),$$

where $u\in\mathscr{U}$, a totally bounded metric space, and φ_u is an operator that maps $\mathbb{D}_\Delta$ to the set $\ell^\infty(\mathscr{U})$, where $\Delta_\cdot$ takes values. Here $\mathbb{D}_\Delta$ denotes the space for the parameter tuple $(\pi,\theta_\cdot,F_Z)$; this space is not stated here explicitly, but is restricted by the regularity conditions of the previous section. Here we identify F_Z with an integral operator $f\mapsto\int f(z)dF_Z(z)$ taking values in $\ell^\infty(\mathscr{F})$ that acts on a Donsker set of bounded measurable functions $\mathscr{F}$, which includes indicators of rectangular sets; see Chernozhukov, Fernández-Val, and Melly (2013) and examples below. The parameter space $\mathbb{D}_\Delta$ is a subset of a normed space $\mathbb{D} := \mathbb{R}^{d_\pi}\times\ell^\infty(\mathscr{Y})^{d_\theta}\times\ell^\infty(\mathscr{F})$. In this notation, the plug-in estimator of the functional Δ_u is

$$\widehat{\Delta}_u = \varphi_u(\widehat{\pi},\widehat{\theta}_y,\widehat{F}_Z),$$

where $\widehat{\pi}$ and $\widehat{\theta}_y$ are the estimators of the parameters defined in Algorithm 3.1 and $\widehat{F}_Z$ is the empirical distribution of Z.

We provide some examples. The distribution of the latent outcome is given by:

$$F_{Y^*}(y) = \varphi_y(\pi,\theta_y,F_Z) = \int\Phi(-x'\beta_y)dF_Z(z),$$

$\mathscr{F}$ contains $\{\Phi(-\cdot'\beta_y) : y \in \mathscr{Y}\}$ as well as the indicators of all rectangles in $\overline{\mathbb{R}}^{d_z}$, $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, $d_z = \dim Z$, and $\mathscr{U} = \mathscr{Y}$. The quantile function of the latent outcome is

$$Q_{Y^*}(\tau) = \varphi_\tau(\pi, \theta_y, F_Z) = \mathbf{Q}_\tau \mathbf{R} F_{Y^*},$$

$\mathscr{F}$ is the same as for the distribution of the latent outcome, $\mathscr{U}$ is a closed subset of $(0,1)$ including the quantile indexes of interest, $\mathbf{R}$ is the non-decreasing rearrangement operator, and $\mathbf{Q}_\tau$ is the left-inverse (quantile) operator. The distribution of the observed outcome is given by:

$$F_Y(y \mid D = 1) = \varphi_y(\pi, \theta_y, F_Z) = \frac{\int \Phi_2\left(-x'\beta(y), z'\pi; -\rho(x'\delta(y))\right) dF_Z(z)}{\int \Phi(z'\pi) dF_Z(z)},$$

$\mathscr{F}$ contains $\{\Phi_2\left(-\cdot'\beta(y), \cdot'\pi; -\rho(\cdot'\delta(y))\right) : y \in \mathscr{Y}\}$ as well as the indicators of all rectangles in $\overline{\mathbb{R}}^{d_z}$, and $\mathscr{U} = \mathscr{Y}$.

The following result is a corollary of Theorem D.1 by the functional delta method. Let $UC(\mathscr{Y}, \xi)$ be the set of functions on $\mathscr{Y}$ that are uniformly continuous with respect to ξ, a standard metric on $\mathbb{R}$, and $UC(\mathscr{F}, \lambda)$ be the set of functionals on $\mathscr{F}$ that are uniformly continuous with respect to λ, where $\lambda(f, \tilde{f}) = [\mathrm{P}(f - \tilde{f})^2]^{1/2}$ for any $f, \tilde{f} \in \mathscr{F}$.

Corollary D.1 (FCLT for $\widehat{\Delta}_u$)**.** *Suppose that Assumption 3 holds, and $(p, t_y, F) \mapsto \varphi_\cdot(p, t_y, F)$, from $\mathbb{D}_\Delta \subset \mathbb{D}$ to $\ell^\infty(\mathscr{U})$ is Hadamard differentiable at (π, θ_y, F_Z), tangentially to $\mathbb{R}^{d_\pi} \times UC(\mathscr{Y}, \xi)^{d_\theta} \times UC(\mathscr{F}, \lambda)$ with derivative $(p, t_y, F) \mapsto \varphi'_\cdot(p, t_y, F)$ that is defined and continuous on $\mathbb{R}^{d_\pi} \times \ell^\infty(\mathscr{Y})^{d_\theta} \times \ell^\infty(\mathscr{F})$. Then,*

$$\sqrt{n}(\widehat{\Delta}_u - \Delta_u) \rightsquigarrow Z_{\Delta_u} := \varphi'_u(Z_\pi, Z_{\theta_y}, Z_F) \text{ in } \ell^\infty(\mathscr{U}),$$

where Z_π and Z_{θ_y} are the random limits in Theorem D.1, Z_F is a tight F_Z-Brownian bridge, and $u \mapsto Z_{\Delta_u}$ is a tight zero-mean Gaussian process.

Remark D.2 (Hadamard Differentiable Functionals)**.** The distributions of the latent and observed outcome together with counterfactual distributions constructed thereof are examples of Hadamard differentiable functions. In the case of the latent outcome, the result follows from the Hadamard differentiability of the counterfactual operator in Chernozhukov, Fernández-Val, and Melly (2013). In the case of the observed outcome, the result follows from the differentiability of the counterfactual operator and the composition rule for Hadamard derivatives applied to the ratio of two functions. Quantile (left-inverse) functionals of these distributions are Hadamard differentiable under additional conditions that guarantee that the quantile operator is Hadamard differentiable. These include that the outcome variable be continuous with density bounded above and away from zero (Chernozhukov, Fernández-Val, and Galichon, 2010). Then the Hadamard differentiability of the quantile function follows from the composition rule for Hadamard derivatives. □

Remark D.3 (Inference on Quantile Functions)**.** There are two alternatives to construct confidence bands for quantile functions. The first approach is the standard method based on characterizing the limit distribution of the estimator of the quantile function using the delta method, which relies on the Hadamard differentiability of the inverse operator. As we mention in Remark D.2, this differentiability

requires additional conditions including that the outcome variable be continuous. The second approach applies to any type of outcome variable. It is based on the generic method of Chernozhukov, Fernández-Val, Melly, and Wüthrich (2016) that inverts confidence bands for distribution functions into confidence bands for quantile function. This method does not rely on the delta method and is therefore more robust to modeling assumptions and widely applicable. It has the shortcoming, however, that the bands might not be centered at the point estimate of the quantile function. We apply the second method to obtain most of the results in the empirical application. □

D.2. **Multiplier Bootstrap.** We make the following assumption about the bootstrap multipliers of Algorithm 3.2:

Assumption 4 (Multiplier Bootstrap)**.** *The multipliers* $(\omega_1, ..., \omega_n)$ *are i.i.d. draws from a random variable* $\omega \sim \mathcal{N}(0,1)$*, and are independent of* $\{(D_i^*, Y_i^*, Z_i)\}_{i=1}^n$ *for all* n*.*

Let

$$\widehat{\theta}_y^b = \widehat{\theta}_y + n^{-1} \sum_{i=1}^n \omega_i \ \widehat{\psi}_i(\widehat{\theta}_y, \widehat{\pi})$$

be the multiplier bootstrap version of $\widehat{\theta}_y$. We establish a functional central limit theorem for the bootstrap for $\widehat{\theta}_y$. Here we use $\rightsquigarrow_{\mathrm{P}}$ to denote bootstrap consistency, i.e. weak convergence conditional on the data in probability, which is formally defined in Appendix E.1.

Theorem D.2 (Bootstrap FCLT for $\widehat{\theta}_y$)**.** *Under the conditions of Theorem D.1 and Assumption 4,*

$$\sqrt{n}(\widehat{\theta}_y^b - \widehat{\theta}_y) \rightsquigarrow_{\mathrm{P}} Z_{\theta_y} \text{ in } \ell^\infty(\mathscr{Y})^{d_\theta},$$

where $y \mapsto Z_{\theta_y}$ *is the same Gaussian process as in Theorem D.1.*

The following result is a corollary of Theorem D.2 by the functional delta method for the bootstrap (van der Vaart and Wellner, 1996, Chapter 3.9). Let $\widehat{\Delta}_u^b = \varphi_u(\widehat{\pi}^b, \widehat{\theta}_y^b, \widehat{F}_Z^b)$, be the multiplier bootstrap version of $\widehat{\Delta}_u$ where

$$\widehat{\pi}^b = \widehat{\pi} - n^{-1} \sum_{i=1}^n \omega_i \ \widehat{H}_1(\widehat{\pi})^{-1} S_{1i}(\widehat{\pi}),$$

and $\widehat{F}_Z^b$ is the weighted empirical distribution of Z that uses $(1+\omega_1, \ldots, 1+\omega_n)$ as sampling weights.

Corollary D.2 (Bootstrap FCLT for $\widehat{\Delta}_u$)**.** *Suppose that the conditions of Corollary D.1 and Assumption 4 hold. Then,*

$$\sqrt{n}(\widehat{\Delta}_u^b - \widehat{\Delta}_u) \rightsquigarrow_{\mathrm{P}} Z_{\Delta_u} \text{ in } \ell^\infty(\mathscr{U}),$$

where Z_{Δ_u} *is the same process as in Corollary D.1.*

APPENDIX E. PROOFS OF SECTION D

E.1. **Notation.** We adopt the standard notation in the empirical process literature, e.g. van der Vaart and Wellner (1996),

$$\mathbb{E}_n[f] = \mathbb{E}_n[f(A)] = n^{-1}\sum_{i=1}^{n} f(A_i),$$

and

$$\mathbb{G}_n[f] = \mathbb{G}_n[f(A)] = n^{-1/2}\sum_{i=1}^{n}(f(A_i) - \mathrm{E}[f(A)]).$$

When the function $\widehat{f}$ is estimated, the notation should interpreted as:

$$\mathbb{G}_n[\widehat{f}\,] = \mathbb{G}_n[f]\,|_{f=\widehat{f}} \text{ and } \mathrm{E}[\widehat{f}\,] = \mathrm{E}[f]\,|_{f=\widehat{f}}\,.$$

We also follow the notation and definitions in van der Vaart and Wellner (1996) of bootstrap consistency. Let D_n denote the data vector and E_n be the vector of bootstrap weights. Consider the random element $Z_n^b = Z_n(D_n, E_n)$ in a normed space $\mathbb{Z}$. We say that the bootstrap law of Z_n^b consistently estimates the law of some tight random element Z and write $Z_n^b \rightsquigarrow_{\mathrm{P}} Z$ in $\mathbb{Z}$ if

$$\sup_{h\in \mathrm{BL}_1(\mathbb{Z})} \left|\mathrm{E}^b h\left(Z_n^b\right) - \mathrm{E}h(Z)\right| \to_{\mathrm{P}^*} 0, \tag{E.25}$$

where $\mathrm{BL}_1(\mathbb{Z})$ denotes the space of functions with Lipschitz norm at most 1, E^b denotes the conditional expectation with respect to E_n given the data D_n, and $\to_{\mathrm{P}^*}$ denotes convergence in (outer) probability.

We use the Z-process framework described in Appendix E.1 of Chernozhukov, Fernández-Val, and Melly (2013). To set-up the problem in terms of this framework, we need to introduce some notation. Let $W := (Z, D, YD)$ denote all the observed variables and $\xi_y := (\pi', \theta_y')'$ be a vector with the model parameters of the first and second steps. Let

$$\varphi_{y,\xi}(W) := \begin{bmatrix} S_{1,\xi}(W) \\ S_{2y,\xi}(W) \end{bmatrix} = \begin{bmatrix} \frac{\partial \ell_{1,\xi}(W)}{\partial \pi} \\ \frac{\partial \ell_{2y,\xi}(W)}{\partial \theta_y} \end{bmatrix}$$

where

$$\begin{aligned} \ell_{1,\xi}(W) &:= D\log\Phi(Z'\pi) + (1-D)\log\Phi(-Z'\pi), \\ \ell_{2y,\xi}(W) &:= D[I_y \log\Phi_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right) + (1-I_y)\log\Phi_2\left(X'\beta(y), Z'\pi; \rho(X'\delta(y))\right)], \end{aligned}$$

be the scores of the first and second steps; and

$$J(y) = \mathrm{E}\left[\frac{\partial \varphi_{y,\xi}(W)}{\partial \xi'}\right] = \begin{bmatrix} H_1 & 0 \\ J_{21y} & H_{2y} \end{bmatrix} \tag{E.26}$$

be the expected Hessian evaluated at the true value of ξ_y. We provide more explicit expressions for the score and expected Hessian in Appendix E.5. Note that

$$J^{-1}(y) = \begin{bmatrix} H_1^{-1} & 0 \\ -H_1^{-1}J_{21y}H_{2y}^{-1} & H_{2y}^{-1} \end{bmatrix} \tag{E.27}$$

by the inverse of the partitioned inverse formula, and

$$\mathrm{E}[\varphi_{y,\xi}(W)\varphi_{\tilde{y},\xi}(W)'] = \begin{bmatrix} \mathrm{E}[S_{1,\xi}(W)S_{1,\xi}(W)'] & 0 \\ 0 & \mathrm{E}[S_{2y,\xi}(W)S_{2\tilde{y},\xi}(W)'] \end{bmatrix} \tag{E.28}$$

because $\mathrm{E}[S_{1,\xi}(W)S_{2y,\xi}(W)'] = 0$ for all $y \in \mathscr{Y}$.

E.2. **Auxiliary Results.** We start by providing sufficient conditions that are useful to verify Condition Z in Chernozhukov, Fernández-Val, and Melly (2013). They are an alternative to Lemma E.1 of Chernozhukov, Fernández-Val, and Melly (2013), where we replace the requirement that the function $\xi \mapsto \Psi(\xi, y) := \mathrm{E}[\varphi_{y,\xi}(W)]$ is the gradient of a convex function by compactness of the parameter space for ξ_y and an identification condition.[8]

Lemma E.1 (Simple sufficient condition for Z)**.** *Suppose that Ξ is a compact subset of $\mathbb{R}^{d_\xi}$, and $\mathscr{Y}$ is a compact interval in $\mathbb{R}$. Let $\mathscr{I}$ be an open set containing $\mathscr{Y}$. Suppose that (a) $\Psi : \Xi \times \mathscr{I} \mapsto \mathbb{R}^{d_\xi}$ is continuous, and $\xi \mapsto \Psi(\xi, y)$ possesses a unique zero at ξ_y that is in the interior of Ξ for each $y \in \mathscr{Y}$, (b) for each $y \in \mathscr{Y}$, $\Psi(\xi_y, y) = 0$, (c) $\frac{\partial}{\partial(\xi', y)}\Psi(\xi, y)$ exists at (ξ_y, y) and is continuous at (ξ_y, y) for each $y \in \mathscr{Y}$, and $\dot{\Psi}_{\xi_y,y} := \frac{\partial}{\partial \xi'}\Psi(\xi, y)|_{\xi_y}$ obeys $\inf_{y\in\mathscr{Y}} \inf_{\|h\|=1} \|\dot{\Psi}_{\xi_y,y}h\| > c_0 > 0$. Then Condition Z of Chernozhukov, Fernández-Val, and Melly (2013) holds and $y \mapsto \xi_y$ is continuously differentiable.*

Proof of Lemma E.1. We restate the statement of Condition Z of Chernozhukov, Fernández-Val, and Melly (2013) with our notation for the reader's reference.

CONDITION Z. *Let $\mathscr{Y}$ be a compact set of some metric space, and Ξ be an arbitrary subset of $\mathbb{R}^{d_\xi}$. Assume (i) for each $y \in \mathscr{Y}$, $\Psi(\cdot, y) : \Xi \mapsto \mathbb{R}^{d_\xi}$ possesses a unique zero at ξ_y, and, for some $\delta > 0$, $\mathscr{N} := \cup_{y\in\mathscr{Y}} B_\delta(\xi_y)$ is a compact subset of $\mathbb{R}^{d_\xi}$ contained in Ξ, (ii) the inverse of $\Psi(\cdot, y)$ defined as $\Psi^{-1}(x, y) := \{\xi \in \Xi : \Psi(\xi, y) = x\}$ is continuous at $x = 0$ uniformly in $y \in \mathscr{Y}$ with respect to the Hausdorff distance, (iii) there exists $\dot{\Psi}_{\xi_y,y}$ such that $\lim_{t\searrow 0} \sup_{y\in\mathscr{Y},\|h\|=1} |t^{-1}[\Psi(\xi_y + th, y) - \Psi(\xi_y, y)] - \dot{\Psi}_{\xi_y,y}h| = 0$, where $\inf_{y\in\mathscr{Y}} \inf_{\|h\|=1} \|\dot{\Psi}_{\xi_y,y}h\| > 0$, and (iv) the maps $y \mapsto \xi_y$ and $y \mapsto \dot{\Psi}_{\xi_y,y}$ are continuous.*

The first part of Z(i) follows immediately from condition (a). The verifications of the second part of Z(i), Z(iii) and Z(iv) are omitted because they follow by the same argument as in the proof of Lemma E.1 of Chernozhukov, Fernández-Val, and Melly (2013).

