Chernozhukov, Victor; Fernández-Val, Iván; Luo, Siyi

Working Paper

Distribution regression with sample selection, with an application to wage decompositions in the UK

cemmap working paper, No. CWP68/18

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Fernández-Val, Iván; Luo, Siyi (2018) : Distribution regression with sample selection, with an application to wage decompositions in the UK, cemmap working paper, No. CWP68/18, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2018.6818

This Version is available at:
https://hdl.handle.net/10419/189818

cemmap
centre for microdata methods and practice

Distribution regression with sample selection, with an application to wage decompositions in the UK

Victor Chernozhukov
Iván Fernández-Val
Siyi Luo

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP68/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

DISTRIBUTION REGRESSION WITH SAMPLE SELECTION, WITH AN APPLICATION TO WAGE DECOMPOSITIONS IN THE UK

VICTOR CHERNOZHUKOV, IVÁN FERNÁNDEZ-VAL, AND SIYI LUO

Abstract. We develop a distribution regression model under endogenous sample selection. This model is a semiparametric generalization of the Heckman selection model that accommodates much rich patterns of heterogeneity in the selection process and effect of the covariates. The model applies to continuous, discrete and mixed outcomes. We study the identification of the model, and develop a computationally attractive two-step method to estimate the model parameters, where the first step is a probit regression for the selection equation and the second step consists of multiple distribution regressions with selection corrections for the outcome equation. We construct estimators of functionals of interest such as actual and counterfactual distributions of latent and observed outcomes via plug-in rule. We derive functional central limit theorems for all the estimators and show the validity of multiplier bootstrap to carry out functional inference. We apply the methods to wage decompositions in the UK using new data. Here we decompose the difference between the male and female wage distributions into four effects: composition, wage structure, selection structure and selection sorting. We uncover positive sorting for single men and negative sorting for married women that accounts for a substantial fraction of the gender wage gap at the top of the distribution. These findings can be interpreted as evidence of assortative matching in the marriage market and glass-ceiling in the labor market.

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

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), and product demands because we do not observed quantities purchased by consumers who do not have access to the product. Sample selection biases the estimation of causal or predictive effects when the reasons for not observing the data are related to the

Date: November 29, 2018.

MIT, BU and BU. First discussion of March, 2017. We thank Manuel Arellano, Marianne Bitler, Stephane Bonhomme, James Heckman, Dennis Kristensen and the seminar participants at Banco Central de Chile, Bristol, BU, Cambridge, Chicago, Northwestern, NYU, Oxford, Queen Mary and UCL 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 was visiting UCL while working on this paper, he is grateful for their hospitality.

outcome of interest. For example, there is sample selection bias in the estimation of a wage equation whenever the employment status and offered wage depend on common unobserved variables such as ability, motivation or skills. The most popular solution to the sample selection bias is the Heckman selection model (HSM) introduced in Heckman (1974). This model offers a convenient and parsimonious way to account for sample selection by making parametric assumptions about the outcome and selection processes. Our development is motivated by the observation that, in addition to the parametric structure, this model imposes strong homogeneity assumptions on how covariates affect the outcome and selection processes and how the selection process operates itself. We develop a generalization of the HSM that relaxes all these three homogeneity restrictions. The resulting model is a semiparametric model, where key parameters are function-valued, thereby considerably generalizing the classical selection model.

Following the literature, we model sample selection using two latent variables for the selection and outcome processes and relate the distribution of these variables with the distribution of the corresponding observed variables. Here we find convenient to work with a local Gaussian representation (LGR) of the joint distribution of the latent variables, which we introduce in the paper. This representation is unique for any joint distribution and might be of independent interest in other settings. The identification analysis is very transparent with the LGR. Thus, we show that the parameters of the LGR are partially identified in the presence of endogenous sample selection because there are only two free probabilities to identify three parameters. We rely on exclusion restrictions to point-identify the three parameters nonparametrically. These conditions require of a binary covariate that does not affect the distribution of the latent outcome and dependence between the latent selection and outcome variables.

Once we have established nonparametric identification with the exclusion restrictions, we introduce a flexible semiparametric distribution regression (DR) model with covariates for the LGR. This model generalizes the HSM by adding multiple sources of heterogeneity to the selection and outcome processes. Thus, it allows for observed and unobserved heterogeneity in selection sorting, together with unobserved heterogeneity in the effect of the covariates on the selection sorting and outcome. In the case of the wage equation, the model can capture the presence of heterogenous returns to schooling across the wage distribution, or positive sorting at the top of the wage distribution and negative sorting at the bottom. The model is semiparametric because its parameters are function-valued and can be applied without modification to continuous, discrete and continuous-discrete outcomes. We show how to construct interesting functionals of the model parameters such as actual and counterfactual distributions of latent and observed outcomes, which can be applied to policy evaluation, treatment effects, wage decompositions and discrimination analysis accounting for sample selection. In the case of wage decompositions, we show how to identify two new effects: a selection sorting effect and a selection structure effect. Selection sorting is determined by whether the employed individuals have higher or lower offered or latent wages than unemployed individuals with the same characteristics. Selection structure

is determined by the proportion of employed individuals and how they are selected based on observed characteristics.

We develop a two-step estimator for the model parameters. The first step consists of a probit regression for the selection equation, which is identical to the first step in the Heckman two-step method (Heckman, 1979). The second step estimates multiple DRs with sample selection correction. The difference between these DRs and the standard DRs without sample selection is that we run bivariate probits instead of univariate probits (Foresi and Peracchi, 1995; Chernozhukov, Fernández-Val, and Melly, 2013). We estimate functionals of the parameters using the plug-in method. We derive functional central limit theorems for all the estimators and show how to use these results to perform uniform inference on function-valued parameters. This type of inference is useful to construct confidence bands and test hypotheses such as whether a coefficient or effect is uniformly zero, constant or positive. We implement the inference methods using Kolmogorov-Smirnov type statistics where the critical values are obtained via multiplier bootstrap (Giné and Zinn, 1984). This bootstrap scheme is convenient in our setting because it avoids repeated computation of estimators in constructing the bootstrap draws of the statistic. We prove the validity of multiplier bootstrap by deriving bootstrap functional central limit theorems for all the estimators.

We apply our methods to study the relationship between wage and employment in the U.K. using updated data from 1978 to 2013. To this end we estimate wage equations for men and women and carry out several wage decompositions accounting for endogenous selection into employment. Here, we uncover positive sorting among single men and negative sorting among married women. This difference in selection sorting is consistent with assortative matching in the marriage market. It also explains a significant proportion of the gender wage gap at the top of the distribution, which is consistent with recent explanations based on glass ceiling theory. We also find that most of the gender wage gap in both observed and offered wages is accounted by differences in the wage structure that are often associated with gender discrimination in the labor market. The effect of education is positive and increases along the distribution. All the heterogeneity that we find is inconsistent with the restrictions of the HSM.

Literature review. The sample selection problem has a long history in statistics and econometrics. Classical references can be found in Lee (1982), Goldberger (1983), Amemiya (1985, Section 10.7), Maddala (1986, Section 9.4), Manski (1989), Manski (1994), and Vella (1998). A popular solution to the problem is the HSM developed by Heckman in a sequence of papers (Heckman, 1974; Heckman, 1976; Heckman, 1979; Heckman, 1990). This model has been extended in several dimensions. Thus, Lee (1983), Prieger (2002) and Smith (2003) replaced the bivariate standard normal copula with other parametric copulas, and Marchenko and Genton (2012) replaced the bivariate normal by a bivariate t-distribution to apply the HSM to heavy tailed data. Ahn and Powell (1993), Powell (1994), Andrews and Schafgans (1998), and Newey (1999) developed semiparametric versions of the HSM and Das, Newey, and Vella (2003) a nonparametric version, focusing on location effect versions with homogeneous effects. None of the

models considered in these extensions accommodates all the sources of heterogeneity allowed by our model.

Arellano and Bonhomme (2017a) proposed another extension of the HSM, which like our model allows for multiple sources of heterogeneity.[1] Their method relies on quantile regression to model the marginal distribution of the latent outcome coupled with a parametric model for the copula of the latent selection and outcome variables. They estimate the model parameters using a three-step method where the first step is the same as in our method, but the second and third steps involve an iterative procedure that alternates between quantile regressions to estimate the outcome equation and nonlinear GMM to estimate the parameters of the copula. They also rely on numerical simulation to estimate functionals of the parameters such as actual and counterfactual distributions of the latent and observed outcomes. Compared to our method, they model the covariates effects as direct on the conditional quantile of the latent distribution, whereas we model the covariate effects as direct on the conditional latent distribution – hence in our framework covariates affect the conditional quantiles indirectly. Further, their modeling approach imposes homogeneity on the copula function, which rules out forms of copula heterogeneity across the distribution of the latent outcome, which are permitted in our approach. Moreover, their quantile regression model requires the latent outcome to be continuous, whereas our distribution regression model can deal with any type of outcome and is therefore more widely applicable. Our method is computationally simpler as it does not involve any iteration between methods in the second step. The identification assumptions are also different and not nested: we impose more structure on the dependence between the outcome and selection processes, whereas they require more variation on the excluded covariates. We provide a more detailed comparison of the identifying assumptions in Appendix A. Finally, from a technical point of view, Arellano and Bonhomme (2017a) only derived pointwise limit theory for the estimators of the model parameters, whereas we derive functional limit theory for the estimators of the model parameters and related functionals.

Outline. Section 2 looks at 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 provides asymptotic theory for the estimation and inference methods, and Section 5 reports the results of the empirical application.

2. Another View of the Sample Selection Problem

2.1. **Local Gaussian Representation of a Joint Distribution.** We start by characterizing a local Gaussian representation (LGR) of the joint distribution of two random variables that is convenient to provide a new view of the identification problem with sample selection and motivate our modeling choices later.

[1]See Arellano and Bonhomme (2017b) for a recent survey on sample selection in quantile models.

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. Our first result shows that F_{Y^*,D^*} can be represented via a standard bivariate normal distribution at a point and with a correlation parameter that depend on the evaluation point (y, d).

Lemma 2.1 (LGR Result)**.** *Let F_{Y^*,D^*} be a joint CDF, then, for any $(y, d) \in \mathbb{R}^2$,*

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

where $\mu(y) \in \mathbb{R}$, $\nu(d) \in \mathbb{R}$, $\rho(y, d) \in [-1, 1]$, and $\Phi_2(\cdot, \cdot; \rho)$ is the joint CDF of a standard bivariate normal random variable with parameter ρ. Moreover, the values of $\mu(y)$, $\nu(d)$ and $\rho(y, d)$ are uniquely determined by $\mu(y) = \Phi^{-1}(F_{Y^*}(y))$, $\nu(d) = \Phi^{-1}(F_{D^*}(d))$, and the solution in ρ of

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

where Φ is the standard normal CDF. Hence, the representation is unique.

Proof. By standard properties of the bivariate normal distribution, the marginals corresponding to the LGR are $\Phi(\mu(y))$ and $\Phi(\nu(d))$. Equalizing these marginals to the marginals of $F_{Y^*,D^*}(y, d)$ yields

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

which uniquely determine $\mu(y)$ and $\nu(x)$ as $\mu(y) = \Phi^{-1}(F_{Y^*}(y))$ and $\nu(d) = \Phi^{-1}(F_{D^*}(d))$. Plugging these expressions in the LGR gives

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

Let $\phi_2(\cdot, \cdot; \rho)$ be the joint probability density function (PDF) of a standard bivariate normal random variable with parameter ρ. The previous equation uniquely determines $\rho(y, d)$ 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

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

□

Lemma 2.1 establishes that any joint CDF admits a unique representation as a sequence of standard bivariate normal distributions. 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). Lemma 2.1 easily extends to CDFs conditional on covariates X by making all the parameters dependent on the value of X.

The parameter $\rho(y,d)$ can be interpreted as a measure of local dependence.[2] Thus, when $\rho(y,d)=0$, the distribution F_{Y^*,D^*} factorizes at (y,d):

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

that is, the events $\{Y^*\leqslant y\}$ and $\{D^*\leqslant d\}$ are independent. Hence we can say that Y^* and D^* are "locally independent" at (y,d).[3] In general, the discrepancy

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

measures deviation away from independent factorization, thereby giving meaning to $\rho(y,d)$ as local dependence parameter. Further, in some cases, $\rho(y,d)$ can also be interpreted as capturing local linear dependence. For example, if $\rho(y,d)=\rho(y)$ the LGR factorizes as

$$F_{Y^*,D^*}(y,d)=\int_{-\infty}^{\Phi^{-1}(F_{Y^*}(y))}\Phi\left(\frac{\Phi^{-1}(F_{D^*}(d))-\rho(y)u}{\sqrt{1-\rho(y)^2}}\right)\phi(u)du.$$

Here, $\rho(y)$ is a local correlation between the "standard-normalized" variables $\Phi^{-1}(F_{D^*}(D^*))$ and $\Phi^{-1}(F_{Y^*}(Y^*))$ at $Y^*=y$. Indeed, if $D^*\sim\mathcal{N}(\nu,1)$ and $Y^*\sim\mathcal{N}(\mu,\sigma^2)$,

$$D^*\mid Y^*=y\sim\mathcal{N}(\nu+\rho(y)(y-\mu)/\sigma,1-\rho(y)^2),$$

so that $\rho(y)$ is the coefficient of a local linear regression of D^* on Y^* at $Y^*=y$.[4]

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

$$\begin{aligned}&C_{Y^*,D^*}(u,v)=\Phi_2(\Phi^{-1}(u),\Phi^{-1}(v);\rho(y_u,d_v)),\\&\forall(u,v)\in[0,1]:\exists y_u\in\mathbb{R}:F_{Y^*}(y_u)=u,\quad\exists d_v\in\mathbb{R}:F_{D^*}(d_v)=v.\end{aligned}\tag{2.1}$$

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^*.[5]

Kolev, Anjos, and Mendes (2006) developed a closely related result to (2.1) for the copula. They established that the copula of any bivariate distribution can be represented by the bivariate

[2]See Tjøstheim, Otneim, and Støve (2018) for a recent survey on measures of local dependence.

[3]This concept is related to, but different from the local independence of Doksum, Blyth, Bradlow, Meng, and Zhao (1994), which is local in only one of the variables. Thus, for example, our concept is symmetric in Y^* and D^* whereas the concept in Doksum, Blyth, Bradlow, Meng, and Zhao (1994) is not.

[4]In this case $\rho(y)$ corresponds to the correlation curve of Bjerve and Doksum (1993).

[5]Note that the marginals of F_{Y^*,D^*} do not identify separately the mean and variances of the local Gaussian representation.

Gaussian copula with a local correlation parameter. The LGR is different from the local Gaussian approximation of Tjøstheim and Hufthammer (2013), which approximates the distribution of a continuous bivariate random variable in a neighborhood of the point of interest by a bivariate normal distribution with local vector of means and variance-covariance matrix, see also Hjort and Jones (1996). As Tjøstheim and Hufthammer (2013) noticed, unlike the LGR, a local Gaussian approximation that intersects with the joint distribution at the point of interest is not unique.

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.

The joint CDF of the observed variables can be written in terms of the LGR of F_{Y^*,D^*} as

$$F_{Y,D}(y,d) = \Phi_2(\mu(y), \nu(0); \rho(y,0))1(d \geqslant 1) + [1 - \Phi(\nu(0))]1(d \geqslant 0).$$

As shown below, the parameters of the LGR are partially identified because F_{Y^*,D^*} is only partially identified. We proceed by characterizing the identified set for these parameters and provide exclusion restrictions to achieve point identification. Since there is a one-to-one relationship between F_{Y^*,D^*} and its LGR, the identified set for the parameters of the LGR determine the identified set for F_{Y^*,D^*}. In what follows, we simplify the notation to

$$\nu := \nu(0), \quad \rho(y) := \rho(y,0).$$

We can only hope to identify $\nu(d)$ and $\rho(y,d)$ at $d = 0$ since we only observe whether $D^* \leqslant 0$.

To understand the source of the partial identification in terms of the LGR, note that in the presence of sample selection there are two free probabilities, $\mathrm{P}(D = 1)$ and $\mathrm{P}(Y \leqslant y \mid D = 1)$, to identify three parameters, $\mu(y)$, ν and $\rho(y)$. The parameter ν is pinned down by the selection probability as

$$\nu = \Phi^{-1}(\mathrm{P}(D = 1)).$$

The parameters $\mu(y)$ and $\rho(y)$ are partially identified as the solutions in (μ, ρ) to

$$F_{Y,D}(y,1) - \mathrm{P}(D = 0) = \Phi_2(\mu, \Phi^{-1}(\mathrm{P}(D = 1)); \rho).$$

These solutions form a one-dimensional manifold in $\mathbb{R} \times [-1,1]$ because $\partial\Phi_2(\mu,\cdot;\rho)/\partial\mu > 0$, $\partial\Phi_2(\cdot,\cdot;\rho)/\partial\rho > 0$, and $\partial^2\Phi_2(\cdot,\cdot;\rho)/\partial\mu\partial\rho > 0$ (Spivak, 1965; Munkres, 1991). The identified set of $(\mu(y), \rho(y))$ can be shrunk using additional information such as that $\rho(y)$ is known to lie in a subinterval of $[-1,1]$, e.g. $|\rho(y)| < 0.2$.

We use exclusion restrictions to achieve point identification of the parameters of the LGR. To state these restrictions in terms of the LGR, we start by extending the LGR to CDFs conditional on covariates. Let Z be a random 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:

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

where $\mu(y \mid z) \in \mathbb{R}$, $\nu(d \mid z) \in \mathbb{R}$, and $\rho(y,d \mid z) \in [-1,1]$. This representation can be characterized using the same argument as in Lemma 2.1 after fixing the value of the covariate Z and letting all the parameters of the LGR to depend on this value. The exclusion restrictions are:

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:* $\mu(y \mid z) = \mu(y)$.
(4) *Selection exclusion:* $\rho(y,0 \mid z) = \rho(y,0)$.

The condition that Z is binary is made to emphasize 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. Part (1) requires that there is sample selection and that Z has variation in the selected population. It is used to guarantee that all the probabilities employed in the identification analysis are well-defined. Part (2) requires that Z affects the probability of selection and rules out corner cases. The condition $\mathrm{P}(D=1 \mid Z=1) < 1$ precludes identification at infinity (see Remark 2.1). The sign of the inequality can be reversed by relabelling the values of Z. Part (3) is a standard exclusion restriction, which is not sufficient for point identification in the presence of sample selection (Manski, 1994; Manski, 2003). It holds if Y^* is independent of Z.[6] Part (4) is an exclusion restriction in the local dependence between Y^* and D^*, which holds if selection sorting is independent of Z. We explain this condition more in detail below with an example and compare it with the identification condition in Arellano and Bonhomme (2017a) in Remark 2.2.

We can get some intuition about the outcome and selection exclusion conditions with an example. Consider a heteroskedastic bivariate normal model for the latent variables, i.e.,

$$(Y^*, D^*) \mid Z = z \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_{Y^*}(z) \\ \mu_{D^*}(z) \end{bmatrix}, \begin{bmatrix} \sigma_{Y^*}(z)^2 & \sigma_{Y^*,D^*}(z) \\ \sigma_{Y^*,D^*}(z) & \sigma_{D^*}(z)^2 \end{bmatrix} \right).$$

The outcome exclusion imposes that

$$\frac{y - \mu_{Y^*}(z)}{\sigma_{Y^*}(z)} = \mu(y),$$

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

whereas the selection exclusion imposes that

$$\frac{\sigma_{Y^*,D^*}(z)}{\sigma_{Y^*}(z)\sigma_{D^*}(z)} = \rho.$$

If Y^* is independent of Z, the outcome exclusion holds and the selection exclusion boils down to the condition that $\sigma_{Y^*,D^*}(z)/\sigma_{D^*}(z)$ does not depend on z. In other words, the slopes of the linear regressions of Y^* on D^* conditional on Z are the same when $Z = 0$ and $Z = 1$.

