I Z A Institute of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12044

Decomposing Real Wage Changes in the United States

Iván Fernández-Val
Aico van Vuuren
Francis Vella

DECEMBER 2018

DISCUSSION PAPER SERIES

IZA DP No. 12044

Decomposing Real Wage Changes in the United States

Iván Fernández-Val
Boston University

Aico van Vuuren
Gothenburg University

Francis Vella
Georgetown University and IZA

DECEMBER 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Decomposing Real Wage Changes in the United States*

We employ CPS data to analyze the sources of hourly real wage changes in the United States for 1976 to 2016 at various quantiles of the wage distribution. We account for the selection bias from the annual hours of work decision by developing and implementing an estimator for nonseparable selection models with censored selection rules. We then decompose wage changes into composition, structural and selection effects. Composition effects have increased wages at all quantiles but the patterns of wage changes are generally determined by the structural effects. Evidence of changes in the selection effects only appear at lower quantiles of the female wage distribution. The combination of these various components produce a substantial increase in wage inequality. This increase has been exacerbated by the changes in females' working hours.

JEL Classification: C14, I24, J00

Keywords: wage inequality, wage decomposition, nonseparable model, selection bias

Corresponding author:
Francis Vella
Economics Department
Georgetown University
37th and O Streets, NW
Washington, DC 20057
USA
E-mail: fgv@georgetown.edu

* We are grateful to useful comments and suggestions from Stephane Bonhomme, Chinhui Juhn, Ivana Komunjer, Yona Rubinstein and Sami Stouli. We have also greatly benefitted from many useful discussions with Franco Peracchi. This paper supersedes Fernandez-Val, van Vuuren and Vella (2017) which was circulated under a different title.

1 Introduction

The dramatic increase in earnings inequality in the United States is an intensely studied phenomenon (see, for example, Katz and Murphy, 1992, Murphy and Welch, 1992, Juhn et al., 1993, Katz and Autor, 1999, Lee, 1999, Lemieux, 2006, Autor et al., 2008, Acemoglu and Autor, 2011, and Murphy and Topel, 2016). While this literature employs a variety of measures of earnings and inequality and utilizes different data sets, there is general agreement that earnings inequality has greatly increased since the early 1980s. As an individual's earnings reflects both the number of hours worked and the average hourly wage rate, understanding the determinants of each is important for uncovering the sources of earnings inequality. This paper examines the evolution and sources of changes in hourly real wage rates in the United States for 1976 to 2016.

Figure 1 presents selected quantiles of male and female real hourly wage rates using the U.S. Current Population Survey, or March CPS, data for the survey years 1976 to 2016 based on individuals aged 24 to 65 years who worked a positive number of hours the previous year and are neither in the Armed Forces nor self-employed.[1] The profiles differ by gender and quantile. Consider the median male wage rate. Despite a slight increase between 1987 and 1990, the overall trend between 1976 and 1994 was negative, reaching a minimum of 17.6 percentage points below its initial level. It rebounds between 1994 and 2002 but falls between 2007 and 2013. The decline between 1976 and 2016 is 13.6 percentage points. The female median wage, despite occasional dips, increases by nearly 25 percentage points over the sample period. The male wage profile at the 25th percentile shows the same cyclical behavior as the median but a greater decline, especially between 1976 and 1994, resulting in a 2016 wage 18.2 percentage points below its 1976 level. In contrast, the female wage at the 25th percentile increases by about 17 percentage points. The profiles at the 10th percentile are similar to those of the 25th and 50th.

Descriptions of wage inequality involve contrasts between the lower and upper parts of the wage distribution and thus we also examine the higher quantiles. There

[1] Section 2 provides a detailed discussion of the data employed in our analysis and our sample selection process. Our empirical investigation does not employ the person supplement weights provided by the March CPS to account for the survey design. However, we also conducted our analysis employing these weights and our results are unchanged.

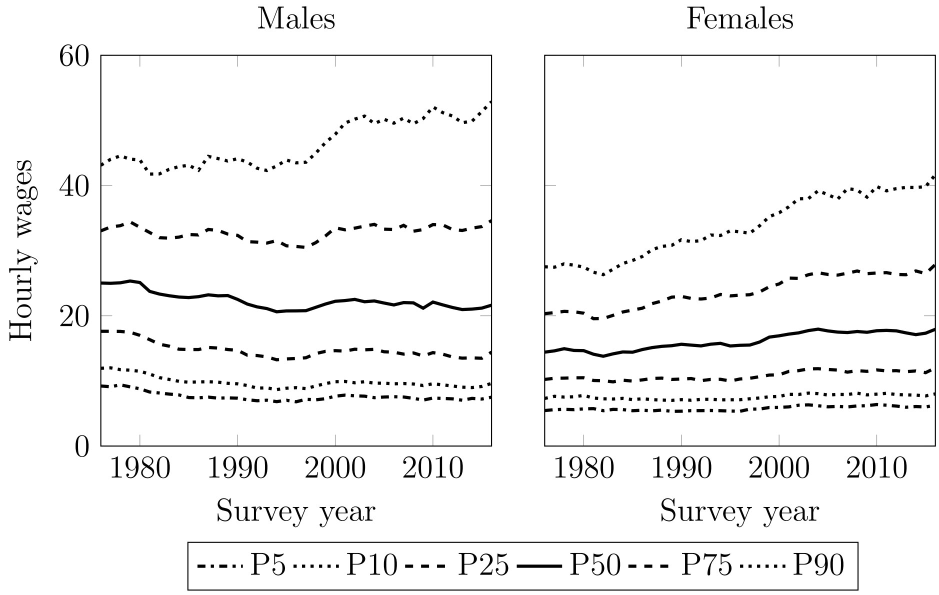

Figure 1: Percentiles of real hourly wages

is a small increase in the male real wage at the 75th percentile over the sample period. However, it was below its 1976 level until the late 1990's before a notable increase in the last 15 years. The profile at the 90th percentile resembles the 75th percentile, although the periods of growth have produced larger increases. For females, there has been strong and steady growth at the 75th and 90th percentiles since 1980 with an increasing gap between each and the median wage.