To show Condition Z(ii), we need to verify that for any $x_t \to 0$ such that $x_t \in \Psi(\Xi, y)$, we have that $d_H(\Psi^{-1}(x_t, y), \Psi^{-1}(0, y)) \to 0$, where d_H is the Hausdorff distance, uniformly in $y \in \mathscr{Y}$. Suppose by contradiction that this is not true, then there is (x_t, y_t) with $x_t \to 0$ and $y_t \in \mathscr{Y}$ such that $d_H(\Psi^{-1}(x_t, y_t), \Psi^{-1}(0, y_t)) \not\to 0$. By compactness of $\mathscr{Y}$, we can select a further subsequence (x_k, y_k) such that $y_k \to y$, where $y \in \mathscr{Y}$. We have that $\Psi^{-1}(0, y) = \xi_y$ is continuous in $y \in \mathscr{Y}$, so we must have

$$d_H(\Psi^{-1}(x_k, y_k), \Psi^{-1}(0, y)) \not\to 0.$$

[8] We adapt the notation of Chernozhukov, Fernández-Val, and Melly (2013) to our problem by using y, $\mathscr{Y}$, ξ_y, d_ξ and Ξ in place of u, $\mathscr{U}$, $\theta_0(u)$, p, and Θ.

Hence, by compactness of Ξ, there is a further subsequence $u_l \in \Psi^{-1}(x_l, y_l)$ with $u_l \to u$ in Ξ, such that $u \neq \Psi^{-1}(0, y) = \xi_y$, and such that $x_l = \Psi(u_l, y_l) \to 0$. But, by continuity $\Psi(u_l, y_l) \to \Psi(u, y) \neq 0$ since $u \neq \Psi^{-1}(0, y)$, yielding a contradiction. □

E.3. **Proof of Theorem D.1.** We only consider the case where $\mathscr{Y}$ is a compact interval of $\mathbb{R}$. The case where $\mathscr{Y}$ is a finite set is simpler. The proof follows the same steps as the proof of Theorem 5.2 of Chernozhukov, Fernández-Val, and Melly (2013) for the DR-estimator without sample selection using Lemma E.1 in place of Lemma E.1 of Chernozhukov, Fernández-Val, and Melly (2013). Let $\Psi(\xi, y) = P[\varphi_{y,\xi}]$ and $\widehat{\Psi}(\xi, y) = P_n[\varphi_{y,\xi}]$, where P_n is the empirical measure and P is the corresponding probability measure. From the first order conditions, the two-step estimator obeys $\widehat{\xi}_y = \phi(\widehat{\Psi}(\cdot, y), 0)$ for each $y \in \mathscr{Y}$, where ϕ is the Z-map defined in Appendix E.1 of Chernozhukov, Fernández-Val, and Melly (2013). The random vector $\widehat{\xi}_y$ is the estimator of $\xi_y = \phi(\Psi(\cdot, y), 0)$ in the notation of this framework. Then, by step 1 below,

$$\sqrt{n}(\widehat{\Psi} - \Psi) \rightsquigarrow Z_\Psi \text{ in } \ell^\infty(\mathscr{Y} \times \mathbb{R}^{d_\xi})^{d_\xi},\ Z_\Psi(y, \xi) = \mathbb{G}\varphi_{y,\xi},$$

where $d_\xi := \dim \xi_y$, $\mathbb{G}$ is a P-Brownian bridge, and Z_Ψ has continuous paths a.s. Step 2 verifies the conditions of Lemma E.1 for $\dot{\Psi}(\xi_y, y) = J(y)$, the Hessian matrix defined in (E.26), which also implies that $y \mapsto \xi_y$ is continuously differentiable in the interval $\mathscr{Y}$. Then, by Lemma E.2 of Chernozhukov, Fernández-Val, and Melly (2013), the map ϕ is Hadamard differentiable with derivative map $(\psi, 0) \mapsto -J^{-1}\psi$ at $(\Psi, 0)$. Therefore, we can conclude by the functional delta method that

$$\sqrt{n}(\widehat{\xi}_y - \xi_y) \rightsquigarrow Z_{\xi_y} := -J^{-1}(y) Z_\Psi(y, \xi_y) \text{ in } \ell^\infty(\mathscr{Y})^{d_\xi}, \tag{E.29}$$

where $y \mapsto Z_{\xi_y}$ has continuous paths a.s.

Step 1 (Donskerness)**.** We verify that $\mathscr{G} = \{\varphi_{y,\xi}(W) : (y, \xi) \in \mathscr{Y} \times \mathbb{R}^{d_\xi}\}$ is P-Donsker with a square-integrable envelope. By inspection of the expression of $\varphi_{y,\xi}(W) = [S_{1,\xi}(W)', S_{2y,\xi}(W)']'$ in Appendix E.5, $\varphi_{y,\xi}(W)$ is a Lipschitz transformation of VC functions with Lipschitz coefficient bounded by $c\|Z\|$ for some constant c and envelope function $c\|Z\|$, which is square-integrable. Hence $\mathscr{G}$ is P-Donsker by Example 19.9 in van der Vaart (1998).

Step 2 (Verification of the Conditions of Lemma E.1)**.** Conditions (a) and (b) are immediate by Assumption 3. To verify (c), note that for $(\tilde{\xi}, \tilde{y})$ in the neighborhood of (ξ_y, y),

$$\frac{\partial \Psi(\tilde{\xi}, \tilde{y})}{\partial(\tilde{\xi}', \tilde{y})} = [J(\tilde{\xi}, \tilde{y}), R(\tilde{\xi}, \tilde{y})],$$

where

$$R(\tilde{\xi}, \tilde{y}) = -\mathrm{E}\left\{ \begin{array}{c} 0 \\ f_{Y|Z,D}(\tilde{y} \mid Z, 1)\Phi_\pi(Z)\Phi_{\tilde{\pi}}(Z) \begin{bmatrix} G_{2,\tilde{\xi}}(Z) \\ G_{3,\tilde{\xi}}(Z) \end{bmatrix} \otimes X \end{array} \right\},$$

for $\tilde{\xi} = (\tilde{\pi}', \tilde{\beta}', \tilde{\rho}')'$, and

$$J(\tilde{\xi}, \tilde{y}) = \begin{bmatrix} J_{11}(\tilde{\xi}, \tilde{y}) & J_{12}(\tilde{\xi}, \tilde{y}) \\ J_{21}(\tilde{\xi}, \tilde{y}) & J_{22}(\tilde{\xi}, \tilde{y}) \end{bmatrix},$$

for

$$J_{11}(\tilde{\xi},\tilde{y}) = \mathrm{E}\left[\{g_1(Z'\tilde{\pi})(D-\Phi_{\tilde{\pi}}(Z)) - G_1(Z'\tilde{\pi})\phi(Z'\tilde{\pi})\}ZZ'\right],$$

with $g_1(u) = dG_1(u)/du$; $J_{12}(\tilde{\xi},\tilde{y}) = 0$;

$$J_{21}(\tilde{\xi},\tilde{y}) = \mathrm{E}\left\{[\Phi_\pi(Z)\Phi^{\nu}_{2,\tilde{\xi}}(Z) - \phi(Z'\pi)\Phi_{2,\xi_{\tilde{y}}}(Z)]\begin{bmatrix} G_{2,\tilde{\xi}}(Z) \\ G_{3,\tilde{\xi}}(Z)\end{bmatrix}\otimes XZ'\right\}$$
$$+\mathrm{E}\left\{(\Phi_\pi(Z)\Phi_{2,\tilde{\xi}}(Z) - \Phi_{\tilde{\pi}}(Z)\Phi_{2,\xi_{\tilde{y}}}(Z))\begin{bmatrix} G^{\nu}_{2,\tilde{\xi}}(Z) \\ \rho'(X'\tilde{\delta})G^{\nu}_{3,\tilde{\xi}}(Z)\end{bmatrix}\otimes XZ'\right\},$$

with $G^{\nu}_{j,\tilde{\xi}}(Z) := G^{\nu}_j\left(-X'\tilde{\beta}, Z'\tilde{\pi}; -\rho(X'\tilde{\delta})\right)$ and $G^{\nu}_j(\mu,\nu;\rho) = \partial G_j(\mu,\nu;\rho)/\partial\nu$ for $j\in\{2,3\}$; and

$$J_{22}(\tilde{\xi},\tilde{y}) = -\mathrm{E}\left\{\Phi_\pi(Z)\begin{bmatrix} \Phi^{\mu}_{2,\tilde{\xi}}(Z)G_{2,\tilde{\xi}}(Z) & \Phi^{\rho}_{2,\tilde{\xi}}(Z)G_{2,\tilde{\xi}}(Z) \\ \Phi^{\mu}_{2,\tilde{\xi}}(Z)\rho'(X'\tilde{\delta})G_{3,\tilde{\xi}}(Z) & \Phi^{\rho}_{2,\tilde{\xi}}(Z)\rho'(X'\tilde{\delta})G_{3,\tilde{\xi}}(Z)\end{bmatrix}\otimes XX'\right\}$$
$$+\mathrm{E}\left\{(\Phi_\pi(Z)\Phi_{2,\tilde{\xi}}(Z) - \Phi_{\tilde{\pi}}(Z)\Phi_{2,\xi_{\tilde{y}}}(Z))\begin{bmatrix} G^{\mu}_{2,\tilde{\xi}}(Z) & G^{\rho}_{2,\tilde{\xi}}(Z) \\ \rho'(X'\tilde{\delta})G^{\mu}_{3,\tilde{\xi}}(Z) & \rho'(X'\tilde{\delta})^2G^{\rho}_{3,\tilde{\xi}}(Z) + \rho''(X'\tilde{\delta})G_{3,\tilde{\xi}}(Z)\end{bmatrix}\otimes XX'\right\},$$

with $G^{a}_{j,\tilde{\xi}}(Z) := G^{a}_j\left(-X'\tilde{\beta}, Z'\tilde{\pi}; -\rho(X'\tilde{\delta})\right)$ and $G^{a}_j(\mu,\nu;\rho) = \partial G_j(\mu,\nu;\rho)/\partial a$ for $j\in\{2,3\}$ and $a\in\{\mu,\rho\}$. In the previous expressions we use some notation defined in Appendix E.5.

Both $(\tilde{\xi},\tilde{y}) \mapsto R(\tilde{\xi},\tilde{y})$ and $(\tilde{\xi},\tilde{y}) \mapsto J(\tilde{\xi},\tilde{y})$ are continuous at (ξ_y, y) for each $y\in\mathscr{Y}$. The computation above as well as the verification of the continuity follow from using the expressions of $\varphi_{y,\xi}$ in Appendix E.5, the dominated convergence theorem, and the following ingredients: (i) a.s. continuity of the map $(\tilde{\xi},\tilde{y}) \mapsto \partial\varphi_{\tilde{y},\tilde{\xi}}(W)/\partial\tilde{\xi}'$, (ii) domination of $\|\partial\varphi_{y,\xi}(W)/\partial\xi'\|$ by a square-integrable function $\|cZ\|$ for some constant c, (iii) a.s. continuity and uniform boundedness of the conditional density function $y \mapsto f_{Y|X,D}(y \mid X, 1)$ by Assumption 3, and (iv) $G_1(Z'\tilde{\pi})$, $G_{2,\tilde{\xi}}(Z)$ and $G_{3,\tilde{\xi}}(Z)$ being bounded uniformly on $\tilde{\xi}\in\mathbb{R}^{d_\xi}$, a.s. By assumption, $J(y) = J(\xi_y, y)$ is positive-definite uniformly in $y\in\mathscr{Y}$.

The expressions of the limit processes given in the theorem follow by partitioning $Z_{\xi_y} = (Z'_\pi, Z'_{\theta_y})'$ and using the expressions of $J^{-1}(y)$ and $\mathrm{E}[\varphi_{y,\xi}(W)\varphi_{\tilde{y},\xi}(W)']$ given in (E.27) and (E.28). □

E.4. **Proof of Theorem D.2.** Let $\widehat{\xi}^b_y := (\widehat{\pi}^{b\prime}, \widehat{\theta}^{b\prime}_y)'$. By definition of the multiplier bootstrap draw of the estimator

$$\sqrt{n}(\widehat{\xi}^b_y - \widehat{\xi}_y) = \mathbb{G}_n\omega^b\varphi_{y,\widehat{\xi}} = \mathbb{G}_n\omega^b\varphi_{y,\xi} + r_y,$$

where $\omega^b \sim N(0,1)$ independently of the data and $r_y := \mathbb{G}_n\omega^b(\varphi_{y,\widehat{\xi}} - \varphi_{y,\xi})$. Then the result follows from $\mathbb{G}_n\omega^b\varphi_{y,\xi} \rightsquigarrow_{\mathrm{P}} Z_{\xi_y}$ in step 3 and $r_y \rightsquigarrow_{\mathrm{P}} 0$ in step 4.

Step 3. Recall that $\varphi_{y,\xi}$ is P-Donsker by step 1 of the proof of Theorem D.1. Then, by $\mathrm{E}\omega^b = 0$, $\mathrm{E}(\omega^b)^2 = 1$ and the Conditional Multiplier Functional Central Limit Theorem (van der Vaart and Wellner, 1996, Theorem 2.9.6),

$$\mathbb{G}_n\omega^b\varphi_{y,\xi} \rightsquigarrow_{\mathrm{P}} Z_{\xi_y},$$

where Z_{ξ_y} is the same limit process as in (E.29).

Step 4. Note that $r_y \rightsquigarrow 0$ because $\varphi_{y,\xi}$ is P-Donsker and $\sqrt{n}(\widehat{\xi}_y - \xi_y) = O_{\mathrm{P}}(1)$ uniformly in $y \in \mathscr{Y}$ by Theorem D.1. To show that $r_y \rightsquigarrow_{\mathrm{P}} 0$, we use that this statement means that for any $\epsilon > 0$, $\mathrm{E}^b 1(\|r_y\|_2 > \epsilon) = o_{\mathrm{P}}(1)$ uniformly in $y \in \mathscr{Y}$. Then, the result follows by the Markov inequality and

$$\mathrm{E}\mathrm{E}^b 1(\|r_y\|_2 > \epsilon) = \mathrm{P}(\|r_y\|_2 > \epsilon) = o(1),$$

uniformly in $y \in \mathscr{Y}$, where the latter holds by the Law of Iterated Expectations and $r_y \rightsquigarrow 0$.

□

E.5. **Expressions of the Score and Expected Hessian.**

E.6. **Score.** Let $\Phi_\pi(Z) := \Phi(Z'\pi)$ and $\Phi_{2,\xi_y}(Z) := \Phi_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$. Note that by the properties of the standard bivariate normal distribution $\Phi_2\left(X'\beta(y), Z'\pi; \rho(X'\delta(y))\right) = \Phi_\pi(Z) - \Phi_{2,\xi_y}(Z)$. Then, straighforward calculations yield

$$S_{1,\xi}(W) = \frac{\partial \ell_{1,\xi}(W)}{\partial \pi} = G_1(Z'\pi)[D - \Phi_\pi(Z)]Z,$$

where $G_1(u) = \phi(u)/[\Phi(u)\Phi(-u)]$, and

$$S_{2y,\xi}(W) = \frac{\partial \ell_{2y,\xi}(W)}{\partial \theta_y} = D(\Phi_{2,\xi_y}(Z) - \Phi_\pi(Z) I_y) \begin{bmatrix} G_{2,\xi_y}(Z) \\ \rho'(X'\delta(y)) G_{3,\xi_y}(Z) \end{bmatrix} \otimes X,$$

where

$$G_{2,\xi_y}(Z) := G_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right), G_{3,\xi_y}(Z) := G_3\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$$

with

$$G_2(\mu, \nu; \rho) = \frac{\Phi_2^\mu(\mu, \nu; \rho)}{\Phi_2(\mu, \nu; \rho)[\Phi(\nu) - \Phi_2(\mu, \nu; \rho)]}, \quad G_3(\mu, \nu; \rho) = \frac{\Phi_2^\rho(\mu, \nu; \rho)}{\Phi_2(\mu, \nu; \rho)[\Phi(\nu) - \Phi_2(\mu, \nu; \rho)]},$$

for

$$\Phi_2^\mu(\mu, \nu; \rho) = \frac{\partial \Phi_2(\mu, \nu; \rho)}{\partial \mu} = \Phi\left(\frac{\nu - \rho\mu}{\sqrt{1-\rho^2}}\right)\phi(\mu), \tag{E.30}$$

and

$$\Phi_2^\rho(\mu, \nu; \rho) = \frac{\partial \Phi_2(\mu, \nu; \rho)}{\partial \rho} = \phi_2(\mu, \nu; \rho). \tag{E.31}$$

To show (E.30) and (E.31), start from the factorization

$$\Phi_2(\mu, \nu; \rho) = \int_{-\infty}^{\mu} \Phi\left(\frac{\nu - \rho v}{\sqrt{1-\rho^2}}\right)\phi(v)\,dv.$$