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

$$F_{Y^*,D^*|Z}(y, 0 \mid z) = \Phi_2(\mu(y), \nu(z); \rho(y)), \quad z \in \{0, 1\}, \tag{2.2}$$

where $\nu(z) := \nu(0 \mid z)$ and $\rho(y) := \rho(y, 0)$. The CDF of the observed variables conditional on Z can be related to this conditional LGR as

$$F_{Y,D|Z}(y, d \mid z) = \Phi_2(\mu(y), \nu(z); \rho(y))1(d \geqslant 1) + [1 - \Phi(\nu(z))]1(d \geqslant 0), \quad z \in \{0, 1\}.$$

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

$$\nu(z) = \Phi^{-1}\left(\mathrm{P}(D = 1 \mid Z = z)\right), \quad z \in \{0, 1\}. \tag{2.3}$$

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

$$F_{Y,D|Z}(y, 1 \mid z) - \mathrm{P}(D = 0 \mid Z = z) = \Phi_2(\mu, \Phi^{-1}\left(\mathrm{P}(D = 1 \mid Z = z)\right); \rho), \quad z \in \{0, 1\}. \tag{2.4}$$

This is a nonlinear system of two equations in two unknowns that has unique solution because the Jacobian is a P-matrix for all $\mu \in \mathbb{R}$ and $\rho \in (-1, 1)$ by Theorem 4 of Gale and Nikaido (1965).

The following theorem provides a detailed identification analysis of the parameters of the conditional LGR in (2.2). It includes boundary cases where the parameters $\mu(y)$ and $\rho(y)$ can be either point or partially identified.

Theorem 2.1 (Identification of LGR with Sample Selection)**.** *Assume that Assumption 1 holds. Then, $\nu(z)$ is identified by* (2.3) *and there are the following cases for the identification of $\mu(y)$ and $\rho(y)$:*

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

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

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

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

(3) *If* $F_{Y,D|Z}(y, 1 \mid 1) = F_{Y,D|Z}(y, 1 \mid 0) = 1$,

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

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

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

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

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

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

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

(7) *Otherwise,* $\mu(y)$ *and* $\rho(y)$ *are point identified as the solution in* (μ,ρ) *to* (2.4)*. This solution exists and is unique.*

Proof. The identification of $\nu(z)$ follows from equalizing the marginals with respect to D^* of $F_{Y^*,D^*|Z}$ and the conditional LGR at $D^* = 0$. Since $\nu(z)$ is identified, we shall use $\Phi(\nu(z))$ in place of $\mathrm{P}(D = 1 \mid Z = z)$ and $\bar{\Phi}(\nu(z))$ in place of $\mathrm{P}(D = 0 \mid Z = z)$ in the rest of the proof to lighten the notation.

Cases (1)–(3) correspond to $\rho(y) = 1$. This boundary case is identified because $\rho(y) = 1$ if and only if $F_{Y,D|Z}(y,1 \mid 1) - \bar{\Phi}(\nu(1)) = F_{Y,D|Z}(y,1 \mid 0) - \bar{\Phi}(\nu(0)) > 0$ or $F_{Y,D|Z}(y,1 \mid 0) = 1$. The if part follows from the Frechet-Hoeffding bounds

$$F_{Y,D|Z}(y,1 \mid z) - \bar{\Phi}(\nu(z)) = F_{Y^*,D^*|Z}(y,0 \mid z) = \min[\Phi(\nu(z)), F_{Y^*}(y)],\quad z \in \{0,1\}, \tag{2.5}$$

and Assumption 1(2). For the case $F_{Y,D|Z}(y,1 \mid 1) - \bar{\Phi}(\nu(1)) = F_{Y,D|Z}(y,1 \mid 0) - \bar{\Phi}(\nu(0)) > 0$, the only if part follows because $\nu \mapsto \Phi_2(\cdot,\nu;\rho)$ is strictly monotonic when $\rho \in (-1,1)$ and $\nu(1) > \nu(0)$ by Assumption 1(2). This shows that $\rho \notin (-1,1)$. Moreover, this case is ruled out when $\rho = -1$ by the Frechet-Hoeffding bounds

$$F_{Y,D|Z}(y,1 \mid z) - \bar{\Phi}(\nu(z)) = F_{Y^*,D^*|Z}(y,0 \mid z) = \max[\Phi(\nu(z)) + F_{Y^*}(y) - 1, 0],\quad z \in \{0,1\}, \tag{2.6}$$

and Assumption 1(2). The case $F_{Y,D|Z}(y,1 \mid 0) = 1$ implies that $\Phi_2(\mu(y),\nu(0);\rho(y)) = \Phi(\nu(0))$, which is only possible when $\rho(y) = 1$.

Now, we can analyze the identification of $\mu(y)$ using (2.5) with $F_{Y^*}(y) = \Phi(\mu(y))$. Case (1) corresponds to $F_{Y,D|Z}(y,1 \mid 1) - \bar{\Phi}(\nu(1)) = \Phi(\mu(y))$, which identifies $\mu(y)$. Case (2) corresponds to $F_{Y,D|Z}(y,1 \mid 0) = 1$ and $F_{Y,D|Z}(y,1 \mid 1) - \bar{\Phi}(\nu(1)) = \Phi(\mu(y))$. The second equation identifies $\mu(y)$. Case (3) corresponds to $F_{Y,D|Z}(y,1 \mid z) = 1$, $z \in \{0,1\}$, which partially identify the parameter from $\Phi(\mu(y)) \geqslant \max[\Phi(\nu(0)), \Phi(\nu(1))] = \Phi(\nu(1))$ by Assumption 1(2).

Cases (4)–(6) correspond to $\rho(y) = -1$. This boundary case is identified because $\rho(y) = -1$ if and only if $F_{Y,D|Z}(y,1 \mid 0) = \bar{\Phi}(\nu(0))$ or $F_{Y,D|Z}(y,1 \mid 0) = F_{Y,D|Z}(y,1 \mid 1)$. Symmetrically to $\rho(y) = 1$, the if part follows from the Frechet-Hoeffding bounds (2.6) and Assumption 1(2), whereas the only if part for the case $F_{Y,D|Z}(y,1 \mid 0) = F_{Y,D|Z}(y,1 \mid 1)$ follows from the Frechet-Hoeffding bounds (2.5) and Assumption 1(2). The only if part for $F_{Y,D|Z}(y,1 \mid 0) = \bar{\Phi}(\nu(0))$ follows because this case implies that $\Phi_2(\mu(y),\nu(0);\rho(y)) = 0$, which is only possible when $\rho(y) = -1$.

Now, we can analyze the identification of $\mu(y)$ using (2.6) with $F_{Y^*}(y) = \Phi(\mu(y))$. Case (4) corresponds to $F_{Y,D|Z}(y, 1 \mid 1) = \Phi(\mu(y))$, which identifies $\mu(y)$. Case (5) corresponds to $F_{Y,D|Z}(y, 1 \mid 0) = \Phi(\mu(y))$ and $F_{Y,D|Z}(y, 1 \mid 1) = \Phi(\mu(y))$. Both of these equations have the same solution that identifies $\mu(y)$. Case (6) corresponds to $F_{Y,D|Z}(y, 1 \mid z) = \bar{\Phi}(\nu(z))$, $z \in \{0, 1\}$, which partially identify the parameter from $\Phi(\mu(y)) \leqslant \min[\bar{\Phi}(\nu(0)), \bar{\Phi}(\nu(1))] = \bar{\Phi}(\nu(1)) = F_{Y,D|Z}(y, 1 \mid z)$ by Assumption 1(2).

Consider now the non-boundary case (7) where $\rho(y) \in (-1, 1)$. The parameters $\mu(y)$ and $\rho(y)$ are identified as the solution in (μ, ρ) to (2.4). 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 (2.4) is a P-matrix when $\rho(y) \in (-1, 1)$.

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

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

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

The boundary cases in Theorem 2.1 are easy to detect. In practice, partial identification usually occurs at extreme values of y. For example, case (3) arises for values of y such that $Y > y$ a.s., and case (6) for values of y such that $Y < y$ a.s.

Remark 2.1 (Identification at Infinity)**.** When $\mathrm{P}(D = 1 \mid Z = 1) = 1$ and $\mu(y \mid z) = \mu(y)$, the conditional LGR at $z = 1$ gives $F_{Y,D|Z}(y, 1 \mid 1) = \lim_{\nu \nearrow +\infty} \Phi_2(\mu(y), \nu; \rho(y \mid 1)) = \Phi(\mu(y))$, which identifies $\mu(y)$ by

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

without the selection exclusion restriction. This result is analogous to the identification at infinity of Chamberlain (1986) where Z is continuous with unbounded support and

$$\lim_{z \nearrow +\infty} \mathrm{P}(D = 1 \mid Z = z) = 1.$$

Note that $\rho(y \mid z)$ is not point identified without further restrictions.

Remark 2.2 (Comparison with Arellano and Bonhomme (2017a), AB17)**.** Assumption 1 is not nested with the conditions that AB17 used to show nonparametric identification of their model. We impose stronger restrictions in the dependence of the latent selection and outcome variables, but require less variation in the excluded covariate Z. We provide a more detailed comparison in Appendix A □

3. Distribution Regression Model with Sample Selection

3.1. **The Model.** We consider a semiparametric version of the LGR with covariates:

$$F_{Y^*,D^*}(y, 0 \mid Z = z) = \Phi_2(-x'\beta(y), -z'\pi; \rho(x'\delta(y))), \tag{3.7}$$

where Y^* is the latent outcome of interest, which can be continuous, discrete or mixed continuous-discrete; 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. The excluded covariates avoid reliance on functional form assumptions to achieve identification. The model for the LGR consists of three indexes. 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 $D = 1$.[7] In the empirical application that we consider below, Y^* is offered wage, D^* is 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 5.

The model (3.7) is semiparametric 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 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. The function $u \mapsto \rho(u)$ is a known link with range $[-1, 1]$, e.g.

[7]The minus signs in (3.7) 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 Heckman selection model; see Example 1.

the Fisher transformation (Fisher, 1915), $\rho(u) = \tanh(u)$. The corresponding distribution of Y^* conditional on Z is

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

The selection bias arises because this distribution is different from the distribution of the observed outcome Y, i.e.

$$F_{Y^*}(y \mid Z = z) \neq F_Y(y \mid Z = z, D = 1) = \frac{\Phi_2(-x'\beta(y), z'\pi; -\rho(x'\delta(y)))}{\Phi(z'\pi)}.$$

Example 1 (HSM)**.** Consider the Heckman (1974) sample selection model (HSM):

$$\begin{aligned} D^* &= Z'\pi + V, \\ Y^* &= X'\beta + \sigma U, \end{aligned}$$

where (U, V) is independent of Z and has standard bivariate normal distribution with parameter ρ, such that

$$F_{Y^*,D^*}(y, 0 \mid Z = z) = \Phi_2\left(\frac{y - x'\beta}{\sigma}, -z'\pi; \rho\right).$$

This is a special case of model (3.7) with

$$\beta_1(y) = (\beta_1 - y)/\sigma, \quad \beta_{-1}(y) = \beta_{-1}/\sigma, \quad \rho(x'\delta(y)) = \rho.$$

The HSM therefore imposes strong homogeneity restrictions in the selection process and effect of the covariates on the outcome and selection sorting. Thus, only the intercept of $\beta(y)$ varies with y, and $\rho(x'\delta(y))$ is invariant to both x and y.

The model (3.7) has multiple data generating process representations as nonseparable systems. One example is

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

where U and V are independent. For example, in the wage application V can be interpreted as unobserved net benefit of working, and U as unobserved skills or innate ability net of V. This representation is similar to the HSM in Example 1 with the difference that the equation for Y^* is nonseparable.[8]

[8]Note that in Example 1 the equation for Y^* can be written as $0 = (X'\beta - Y^*)/\sigma + \rho V + \sqrt{1 - \rho^2}\widetilde{U}$, where $\widetilde{U}$ is standard normally distributed and independent of V and Z.

3.2. **Functionals.** There are several functionals of the parameters of the model (3.7) that can be of interest. One is the marginal distribution of the latent outcome Y^*

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

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. We can also construct counterfactual distributions by combining coefficients $\beta(y)$ and distributions F_X from different populations or groups. These distributions are useful to decompose the distribution of offered wages between females and males or between blacks and whites, which can be used to uncover discrimination in the labor market.

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

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

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.

Remark 3.1 (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 probability density function (PDF), and $\phi_2(\cdot, \cdot; \rho)$ be the joint PDF of a standard bivariate normal random variable with parameter ρ.[9] Increasing ρ therefore shifts the distribution to the right (increases quantiles) because it makes selection sorting more positive while the size of the 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

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

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 variation in the parameters changes the composition of the selected population. Consider the following extreme 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 to 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. □

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 $\mathcal{Y}$ 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.[10]

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

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

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 $y \in \mathcal{Y}$. 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 $y \in \mathcal{Y}$. 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:*

$$\widehat{\pi} = \underset{c \in \mathbb{R}^{d_\pi}}{\arg\max}\, L_1(c) = \frac{1}{n}\sum_{i=1}^{n}\left[D_i \log \Phi(Z_i'c) + (1 - D_i)\log \Phi(-Z_i'c)\right], \quad d_\pi := \dim \pi.$$

(2) Run multiple DRs with sample selection correction to estimate θ_y: for each $y \in \mathcal{Y}$

$$\begin{aligned}\widehat{\theta}_y = \underset{t=(b,d)\in\Theta}{\arg\max}\, L_2(t, \widehat{\pi}) = \frac{1}{n}\sum_{i=1}^{n} D_i \big[I_{yi} \log \Phi_2\left(-X_i'b, Z_i'\widehat{\pi}; -\rho(X_i'd)\right) \\ + (1 - I_{yi}) \log \Phi_2\left(X_i'b, Z_i'\widehat{\pi}; \rho(X_i'd)\right)\big],\end{aligned}$$

where $\Theta \in \mathbb{R}^{d_\theta}$ is a compact parameter set, and

$$d_\theta := \dim \theta_u, \quad \rho(u) = \tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \in [-1, 1], \quad \frac{\partial \rho(u)}{\partial u} > 0.$$

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

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

outcome is

$$\widehat{F}_{Y^*}(y) = \frac{1}{n}\sum_{i=1}^{n} \Phi(-X_i'\widehat{\beta}(y)), \tag{3.8}$$

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

$$\widehat{F}_Y(y \mid D=1) = \frac{\sum_{i=1}^n \Phi_2(-X_i'\widehat{\beta}(y), Z_i'\widehat{\pi}; -\rho(X_i'\widehat{\delta}(y)))}{\sum_{i=1}^n \Phi(Z_i'\widehat{\pi})}, \tag{3.9}$$

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. For example, the estimator of the τ-quantile of the latent outcome is $\widehat{Q}_{Y^*}(\tau) = \mathbf{Q}_\tau(\widehat{F}_{Y^*})$.

3.4. **Uniform Inference.** 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 be 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 $y \mapsto c'\theta_y$, $y \in \mathcal{Y}$, 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

$$\mathrm{P}\left[c'\theta_y \in CB_p(c'\theta_y), \forall y \in \mathcal{Y}\right] \to p, \text{ as } n \to \infty.$$

We form $CB_p(c'\theta_y)$ as

$$CB_p(c'\theta_y) = c'\widehat{\theta}_y \pm cv(p) SE(c'\widehat{\theta}_y),$$

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 $cv(p)$ is a critical value, i.e. a consistent estimator of the p-quantile of the statistic

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

We obtain the standard error and critical value from the limit distribution of $\widehat{\theta}_y$ derived in Section 4.

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 $b \in 1, \ldots, B$ and the finite grid $\bar{\mathcal{Y}} \subseteq \mathcal{Y}$, 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,*

$$\omega_i^b = \tilde{\omega}_i^b - \sum_{i=1}^n \tilde{\omega}_i^b/n, \quad \tilde{\omega}_i^b \sim \ \textit{i.i.d.}\ \mathcal{N}(0,1).$$

(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 estimators of the influence function of* $\widehat{\theta}_y$ *given in* (4.12). *(2) Construct bootstrap realization of maximal t-statistic* $t_{\mathcal{Y}}$ *for the functional of interest,*

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

where $\widehat{\Sigma}_{\theta_y\theta_y}$ *is an estimator of the asymptotic variance-covariance matrix of* $\widehat{\theta}_y$ *given in* (4.11). *(ii) Compute the critical value* $cv(p)$ *as the simulation p-quantile of* $t_{\bar{\mathcal{Y}}}^b$,

$$cv(p) = p - \textit{quantile of } \{t_{\mathcal{Y}}^b : 1 \leqslant b \leqslant B\}$$

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. Asymptotic Theory

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

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

$$\Sigma_{\theta_y\theta_{\tilde{y}}} := H_{2y}^{-1}\left\{n\mathrm{E}\left[\widetilde{S}_{2y}\widetilde{S}_{2\tilde{y}}'\right] - J_{21y}H_1^{-1}J_{21\tilde{y}}'\right\}H_{2\tilde{y}}^{-1}, \tag{4.10}$$

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 2 (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* (3.7). *(3) The support of* Z, $\mathcal{Z}$, *is a compact set. (4) The support of* 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 the support of* (Y, 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* $y \in \mathcal{Y}$; *and the matrices* H_1, H_{2y} *and* $\Sigma_{\theta_y\theta_y}$ *are nonsingular for each* $y \in \mathcal{Y}$.

Part (1) is a standard condition about the sampling and selection process, which is designed for cross sectional data. Part (2) imposes the semiparametric 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 $\ell^\infty(\mathcal{Y})$ be the set of bounded functions on $\mathcal{Y}$, and $\rightsquigarrow$ denote weak convergence (in distribution).

Theorem 4.1 (FCLT for $\widehat{\theta}_y$)**.** *Under Assumption 2,*

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

and

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

where $y \mapsto Z_{\theta_y}$ is a zero-mean Gaussian process with uniformly continuous sample paths and covariance function $\Sigma_{\theta_y\theta_{\tilde{y}}}$, $y, \tilde{y} \in \mathcal{Y}$, defined in (4.10)*.*

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

$$\widehat{\Sigma}_{\theta_y\theta_{\tilde{y}}} = n^{-2}\sum_{i=1}^{n}\widehat{\psi}_i(\widehat{\theta}_y, \widehat{\pi})\widehat{\psi}_i(\widehat{\theta}_{\tilde{y}}, \widehat{\pi})'. \tag{4.11}$$

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

$$\widehat{\psi}_i(t, c) = -\widehat{H}_{2y}(t, c)^{-1}\left(S_{2yi}(t, c) - \widehat{J}_{21y}(t, c)\widehat{H}_1(c)^{-1}S_{1i}(c)\right), \tag{4.12}$$

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 $u \in \mathcal{U}$, a totally bounded metric space, and φ_u is an operator that maps $\mathbb{D}_\Delta$ to the set $\ell^\infty(\mathcal{U})$, 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 $\ell^\infty(\mathcal{F})$ that acts on a Donsker set of bounded measurable functions $\mathcal{F}$, 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 $\mathbb{D} := \mathbb{R}^{d_\pi} \times \ell^\infty(\mathcal{Y})^{d_\theta} \times \ell^\infty(\mathcal{F})$. 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:

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

$\mathcal{F}$ contains $\{\Phi(- \cdot' \beta_y) : y \in \mathcal{Y}\}$ 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 $\mathcal{U} = \mathcal{Y}$. The quantile function of the latent outcome is

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

$\mathcal{F}$ is the same as for the distribution of the latent outcome, and $\mathcal{U}$ 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:

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

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

The following result is a corollary of Theorem 4.1 by the functional delta method. Let $UC(\mathcal{Y}, \xi)$ be the set of functions on $\mathcal{Y}$ that are uniformly continuous with respect to ξ, a standard metric on $\mathbb{R}$, and $UC(\mathcal{F}, \lambda)$ be the set of functionals on $\mathcal{F}$ that are uniformly continuous with respect to λ, where $\lambda(f, \tilde{f}) = [\mathrm{P}(f - \tilde{f})^2]^{1/2}$ for any $f, \tilde{f} \in \mathcal{F}$.