Figure 2 reports the time series behavior of the 90th/10th percentile ratios and confirms the widening wage gap for both males and females with increases over the sample period of 54.6 and 37.9 percent respectively. Rather than immediately focusing on these ratios, we first identify the determinants of the observed wage changes. The cyclicality of the profiles in Figure 1 suggests that each responds to business cycle forces although the strength of this relationship varies by quantile. The behavior of the various profiles suggests that the workers in their associated labor markets are undergoing contrasting experiences.

Previous empirical work has identified some possible determinants of the growth in wage inequality. First, the labor economics literature (see for example, Juhn, et al., 1993, Katz and Autor, 1999, Welch, 2000, Autor et al., 2008, Acemoglu and Autor, 2011, and Murphy and Topel, 2016) has highlighted the increasing skill premia and returns to higher education. The manner in which the prices of individ-

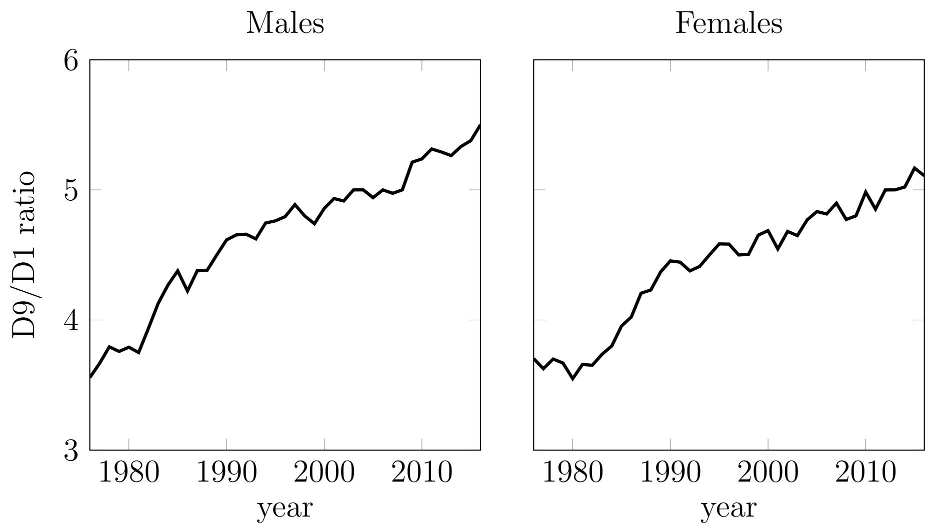

Figure 2: D9-D1 ratio.

ual's characteristics contribute to wages is known as the "structural effect". This also captures factors such as declining real minimum wages and the decrease in the union premium (see, for example, DiNardo et al., 1996 and Lee, 1999). Krueger and Posner (2018) argue that employers have also reduced the bargaining power of relatively low paid workers by including non-compete clauses in employment contracts. Second, the workforce has changed over the past 41 years, suggesting that changes in workers' characteristics also contributed to wage movements. This reflects increases in educational attainment, decreases in unionization rates, and changes in the age structure. The large increase in female labor force participation may have also produced changes in the labor force composition. The contribution to changes in the distribution of hourly wages attributable to these observed characteristics is known as the "composition effect".

Earlier papers (see, for example, Angrist et al., 2006, and Chernozhukov et al., 2013) estimated structural and composition effects under general conditions. However, they ignore the potential selection bias from the employment outcome (see Heckman, 1974, 1979). This "selection effect" is potentially important as the movements in the employment rates and the average number of annual hours worked of both males and females, shown in Figure 3, suggest that unobserved features of the workforce may have changed over the sample period. Just as the returns to observed characteristics may evolve over time and differ across the wage distribution, the returns to unobservables may also vary. Mulligan and Rubinstein (2008) find that selection by females into full-time employment played an important role in explaining the variation in inequality and that the selected sample of working females became increasingly more productive in terms of unobservables during the last three decades of the twentieth century. Moreover, this contributed to a reduction in the gender wage gap and an increase in wage inequality for females. Maasoumi and Wang (2018) support these findings.