Then, (E.30) follows from taking the partial derivative with respect to μ using the Leibniz integral rule. Taking the partial derivative with respect to ρ yields

$$\begin{aligned}\frac{\partial\Phi_2(\mu,\nu;\rho)}{\partial\rho} &= \int_{-\infty}^{\mu}\phi\left(\frac{\nu-\rho\upsilon}{\sqrt{1-\rho^2}}\right)\frac{\rho\nu-\upsilon}{(1-\rho^2)^{\frac{3}{2}}}\phi(\upsilon)d\upsilon \\ &= \int_{-\infty}^{\mu}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(\nu-\rho\upsilon)^2}{2(1-\rho^2)}\right]\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{\upsilon^2}{2}\right]\frac{\rho\nu-\upsilon}{(1-\rho^2)^{\frac{3}{2}}}d\upsilon \\ &= \int_{-\infty}^{\mu}\frac{\rho\nu-\upsilon}{2\pi(1-\rho^2)^{\frac{3}{2}}}\exp\left[-\frac{\nu^2-2\rho\upsilon\nu+\upsilon^2}{2(1-\rho^2)}\right]d\upsilon \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{\nu^2-2\rho\mu\nu+\mu^2}{2(1-\rho^2)}\right] = \phi_2(\mu,\nu;\rho)\end{aligned}$$

E.7. **Expected Hessian.** Straighforward calculations yield

$$H_1 = \mathrm{E}\left[\frac{\partial\ell_{1,\xi}(W)}{\partial\pi\partial\pi'}\right] = -\mathrm{E}\left[G_1(Z'\pi)\phi(Z'\pi)ZZ'\right],\ \ \mathrm{E}\left[\frac{\partial\ell_{1,\xi}(W)}{\partial\pi\partial\theta_y'}\right] = 0,$$

$$J_{21y} = \frac{\partial\ell_{2y,\xi}(W)}{\partial\theta_y\partial\pi'} = \mathrm{E}\left\{[\Phi_\pi(Z)\Phi^{\nu}_{2,\xi_y}(Z)-\phi(Z'\pi)\Phi_{2,\xi_y}(Z)]\begin{bmatrix} G_{2,\xi_y}(Z) \\ \rho'(X'\delta(y))G_{3,\xi_y}(Z)\end{bmatrix}\otimes XZ'\right\},$$

where $\Phi^{\nu}_{2,\xi_y}(Z) = \Phi^{\nu}_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$ with

$$\Phi^{\nu}_2(\mu,\nu;\rho) = \frac{\partial\Phi_2(\mu,\nu;\rho)}{\partial\nu} = \Phi\left(\frac{\mu-\rho\nu}{\sqrt{1-\rho^2}}\right)\phi(\nu),$$

by a symmetric argument to (E.30), and

$$H_{2y} = \frac{\partial\ell_{2y,\xi}(W)}{\partial\theta_y\partial\theta_y'} = -\mathrm{E}\left\{\Phi_\pi(Z)\begin{bmatrix}\Phi^{\mu}_{2,\xi_y}(Z)G_{2,\xi_y}(Z) & \Phi^{\rho}_{2,\xi_y}(Z)G_{2,\xi_y}(Z) \\ \Phi^{\mu}_{2,\xi_y}(Z)\rho'(X'\delta(y))G_{3,\xi_y}(Z) & \Phi^{\rho}_{2,\xi_y}(Z)\rho'(X'\delta(y))G_{3,\xi_y}(Z)\end{bmatrix}\otimes XX'\right\},$$

where

$$\begin{aligned}\Phi^{\mu}_{2,\xi_y}(Z) &:= \Phi^{\mu}_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right) \\ \Phi^{\rho}_{2,\xi_y}(Z) &:= \Phi^{\rho}_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right).\end{aligned}$$

Appendix F. Monte Carlo Simulation

We conduct a Monte Carlo simulation calibrated to the empirical application to study the properties of the estimation and inference methods in small samples. The data generating process is the HSM of Example 3.1 with the values of the covariates and parameters calibrated to the data for women in the last ten years of the sample (2004–2013). We do not use the entire dataset to speed up computation. We generate 500 artificial datasets and estimate the DR-model with the same specifications for the selection and outcome equations as in the empirical application and specification 1 for the selection sorting function, i.e. $\rho(x'\delta(y)) = \rho(y)$.

Figures 2, 3 and 4 report the biases, standard deviations and root mean square errors for the estimators of the coefficients of the college (age when ceasing school 21–22) and marital status indicators in the outcome equation, and $\rho(y)$ in the selection sorting function, as a function of the quantile indexes of the values of log real hourly wage in the data used in the calibration.[9] Although these coefficients are constant in the HSM, we do not impose this condition in the estimation. The estimates are obtained with Algorithm 3.1 replacing $\mathscr{Y}$ by a finite grid containing the sample quantiles of log real hourly wage with indexes $\{0.10, 0.11, \ldots, 0.90\}$ in the original subsample of women in the last ten years of the sample. All the results are in percentage of the true value of the parameter. As predicted by the asymptotic theory, the biases are all small relative to the standard deviations and root mean squared errors. The estimation error increases for all the coefficients as we move away from the median towards tail values of the outcome.

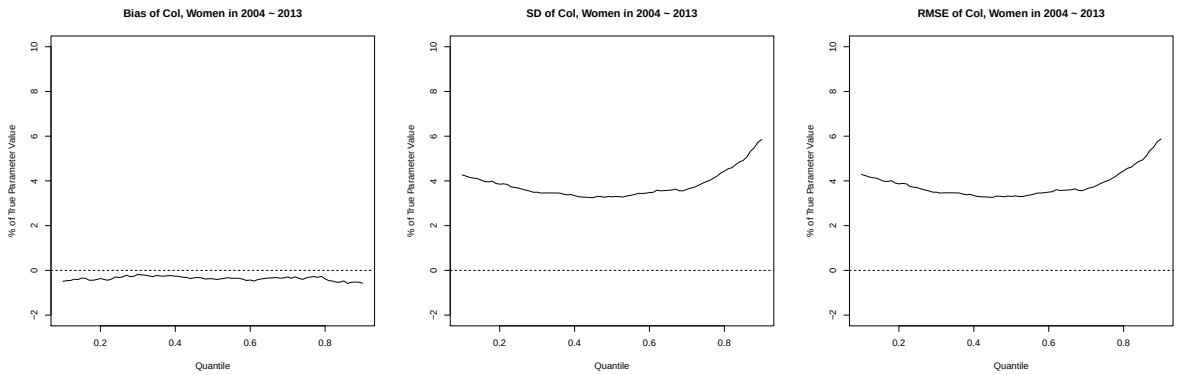

FIGURE 2. Bias, SD and RMSE for the coefficient of the college indicator in the outcome equation

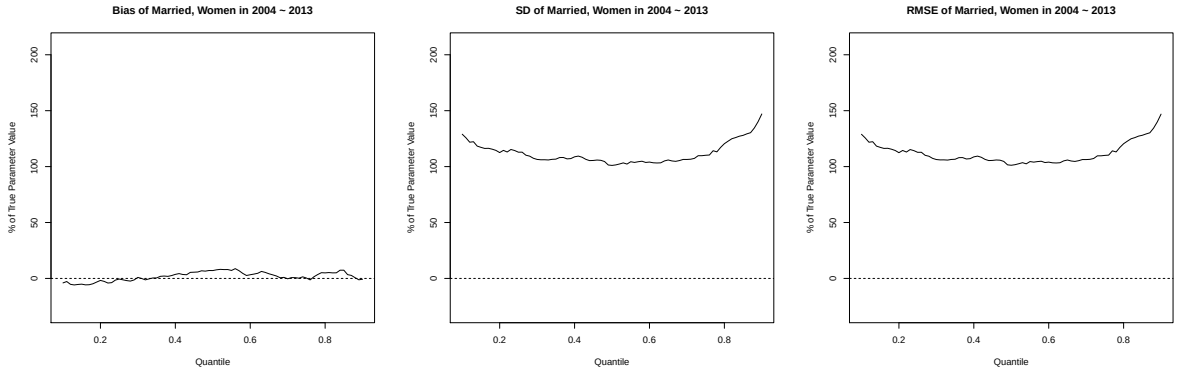

FIGURE 3. Bias, SD and RMSE for the coefficient of the marital status indicator in the outcome equation

[9]We find similar results for the other coefficients of the outcome equation. We do not report these reports for the sake of brevity.

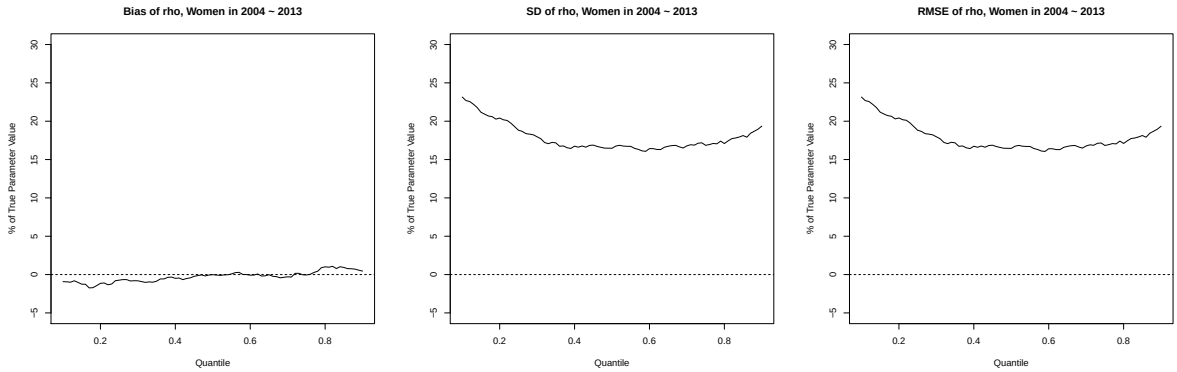

FIGURE 4. Bias, SD and RMSE for coefficient $\rho(y)$ in the selection sorting equation

Table 1 shows results on the finite sample properties of 95% confidence bands for the coefficients of the indicators of college and marital status in the outcome equation and $\rho(y)$ of the selection sorting function. The confidence bands are constructed by Algorithm 3.2 with $B = 200$ bootstrap repetitions and the same grid of values $\bar{\mathscr{Y}}$ as for the estimators. We report the average length of the confidence bands integrated over threshold values, the average value of the estimated critical values, and the empirical coverages of the confidence bands. For comparison, we also report the coverage of pointwise confidence bands using the normal distribution, i.e. with critical value equal to 1.96. The last row computes the ratio of the standard error averaged across simulations to the simulation standard deviation, integrated over threshold values. We find that the bands have coverages close to the nominal level. As expected, pointwise bands severely undercover the entire functions. The standard errors based on the asymptotic distribution provide a fair approximation to the sampling variability of the estimator.

TABLE 1. Properties of 95% Confidence Bands

	College	Married	$\rho(y)$
Average Length	0.38	0.16	0.35
Average Critical Value	2.91	2.89	2.88
Coverage uniform band (%)	96	98	96
Coverage pointwise band (%)	68	64	67
Average SE/SD	1.04	1.05	1.07

Notes: Nominal level of critical values is 95%. 500 simulations with 200 bootstrap draws.

APPENDIX G. WAGE DECOMPOSITIONS IN THE UK: ADDITIONAL RESULTS

G.1. **Descriptive statistics and background on U.K. labor market.** Table 2 reports means and standard deviations of all the variables used in the analysis. We report these statistics for the entire sample, and by employment status and gender. The overall employment rate is 74%. Women are 17% less likely

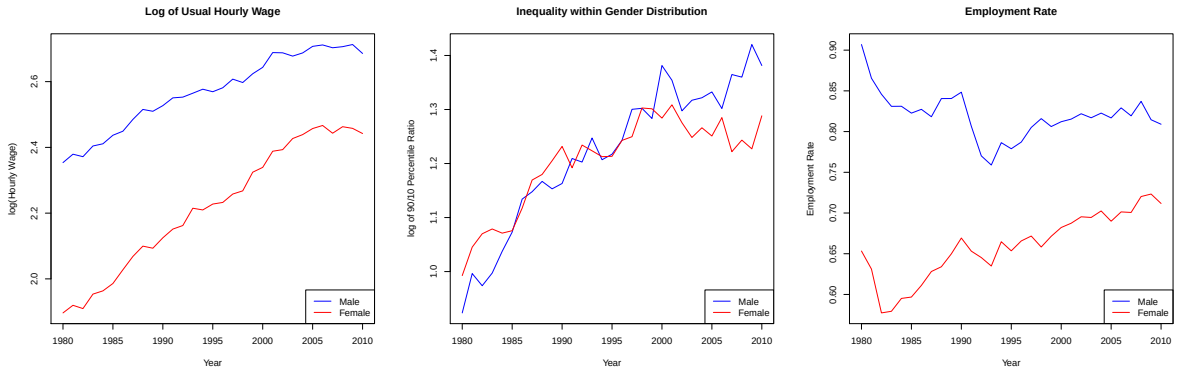

FIGURE 5. Trends in U.K. labor market 1978-2013 by gender: left panel reports the average of the log wage rate, the middle panel reports the 90-10 percentile spread of the log wage rate, and the right panel reports the employment rate

to be employed than men, and the unconditional gender wage gap is 33%. Overall, women and men are similar in terms of covariates. Both working men and women are relatively more highly educated, younger, and more likely to be married than their non-working counterparts. Having young children and high out-of-work benefits is negatively associated with employment for women but not for men.

Figure 5 provides some background on the U.K. labor market using our data. The left panel shows that over 36 years the average wages of working men and women have continuously grown and the unconditional gender wage gap has progressively narrowed from 46% to 24%. The middle panel indicates that the growth of average wage has come together with an increase in wage inequality for both working men and women until 2000. The positive trend in wage inequality has continued for men after 2000, but not for women. The right panel shows opposite trends in the employment rate for men and women, where the gender employment gap has steadily and sharply reduced from 34% to 8%.

G.2. **Additional Empirical Results.** This section includes the following empirical results omitted from the main text:

- Table 3: estimates and standard errors of the coefficients of the employment equation.
- Table 4: estimates and 95% confidence intervals for components of employment rate decomposition between women and men.
- Figures 6–13: estimates and 95% confidence bands of the coefficients of education and marital status indicators in the wage equation for men and women in all the specifications.[10]
- Figures 14 and 15: estimates and 95% confidence bands of the coefficients of fertility in the wage equation for men and women in specification 4.
- Figures 16-18: estimates and 95% confidence bands of the sorting equation.

[10]The dashed lines report estimates of β/σ in the outcome equation of the HSM.

TABLE 2. Summary Statistics

	Full		Male		Female	
	All	Employed	All	Employed	All	Employed
Log Hourly Wage		2.38		2.54		2.21
		(0.54)		(0.51)		(0.52)
Employed	0.74		0.83		0.66	
	(0.44)		(0.38)		(0.47)	
Ceased School at						
$\leqslant$ 15	0.33	0.30	0.33	0.31	0.33	0.29
	(0.47)	(0.46)	(0.47)	(0.46)	(0.47)	(0.45)
16	0.31	0.30	0.32	0.32	0.30	0.29
	(0.46)	(0.46)	(0.47)	(0.47)	(0.46)	(0.45)
17-18	0.18	0.19	0.16	0.17	0.20	0.22
	(0.38)	(0.39)	(0.37)	(0.37)	(0.40)	(0.41)
19-20	0.04	0.05	0.04	0.04	0.04	0.05
	(0.20)	(0.21)	(0.20)	(0.20)	(0.21)	(0.22)
21-22	0.09	0.11	0.09	0.10	0.09	0.12
	(0.29)	(0.31)	(0.29)	(0.30)	(0.29)	(0.32)
$\geqslant$23	0.05	0.05	0.06	0.06	0.04	0.04
	(0.21)	(0.22)	(0.23)	(0.24)	(0.19)	(0.20)
Age	40.13	39.84	40.22	39.76	40.06	39.92
	(10.43)	(10.10)	(10.40)	(10.11)	(10.45)	(10.08)
Married	0.76	0.79	0.78	0.81	0.75	0.76
	(0.43)	(0.41)	(0.42)	(0.39)	(0.43)	(0.43)
Number of children with age						
0–1	0.06	0.05	0.06	0.06	0.06	0.03
	(0.24)	(0.22)	(0.24)	(0.25)	(0.24)	(0.18)
2	0.05	0.04	0.05	0.05	0.05	0.03
	(0.23)	(0.21)	(0.23)	(0.23)	(0.23)	(0.18)
3–4	0.10	0.09	0.10	0.10	0.11	0.07
	(0.32)	(0.29)	(0.31)	(0.32)	(0.32)	(0.27)
5–10	0.32	0.29	0.30	0.30	0.33	0.28
	(0.64)	(0.61)	(0.62)	(0.62)	(0.65)	(0.59)
11–16	0.30	0.30	0.28	0.29	0.32	0.32
	(0.63)	(0.62)	(0.61)	(0.61)	(0.64)	(0.63)
17–18	0.03	0.04	0.03	0.03	0.04	0.04
	(0.19)	(0.19)	(0.18)	(0.18)	(0.19)	(0.20)
Benefit Income	5.44	5.50	5.25	5.29	5.60	5.73
	(0.74)	(0.78)	(0.70)	(0.72)	(0.73)	(0.78)
Observations	258,900	190,765	119,396	98,764	139,504	92,001

Notes: all the entries are means with standard deviations in parentheses.