Corollary 4.1 (FCLT for $\widehat{\Delta}_u$)**.** *Suppose that Assumption 2 holds, and $(p, t_y, F) \mapsto \varphi_\cdot(p, t_y, F)$, from $\mathbb{D}_\Delta \subset \mathbb{D}$ to $\ell^\infty(\mathcal{U})$ is Hadamard differentiable at (π, θ_y, F_Z), tangentially to $\mathbb{R}^{d_\pi} \times UC(\mathcal{Y}, \xi)^{d_\theta} \times$*

$UC(\mathcal{F},\lambda)$ with derivative $(p,t_y,F)\mapsto\varphi'_\cdot(p,t_y,F)$ that is defined and continuous on $\mathbb{R}^{d_\pi}\times\ell^\infty(\mathcal{Y})^{d_\theta}\times\ell^\infty(\mathcal{F})$. Then,

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

where Z_π and Z_{θ_y} are the random limits in Theorem 4.1, Z_F is a tight F_Z-Brownian bridge, and $u\mapsto Z_{\Delta_u}$ is a tight zero-mean Gaussian process.

Remark 4.1 (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 Hadarmard derivatives.

Remark 4.2 (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 4.1, this differentiability requires of 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 of 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.

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

Assumption 3 (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 B.

Theorem 4.2 (Bootstrap FCLT for $\widehat{\theta}_y$)**.** *Under the conditions of Theorem 4.1 and Assumption 3,*

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

where $y \mapsto Z_{\theta_y}$ is the same Gaussian process as in Theorem 4.1.

The following result is a corollary of Theorem 4.2 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 4.2 (Bootstrap FCLT for $\widehat{\Delta}_u$)**.** *Suppose that the conditions of Corollary 4.1 and Assumption 3 hold. Then,*

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

where Z_{Δ_u} is the same process as in Corollary 4.1.

5. 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.

5.1. **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 (2017a) 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.[11] The three surveys contain repeated cross-sectional observations for women and men. The selection of the sample is

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

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 119,765 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.[12] 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 7.94%, Wales 4.99%, Scotland 8.92%, and Northern Ireland 4.23%).[13]

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 considered were out of work.[14] 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 excluded covariate. In this case the outcome and exclusion restrictions 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.

[12]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.

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

[14]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.

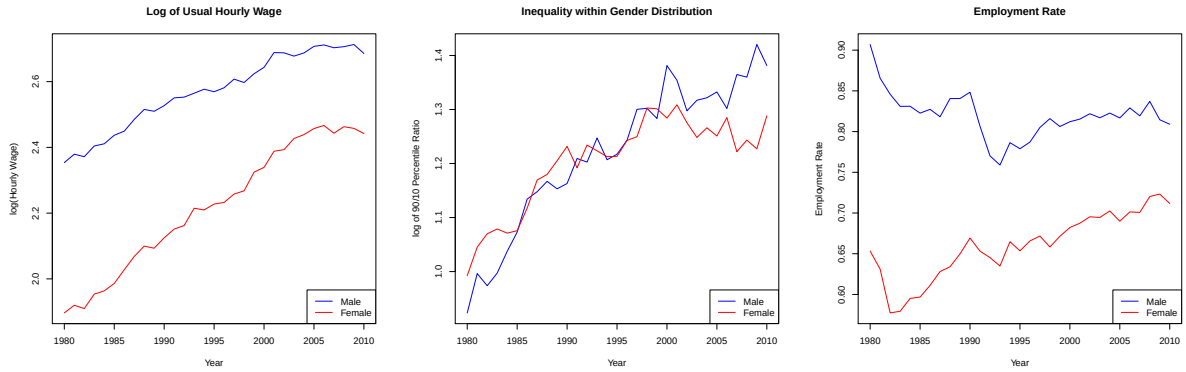

FIGURE 1. Trends in U.K. labor market 1978-2013 by gender: left panel reports the average of the log wage rate, the middle panel reports the 90-10 percentile spread of the log wage rate, and the right panel reports the employment rate

Table 1 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 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 1 provides some background on the U.K. labor market. 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%.

5.2. **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 $X'\delta(y)$ include:

- Specification 1: a constant.

TABLE 1. 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.

- 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.

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 due to imprecision in the estimation of the parameter $\delta(y)$.

5.3. **Model Parameters.** We report point estimates and 95% confidence bands for the coefficients of the education and marital status indicators in the outcome equation and the correlation function in the selection sorting. Estimates and 95% confidence bands for the coefficients of the selection equation, coefficients of the fertility indicators in the outcome equation and coefficients in the selection sorting function are given in the Appendix E. The estimates are obtained with Algorithm 3.1 replacing $\mathcal{Y}$ 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 $B = 200$ bootstrap repetitions and the same finite grid as for the estimates. We also report estimates from the HSM of Example 1 with dash lines as a benchmark of comparison.[15]

The estimates of the coefficients of the education and marital status indicators in the outcome equation are reported in Figure 2 for men and Figure 3 for women. These estimates correspond to specification 1. Estimates for specifications 2–4 are given in Appendix E. For all the specifications and genders, we find that the returns to education are heterogenous across the distribution and broadly increasing in the years of education (age leaving school). The HSM completely misses the heterogeneity and estimates averaged coefficients. The coefficient of the marital status indicator is uniformly positive for men, whereas is negative but mainly statistically not different from zero for women. We cannot reject that this coefficient is homogeneous across the distribution for both men and women.[16]

Figures 4–7 display the estimates of the sorting effect functions for specifications 1–4, respectively. The estimates of the coefficients of these functions for specifications 3 and 4 are given in Appendix E. Figure 4 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 5, 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

[15]We report estimates of β/σ in the outcome equation of the HSM for comparability; see Example 1.

[16]We find more heterogeneity in the coefficient of the marital status indicator in the specifications 2 and 4 that include marital status in the selection sorting function.

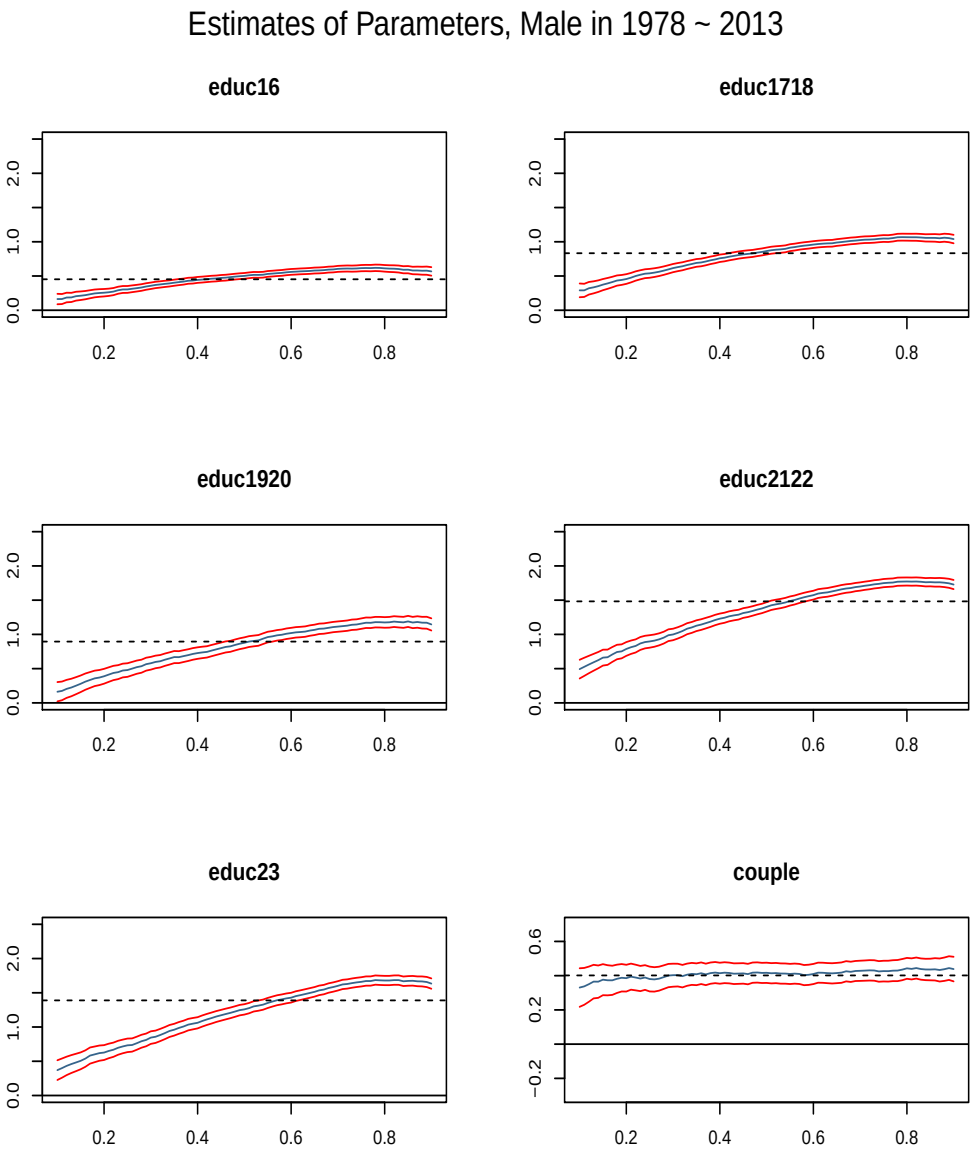

FIGURE 2. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 1 for men

working men and decide not to work (Neal, 2004). Figure 6 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

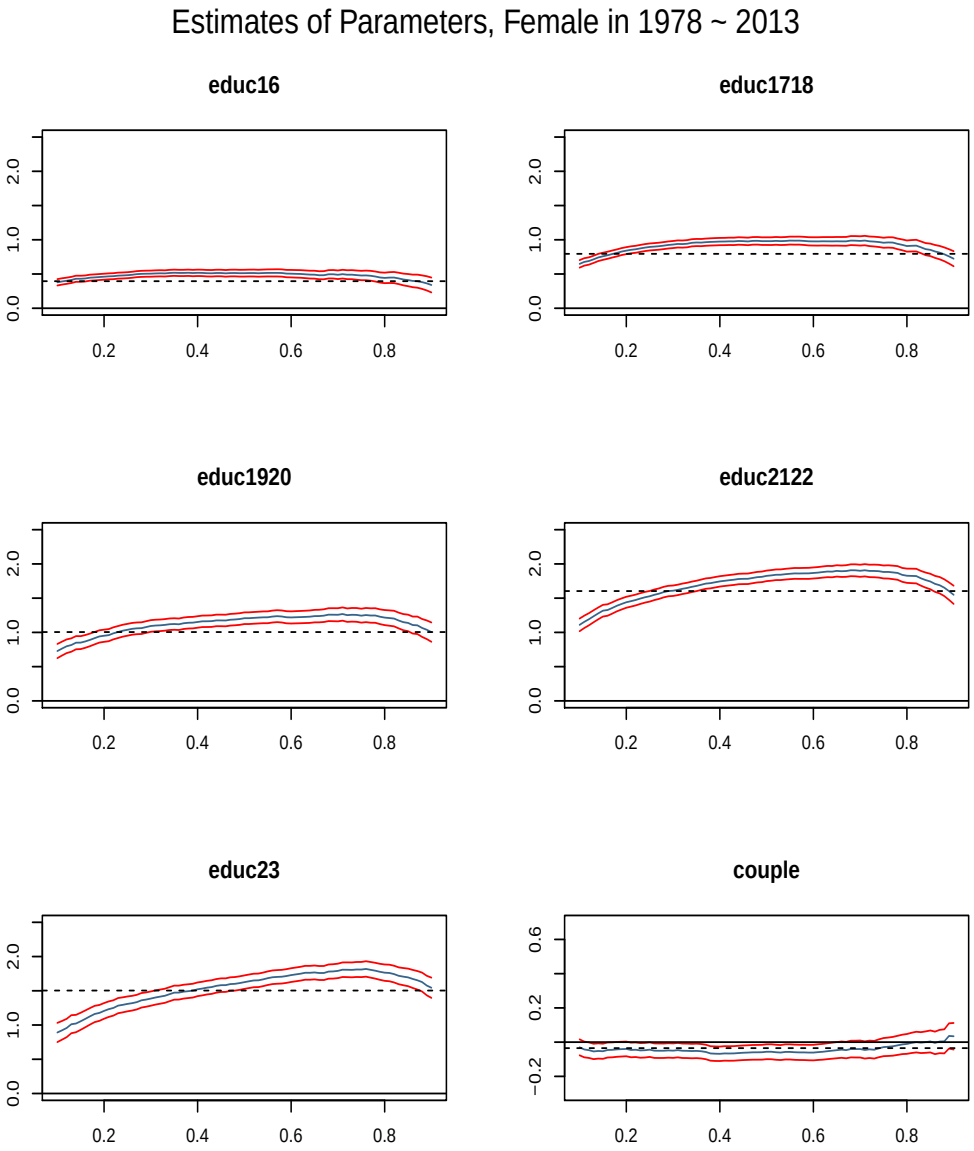

FIGURE 3. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 1 for women

over time, but also displays a positive trend, specially at the bottom of the distribution. Figure 7 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.[17]

[17]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 in Appendix E show that the results on the trends are statistically significant.

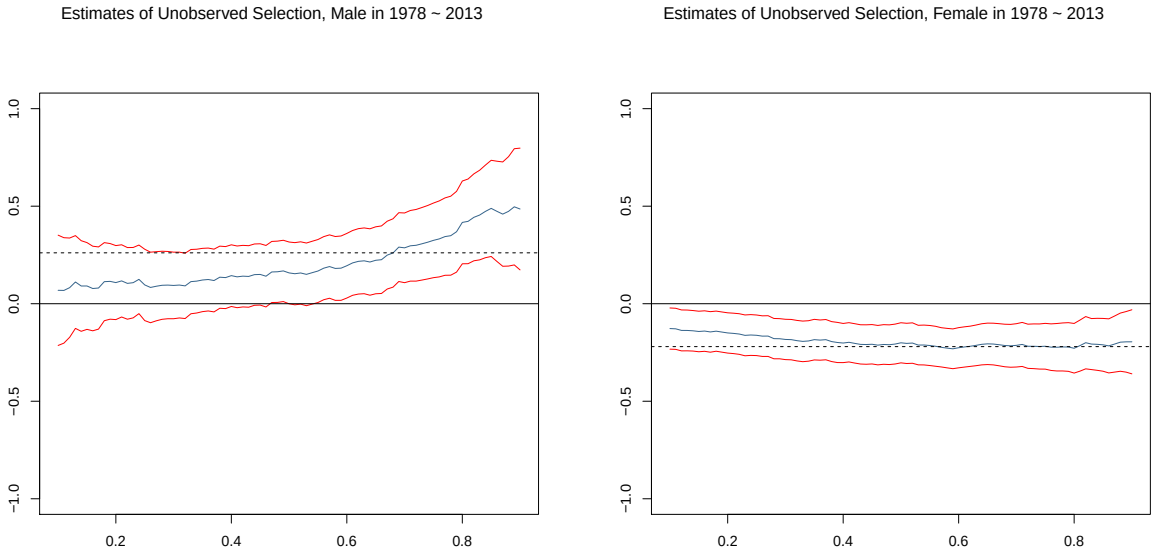

FIGURE 4. Estimates and 95% confidence bands for the selection sorting function: specification 1

5.4. **Distributions of Offered and Observed Wages.** Figure 8 shows point estimates of the quantiles of offered and observed wages for men and women based on specification 1. Estimates for the other specifications and confidence bands for all the specifications are given in Appendix E. 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 (3.8) and (3.9).[18] We find opposite signs in the sample selection bias for men and women. The quantiles of the observed wages are below the quantiles of latent wages for men, but the opposite holds for women. This pattern is consistent with the sign of the estimates of the selection sorting function, where we found positive sorting for men and negative sorting for women. In results reported in Appendix E, we find that the majority of the difference between the distribution of offered wages between women and men is explained by differences in the wage structure, $\beta(y)$, whereas differences in composition, F_Z, have very little explanatory power. This result can be interpreted as evidence of gender discrimination in the labor market.

5.5. **Wage Decompositions.** We 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:

[18]The model-based estimator of the observed distribution in (3.9) produces almost identical estimates to the empirical distribution of the observed wages.

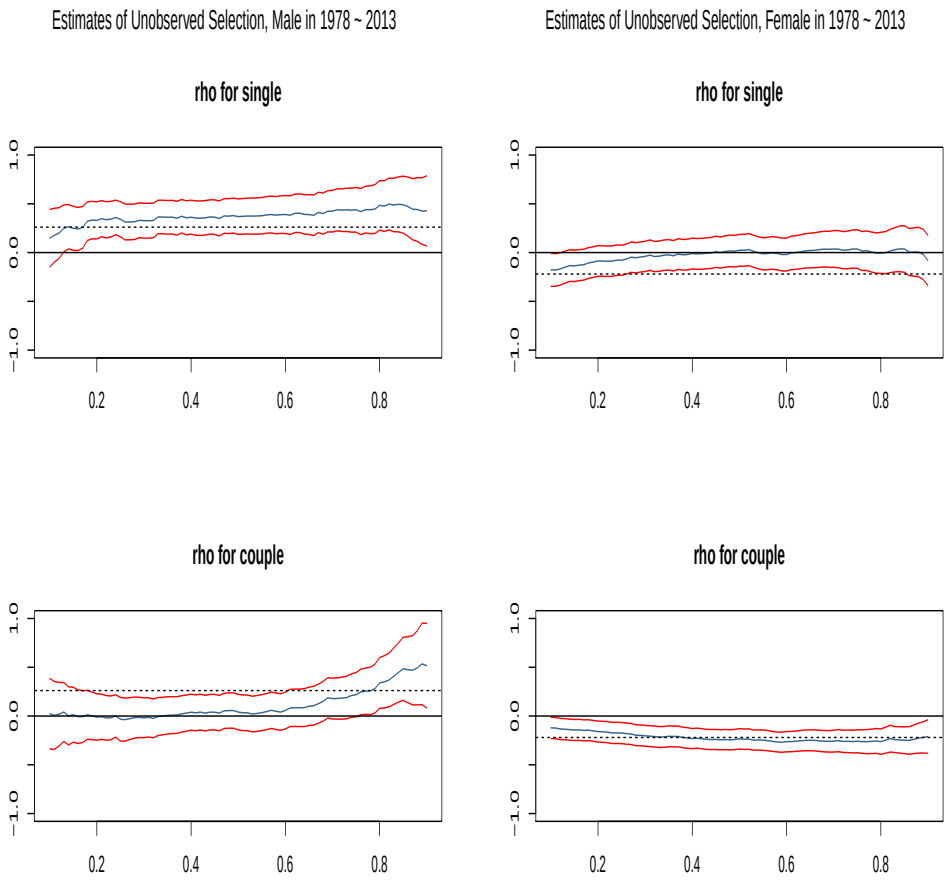

FIGURE 5. Estimates and 95% confidence bands for the selection sorting function: specification 2

(1) Selection (employment) sorting: $\delta(y)$.
(2) Selection (employment) structure: π.
(3) Outcome (wage) structure: $\beta(y)$.
(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

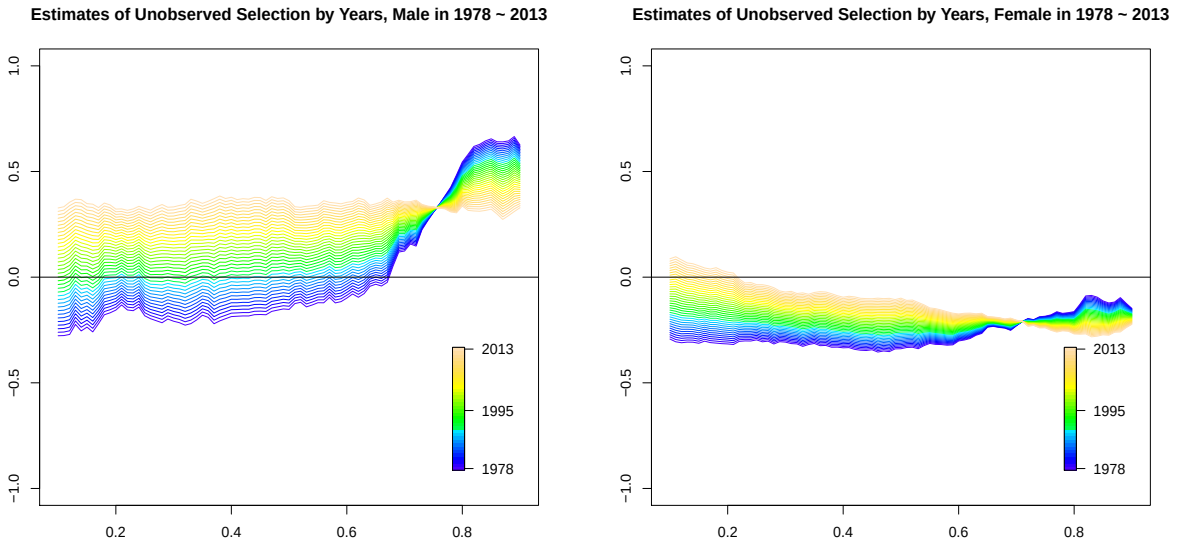

FIGURE 6. Estimates and 95% confidence bands for the selection sorting function: specification 3

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

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

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.