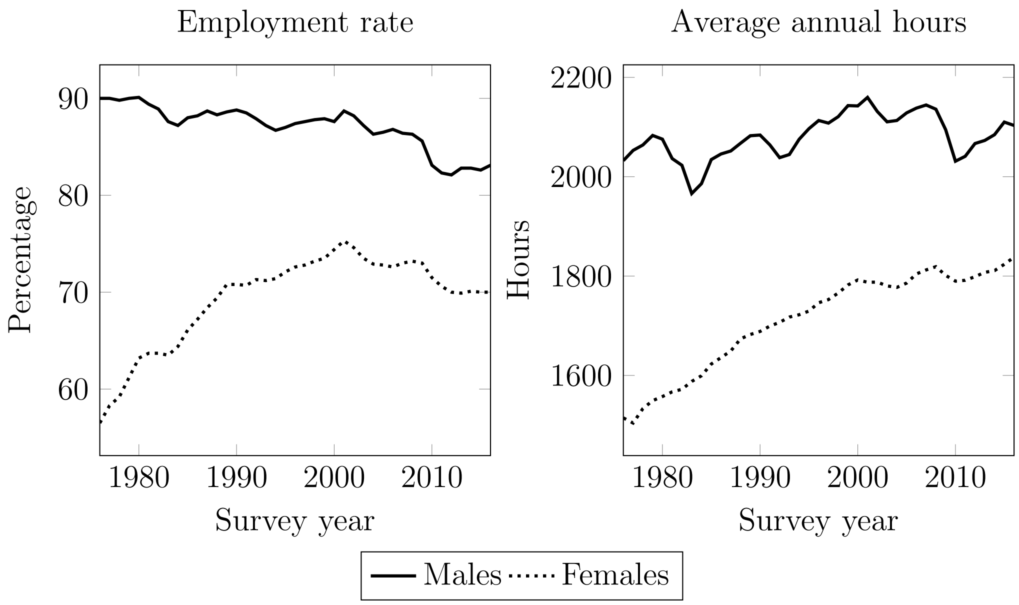

Figure 3: Employment rate and average annual hours worked of employees

Mulligan and Rubinstein employ the Heckman (1979) selection model to evaluate the role of selection in changes in the conditional mean wage. This is straightforward given the separable nature of the model. However, to evaluate the impact of selection at different quantiles of the wage distribution requires a nonseparable model. Maasoumi and Wang (2018) do so by employing the Arellano and Bonhomme (2017) procedure which is based on a copula approach to quantile selection models. These studies allocate observations into categories denoting whether or not the individual worked full time. This facilitates the use of the Heckman (1979) and the Arellano and Bonhomme procedures as each employs a binary selection rule. One is then restricted to examining full time wages. However, to obtain a fuller understanding of the evolution of real wages it is useful to examine the wages of all those working across a range of hours and account for the selection bias from the hours of work decision. Figure 4 presents the fraction of workers who are full time and full year (FTFY) defined as working at least 35 hours a week for a minimum of 50 weeks. There is a large increase in the fraction of FTFY females, from 45 percent in 1976 to 65 percent in 2016, while the percentage of males increases from 75 to 81. However, while the proportions working full time are increasing, there are a substantial number of individuals working less than full time. Moreover, the real wage changes of the corresponding groups, presented in Figure 5, feature important differences. While there are instances of specific quantiles and time periods when the non-FTFY wage is higher than the FTFY category, the opposite is generally true and the differences are large in some periods. The general trends also appear to sometimes differ across categories. Note that while we illustrate these differences by examining these two categories, it is possible that there is a continuum of differences along the hours distribution.

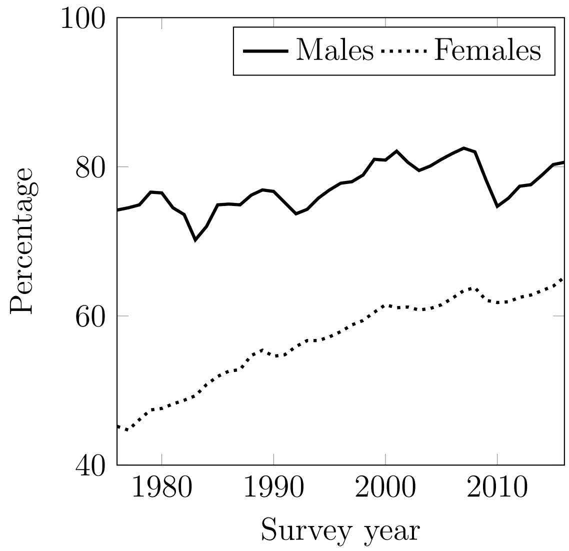

Figure 4: Fraction of full-time-full-year workers

This evidence suggests that considering only the wages of the FTFY workers may not provide an accurate representation of the evolution of wages for all workers. Also, while the differences in the wages across these two working categories might capture structural and composition effects, they may also reflect selection effects. The failure to include non-FTFY workers means that the selection effects in earlier studies may reflect movement from the non-FTFY to FTFY rather than from