Source: FES/EFS/LCFS Data.

TABLE 3. Estimates of Coefficients of the Employment Equation

Variable	Male	Female	Variable	Male	Female
educ16	0.25	0.06	numch34	-0.18	-0.63
	(0.01)	(0.01)		(0.02)	(0.01)
educ1718	0.46	0.20	numch510	-0.18	-0.33
	(0.02)	(0.01)		(0.01)	(0.01)
educ1920	0.42	0.16	numch1116	-0.16	-0.15
	(0.03)	(0.02)		(0.01)	(0.01)
educ2122	0.74	0.28	numch1718	-0.02	-0.11
	(0.02)	(0.02)		(0.03)	(0.02)
educ23	0.51	0.15	benefit	-0.35	-0.42
	(0.02)	(0.02)		(0.01)	(0.01)
couple	-4.02	-8.14	benefit×couple	0.87	1.40
	(0.09)	(0.08)		(0.02)	(0.02)
numch1	-0.16	-0.90	constant	2.50	2.75
	(0.02)	(0.02)		(0.08)	(0.07)
numch2	-0.18	-0.77			
	(0.02)	(0.02)			

Notes: standard errors in parentheses. The specification includes a quartic polynomial in age, region indicators and survey year indicators.

- Figure 19: estimates and 95% confidence bands for the quantiles of the observed and latent wages in specification 4.
- Figure 21-20: estimates and 95% confidence bands for the decomposition of offered wages between women and men for specifications 1–3.
- Figure 23: estimates and 95% confidence bands for decomposition between men and women with aggregated selection effects.
- Figure 24: estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women in specifications 1–3.
- Figures 25–28: estimates and 95% confidence bands for components of wage decomposition between women and men.
- Figures 29 and 30: estimates and 95% confidence bands for components of wage decomposition between first and second half of sample period for men and women in specification 2.

TABLE 4. Employment rate decomposition between men and women

Employment (%)		Structure (π)	
		Male	Female
Composition (F_Z)	Male	83	59
		(82, 83)	(59, 59)
	Female	83	66
		(83, 83)	(66, 66)