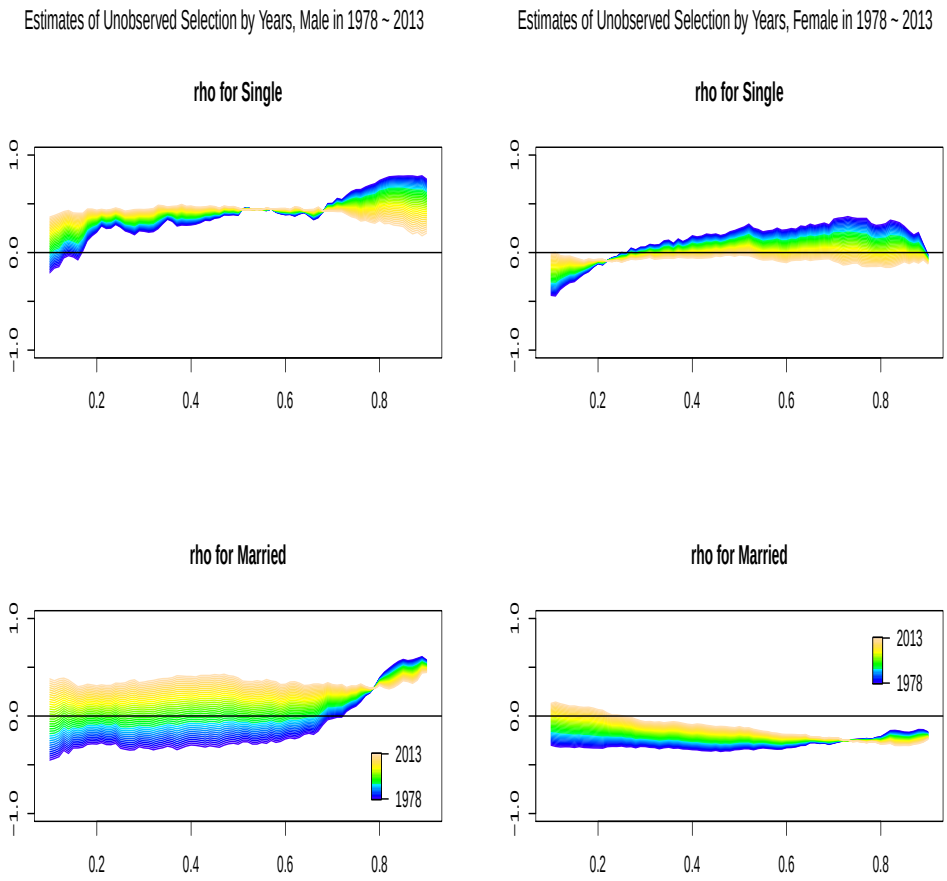

FIGURE 7. Estimates and 95% confidence bands for the selection sorting function: specification 4

Figure 9 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 1. The bands for the contributions are joint for all the components and rely on the delta method; see Remark 4.2. Estimates of the components of the decomposition and the analysis based on specifications 2–4 are given in Appendix E. 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 that might be associated to gender discrimination. However, differences in sorting and employment structure also account for an important percentage of the gap, specially 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

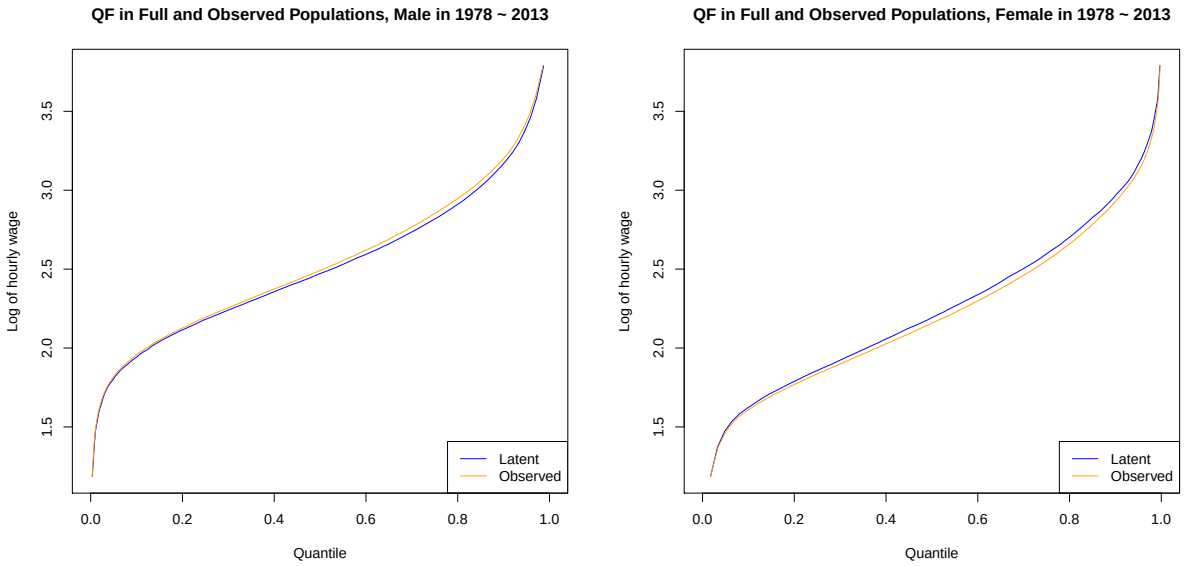

FIGURE 8. Estimates of the quantiles of observed and offered (latent) wages: specification 1

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. 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 3.1 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 E.15 in Appendix E. Differences in the composition of the characteristics contribute very little to explain the gender gap. Finally, the estimates from the HSM in dash lines pick up the average contributions of the components, but miss all the heterogeneity across the distribution.

Figures 10 and 11 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 1 for women and men, respectively. Estimates of the components of the decompositions are given in Appendix E. 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 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

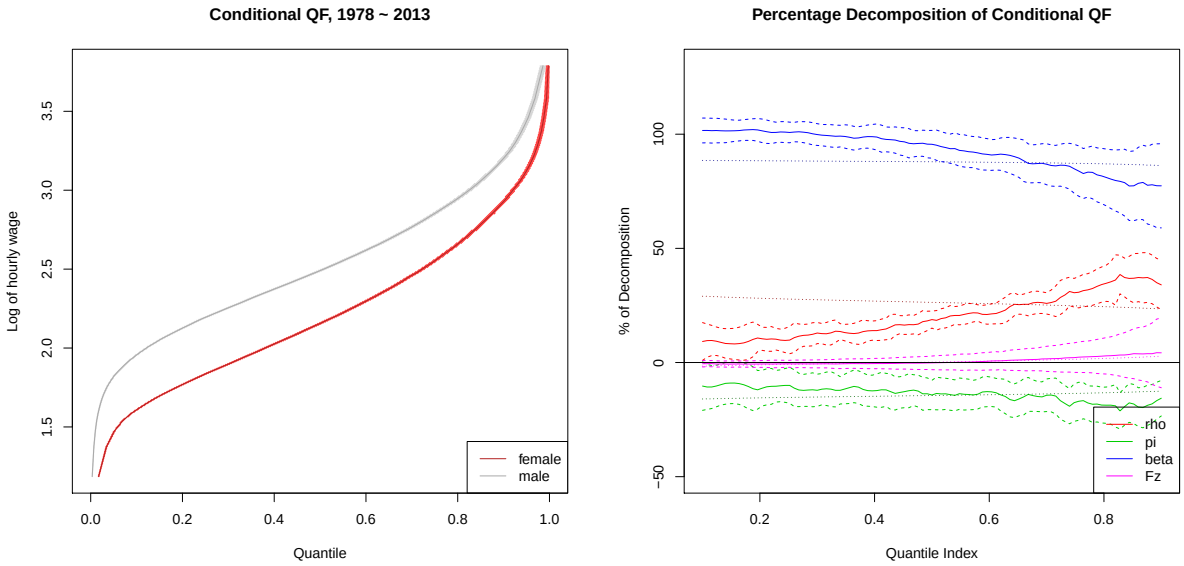

FIGURE 9. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women in specification 1

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.

5.6. **Discussion.** The main findings can be summarized as: (1) heterogeneous positive effect of education and homogeneous effect of being married on offered wages for both genders; (2) 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; (3) heterogeneity in selection sorting decreases gradually over time; (4) differences in returns to characteristics in the wage equation, which might be associated to gender discrimination in the labor market, account for most of the gender wage gap; (5) 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 (6) 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 assumed positive sorting for men and women in some of their estimates

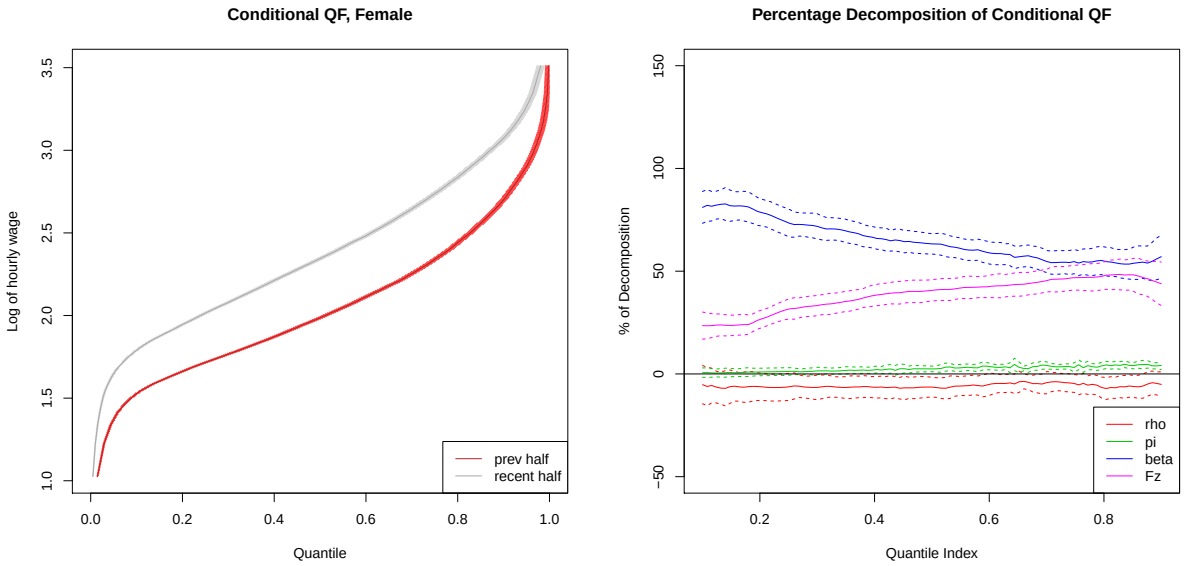

FIGURE 10. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between first and second half of the sample period for women in specification 1

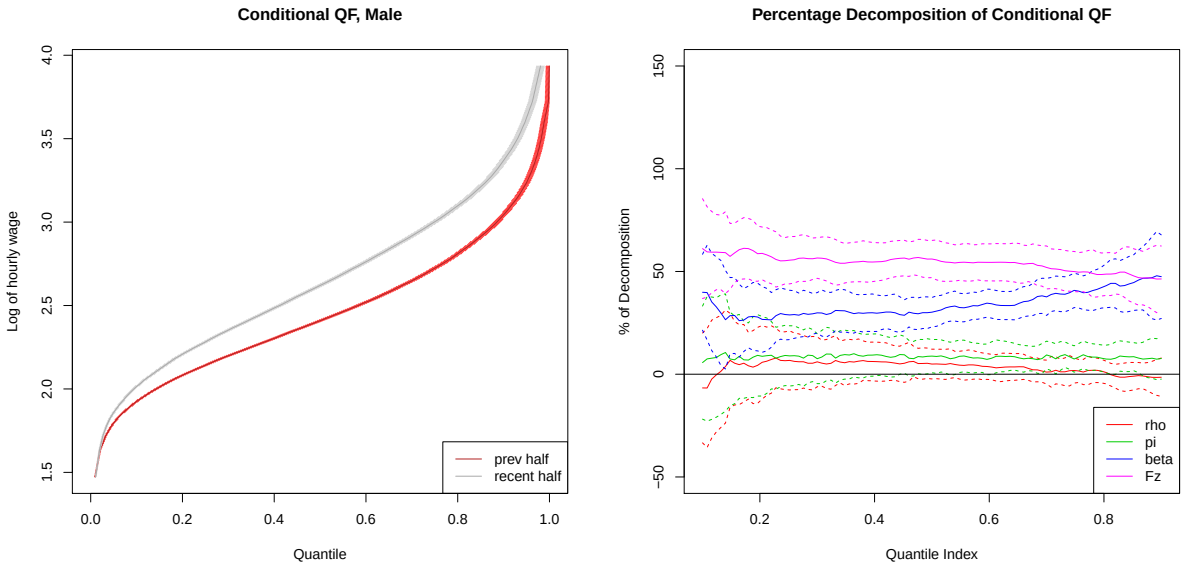

FIGURE 11. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between first and second half of the sample period for men in specification 1

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. They also found evidence against the validity of out-of-work benefit income as a valid excluded covariate for men. Arellano and Bonhomme (2017a) using the same data from the FES, also found positive sorting for men, stronger for single than for married men, using 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 (Forthcoming) applied the methodology of Arellano and Bonhomme (2017a) 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.

6. 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 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 12, 13 and 14 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.[19] 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 $\mathcal{Y}$ 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

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

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.

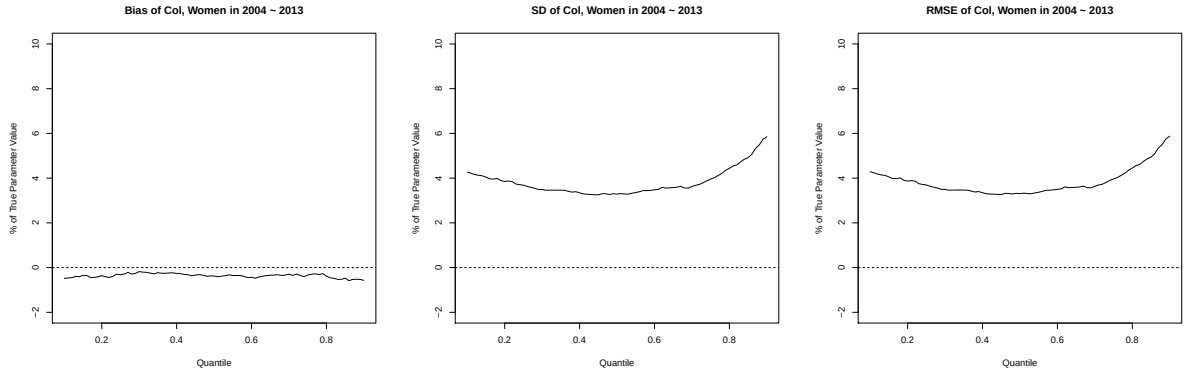

FIGURE 12. Bias, SD and RMSE for the coefficient of the college indicator in the outcome equation

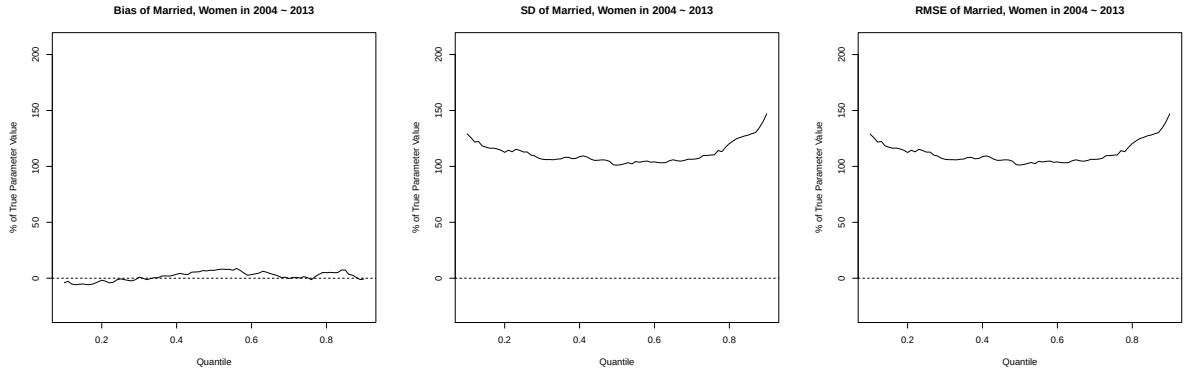

FIGURE 13. Bias, SD and RMSE for the coefficient of the marital status indicator in the outcome equation

Table 2 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 $\bar{\mathcal{Y}}$ 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

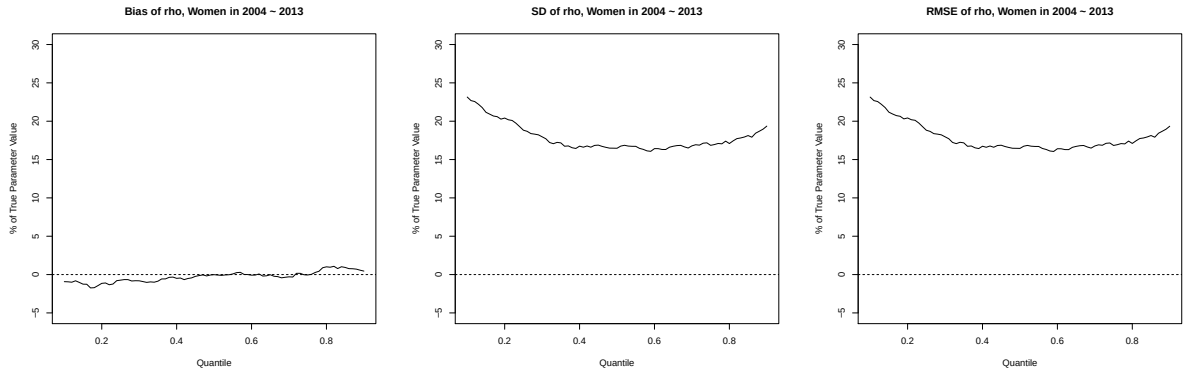

FIGURE 14. Bias, SD and RMSE for coefficient $\rho(y)$ in the selection sorting equation

functions. The standard errors based on the asymptotic distribution provide a fair approximation to the sampling variability of the estimator.

TABLE 2. 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.

7. 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 semiparametric in nature, as it has function-valued parameters, and is able to considerably generalize the classical selection model of Heckman (1974). Furthermore, the model allows for richer covariate effects than the previous semiparametric generalizations which allowed the location effects for covariates. We propose to estimate the model by a process of 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 extensive wage decomposition analysis for the U.K. using new data, generating both new findings and demonstrating the power of the method.

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

Arellano, M., and S. Bonhomme (2017b): "Sample Selection in Quantile Regression: A Survey," in *Handbook of Quantile Regression*, ed. by R. Koenker, V. Chernozhukov, X. He, and L. Peng, chap. 13. CRC Press.

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

Bjerve, S., and K. Doksum (1993): "Correlation Curves: Measures of Association as Functions of Covariate Values," *Ann. Statist.*, 21(2), 890–902.

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.

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.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70(1), 33–58.

Doksum, K., S. Blyth, E. Bradlow, X.-L. Meng, and H. Zhao (1994): "Correlation Curves as Local Measures of Variance Explained by Regression," *Journal of the American Statistical Association*, 89(426), 571–582.

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.

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

Hjort, N. L., and M. C. Jones (1996): "Locally parametric nonparametric density estimation," *Ann. Statist.*, 24(4), 1619–1647.

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," .

Kolev, N., U. d. Anjos, and B. V. d. M. Mendes (2006): "Copulas: A Review and Recent Developments," *Stochastic Models*, 22(4), 617–660.

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.