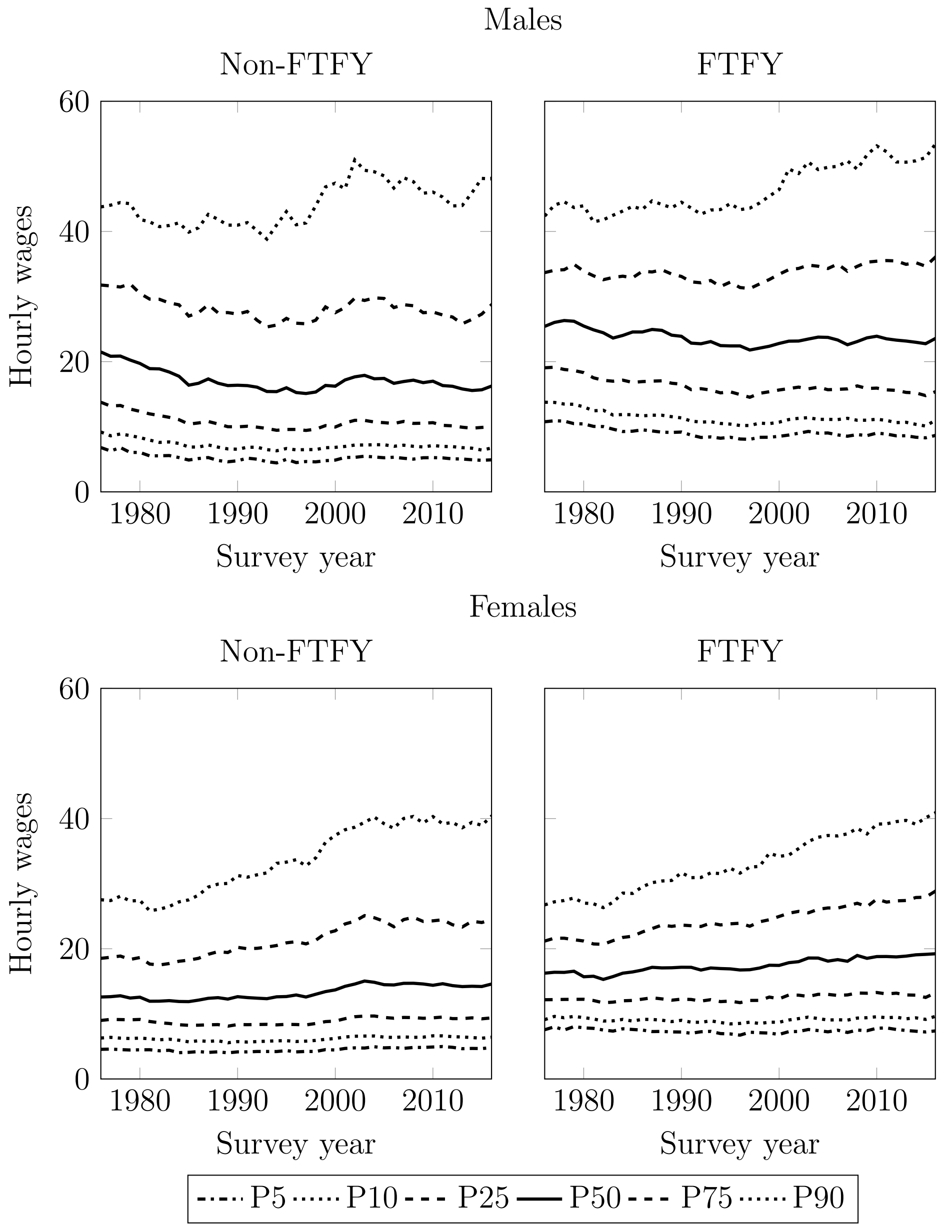

Figure 5: Percentiles of real hourly wages

non-employment to FTFY. A preferred approach would allow for different selection effects at different points of the hours distribution while incorporating varying structural and compositional effects. We incorporate these considerations by providing an estimation strategy for nonseparable models with a censored selection rule. We estimate the relationship between the individual's wage and their characteristics while accounting for selection resulting from the individual's location in the hours distribution. We employ distribution regression methods and account for sample selection via an appropriately constructed control function. Although our control function requires a censored variable as the basis of the selection rule, this is provided via the number of annual working hours.

Our estimator contributes to the literatures on nonseparable models with endogeneity (see, for example, Chesher, 2003, Ma and Koenker, 2006, Florens et al., 2008, Imbens and Newey, 2009, Jun, 2009 and Masten and Torgovitsky, 2014) and nonseparable sample selection models (for example, Newey, 2007 and Arellano and Bonhomme, 2017). We provide statements, applicable for both the selected and non-selected samples, regarding the wage distribution conditional on specific values of the control function. This local approach to identification is popular in many contexts (see, for example, Chesher, 2003 and Heckman and Vytlacil, 2005). We show that for any population observation with a positive probability of being selected, selection is irrelevant for the wage distribution conditional on the control function. Hence, we can estimate certain objects that are appropriate for the whole population conditional on the value of the control function. Our approach, when applicable, is an alternative to Arellano and Bonhomme (2017) and requires weaker identification conditions for the associated decomposition exercises described below.

Our empirical results are striking. Male real wages at the median and below have decreased despite an increasing skill premium and an increase in educational attainment. The reduction primarily reflects the large wage decreases of those with low levels of education. Wages at the upper quantiles have increased drastically due to the increasing skill premium. The combination of these two trends substantially increases male wage inequality. Female wage growth at lower quantiles is modest although the median wage has grown steadily. Gains at the upper quantiles for females are substantial and reflect increasing returns to schooling. These factors also have produced a substantial increase in female wage inequality. Changes in the impact of selection appear at the lower parts of the female wage distribution. Moreover, we find that these changes in selection decrease the observed wage growth and increase observed wage inequality.

The following section discusses the data. Section 3 outlines our empirical model. Section 4 provides the objects of interest and our most important identification results. Section 5 presents the empirical results and Section 6 provides some additional discussion of the empirical results. Concluding comments are offered in Section 7.