95% bootstrap confidence intervals in parentheses

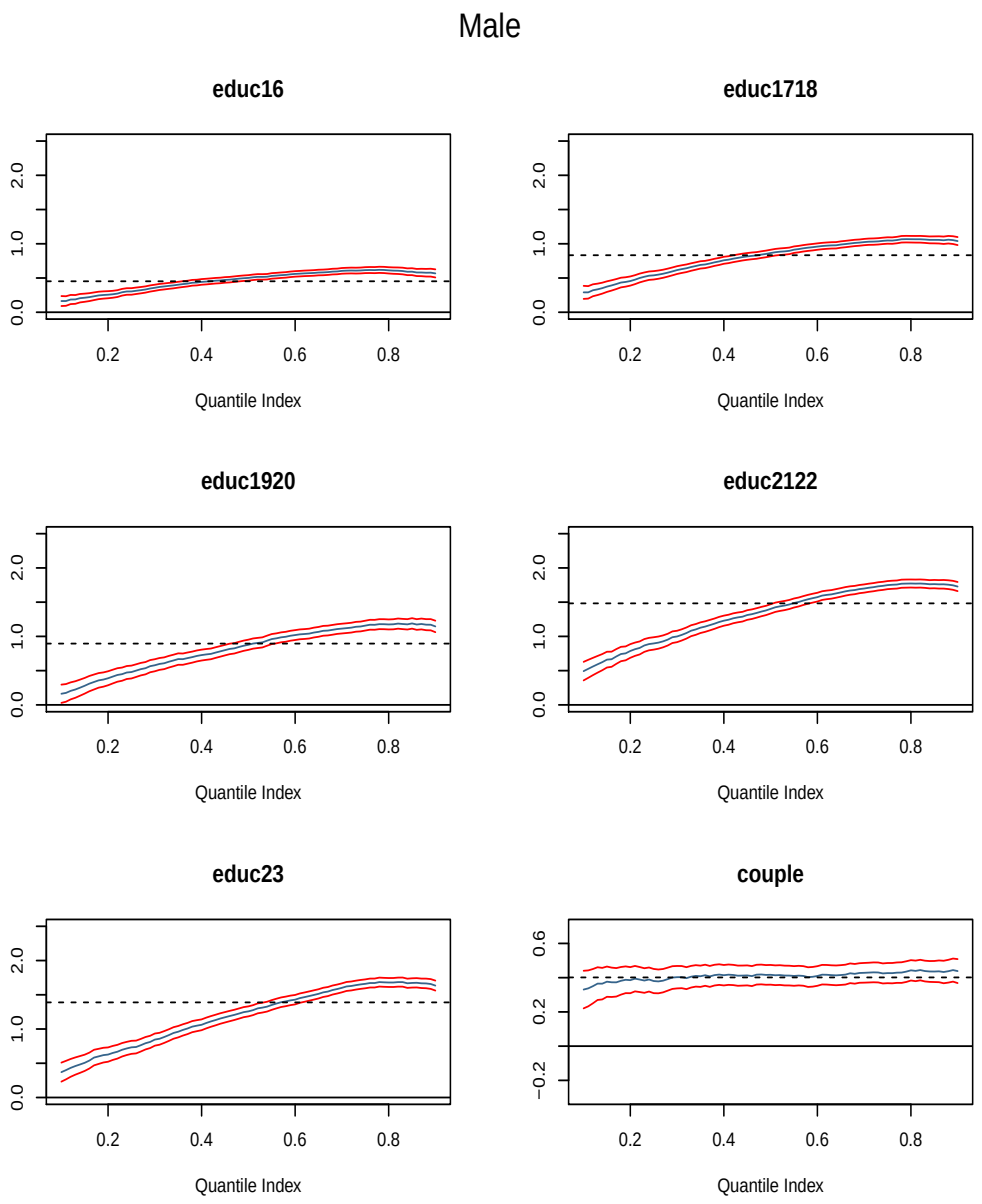

FIGURE 6. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 1 for men

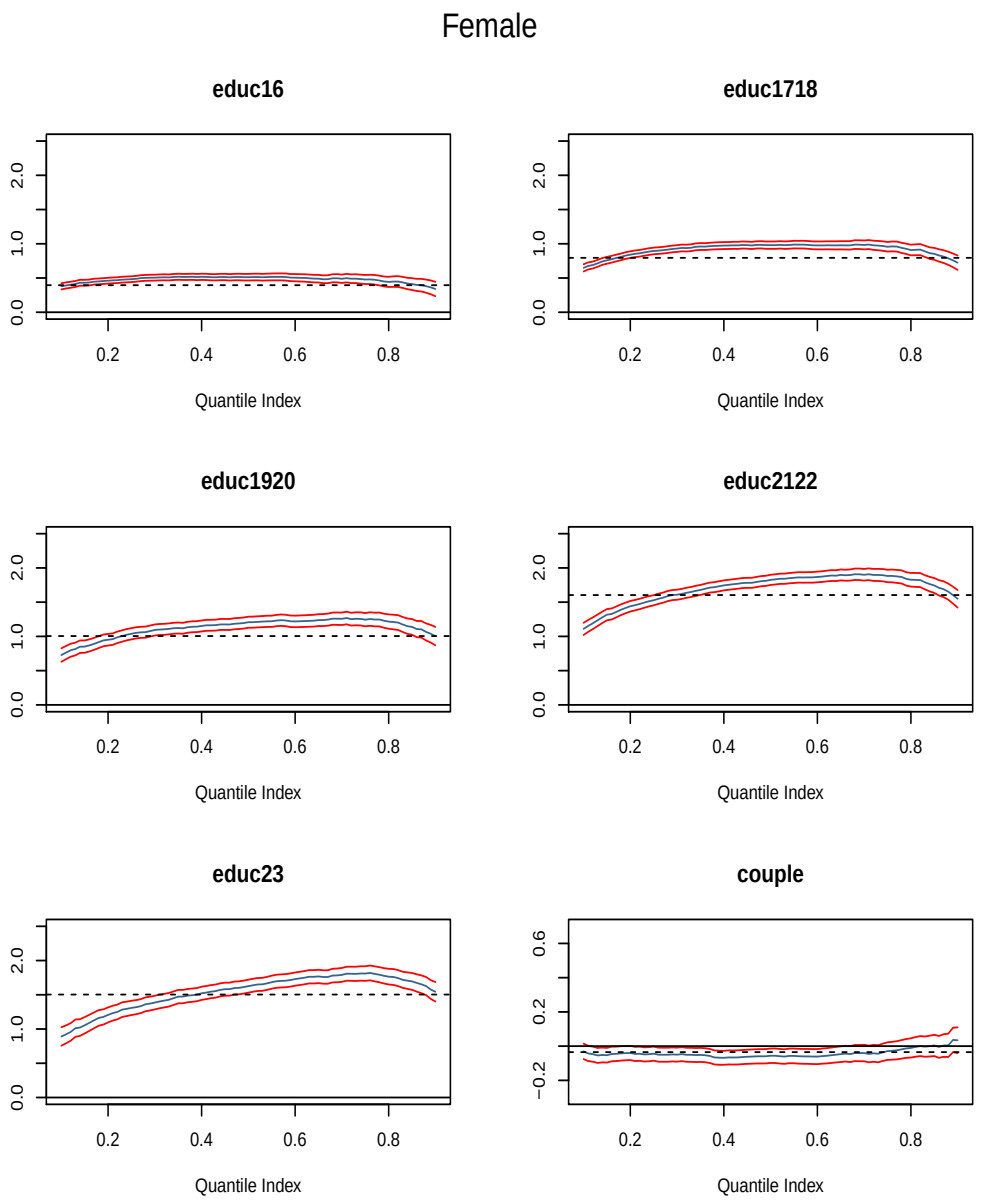

FIGURE 7. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 1 for women

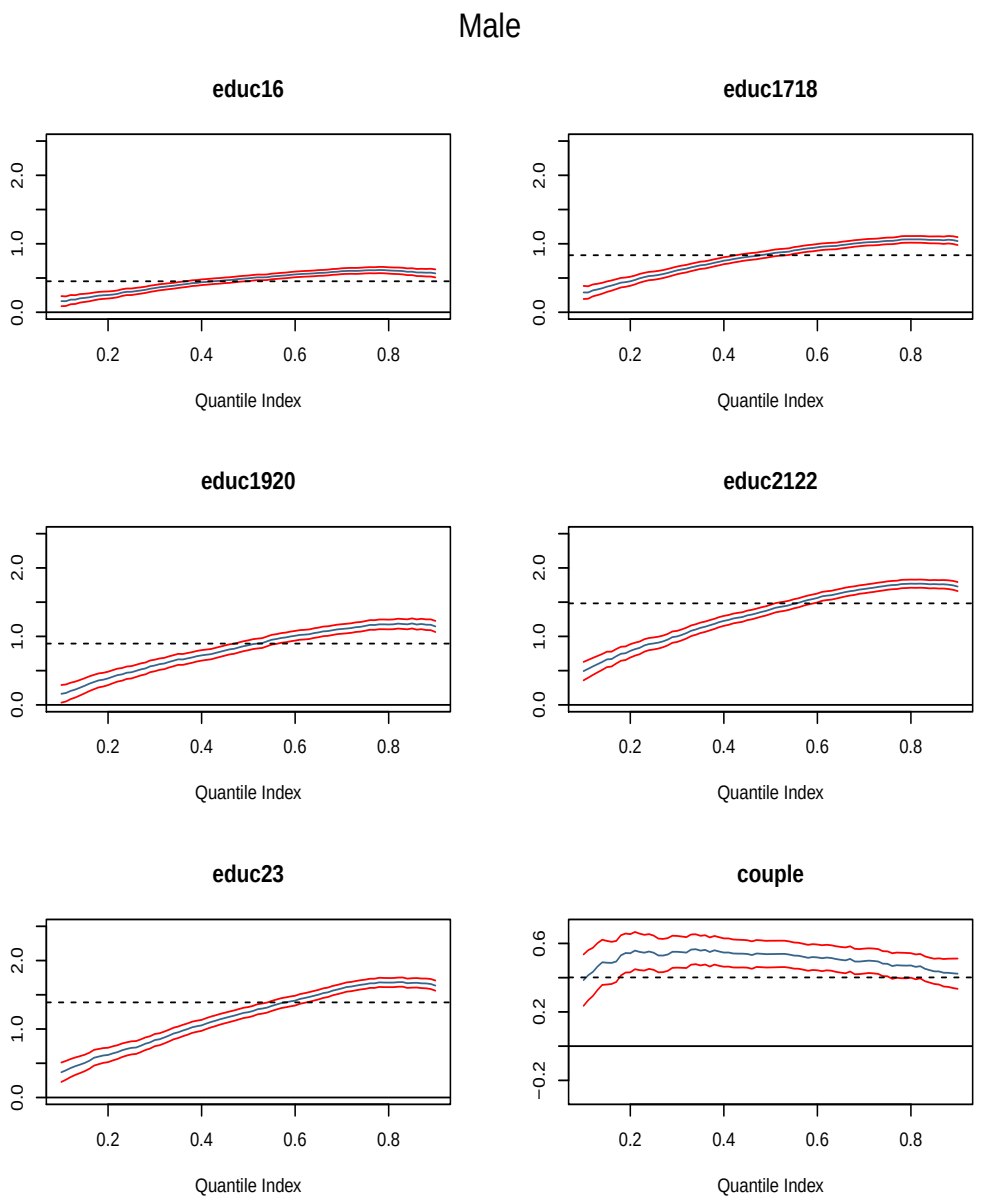

FIGURE 8. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 2 for men

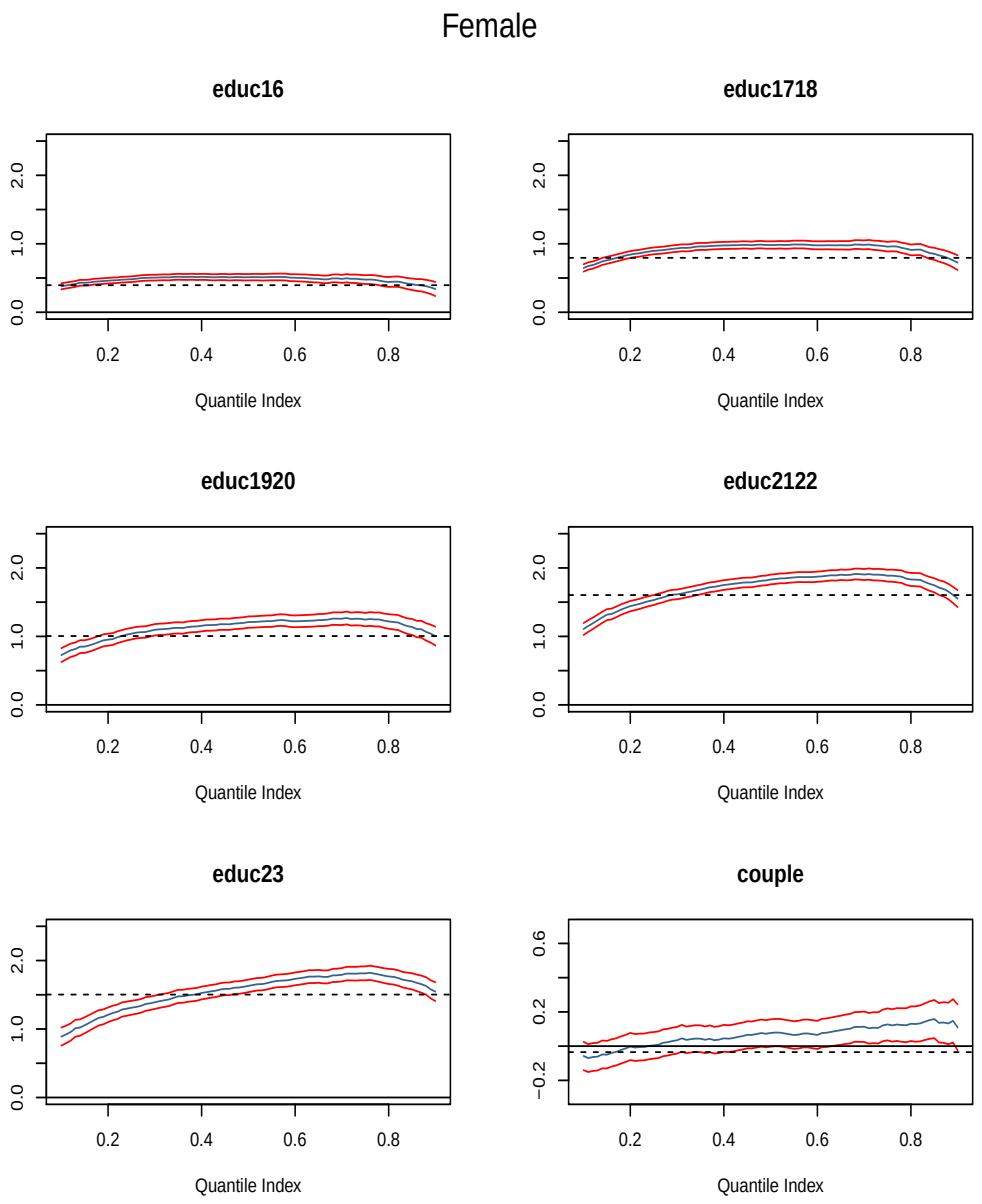

FIGURE 9. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 2 for women

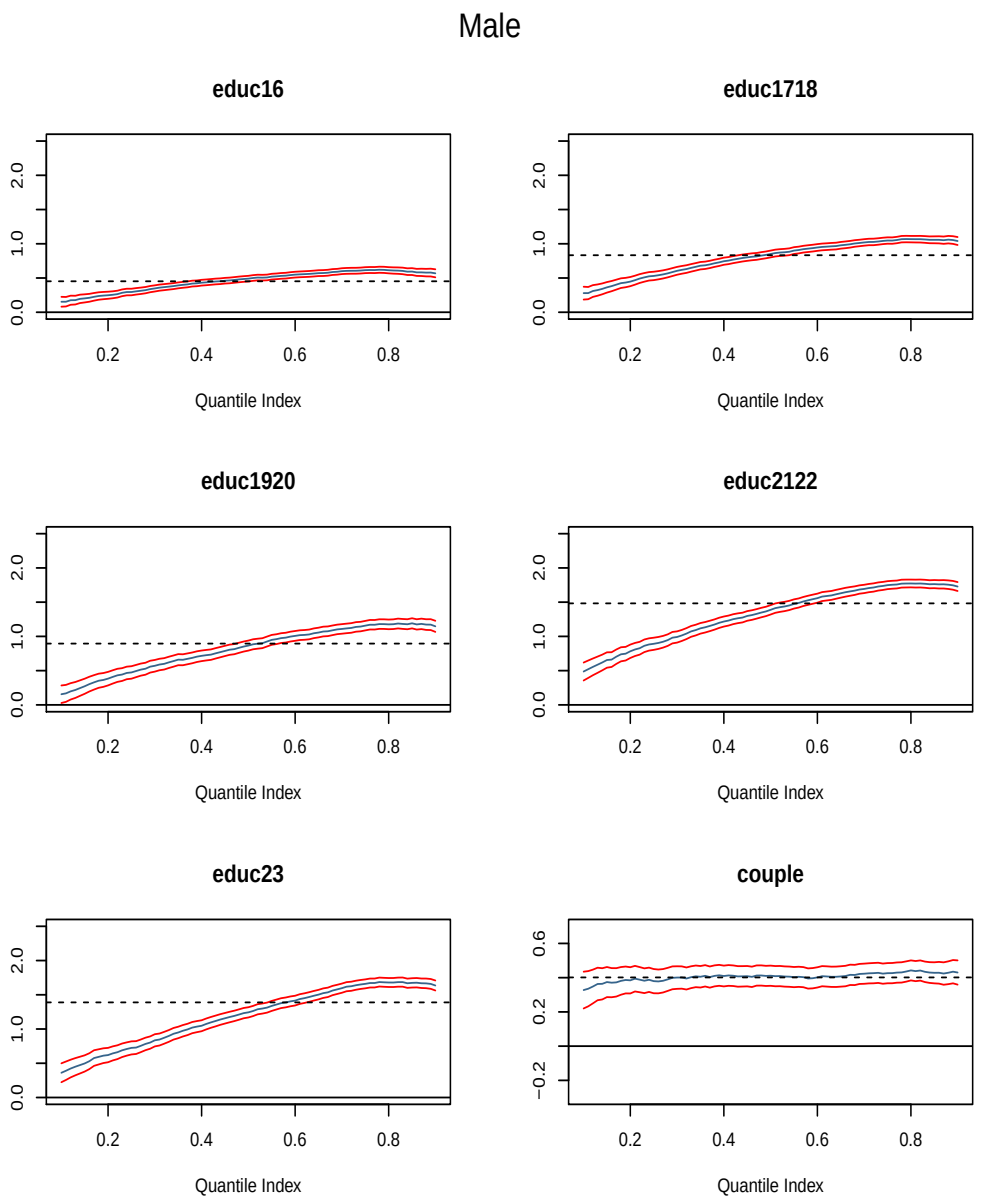

FIGURE 10. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 3 for men

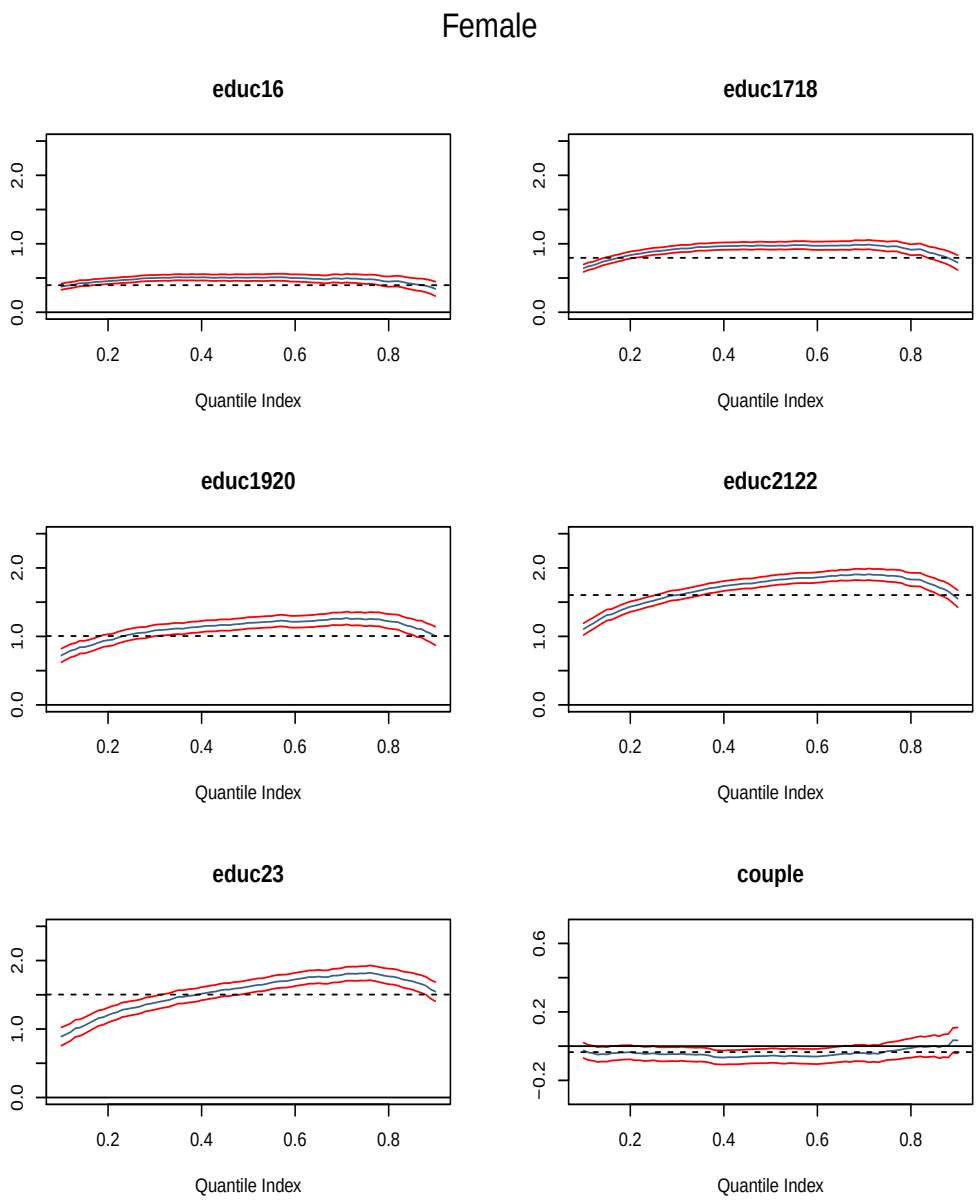

FIGURE 11. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 3 for women

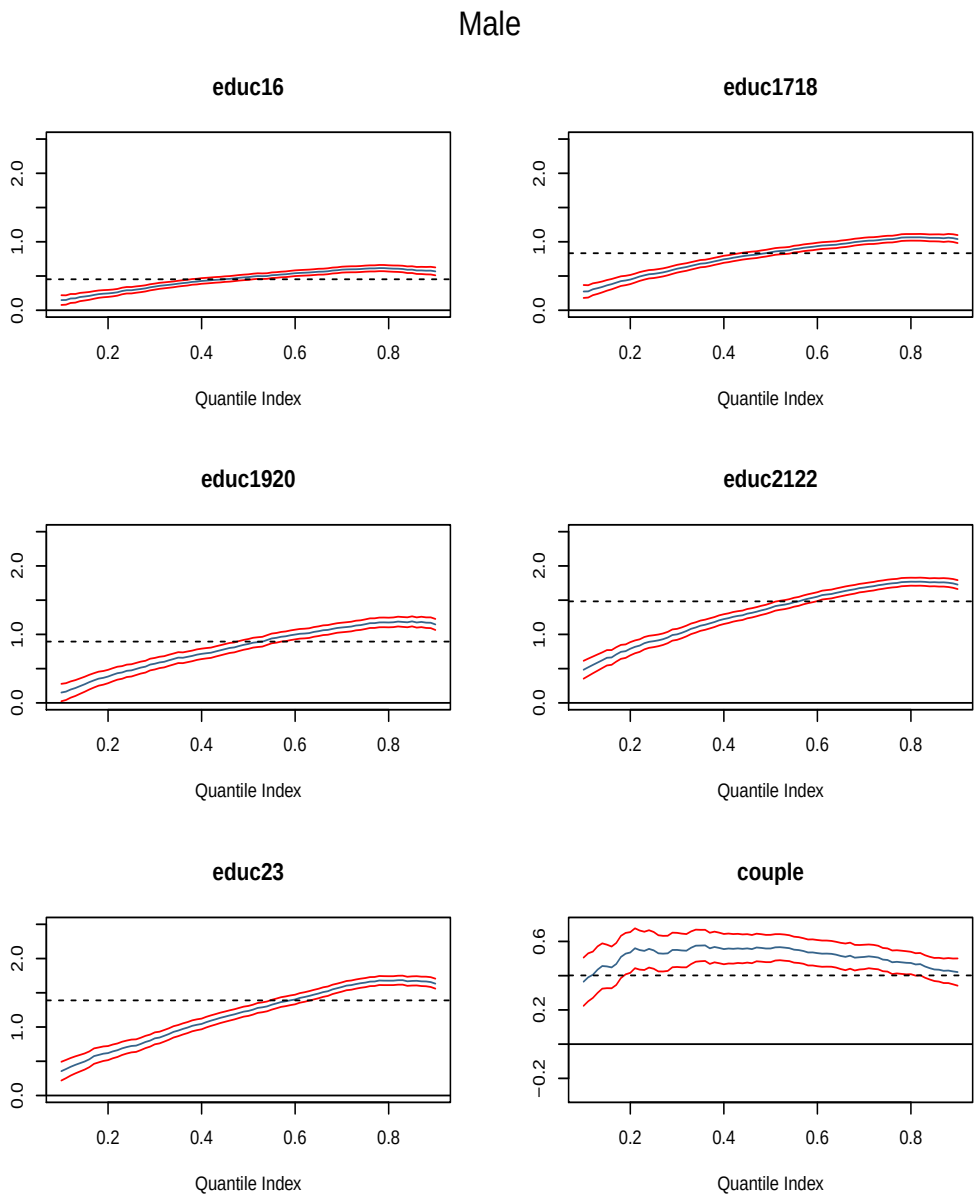

FIGURE 12. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 4 for men

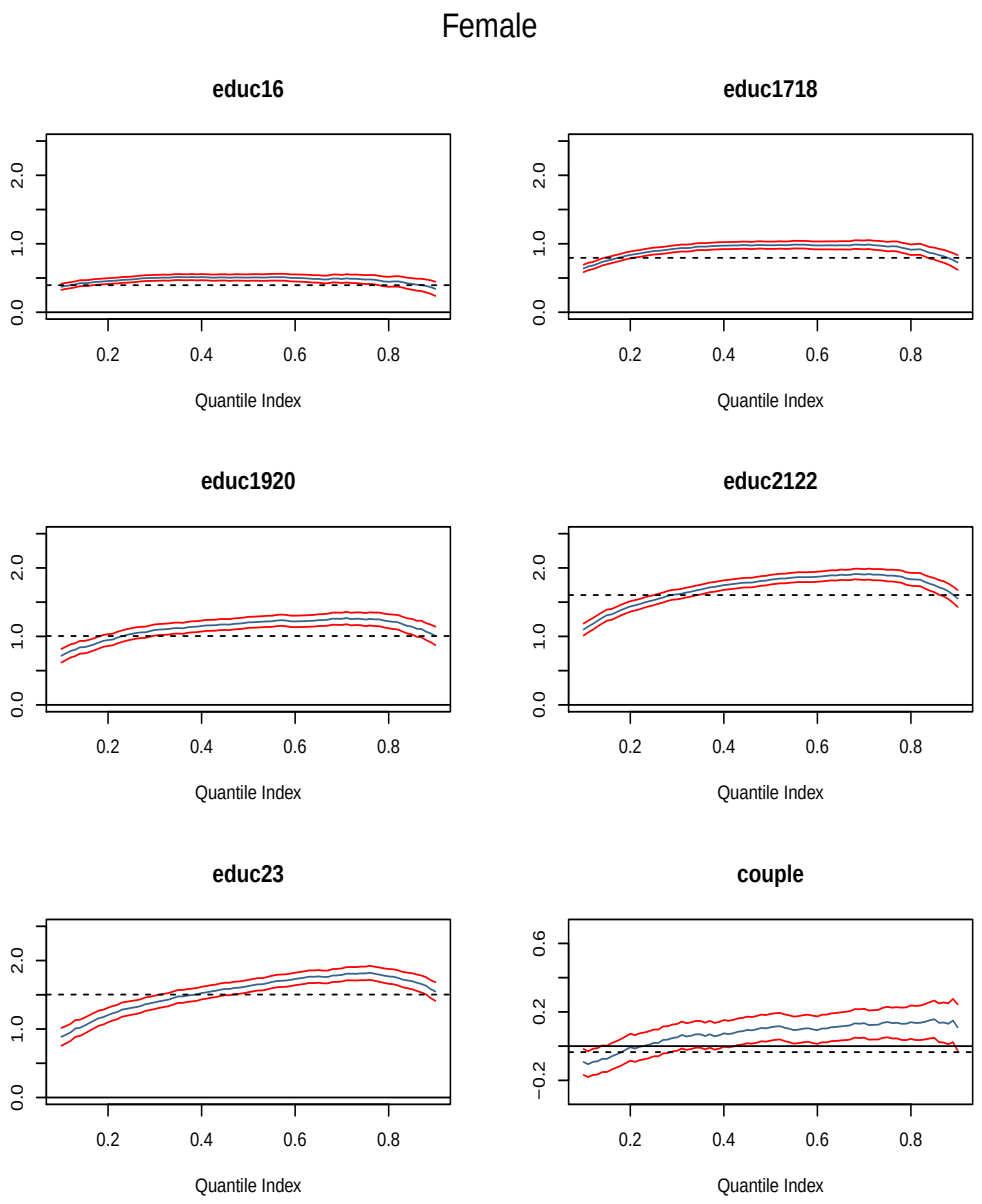

FIGURE 13. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 4 for women

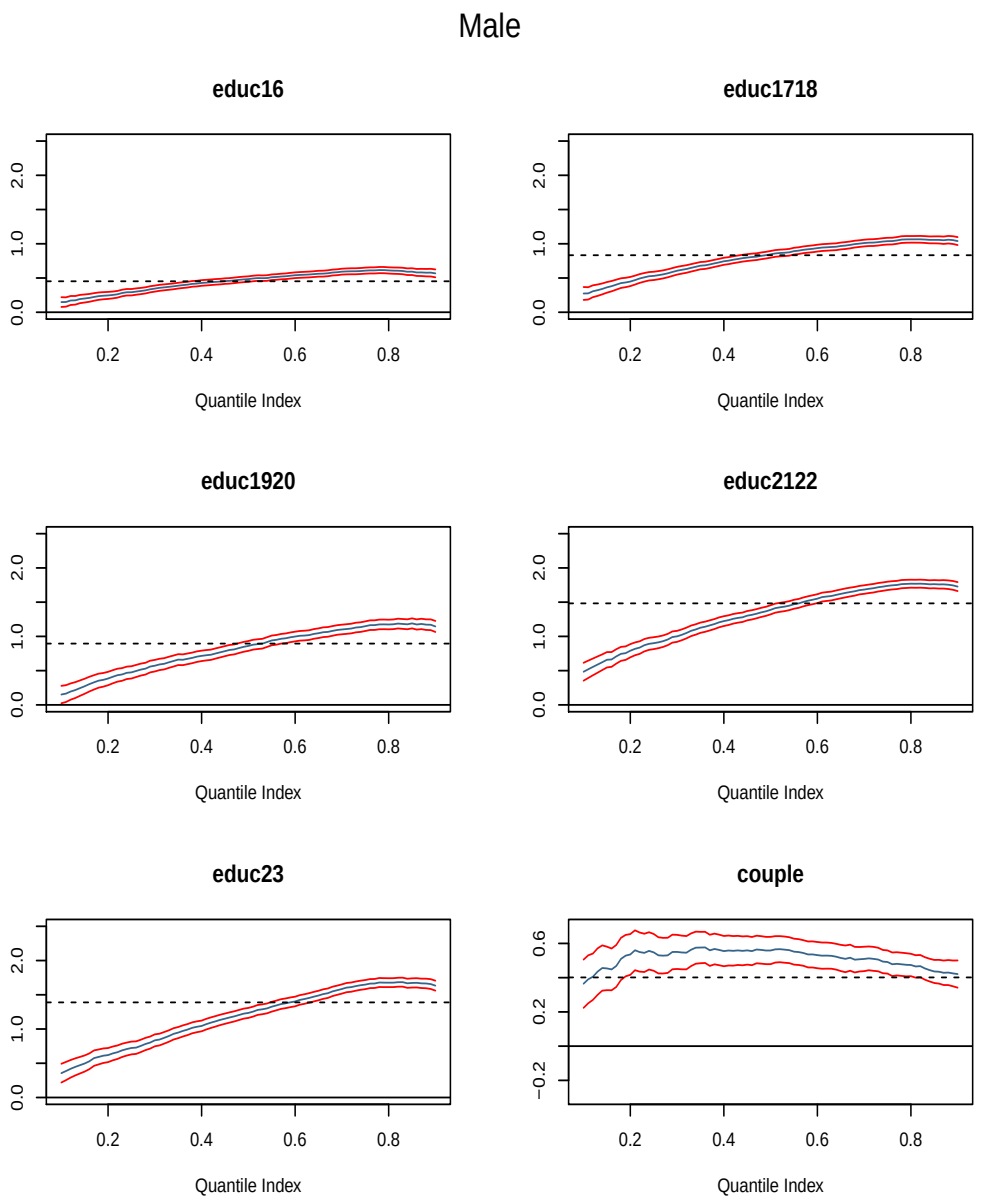

FIGURE 14. Estimates and 95% confidence bands for coefficients of fertility in the outcome equation: specification 4 for men