Maasoumi, E., and L. Wang (Forthcoming): "The Gender Gap between Earnings Distributions," *Journal of Political Economy*.

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. (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.

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. (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.

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.

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*.

Tjøstheim, D., and K. O. Hufthammer (2013): "Local Gaussian correlation: A new measure of dependence," *Journal of Econometrics*, 172(1), 33 – 48.

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.

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

Appendix A. Detailed Comparison with AB17

We need to introduce some notation to state the conditions of AB17. Let $p(z) = \mathrm{P}(D = 1 \mid Z = z)$ and $V = F_{D^*|Z}(D^* \mid Z)$ such that $V \mid Z \sim U(0,1)$.[20] AB17 assumed that (i) (Y^*, V) are 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. The condition (iii) requires Z to have continuous variation and is therefore more restrictive than our assumption that Z can be binary. We now show that our selection exclusion neither implies nor is implied by conditions (i) and (ii). Selection exclusion implies that for any $u \in [0,1]$ that satisfy $F_{Y^*}(y_u) = u$ 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 $p(Z)$.[21] 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

[20] 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 $D^* = 0$.

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

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

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

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.[22] 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.[23]

Appendix B. 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

$$\sup_{h \in \mathrm{BL}_1(\mathbb{Z})} \left| \mathrm{E}^b h\left(Z_n^b\right) - \mathrm{E} h(Z) \right| \to_{\mathrm{P}^*} 0, \tag{B.13}$$

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.

[22] If Y^* is absolutely continuous with strictly increasing distribution, this restriction can be expressed as $\Phi^{-1}(V) \mid Y^* = y \sim N(\tilde{\rho}(y)\Phi^{-1}(F_{Y^*}(y)), 1 - \tilde{\rho}(y)^2)$.

[23] Alternatively, condition (ii) is equivalent to $v \mapsto \tilde{\rho}(\cdot, v)$ being real analytic, which is weaker than $\tilde{\rho}(y, v) = \tilde{\rho}(y)$.

Appendix C. Proofs of Section 4

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

$$J(y) = \mathrm{E}\left[\frac{\partial \varphi_{y,\xi}(W)}{\partial \xi'}\right] = \begin{bmatrix} H_1 & 0 \\ J_{21y} & H_{2y} \end{bmatrix} \tag{C.14}$$

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 D. Note that

$$J^{-1}(y) = \begin{bmatrix} H_1^{-1} & 0 \\ -H_1^{-1} J_{21y} H_{2y}^{-1} & H_{2y}^{-1} \end{bmatrix} \tag{C.15}$$

by the inverse of the partitioned inverse formula, and

$$\mathrm{E}[\varphi_{y,\xi}(W)\varphi_{\tilde{y},\xi}(W)'] = \begin{bmatrix} \mathrm{E}[S_{1,\xi}(W)S_{1,\xi}(W)'] & 0 \\ 0 & \mathrm{E}[S_{2y,\xi}(W)S_{2\tilde{y},\xi}(W)'] \end{bmatrix} \tag{C.16}$$

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

C.1. **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.[24]

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

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

Condition Z of Chernozhukov, Fernández-Val, and Melly (2013) holds and $y \mapsto \xi_y$ is continuously differentiable.

Proof of Lemma C.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 $\mathcal{Y}$ be a compact set of some metric space, and Ξ be an arbitrary subset of $\mathbb{R}^{d_\xi}$. Assume (i) for each $y \in \mathcal{Y}$, $\Psi(\cdot, y) : \Xi \mapsto \mathbb{R}^{d_\xi}$ possesses a unique zero at ξ_y, and, for some $\delta > 0$, $\mathcal{N} := \cup_{y\in\mathcal{Y}} B_\delta(\xi_y)$ 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 $y \in \mathcal{Y}$ with respect to the Hausdorff distance, (iii) there exists $\dot{\Psi}_{\xi_y,y}$ such that $\lim_{t\searrow 0} \sup_{y\in\mathcal{Y},\|h\|=1} |t^{-1}[\Psi(\xi_y + th, y) - \Psi(\xi_y, y)] - \dot{\Psi}_{\xi_y,y}h| = 0$, where $\inf_{y\in\mathcal{Y}} \inf_{\|h\|=1} \|\dot{\Psi}_{\xi_y,y}h\| > 0$, 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 $x_t \in \Psi(\Xi, y)$, $d_H(\Psi^{-1}(x_t, y), \Psi^{-1}(0, y)) \to 0$, where d_H is the Hausdorff distance, uniformly in $y \in \mathcal{Y}$. Suppose by contradiction that this is not true, then there is (x_t, y_t) with $x_t \to 0$ and $y_t \in \mathcal{Y}$ such that $d_H(\Psi^{-1}(x_t, y_t), \Psi^{-1}(0, y_t)) \not\to 0$. By compactness of $\mathcal{Y}$, we can select a further subsequence (x_k, y_k) such that $y_k \to y$, where $y \in \mathcal{Y}$. We have that $\Psi^{-1}(0, y) = \xi_y$ is continuous in $y \in \mathcal{Y}$, so we must have $d_H(\Psi^{-1}(x_k, y_k), \Psi^{-1}(0, y)) \not\to 0$. 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. □

C.2. **Proof of Theorem 4.1.** We only consider the case where $\mathcal{Y}$ is a compact interval of $\mathbb{R}$. The case where $\mathcal{Y}$ 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 C.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 $y \in \mathcal{Y}$, 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,

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

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 C.1 for $\dot{\Psi}(\xi_y, y) = J(y)$, the Hessian matrix defined in (C.14), which also implies that $y \mapsto \xi_y$ is continuously differentiable in the interval $\mathcal{Y}$. 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

$$\sqrt{n}(\widehat{\xi}_y - \xi_y) \rightsquigarrow Z_{\xi_y} := -J^{-1}(y) Z_\Psi(y, \xi_y) \text{ in } \ell^\infty(\mathcal{Y})^{d_\xi}, \tag{C.17}$$

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

Step 1 (Donskerness)**.** We verify that $\mathcal{G} = \{\varphi_{y,\xi}(W) : (y, \xi) \in \mathcal{Y} \times \mathbb{R}^{d_\xi}\}$ 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 D, $\varphi_{y,\xi}(W)$ 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 $\mathcal{G}$ is P-Donsker by Example 19.9 in van der Vaart (1998).

Step 2 (Verification of the Conditions of Lemma C.1)**.** Conditions (a) and (b) are immediate by Assumption 2. 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

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

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

$$J(\tilde{\xi}, \tilde{y}) = \left[\begin{array}{cc} J_{11}(\tilde{\xi}, \tilde{y}) & J_{12}(\tilde{\xi}, \tilde{y}) \\ J_{21}(\tilde{\xi}, \tilde{y}) & J_{22}(\tilde{\xi}, \tilde{y}) \end{array} \right],$$

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

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

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

$$\begin{aligned} J_{22}(\tilde{\xi}, \tilde{y}) = -\mathrm{E}\left\{ \Phi_\pi(Z) \left[\begin{array}{cc} \Phi^\mu_{2,\tilde{\xi}}(Z) G_{2,\tilde{\xi}}(Z) & \Phi^\rho_{2,\tilde{\xi}}(Z) G_{2,\tilde{\xi}}(Z) \\ \Phi^\mu_{2,\tilde{\xi}}(Z)\rho'(X'\tilde{\delta}) G_{3,\tilde{\xi}}(Z) & \Phi^\rho_{2,\tilde{\xi}}(Z)\rho'(X'\tilde{\delta}) G_{3,\tilde{\xi}}(Z) \end{array} \right] \otimes XX' \right\} \\ + \mathrm{E}\left\{ (\Phi_\pi(Z)\Phi_{2,\tilde{\xi}}(Z) - \Phi_{\tilde{\pi}}(Z)\Phi_{2,\xi_{\tilde{y}}}(Z)) \left[\begin{array}{cc} G^\mu_{2,\tilde{\xi}}(Z) & G^\rho_{2,\tilde{\xi}}(Z) \\ \rho'(X'\tilde{\delta}) G^\mu_{3,\tilde{\xi}}(Z) & \rho'(X'\tilde{\delta})^2 G^\rho_{3,\tilde{\xi}}(Z) + \rho''(X'\tilde{\delta}) G_{3,\tilde{\xi}}(Z) \end{array} \right] \otimes XX' \right\}, \end{aligned}$$

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 D.

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 $y \in \mathcal{Y}$. The computation above as well as the verification of the continuity follow from using the expressions of $\varphi_{y,\xi}$ in Appendix D, 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 2, 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 $y \in \mathcal{Y}$.

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 (C.15) and (C.16). □

C.3. **Proof of Theorem 4.2.** 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 4.1. 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 (C.17).

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 $y \in \mathcal{Y}$ by Theorem 4.1. 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 $y \in \mathcal{Y}$. 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 $y \in \mathcal{Y}$, where the latter holds by the Law of Iterated Expectations and $r_y \rightsquigarrow 0$.

□

Appendix D. Expressions of the Score and Expected Hessian

D.1. **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 $G_{2,\xi_y}(Z) := G_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$ and $G_{3,\xi_y}(Z) := G_3\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$ 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

$$\Phi_2^\mu(\mu,\nu;\rho) = \frac{\partial \Phi_2(\mu,\nu;\rho)}{\partial \mu} = \Phi\left(\frac{\nu - \rho\mu}{\sqrt{1-\rho^2}}\right)\phi(\mu), \tag{D.18}$$

and

$$\Phi_2^\rho(\mu,\nu;\rho) = \frac{\partial \Phi_2(\mu,\nu;\rho)}{\partial \rho} = \phi_2(\mu,\nu;\rho). \tag{D.19}$$

To show (D.18) and (D.19), 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, (D.18) follows from taking the partial derivative with respect to μ using the Leibniz integral rule. Taking the partial derivative with respect to ρ yields

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

D.2. **Expected Hessian.** Straighforward calculations yield

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

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

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 (D.18), 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 $\Phi^\mu_{2,\xi_y}(Z) := \Phi^\mu_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$ and $\Phi^\rho_{2,\xi_y}(Z) := \Phi^\rho_2\left(-X'\beta(y), Z'\pi; -\rho(X'\delta(y))\right)$.

Appendix E. Additional Empirical Results

E.1. **Model Parameters.**

E.2. **Distribution of Offered and Observed Wages.**

E.3. **Participation and Wage Decompositions.**

TABLE E.1. Estimates of Coefficients of the Selection Equation

Variable	Male	Female	Variable	Male	Female
educ16	0.25	0.06	agep4	-0.07	-0.05
	(0.01)	(0.01)		(0.02)	(0.02)
educ1718	0.46	0.20	numch1	-0.16	-0.90
	(0.02)	(0.01)		(0.02)	(0.02)
educ1920	0.42	0.16	numch2	-0.18	-0.77
	(0.03)	(0.02)		(0.02)	(0.02)
educ2122	0.74	0.28	numch34	-0.18	-0.63
	(0.02)	(0.02)		(0.02)	(0.01)
educ23	0.51	0.15	numch510	-0.18	-0.33
	(0.02)	(0.02)		(0.01)	(0.01)
couple	-4.02	-8.14	numch1116	-0.16	-0.15
	(0.09)	(0.08)		(0.01)	(0.01)
agep	-0.64	-0.92	numch1718	-0.02	-0.11
	(0.03)	(0.02)		(0.03)	(0.02)
agep2	-0.83	-0.68	tubeninc0	-0.35	-0.42
	(0.03)	(0.02)		(0.01)	(0.01)
agep3	-0.07	-0.08	m_inc0	0.87	1.40
	(0.02)	(0.02)		(0.02)	(0.02)
constant	2.50	2.75			
	(0.08)	(0.07)			

Notes: standard errors in parentheses.