2 Data

We employ micro-level data from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS), or March CPS, for the 41 survey years from 1976 to 2016 which report annual earnings for the previous calendar year.[2] The 1976 survey is the first for which information on weeks worked and usual hours of work per week last year are available. To avoid issues related to retirement and ongoing educational investment we restrict attention to those aged 24–65 years in the survey year. This produces an overall sample of 1,794,466 males and 1,946,957 females, with an average annual sample size of 43,767 males and 47,487 females. The annual sample sizes range from a minimum of 30,767 males and 33,924 females in 1976 to a maximum of 55,039 males and 59,622 females in 2001.

Annual hours worked last year are defined as the product of weeks worked and usual weekly hours of work last year. Most of those reporting zero hours respond that they are not in the labor force (i.e., they report themselves as doing housework, unable to work, at school, or retired) in the week of the March survey. We define hourly wages as the ratio of reported annual earnings and annual hours worked last year. Hourly wages are unavailable for those not in the labor force. However, for the Armed Forces, the self-employed, and unpaid family workers annual earnings or annual hours tend to be poorly measured. Thus we exclude these groups from our sample and focus on civilian dependent employees with positive hourly wages and people out of the labor force last year. This restricted sample contains 1,551,796 males and 1,831,220 females (respectively 86.5 percent and 94.1 percent of the original sample of people aged 24–65), with average annual sample sizes of 37,849 males and 44,664 females. The subsample of civilian dependent employees with positive hourly wages contains 1,346,918 males and 1,276,125 females, with an average annual sample size of 32,852 males and 31,125 females.

[2] We downloaded the data from the IPUM-CPS website maintained by the Minnesota Population Center at the University of Minnesota (Flood et al., 2015), which provides standardized data from 1963 to 2016.

3 Model and Objects of Interest

The model has the following structure:

$$Y = g(X, \epsilon) \quad \text{if } C > 0, \tag{3.1}$$

$$C = \max\left(h(Z, \eta), 0\right), \tag{3.2}$$

where Y and C are observable random variables, and X and Z are vectors of observable explanatory variables. The set of variables in X is a subset of those in Z. We do not need to impose an exclusion restriction on Z with respect to the elements of X, although our identification assumptions will be more plausible with such a restriction. The functions g and h are unknown and ϵ and η are respectively a vector and a scalar of potentially mutually dependent unobservables. We shall impose restrictions on the stochastic properties of these unobservables. Our objective is to estimate functionals related to g noting that Y is only observed when C is above some known threshold normalized to be zero. The non observability of Y for specific values of C induces the possibility of selection bias. For generality, we refer to (3.1) and (3.2) as the outcome and selection equations although below they refer to the wage and hours equations respectively. The model is a nonparametric and nonseparable representation of the tobit type-3 model and is a variant of the Heckman (1979) selection model. It has been studied in more restricted settings than considered here (see Amemiya, 1978, 1979, Vella, 1993, Honoré et al., 1997, Chen, 1997 and Lee and Vella, 2006). This model is ideal for our purposes as it incorporates a conditional wage distribution in the presence of nonparticipants and variation in the number of hours worked. Estimating functions related to g while accounting for selection enables the construction of counterfactual distributions adjusted for selection required for our decomposition exercises.

As the following approach is applicable in many settings we acknowledge that the binary selection model is more frequently employed in empirical work than the censored selection model. However, this partially reflects the popularity of the Heckman (1979) two-step procedure. Many empirical investigations dichotomize censored selection variables in order to employ the Heckman procedure.[3] In the parametric

[3]Examples of such commonly encountered censoring selection variables include unemployment duration, training program length and the magnitude or length of welfare benefit receipt.

setting there are no substantial benefits in using the censored, rather than the dichotomized, form of the selection variable other than that the selection variable can appear in the outcome equation as a regressor and the variation in the selection rule provides an additional form of identification. However, in the nonseparable setting the use of the censored selection rule has the additional advantage that, in comparison with the closely related existing treatment effects literature, we are able to allow for nonseparability in both the outcome and selection equations. Moreover, using the censored selection rule enables inference for both the selected and non-selected populations and the local effects. It also allows the selection effects to vary by C. Thus it could be argued that the censored selection approach should be employed when possible.[4]

4 Identification of objects of interest

We account for selection bias through an appropriately constructed control function. We establish the existence of such a function for this model and then define some objects of interest incorporated in (3.1)-(3.2).

Let $\perp\!\!\!\perp$ denote stochastic independence. We begin with the following assumption:

Assumption 1 (Control Function) $(\epsilon, \eta) \perp\!\!\!\perp Z$, *$\eta$ is a continuously distributed random variable with strictly increasing CDF on the support of η, and $t \mapsto h(Z, t)$ is strictly increasing a.s.*

This assumption allows for endogeneity between X and ϵ in the selected population with $C > 0$, since in general ϵ and η are dependent, *i.e.*, $\epsilon \perp\!\!\!\perp X \mid C > 0$. The monotonicity assumption allows a non-monotonic relationship between ϵ and C because ϵ and η are allowed to be non-monotonically dependent. Under Assumption 1, the

[4]The model can be extended in several directions. For example, the selection variable C could be censored in a number of ways provided there are some region(s) for which it is continuously observed. This allows for top, middle and/or bottom censoring. The approach is also applicable when the Y is censored. For example:

$$Y = \max(g(X, \epsilon), 0) \quad \text{if } C > 0.$$

The model can also be extended to include C in the outcome equation as an explanatory variable provided that there is an exclusion restriction in Z with respect to X. This corresponds to the triangular system of Imbens and Newey (2009) with censoring in the first stage equation.

distribution of η can be normalized to be uniform on $[0,1]$ without loss of generality (Matzkin, 2003).[5] The following result establishes the existence of a control function for the selected population in this setting. That is, there is a function of the observable data such that once it is conditioned upon, the unobservable component is independent of the explanatory variables in the outcome equation for the selected population. Let $V := F_C(C \mid Z)$ where $F_C(\cdot \mid z)$ denotes the CDF of C conditional on $Z = z$.