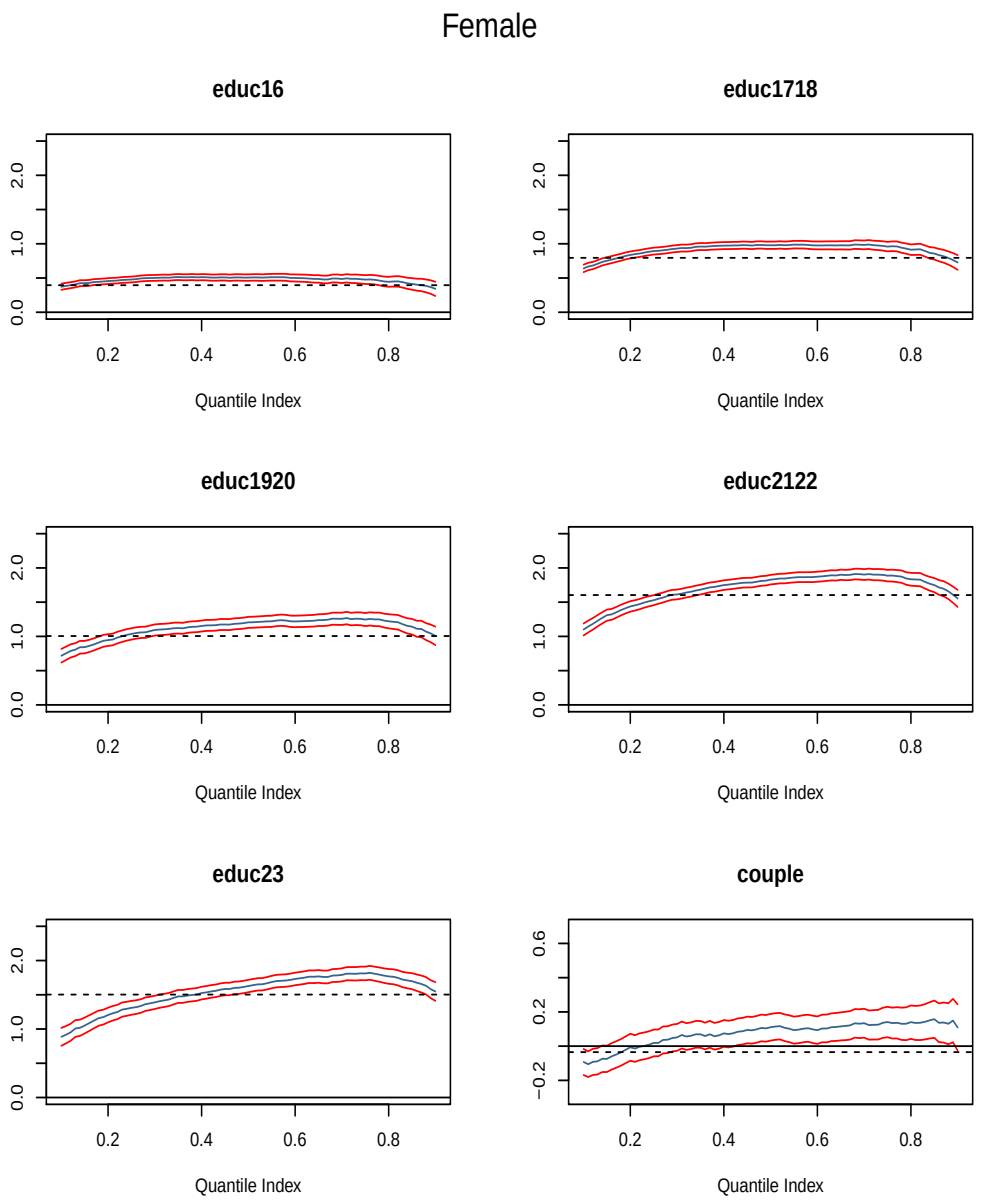

FIGURE 15. Estimates and 95% confidence bands for coefficients of fertility in the outcome equation: specification 4 for women

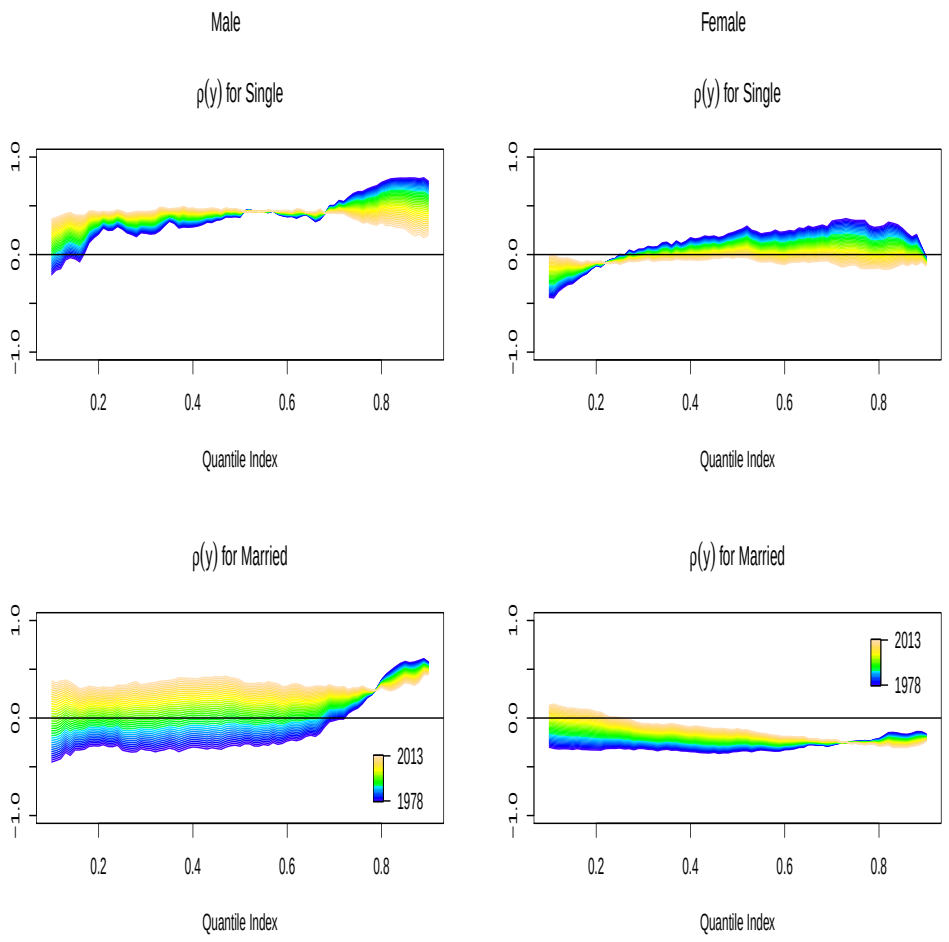

FIGURE 16. Estimates and 95% confidence bands for the selection sorting function: specification 4

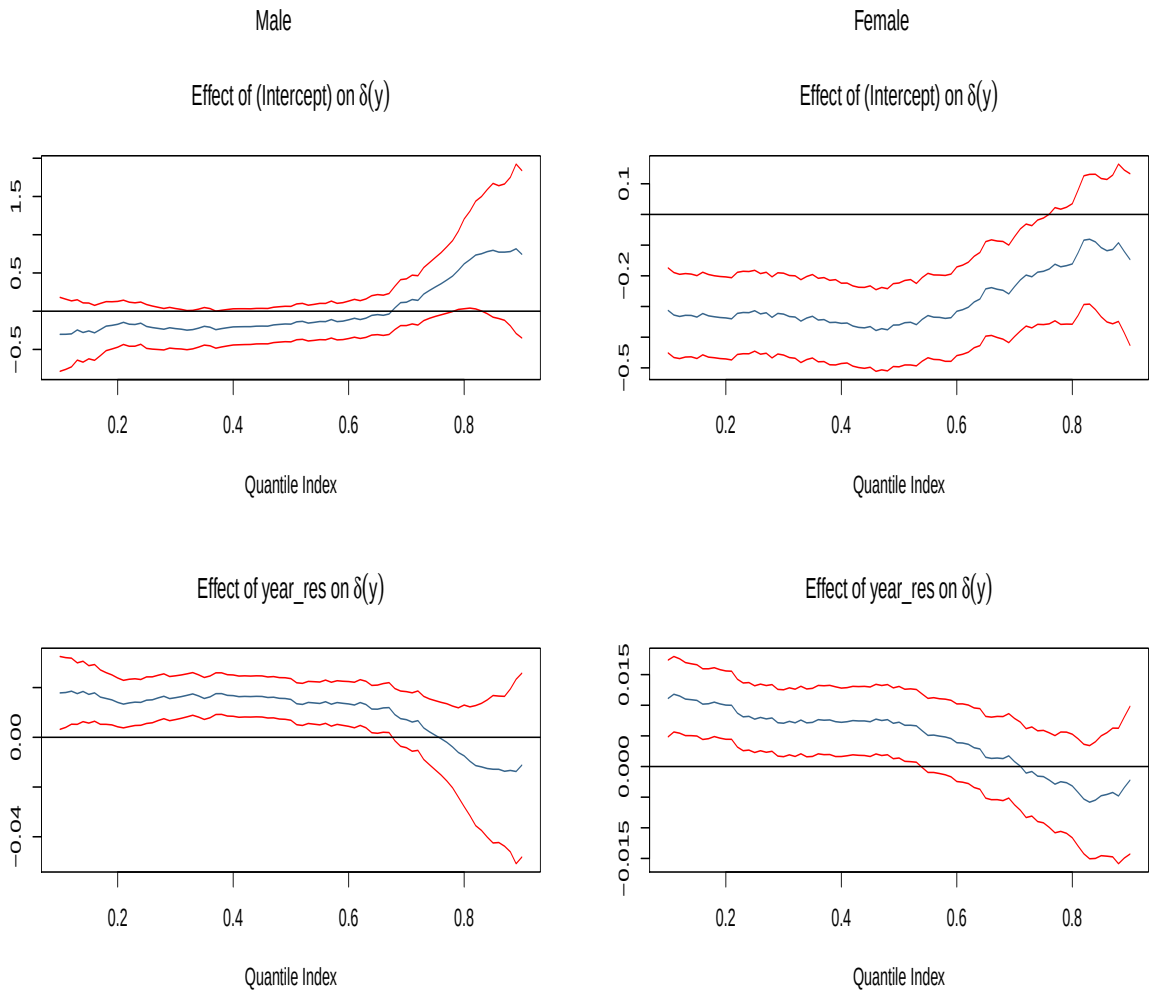

FIGURE 17. Estimates and 95% confidence bands for coefficients of the selection sorting function: specification 3

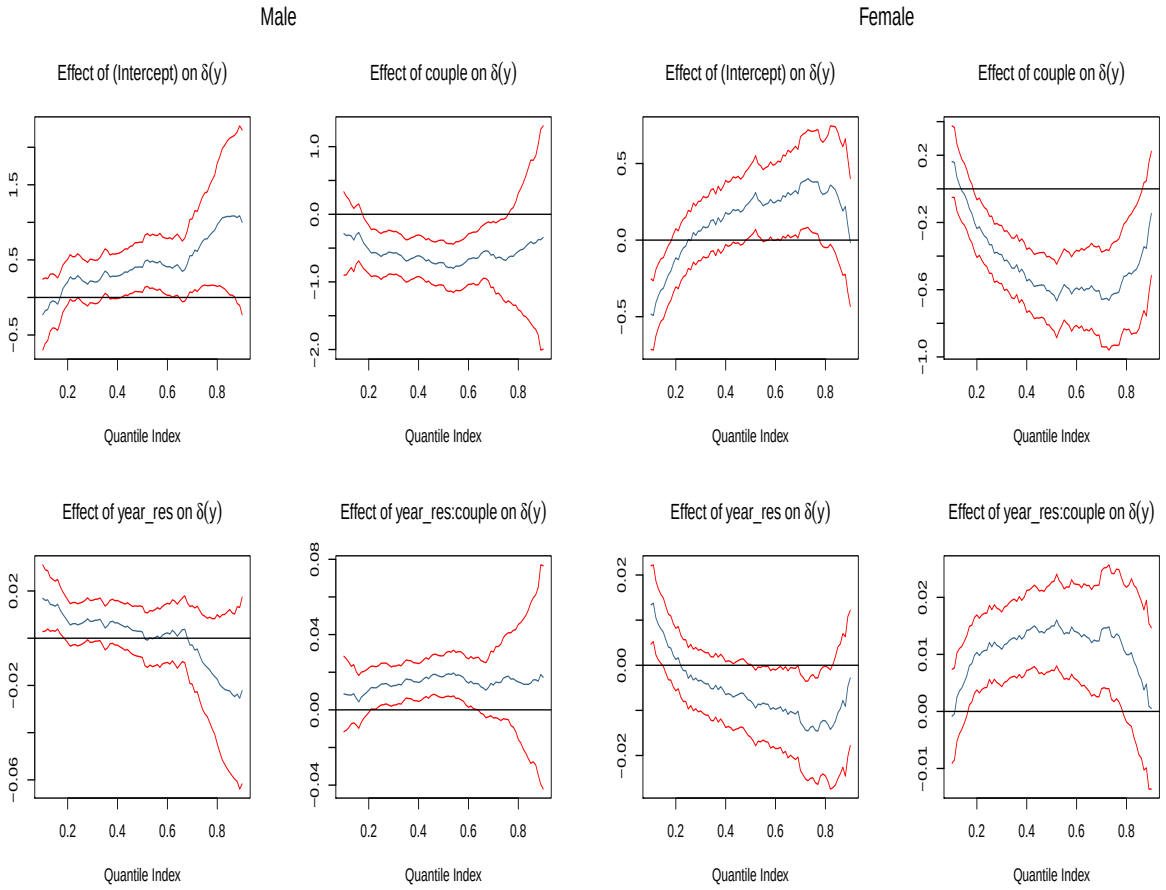

FIGURE 18. Estimates and 95% confidence bands for coefficients of the selection sorting function: specification 4