TABLE E.2. Participation decomposition between men and women

Participation (%)		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

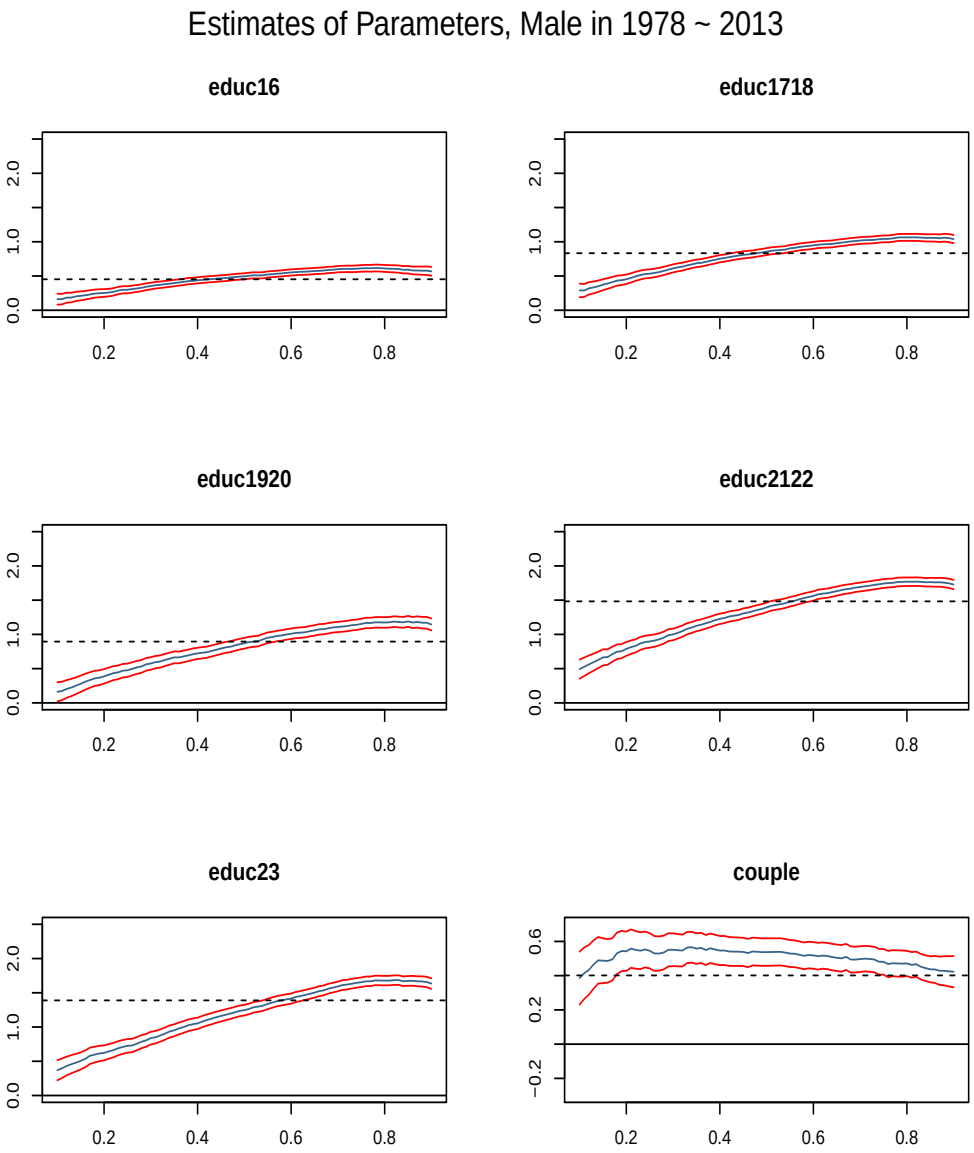

FIGURE E.1. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 2 for men

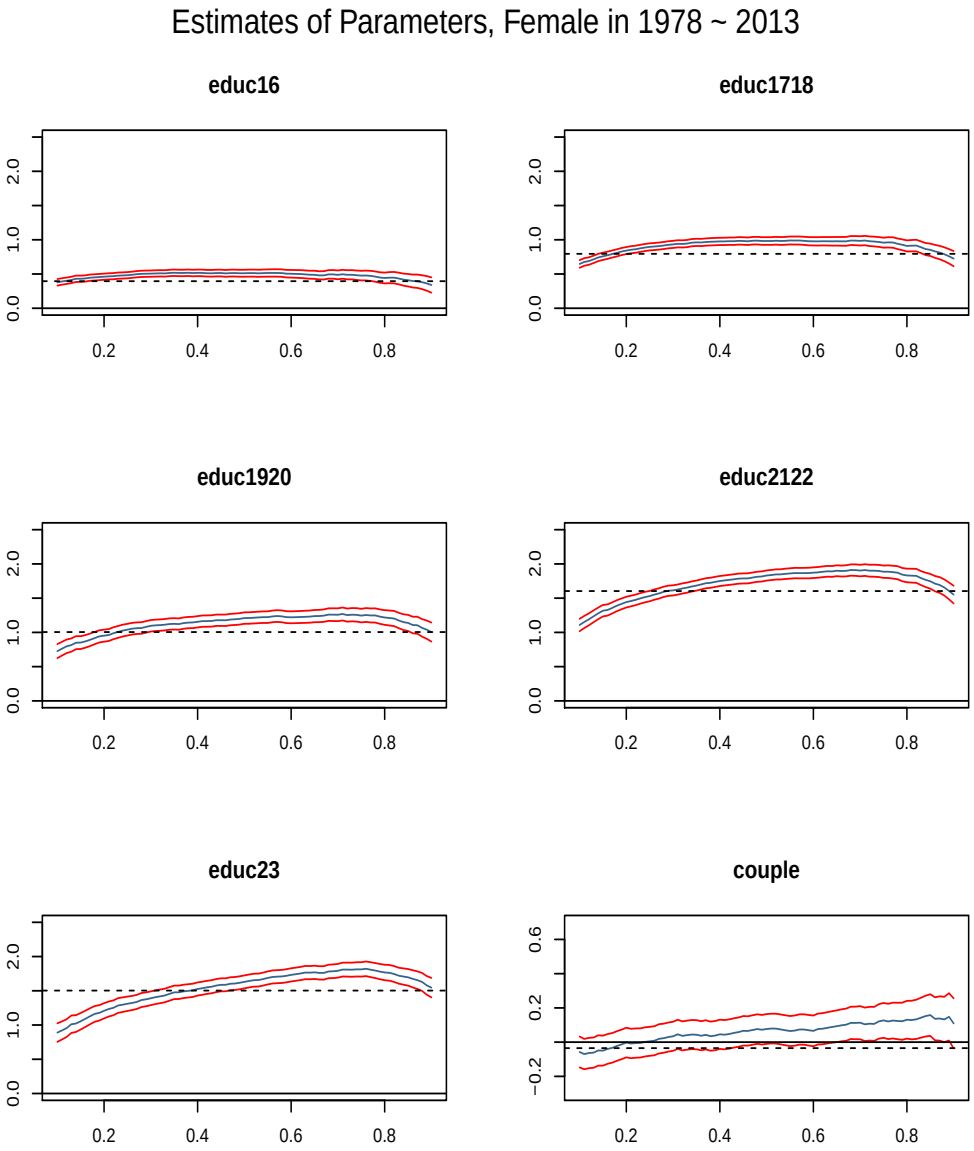

FIGURE E.2. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 2 for women

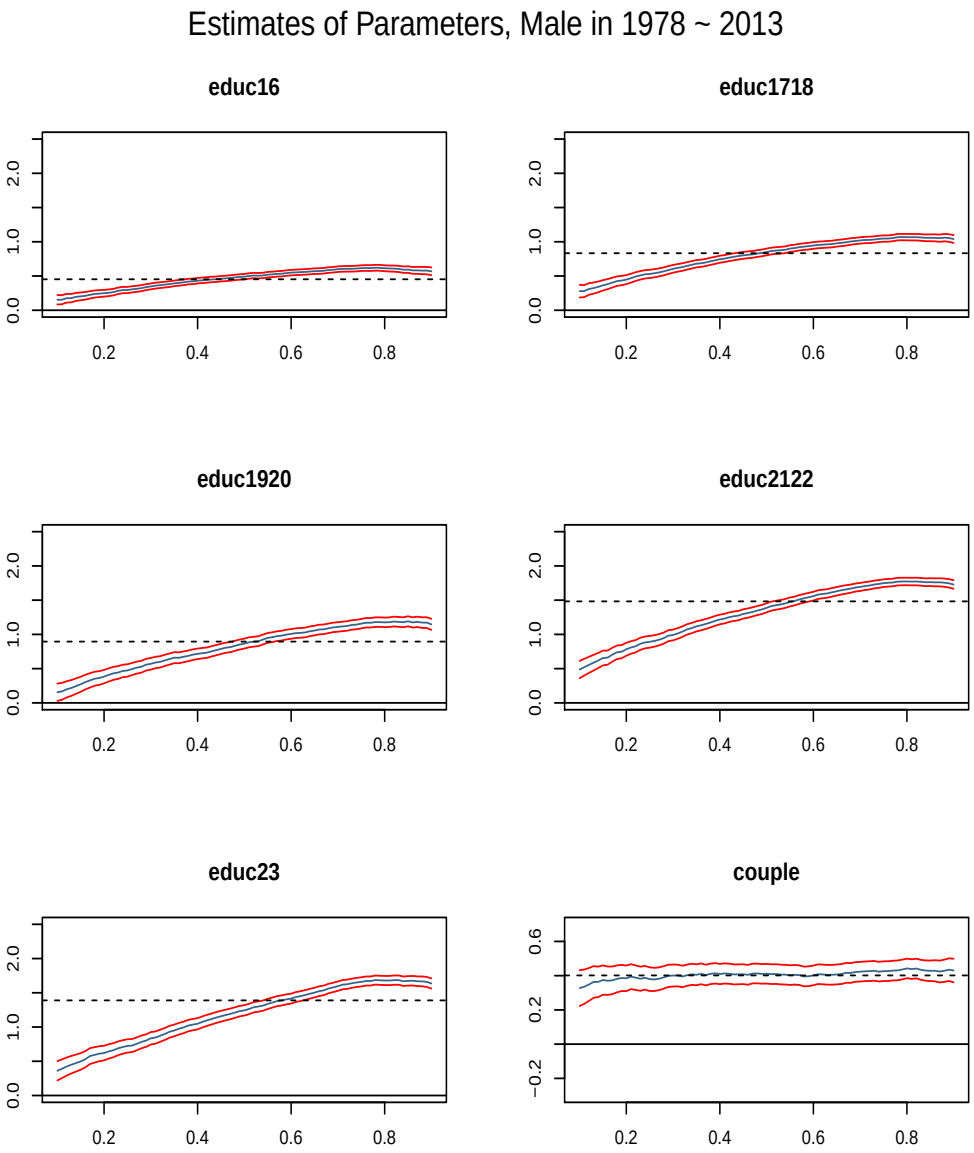

FIGURE E.3. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 3 for men

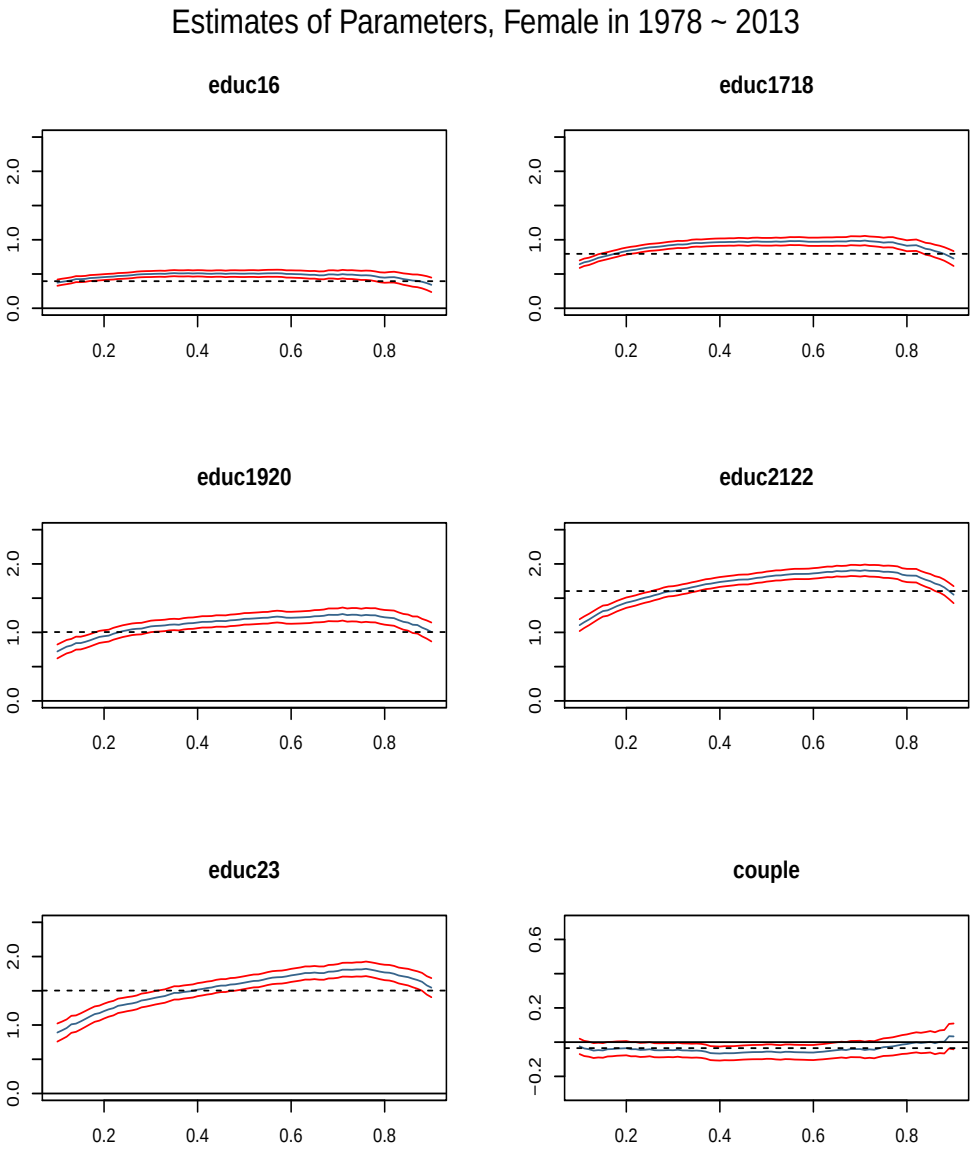

FIGURE E.4. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 3 for women

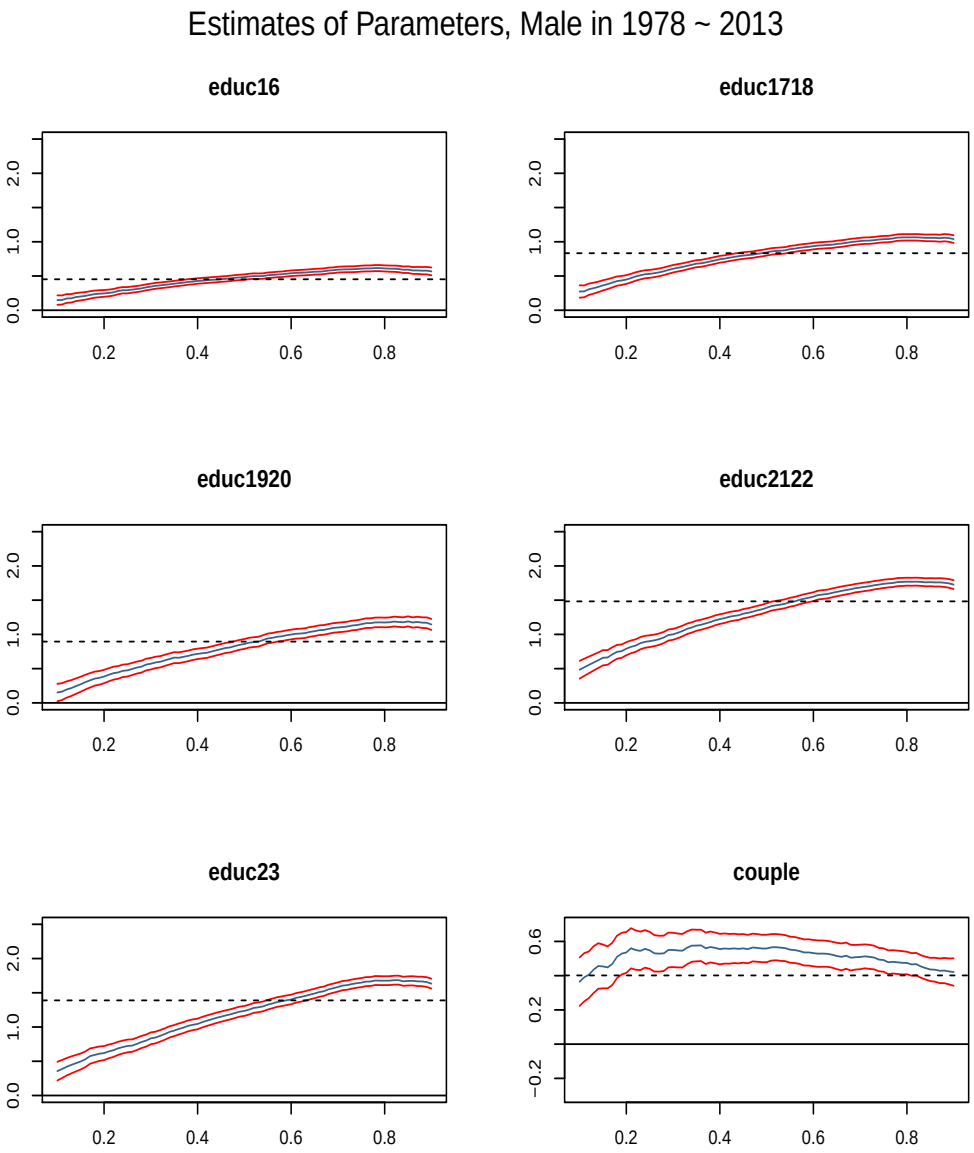

FIGURE E.5. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 4 for men

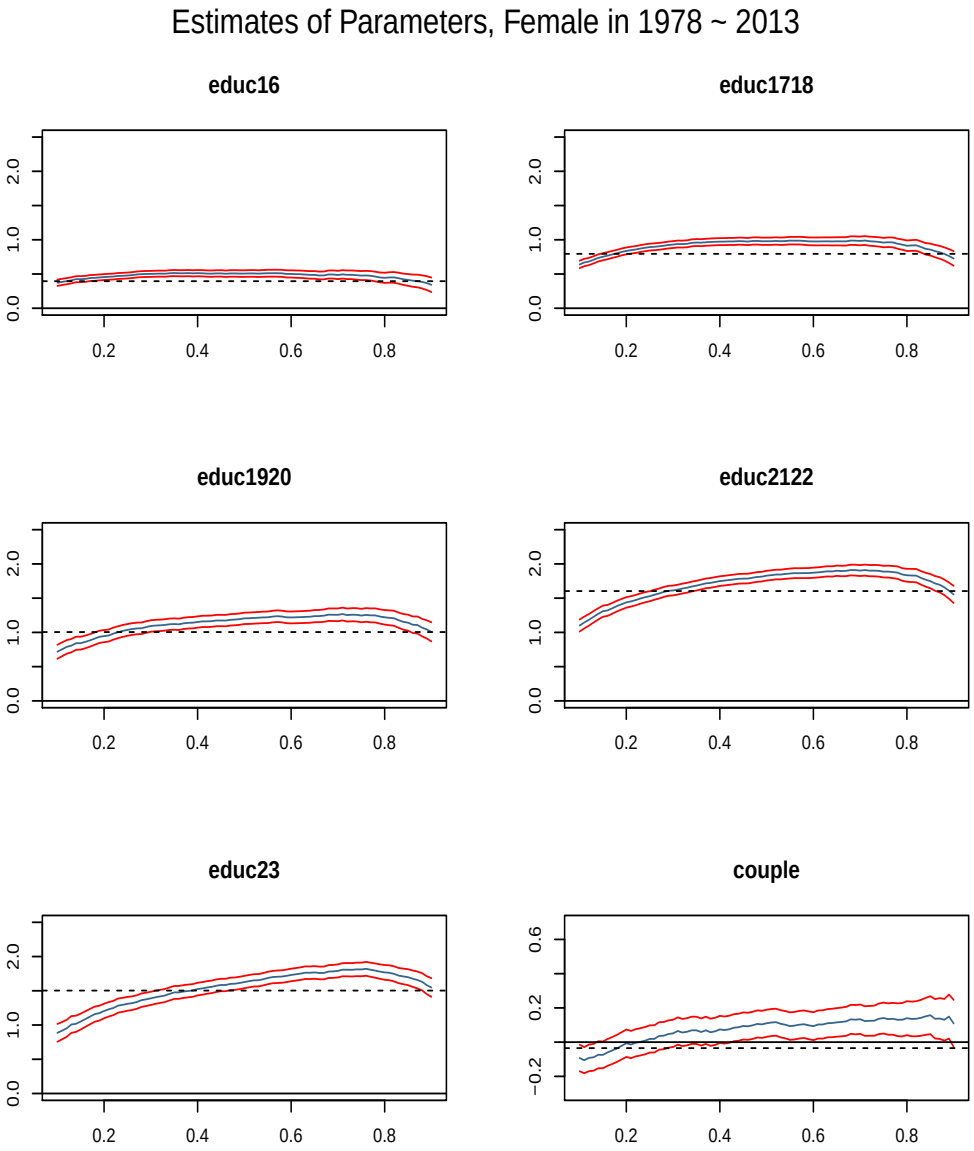

FIGURE E.6. Estimates and 95% confidence bands for coefficients of education and marital status in the outcome equation: specification 4 for women

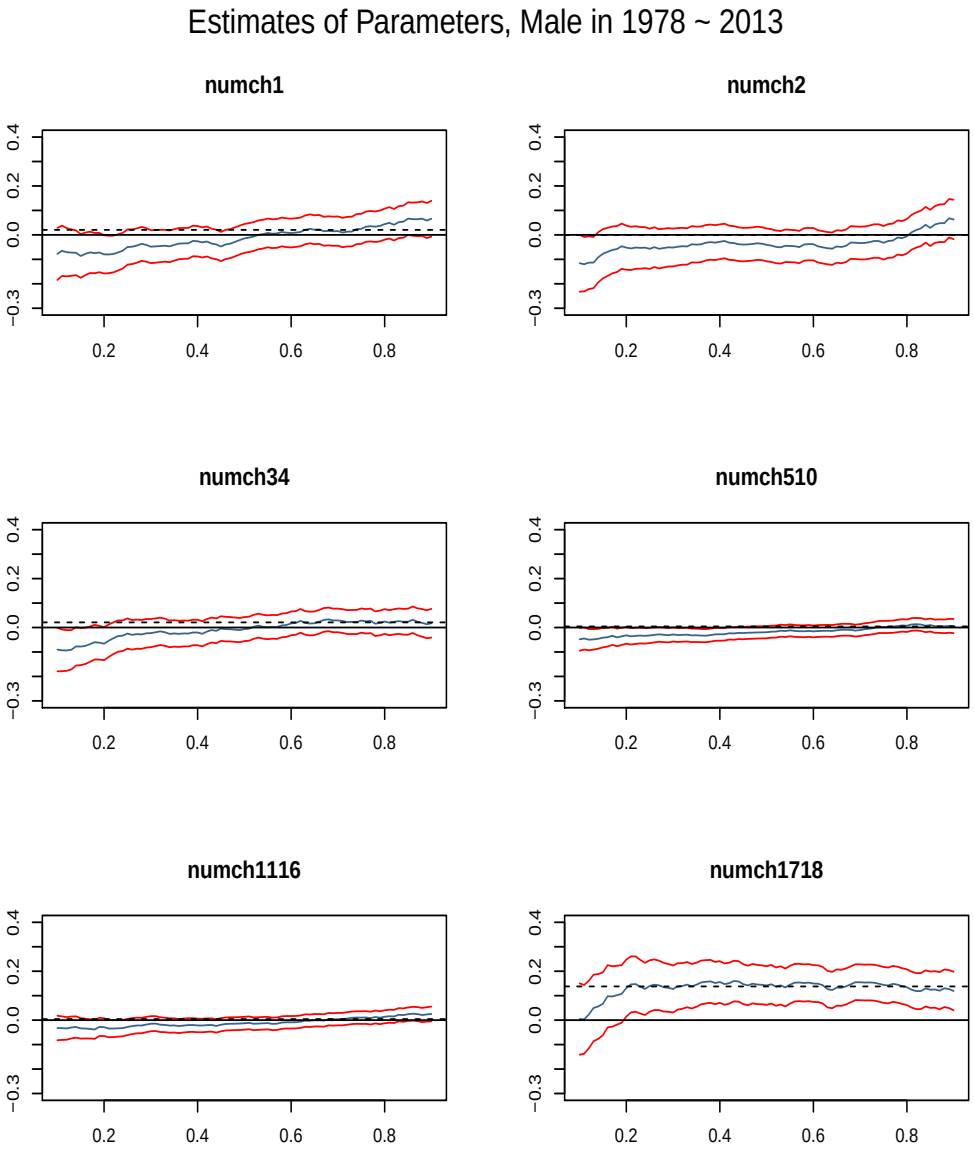

FIGURE E.7. Estimates and 95% confidence bands for coefficients of fertility in the outcome equation: specification 1 for men

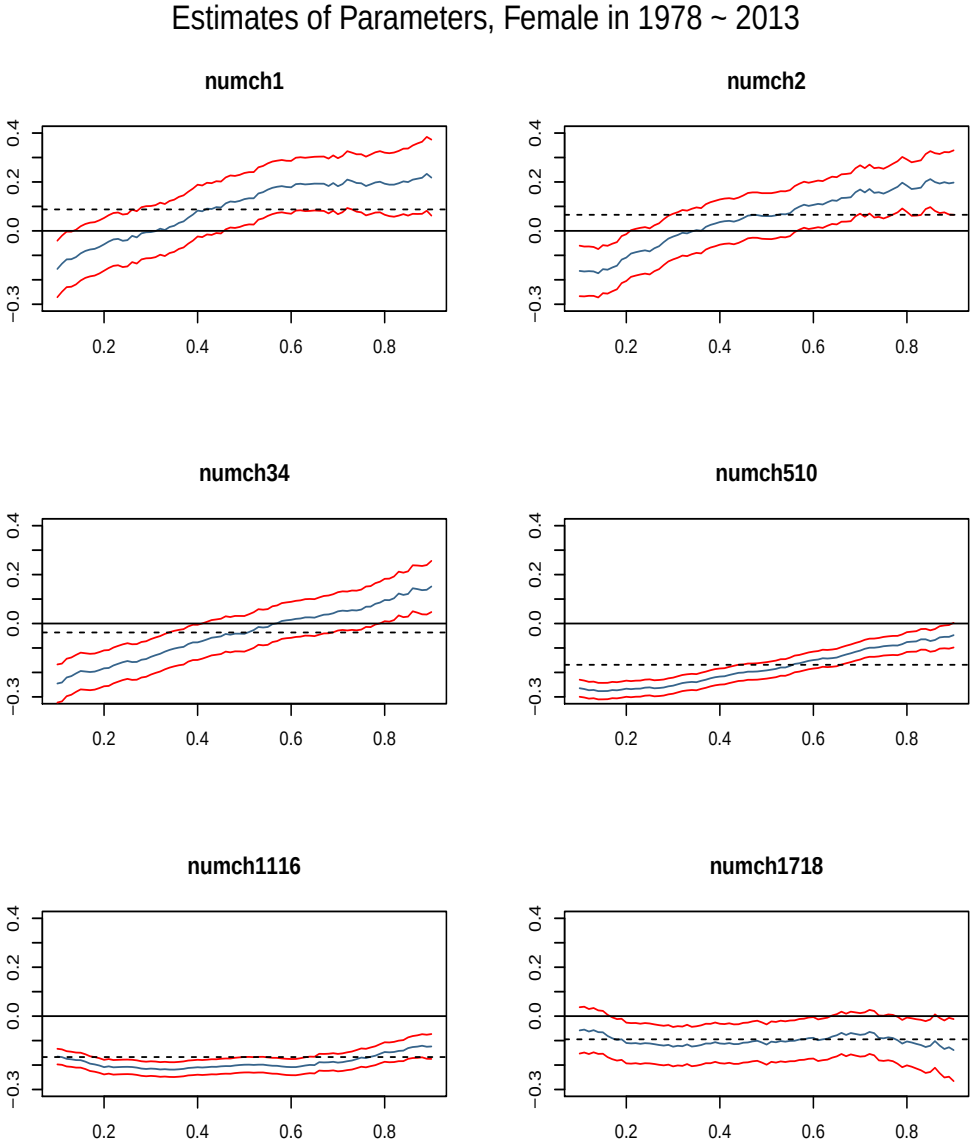

FIGURE E.8. Estimates and 95% confidence bands for coefficients of fertility in the outcome equation: specification 1 for women