Lemma 1 (Existence of Control Function) *Under the model in (3.1)-(3.2) and Assumption 1:*

$$\epsilon \perp\!\!\!\perp Z \mid V, C > 0.$$

Proofs for the results in this section are provided in Appendix B. The intuition behind Lemma 1 is based on three observations. First, $V = \eta$ when $C > 0$, so conditioning on V is identical to conditioning on η in the selected population. Second, conditioning on Z and η makes selection, $i.e., C > 0$, deterministic. Therefore, the distribution of ϵ, conditional on Z and η, does not depend on the condition that $C > 0$. The final observation, namely the assumption that $(\epsilon, \eta) \perp\!\!\!\perp Z$, is sufficient to prove the Lemma.

We consider two classes of objects: (1) local effects conditional on the control function, and (2) global effects based on integration over the control function.

4.1 Local effects

Let $\mathcal{Z}$, $\mathcal{X}$, and $\mathcal{V}$ denote the marginal supports of Z, X, and V in the selected population, respectively. Let the set $\mathcal{XV}$ denote the joint support of X and V in the selected population. That is:

Definition 1 (Identification set) *Define:*

$$\mathcal{XV} := \{(x, v) \in \mathcal{X} \times \mathcal{V} : h(z, v) > 0, z \in \mathcal{Z}(x)\} \qquad ,$$

where $\mathcal{Z}(x) = \{z \in \mathcal{Z} : x \subseteq z\}$, i.e. *the set of values of Z with the component $X = x$.*

[5]Indeed if $t \mapsto h(z,t)$ is strictly increasing, and η is continuously distributed with $\eta \sim F_\eta$, then $\tilde{h}(z, \tilde{\eta}) = h(z, F_\eta(\tilde{\eta}))$ is such that $t \mapsto \tilde{h}(z,t)$ is strictly increasing and $\tilde{\eta} \sim U(0,1)$.

Depending on the values of (X, η), we can classify the units of observation into 3 groups: (1) always selected units when $h(z, t) > 0$ for all $z \in \mathcal{Z}(x)$, (2) switchers when $h(z, t) > 0$ for some $z \in \mathcal{Z}(x)$ and $h(z, t) \leq 0$ for some $z \in \mathcal{Z}(x)$, and (3) never selected units when $h(z, t) \leq 0$ for all $z \in \mathcal{Z}(x)$. The set $\mathcal{XV}$ only includes always selected units and switchers, *i.e.* units with (X, V) such that they are observed for some values of Z. When $X = Z$ there are no switchers because the set $\mathcal{Z}(x)$ is a singleton. Otherwise the size of the set $\mathcal{XV}$ increases with the support of the excluded variables and their strength in the selection equation. The local average structural function is the first local effect we consider.

Definition 2 (LASF) *The local average structural function (LASF) at (x, v) is:*

$$\mu(x, v) = \mathbb{E}(g(x, \epsilon) \mid V = v).$$

The LASF is the expected value of the potential outcome $g(x, \epsilon)$ obtained by fixing X at x conditional on $V = v$ for the entire population. It is useful for measuring the effect of X on the mean of Y. For example, the average treatment effect of changing X from x_0 to x_1 conditional on $V = v$ is:

$$\mu(x_1, v) - \mu(x_0, v).$$

The following result shows that $\mu(x, v)$ is identified for all $(x, v) \in \mathcal{XV}$.

Theorem 1 (Identification of LASF) *Under the model (3.1)-(3.2), Assumption 1 and $\mathbb{E}|Y| < \infty$, for $(x, v) \in \mathcal{XV}$,*

$$\mu(x, v) = \mathbb{E}(Y \mid X = x, V = v, C > 0). \tag{4.1}$$

According to Theorem 1, the LASF is identical to the expected value of the outcome variable conditional on $(X, V) = (x, v)$ in the selected population. The proof of this theorem is based on Assumption 1 that allows for the LASF to be conditional on the outcome of $(Z, V) = (z, v)$. Since $(x, v) \in \mathcal{XV}$, there is a $z \in \mathcal{Z}(x)$ such that $h(z, v) > 0$ and hence the expected mean outcome of $g(x, \epsilon)$ conditional on $V = v$ for the total sample, *i.e.* the LASF, is the same as the mean outcome for the selected population. That is, selection is irrelevant for the distribution of the outcome variable conditional on the control function. This mean outcome is equal to the conditional expectation in the selected population, which is a function of the data distribution and is hence identified. When X is continuous and $x \mapsto g(x, \epsilon)$ is differentiable a.s., we can consider the average derivative of $g(x, \epsilon)$ with respect to x conditional on the control function.

Definition 3 (LADF) *The local average derivative function (LADF) at (x, v) is:*

$$\delta(x, v) = \mathbb{E}[\partial_x g(x, \epsilon) \mid V = v], \qquad \partial_x := \partial/\partial x. \tag{4.2}$$

The LADF is the first-order derivative of the LASF with respect to x, provided that we can interchange differentiation and integration in (4.2). This is made formal in the next corollary which shows that the LADF is identified for all $(x, v) \in \mathcal{XV}$.