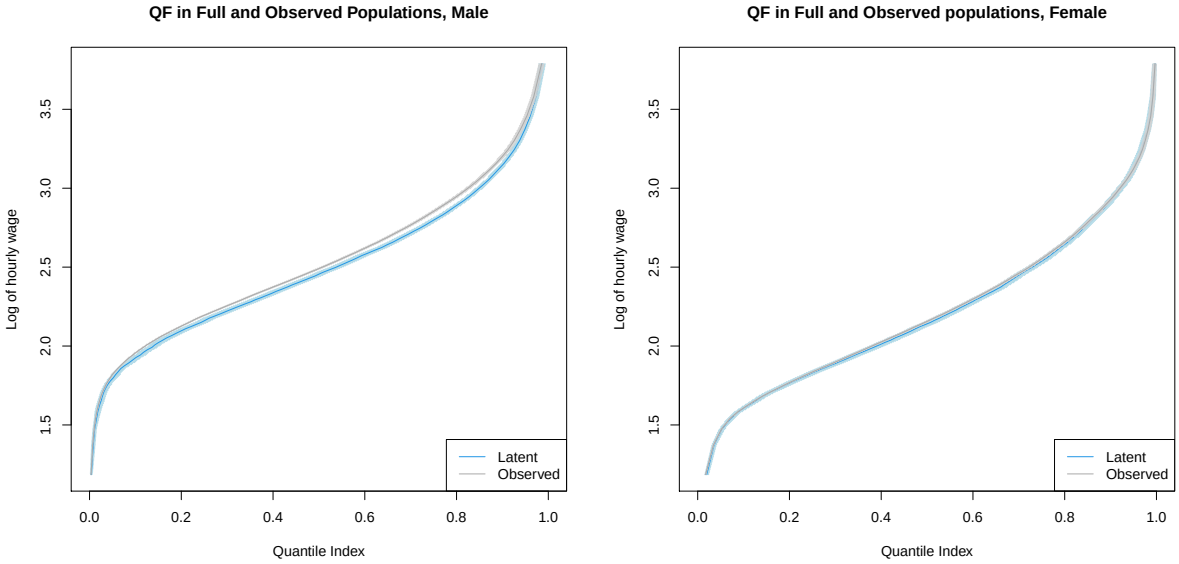

FIGURE 19. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages: specification 4

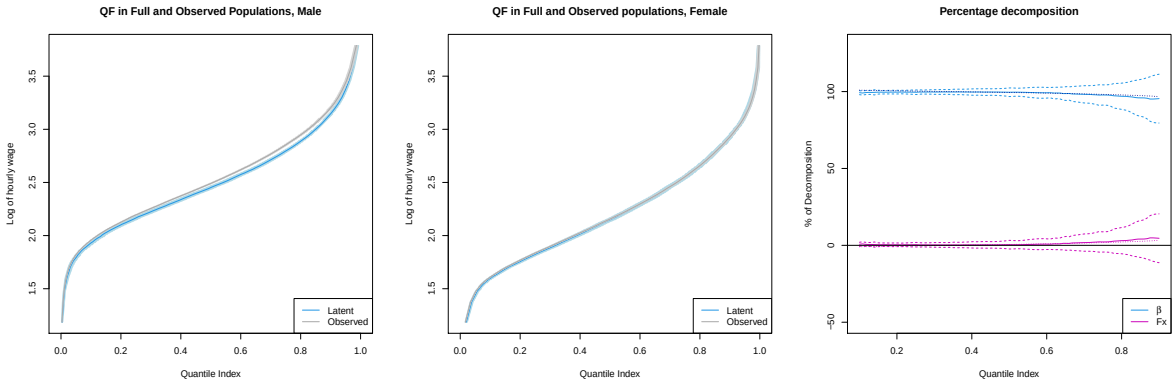

FIGURE 20. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 1

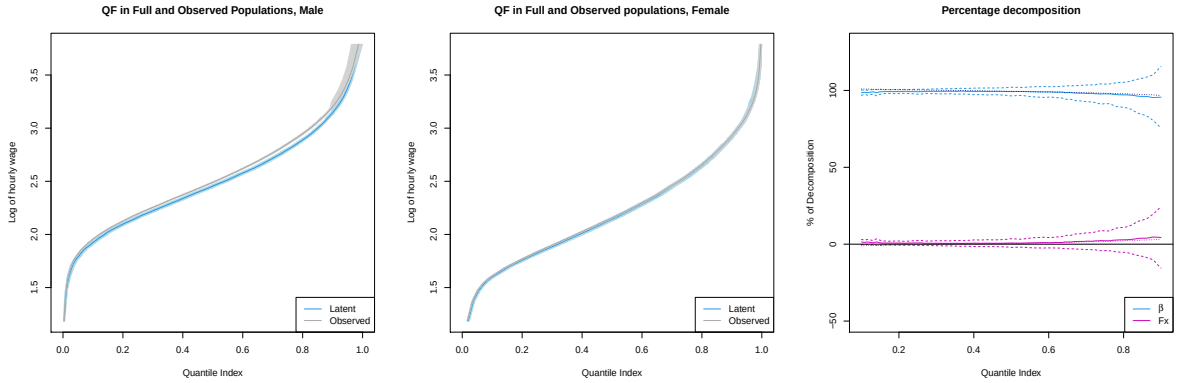

FIGURE 21. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 2

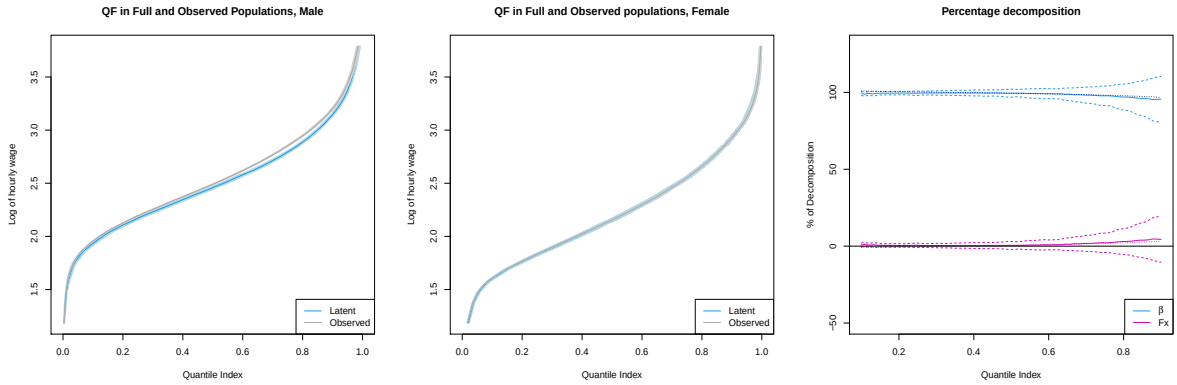

FIGURE 22. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 3

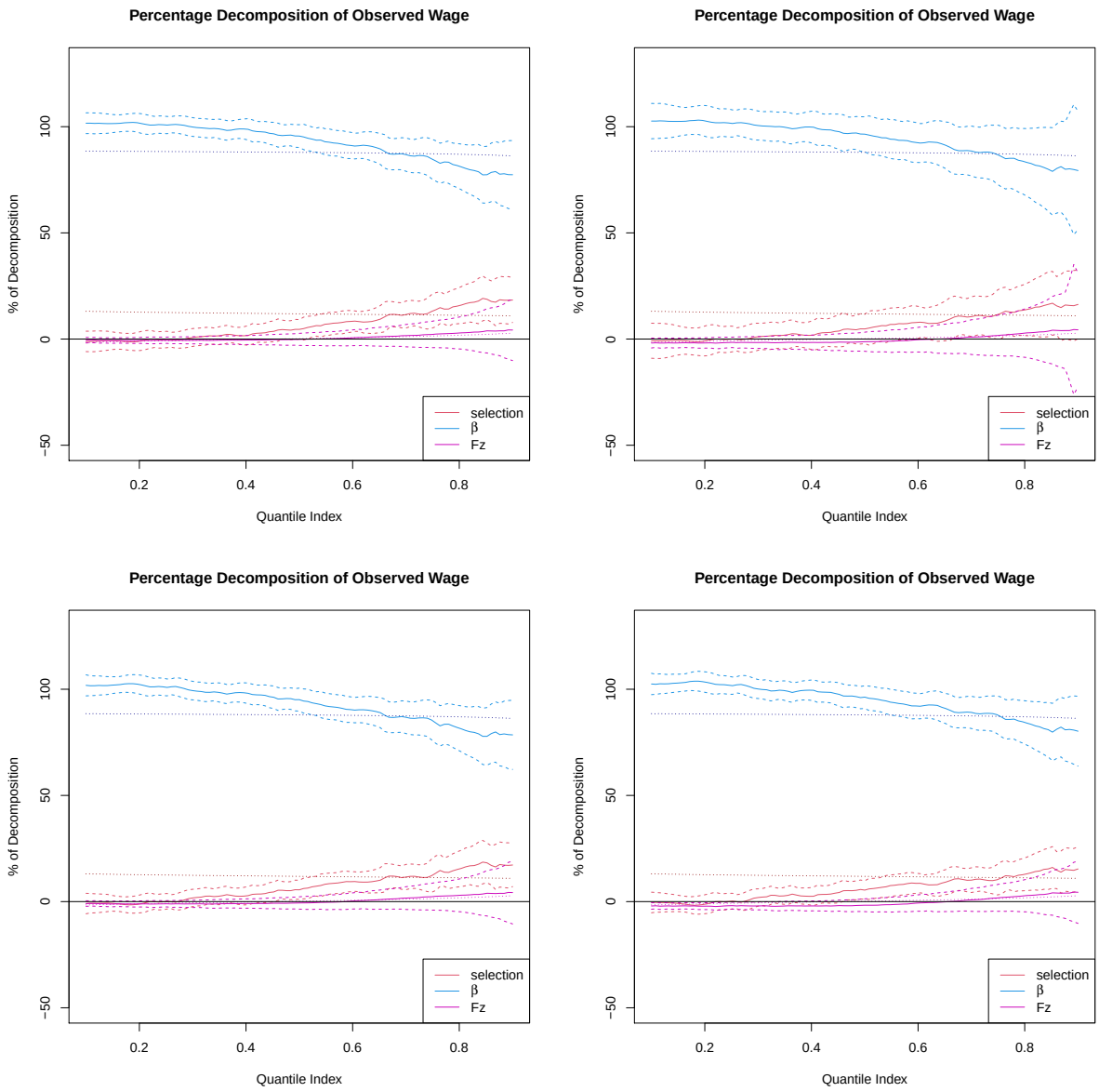

FIGURE 23. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women with aggregated selection effect: (upper left) specification 1, (upper right) specification 2, (bottom left) specification 3, and (bottom right) specification 4

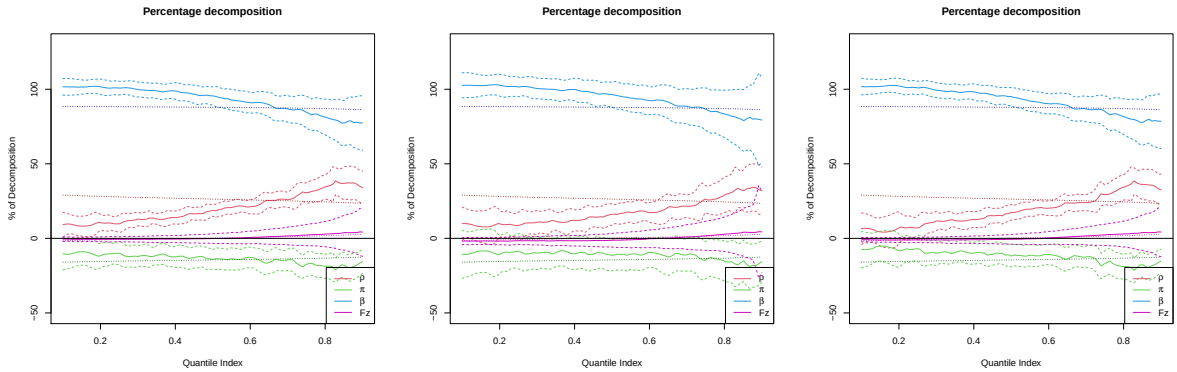

FIGURE 24. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women: (left) specification 1, (middle) specification 2, and (right) specification 3