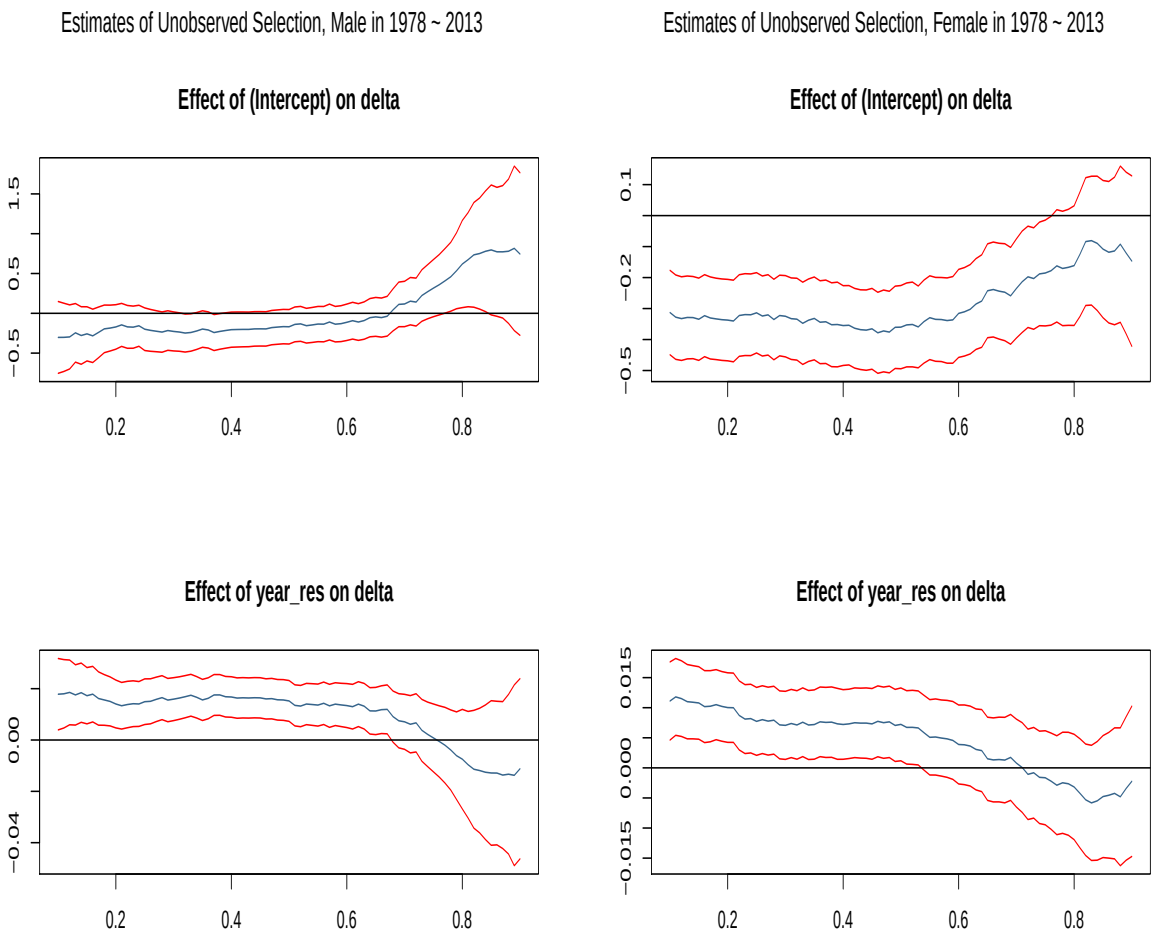

FIGURE E.9. Estimates and 95% confidence bands for coefficients of the selection sorting function: specification 3

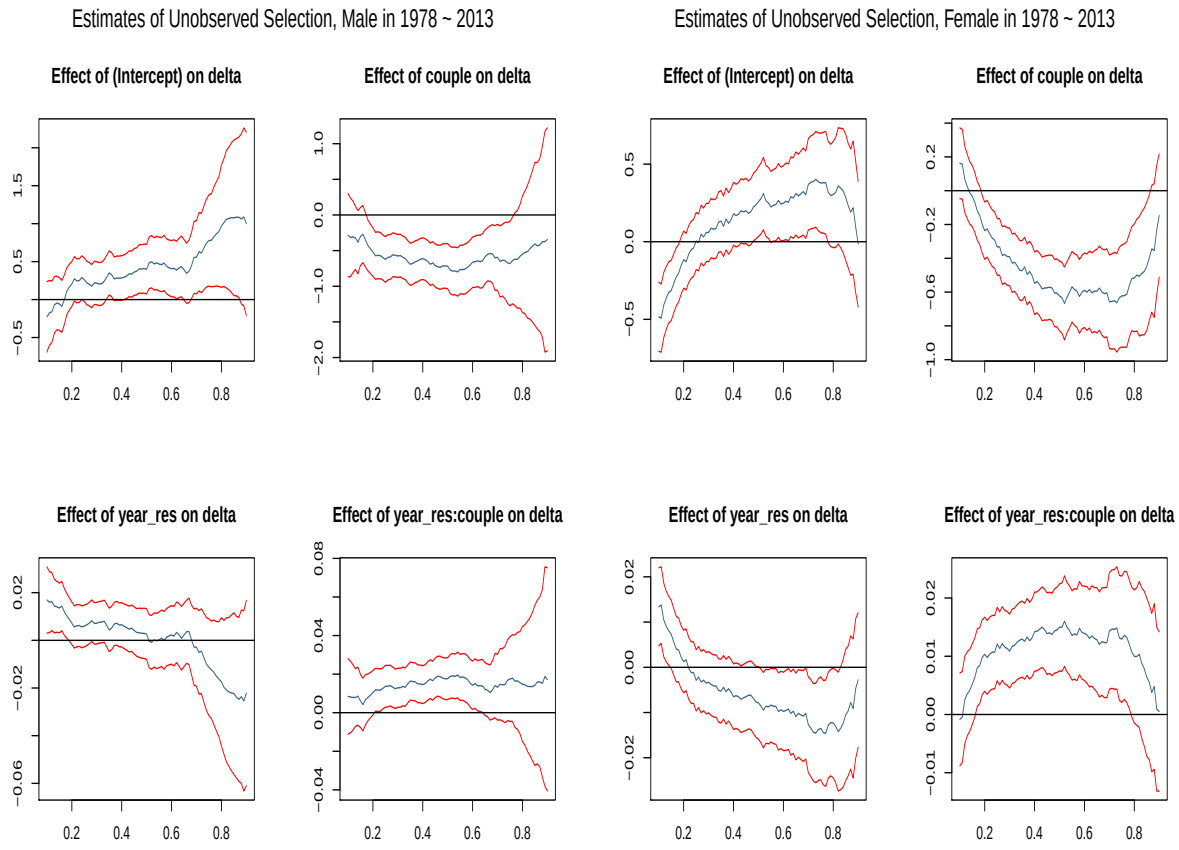

FIGURE E.10. Estimates and 95% confidence bands for coefficients of the selection sorting function: specification 4

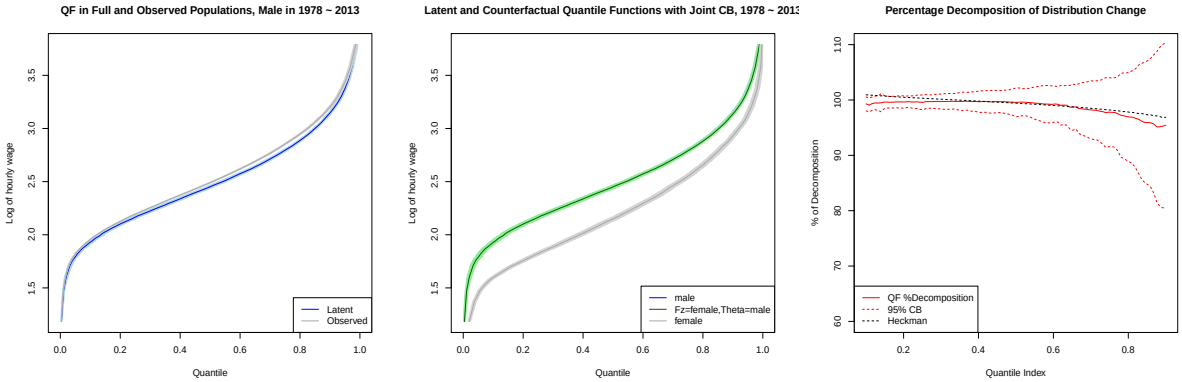

FIGURE E.11. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 1

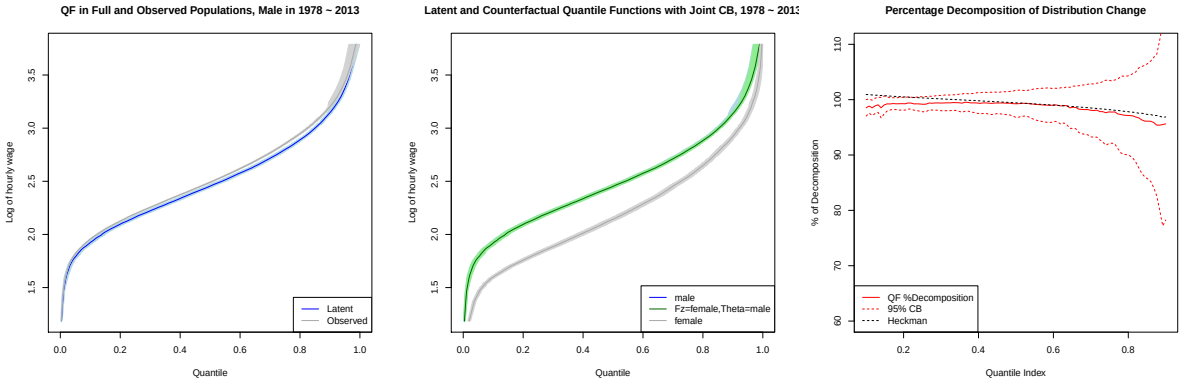

FIGURE E.12. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 2

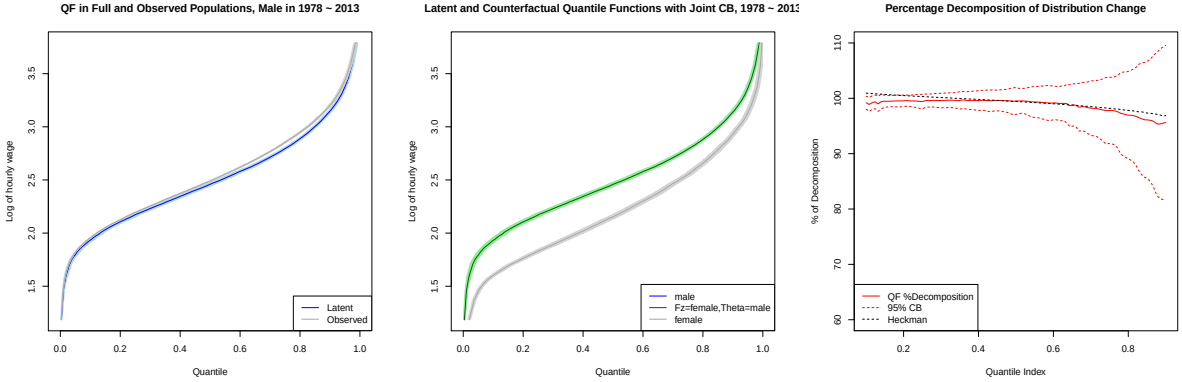

FIGURE E.13. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 3

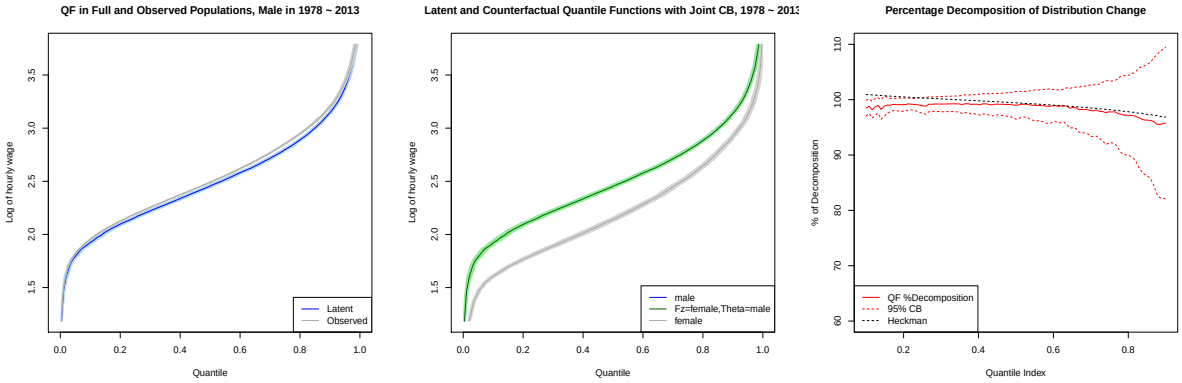

FIGURE E.14. Estimates and 95% confidence bands for the quantiles of observed and offered (latent) wages and decomposition of offered wages between women and men: specification 4

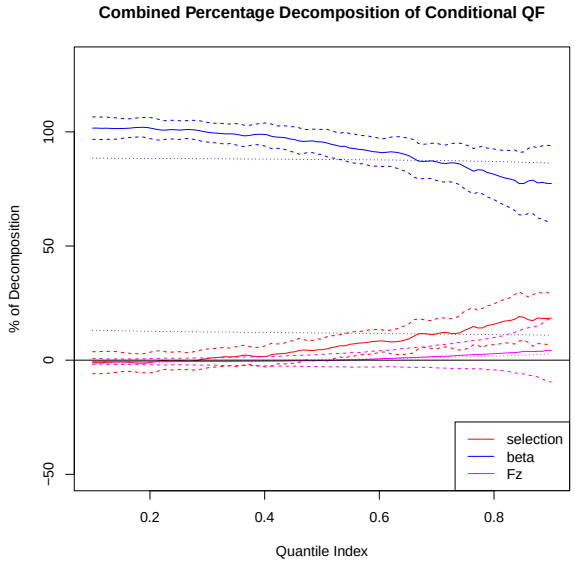

FIGURE E.15. Estimates and 95% confidence bands for decomposition between men and women with aggregated selection effects in specification 1

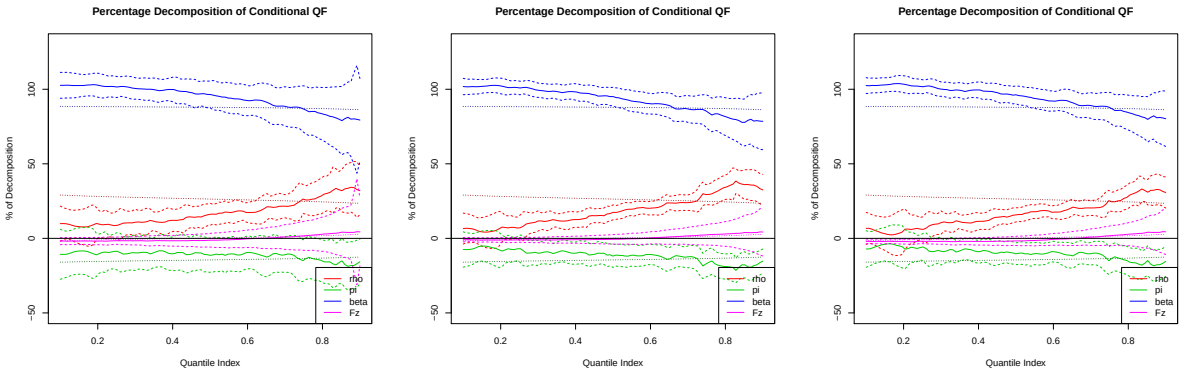

FIGURE E.16. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women: (left) specification 2, (middle) specification 3, and (right) specification 4

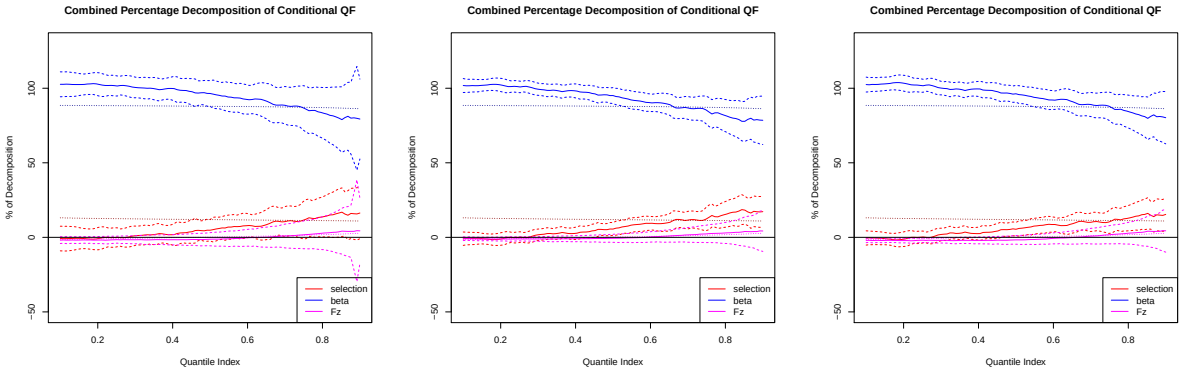

FIGURE E.17. Estimates and 95% confidence bands for the quantiles of observed wages and decomposition between men and women with aggregated selection effect: (left) specification 2, (middle) specification 3, and (right) specification 4