Corollary 1 (Identification of LADF) *Assume that for all $x \in \mathcal{X}$, $g(x, \epsilon)$ is continuously differentiable in x a.s., $\mathbb{E}[|g(x, \epsilon)|] < \infty$, and $\mathbb{E}[|\partial_x g(x, \epsilon)|] < \infty$. Under the conditions of Theorem 1, for $(x, v) \in \mathcal{XV}$,*

$$\delta(x, v) = \partial_x \mu(x, v) = \partial_x \mathbb{E}(Y \mid X = x, V = v, C > 0) \quad .$$

The local effects extend to distributions and quantiles.

Definition 4 (LDSF and LQSF) *The local distribution structural function (LDSF) at (y, x, v) is:*

$$G(y, x, v) = \mathbb{E}[1\{g(x, \epsilon) \leq y\} \mid V = v].$$

The local quantile structural function (LQSF) at (τ, x, v) is:

$$q(\tau, x, v) := \inf\{y \in \mathbb{R} : G(y, x, v) \geq \tau\}.$$

The LDSF is the distribution function of the potential outcome $g(x, \epsilon)$ conditional on the value of the control function for the entire population. The LQSF is the left-inverse function of $y \mapsto G(y, x, v)$ and corresponds to quantiles of $g(x, \epsilon)$. Differences of the LQSF across levels of x correspond to quantile effects conditional on V for the entire population. For example, the τ-quantile treatment effect of changing X

from x_0 to x_1 is:

$$q(\tau, x_1, v) - q(\tau, x_0, v).$$

Remark 1 (Identification of LDSF) *Identification of the LDSF follows by the same argument as the identification of the LASF, replacing $g(x, \epsilon)$ (as in Definition 2) by $1\{g(x, \epsilon) \leq y\}$ and Y (as in equation (4.1)) by $1\{Y \leq y\}$. Thus, under Assumption 1, for $(x, v) \in \mathcal{XV}$:*

$$\mathbb{E}[1\{g(x, \epsilon) \leq y\} \mid V = v] = F_{Y|X,V,C>0}(y \mid x, v).$$

The LQSF is then identified by the left-inverse function of $y \mapsto F_{Y|X,V,C>0}(y \mid x, v)$, the conditional quantile function $\tau \mapsto \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0]$, *i.e.*, for $(x, v) \in \mathcal{XV}$,

$$q(\tau, x, v) = \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0].$$

We also consider the derivative of $q(\tau, x, v)$ with respect to x and call it the local quantile derivative function (LQDF). This corresponds to the average derivative of $g(x, \epsilon)$ with respect to x at the quantile $q(\tau, x, v)$ conditional on $V = v$ under suitable regularity conditions; see Hoderlein and Mammen (2011). Thus, for $(\tau, x, v) \in [0, 1] \times \mathcal{XV}$,

$$\delta_\tau(x, v) := \partial_x q(\tau, x, v) = \mathbb{E}[\partial_x g(x, \epsilon) \mid V = v, g(x, \epsilon) = q(\tau, x, v)].$$

By an analogous argument to Corollary 1, the LQDF is identified at $(\tau, x, v) \in [0, 1] \times \mathcal{XV}$ by:

$$\delta_\tau(x, v) = \partial_x \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0],$$

provided that $x \mapsto \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0]$ is differentiable and other regularity conditions hold.

Remark 2 (Exclusion restrictions) *The identification of local effects does not explicitly require exclusion restrictions in Z with respect to X although the size of the identification set $\mathcal{XV}$ depends on such restrictions. For example, if $h(z, \eta) = z + \Phi^{-1}(\eta)$ where Φ is the standard normal distribution and $X = Z$, then $\mathcal{XV} =$*

$\{(x, v) \in \mathcal{X} \times \mathcal{V}] : x > -\Phi^{-1}(v)\} \subset \mathcal{X} \times \mathcal{V}$; *whereas if* $h(z, \eta) = x + z_1 + \Phi^{-1}(\eta)$ *for* $Z = (X, Z_1)$, *then* $\mathcal{XV} = \{(x, v) \in \mathcal{X} \times \mathcal{V} : x > -\Phi^{-1}(v) - z_1, z_1 \in \mathcal{Z}(x)\}$, *such that* $\mathcal{XV} = \mathcal{X} \times \mathcal{V}$ *if* Z_1 *is independent of* X *and supported in* $\mathbb{R}$.