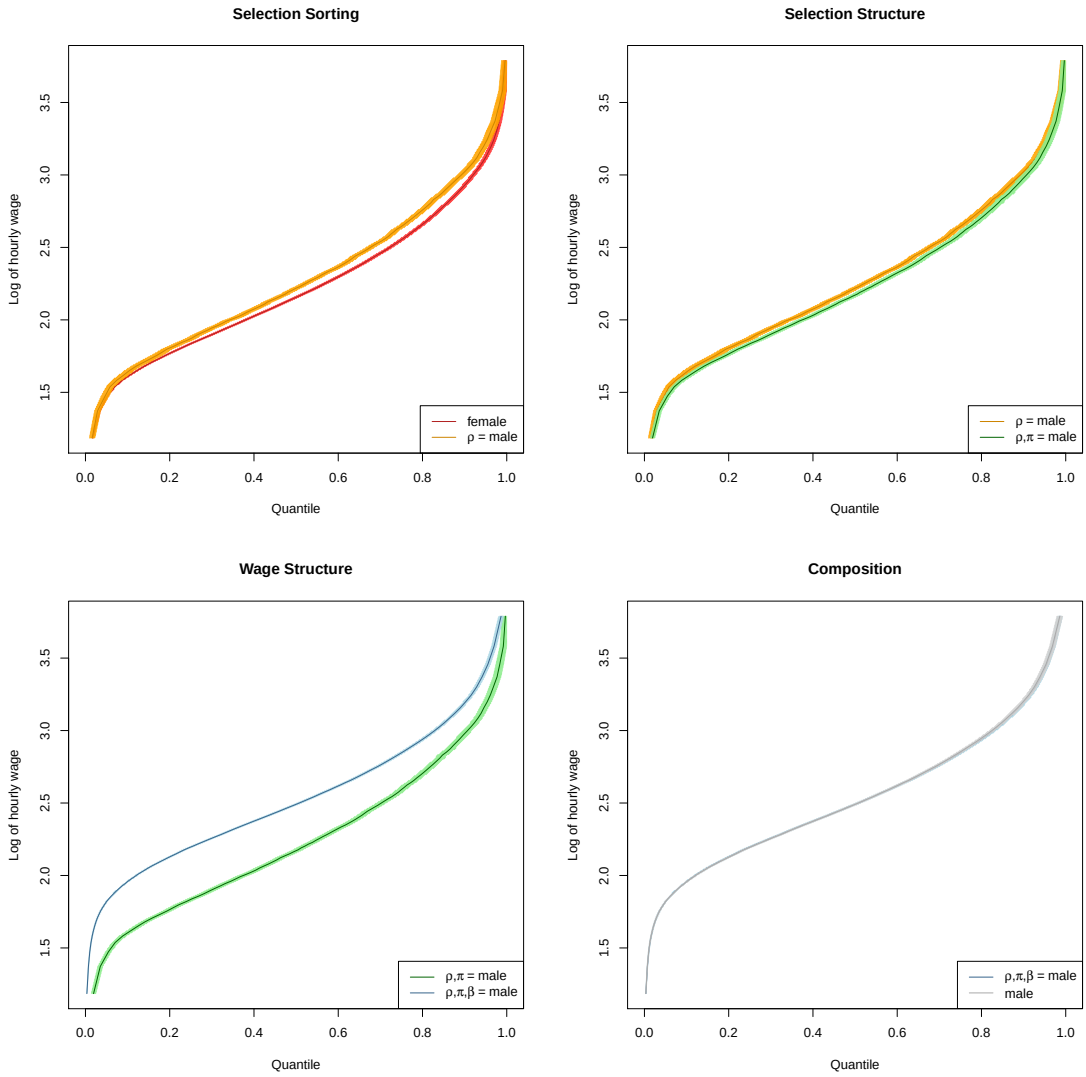

FIGURE 25. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 1

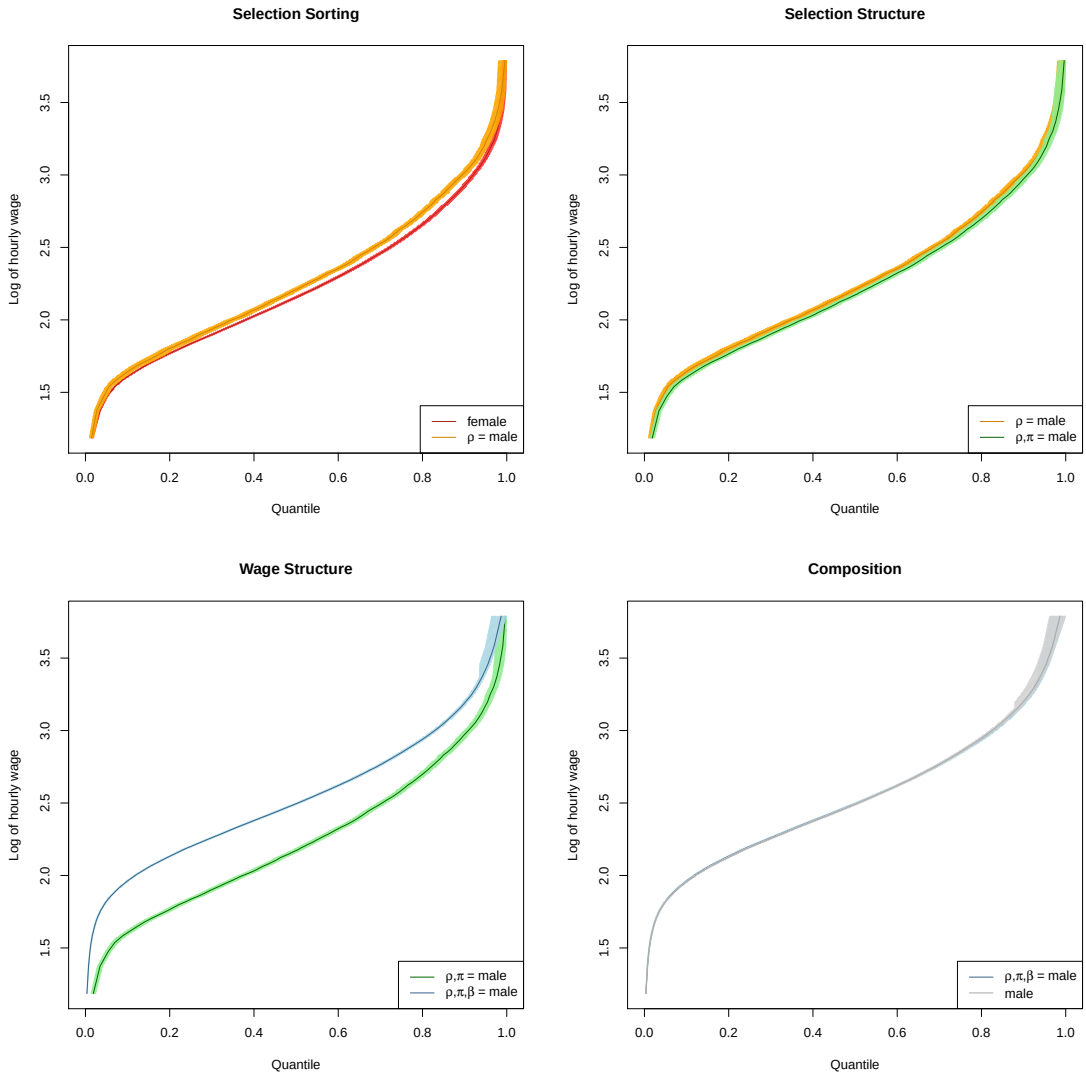

FIGURE 26. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 2

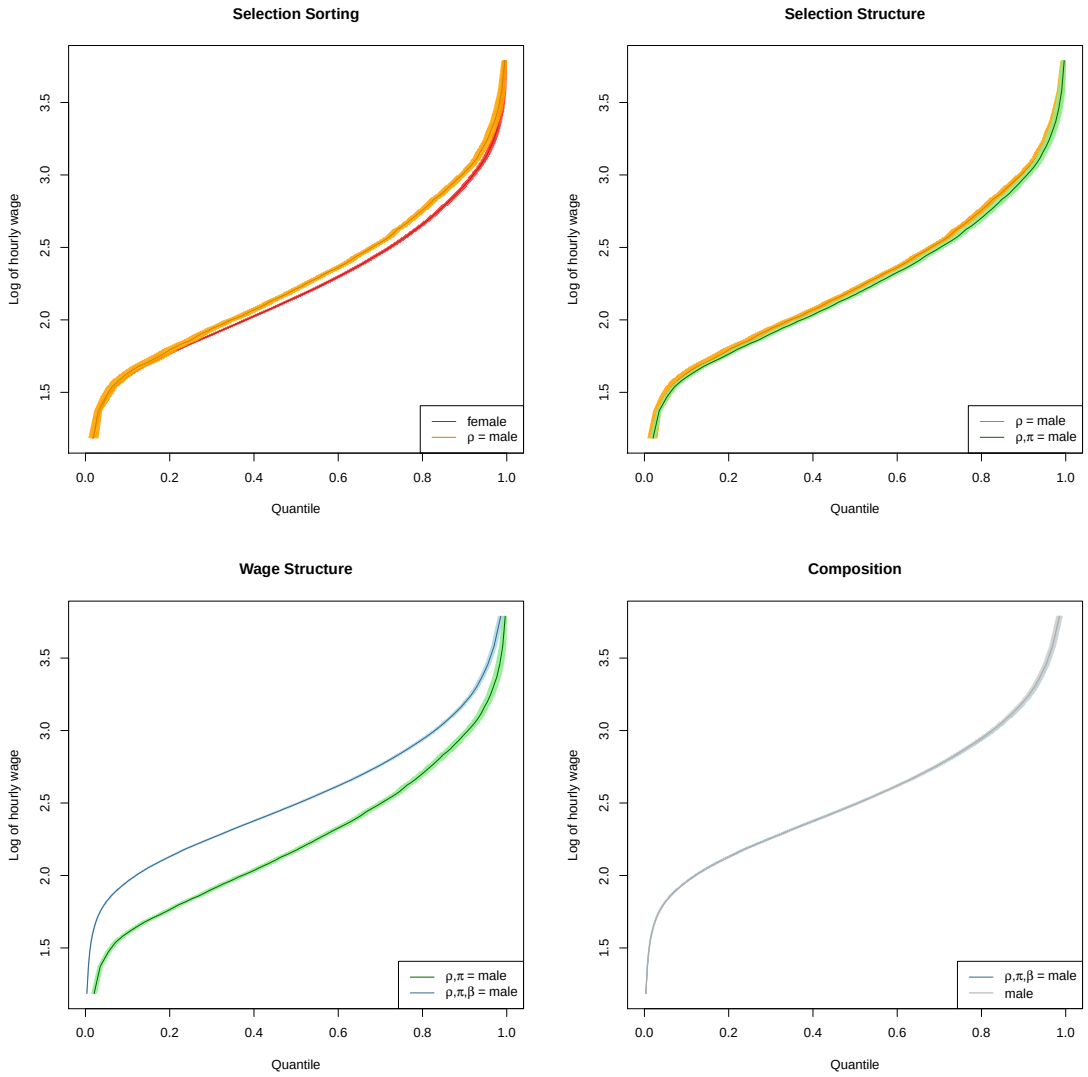

FIGURE 27. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 3

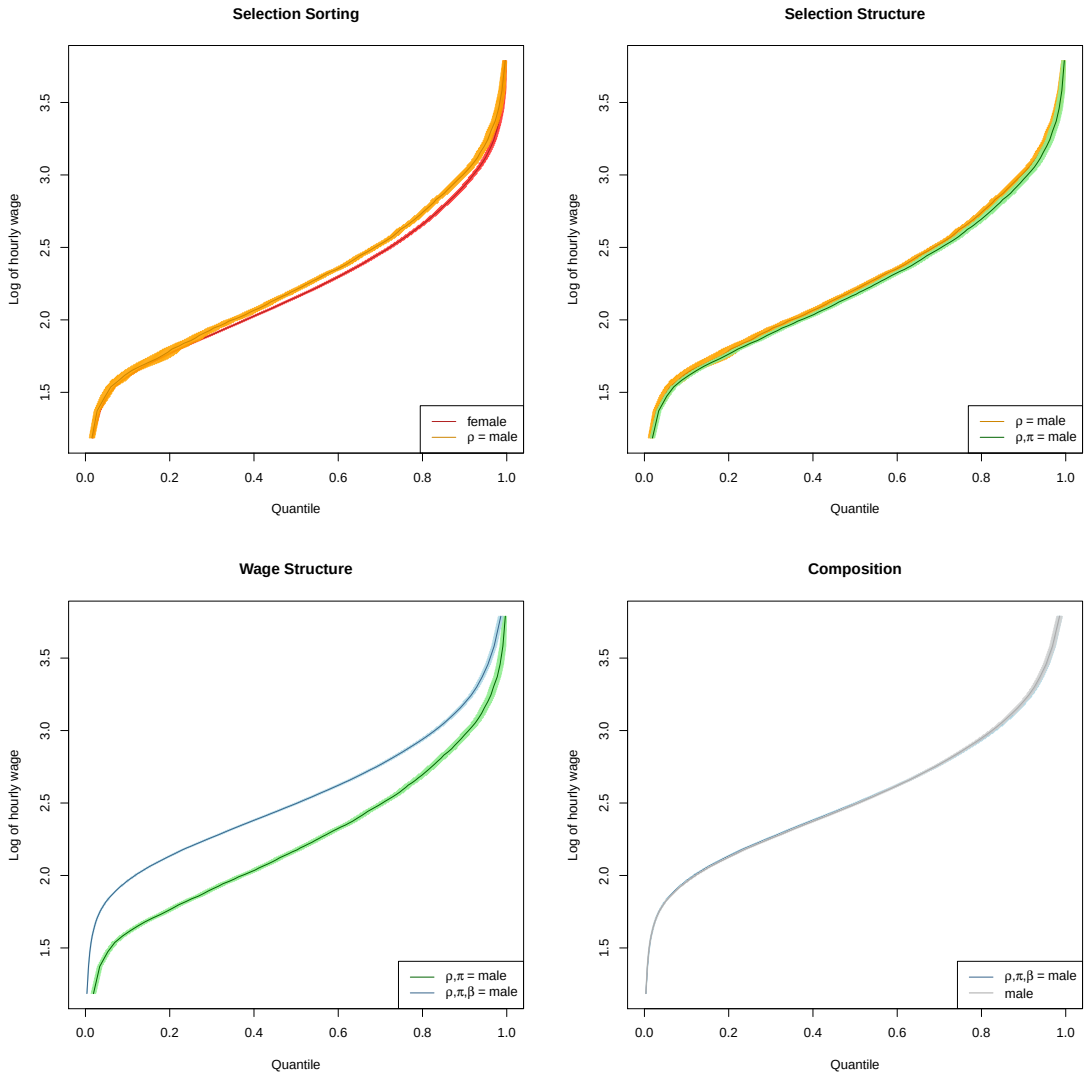

FIGURE 28. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 4

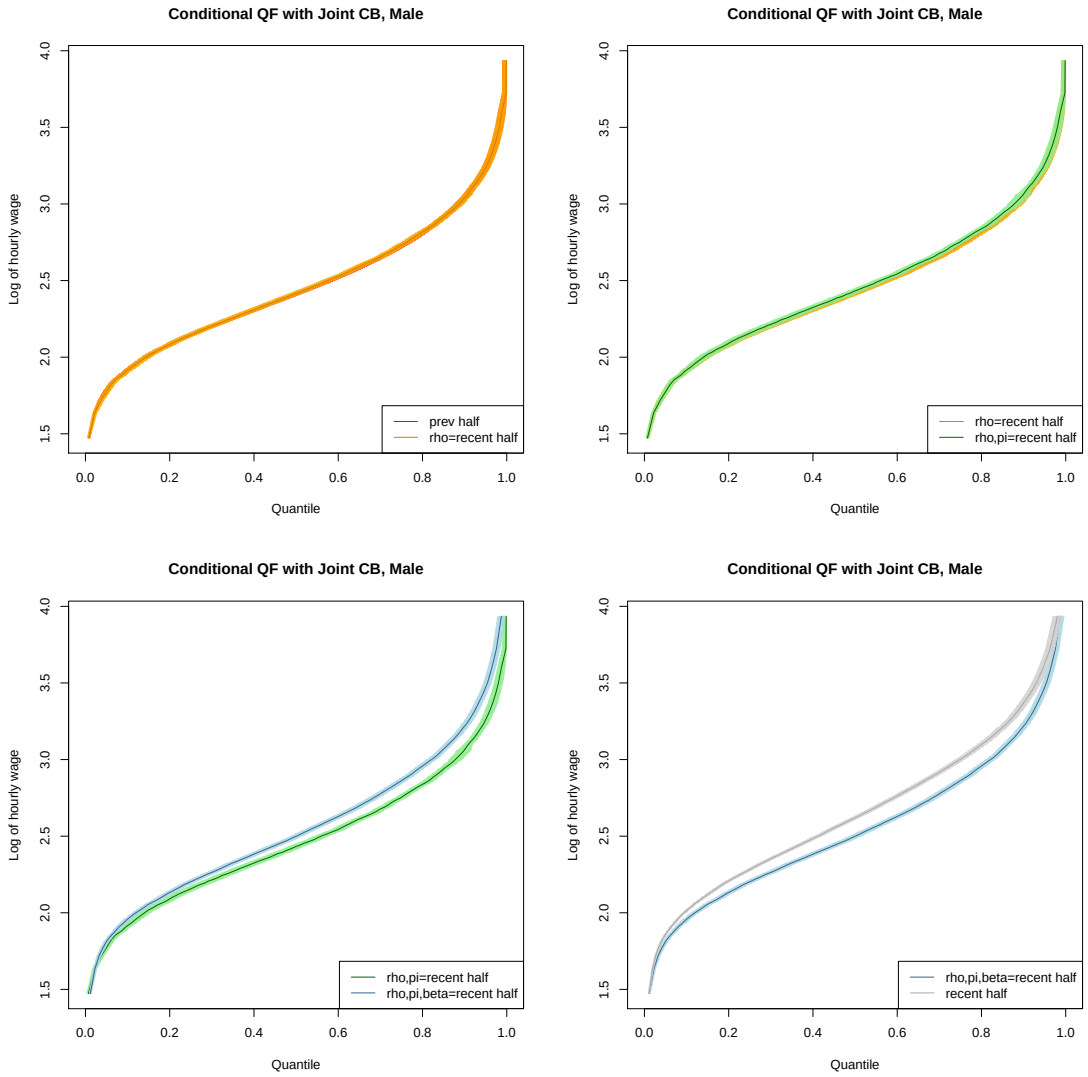

FIGURE 29. Estimates and 95% confidence bands for components of wage decomposition between first and second half of sample period for men in specification 2

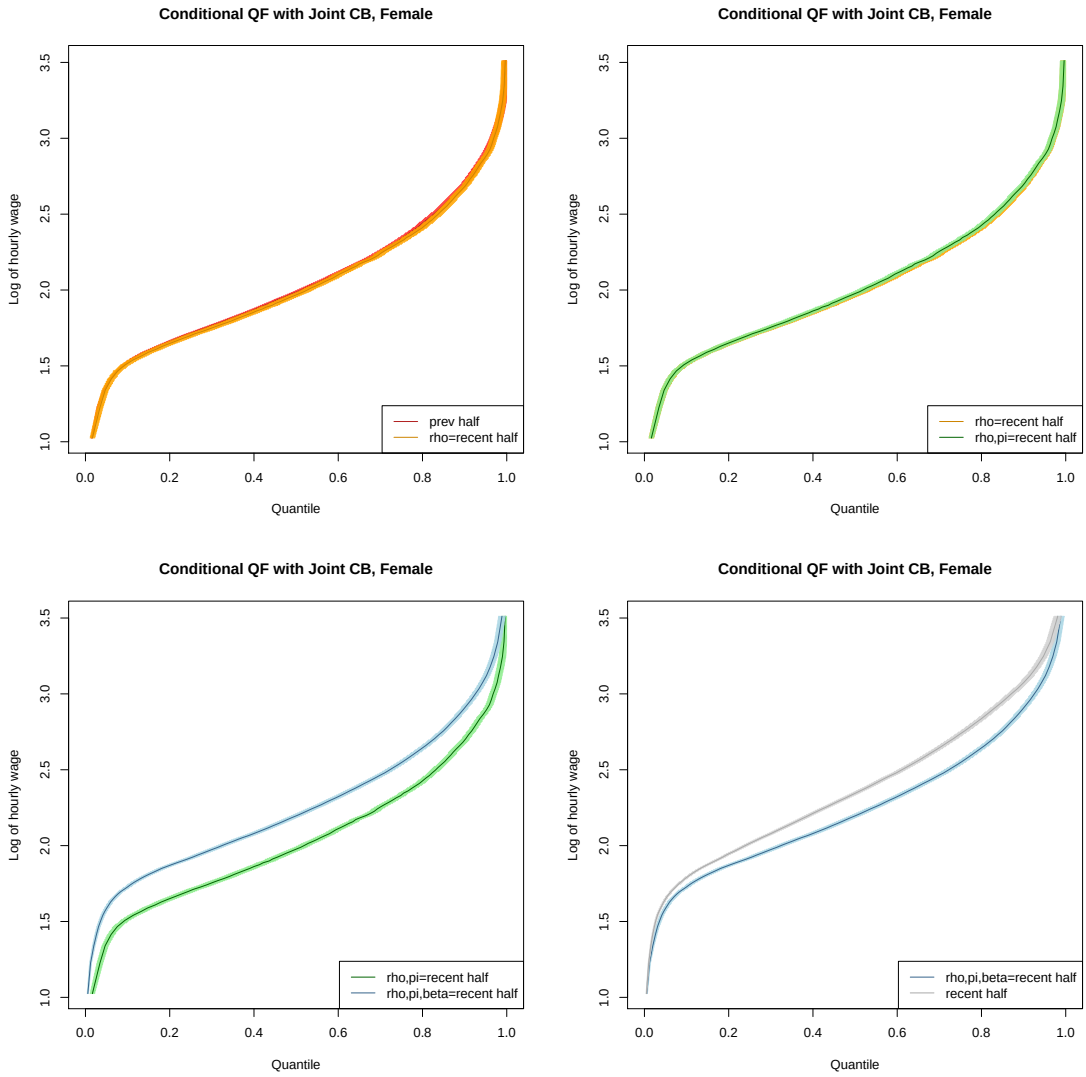

FIGURE 30. Estimates and 95% confidence bands for components of wage decomposition between first and second half of sample period for women in specification 2