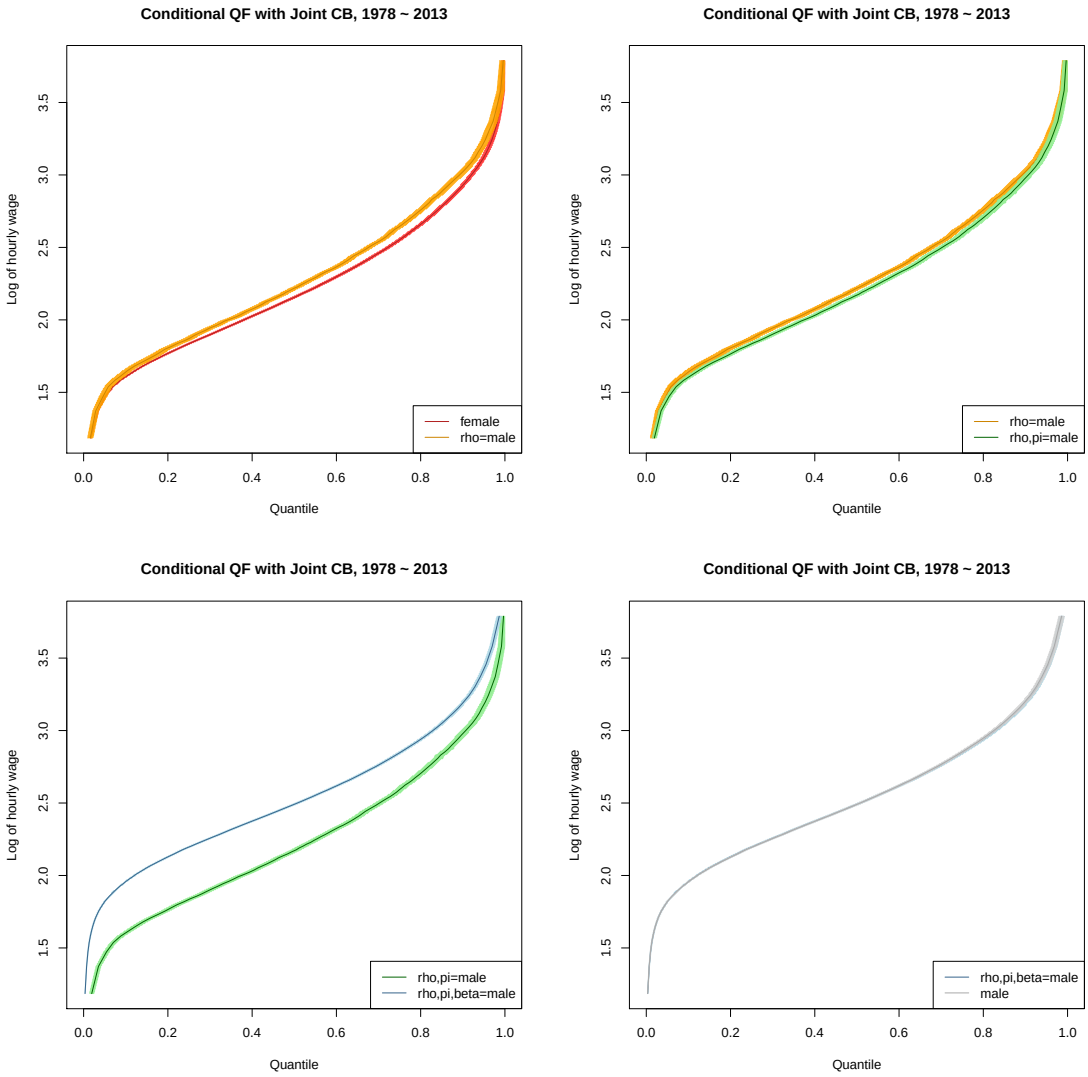

FIGURE E.18. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 1

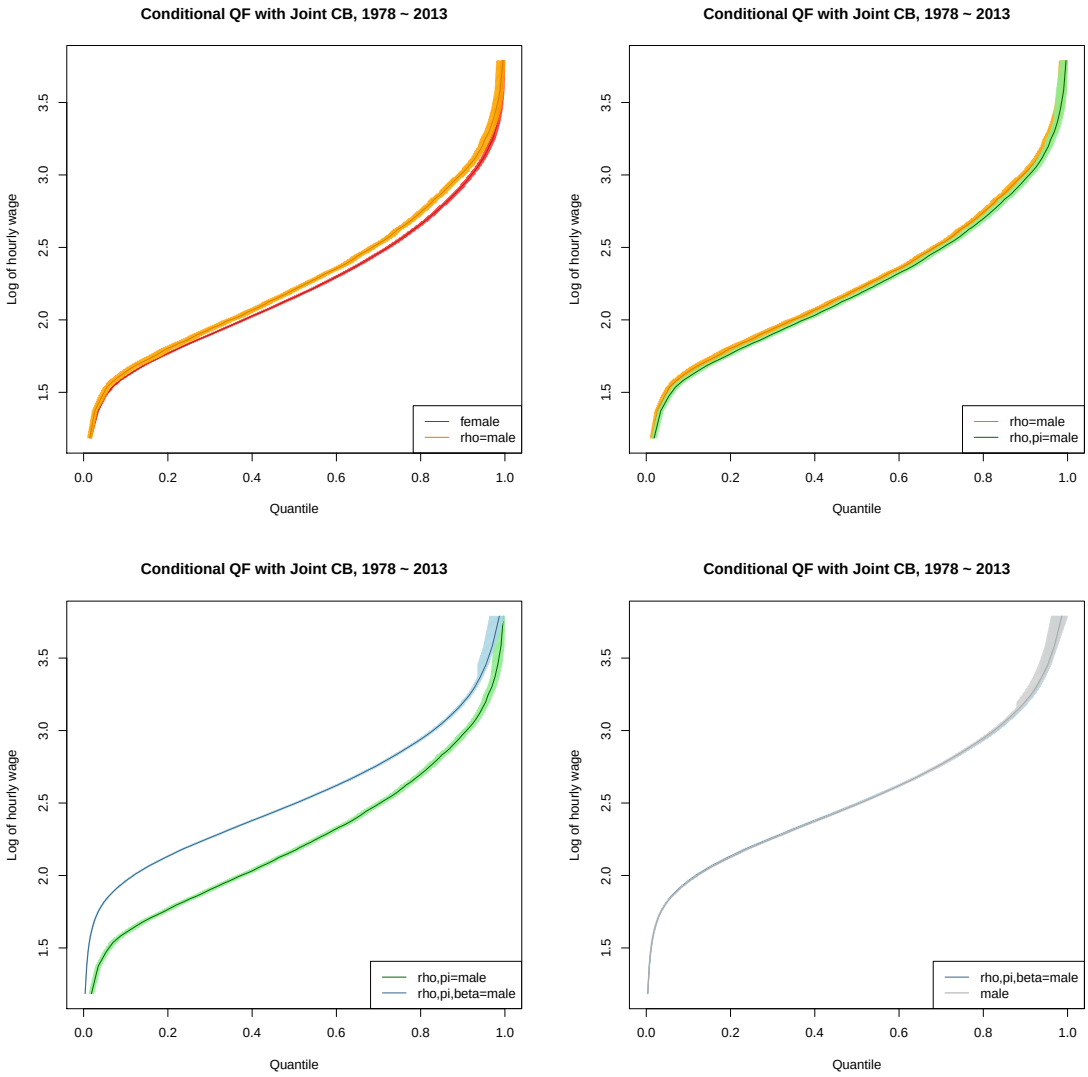

FIGURE E.19. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 2

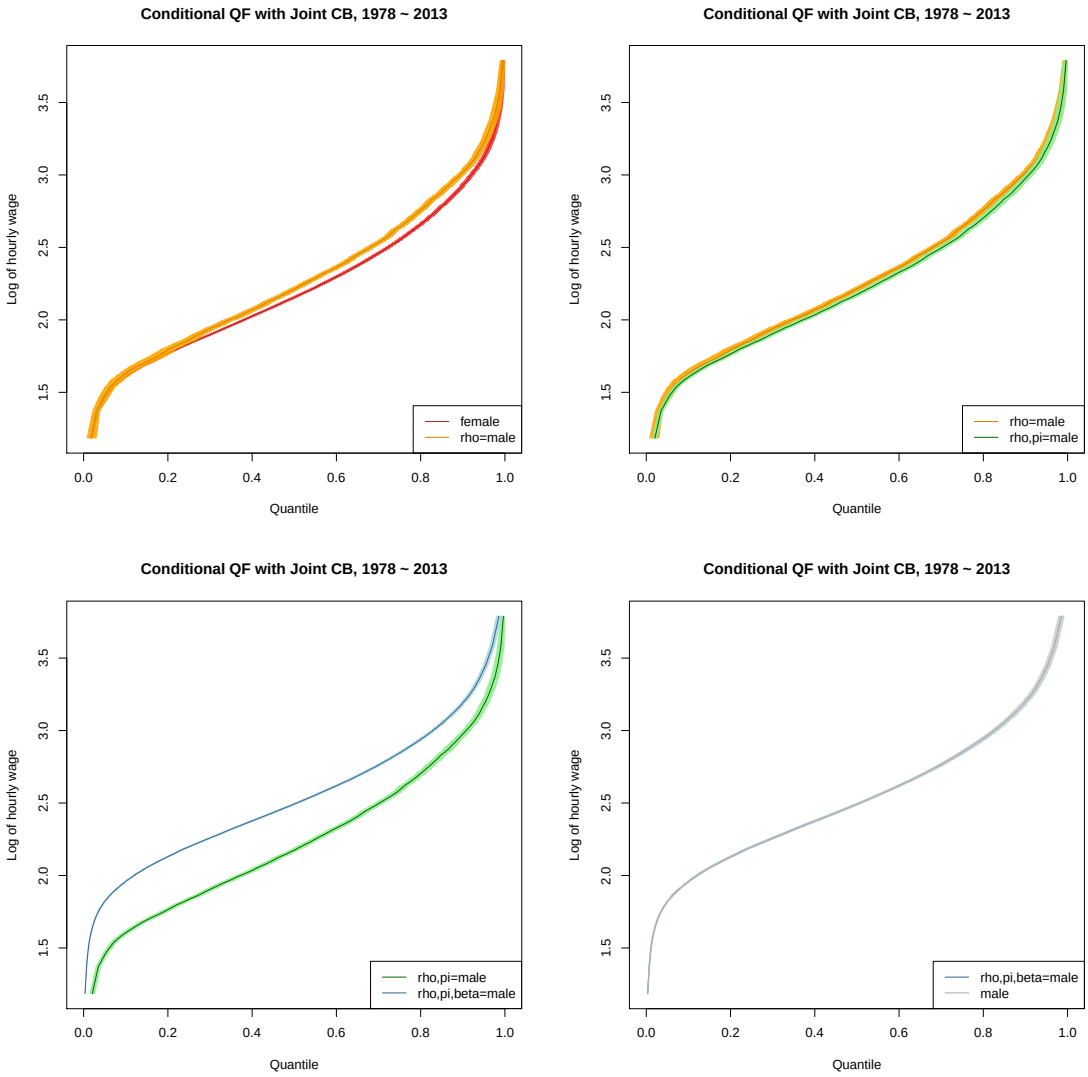

FIGURE E.20. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 3

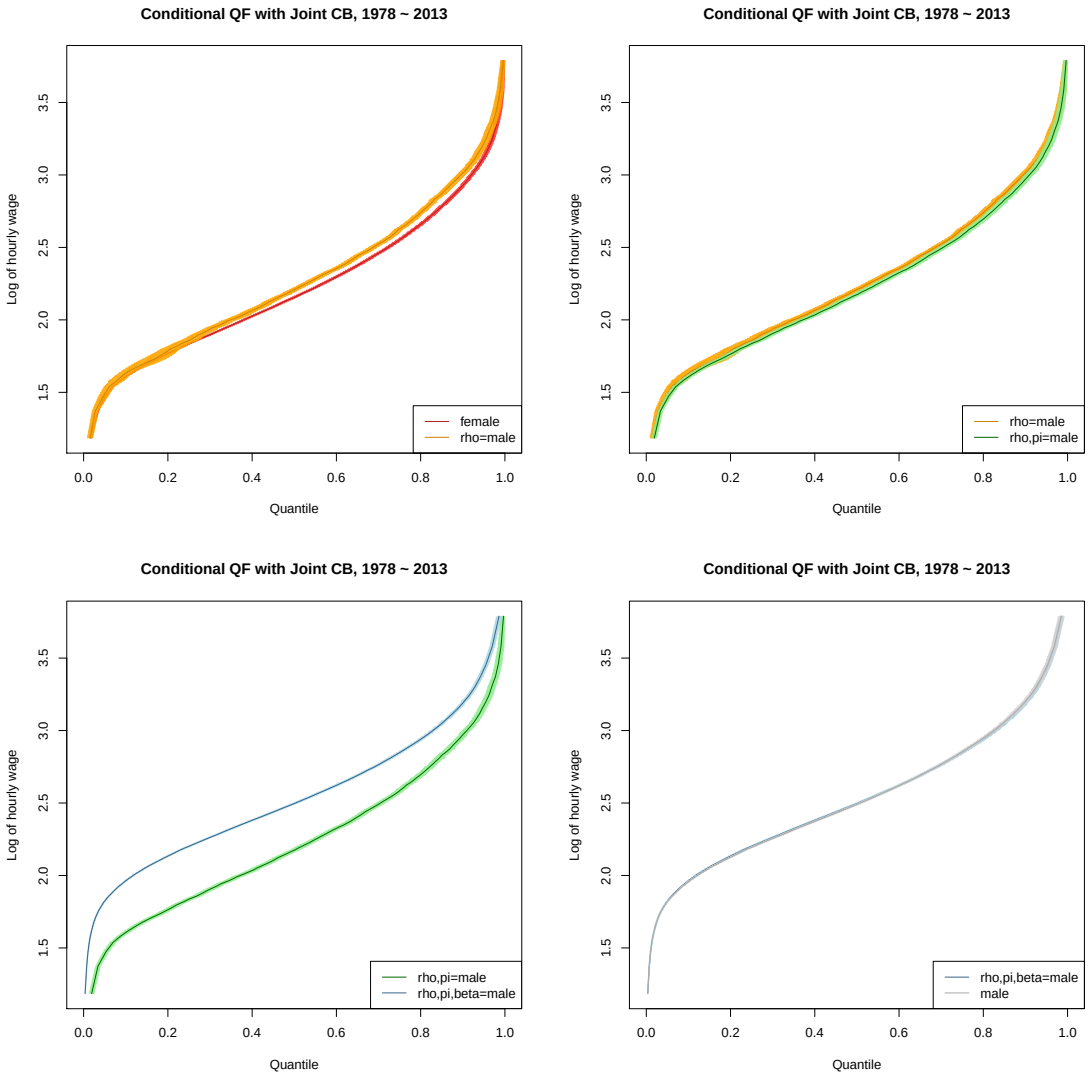

FIGURE E.21. Estimates and 95% confidence bands for components of wage decomposition between women and men in specification 4

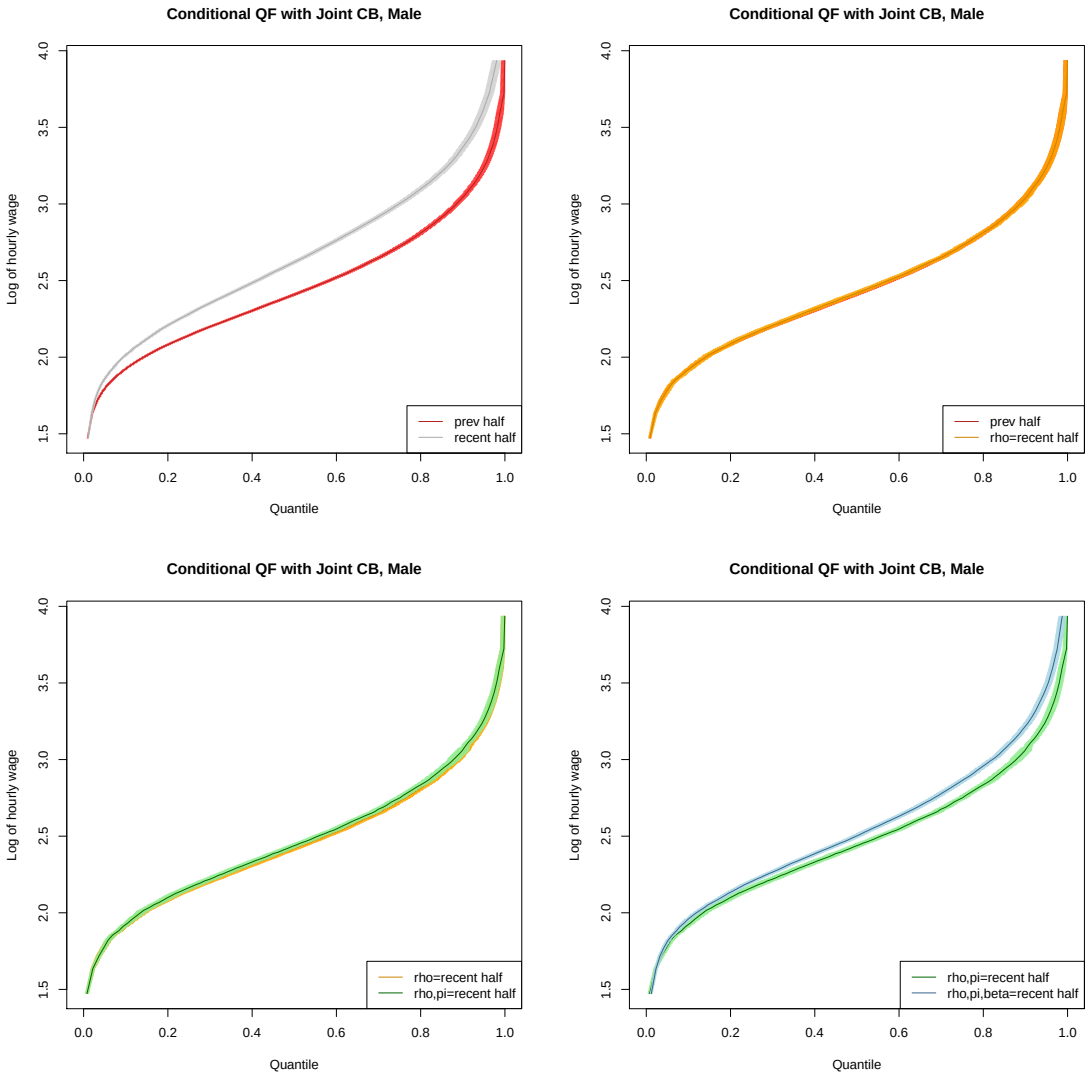

FIGURE E.22. Estimates and 95% confidence bands for components of wage decomposition between first and second half of sample period for men in specification 1

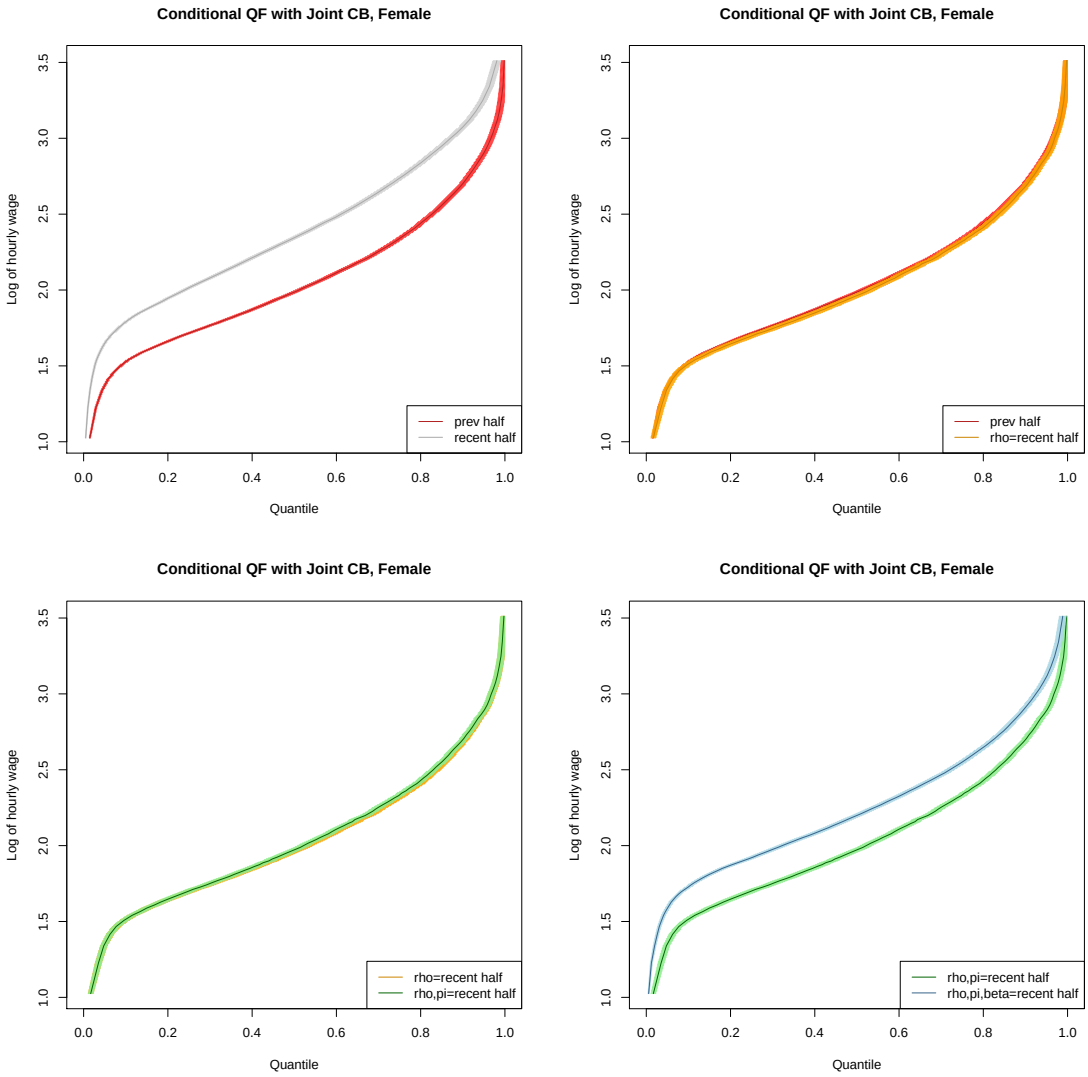

FIGURE E.23. Estimates and 95% confidence bands for components of wage decomposition between first and second half of sample period for women in specification 1