4.2 Global effects

The global counterparts of the local effects are obtained by integration over the control function in the selected population. A typical global effect at $x \in \mathcal{X}$ is:

$$\theta_S(x) = \int \theta(x, v) dF_{V|C>0}(v), \tag{4.3}$$

where $\theta(x, v)$ can be any of the local objects defined above and $F_{V|C>0}(v)$ is the distribution of V in the selected population. Identification of $\theta_S(x)$ requires identification of $\theta(x, v)$ over $\mathcal{V}$, the support of V in the selected population. For example, the average structural function (ASF):

$$\mu_S(x) := \mathbb{E}[g(x, \epsilon) \mid C > 0],$$

gives the average of the potential outcome $g(x, \epsilon)$ in the selected population. By the law of iterated expectations, this is the global effect (4.3) with $\theta(x, v) = \mu(x, v)$, the LASF. The average treatment effect of changing X from x_0 to x_1 in the selected population is:

$$\mu_S(x_1) - \mu_S(x_0).$$

Similarly, one can consider the distribution structural function (DSF) in the selected population as in Newey (2007), *i.e:*

$$G_S(y, x) := \mathbb{E}[1\{g(x, \epsilon) \leq y\} \mid C > 0],$$

which gives the distribution of the potential outcome $g(x, \epsilon)$ at y in the selected population. This is the global effect (4.3) with $\theta(x, v) = G(y, x, v)$. We construct the quantile structural function (QSF) in the selected population as the left-inverse of $y \mapsto G_S(y, x)$. That is:

$$q_S(\tau, x) := \inf\{y \in \mathbb{R} : G_S(y, x) \geq \tau\}.$$

The QSF gives the quantiles of $g(x, \epsilon)$. Unlike $G_S(y, x)$, $q_S(\tau, x)$ cannot be obtained by integration of the corresponding local effect, $q(\tau, x, v)$, because we cannot interchange quantiles and expectations. The τ-quantile treatment effect of changing X from x_0 to x_1 in the selected population is:

$$q_S(\tau, x_1) - q_S(\tau, x_0).$$

Global counterparts of the LADF and LQSF are obtained by taking derivatives of $\mu_S(x)$ and $q_S(\tau, x)$ with respect to x.

As in Newey (2007), identification of the global effects in the selected population requires a condition on the support of the control function. Let $\mathcal{V}(x)$ denote the support of V conditional on $X = x$, *i.e.* $\mathcal{V}(x) := \{v \in \mathcal{V} : (x, v) \in \mathcal{XV}\}$.

Assumption 2 (Common Support) $\mathcal{V}(x) = \mathcal{V}$.

The main implication of common support is the identification of $\theta(x)$ from the identification of $\theta(x, v)$ in $v \in \mathcal{V}(x) = \mathcal{V}$. Assumption 2 is only plausible under exclusion restrictions on Z with respect to X; see the example in Remark 2. We now establish the identification of the typical global effect (4.3).

Theorem 2 (Identification of Global Effects) *If $\theta(x, v)$ is identified for all $(x, v) \in \mathcal{XV}$, then $\theta_S(x)$ is identified for all $x \in \mathcal{X}$ that satisfy Assumption 2.*

We can now apply this result to show identification of global effects in the selected population, because under Assumption 1 the local effects are identified over $\mathcal{XV}$, which is the support of (X, V) in the selected population.

Remark 3 (Global Effects in the Entire Population) *The effects in the selected population generally differ from the effects in the entire population, except under the additional support condition:*

$$\mathcal{V} = (0, 1), \tag{4.4}$$

which imposes that the control function is fully supported in the selected population. This condition requires an excluded variable in Z with sufficient variation to make $h(Z, \eta) > 0$ for any $\eta \in [0, 1]$.

4.3 Global effects for "the treated" and average derivatives

Assumption 2 might be considered implausible in the absence of a variable with large support which can be included in the selection equation but excluded from the outcome equation. Without this assumption the global effects are not point identified, but can be bounded following a similar approach to Imbens and Newey (2009). We consider instead the alternative generic global effect:

$$\theta_S(x \mid x_0) = \int \theta(x, v) dF_{V|X,C>0}(v \mid x_0), \tag{4.5}$$

which is point identified under weaker support conditions than (4.3). Examples of (4.5) include the ASF conditional on $X = x_0$ in the selected population:

$$\mu_S(x \mid x_0) = \mathbb{E}[g(x, \epsilon) \mid X = x_0, C > 0],$$

which is (4.5) with $\theta(x, v) = \mu(x, v)$. This measures the mean of the potential outcome $g(x, \epsilon)$ for the selected individuals with $X = x_0$, and can be used to construct the average treatment effect on the treated of changing X from x_0 to x_1 :

$$\mu_S(x_1 \mid x_0) - \mu_S(x_0 \mid x_0).$$

The object in (4.5) is identified in the selected population under the following support condition:

Assumption 3 (Weak Common Support) $\mathcal{V}(x) \supseteq \mathcal{V}(x_0)$.

Assumption 3 is weaker than Assumption 2 because $\mathcal{V}(x_0) \subseteq \mathcal{V}$. For example, if the selection equation (3.2) is monotone in X, Assumption 3 is satisfied by setting x_0 lower than x.

We define the τ-quantile treatment on the treated as:

$$q_S(\tau, x_1 \mid x_0) - q_S(\tau, x_0 \mid x_0),$$

where $q_S(\tau, x \mid x_0)$ is the left-inverse of the DSF conditional on $X = x_0$ in the selected population,

$$G_S(y, x \mid x_0) := \mathbb{E}[1\{g(x, \epsilon) \leq y\} \mid X = x_0, C > 0],$$

which is (4.5) with $\theta(x, v) = G(y, x, v)$.

We now establish identification of the typical global effect (4.5).

Theorem 3 (Identification of Global effects on "the treated") *If $\theta(x, v)$ is identified for all $(x, v) \in \mathcal{XV}$, then $\theta_S(x \mid x_0)$ is identified for all $x \in \mathcal{X}$ that satisfy Assumption 3.*

We can define global objects in the selected population that are identified without a common support assumption when X is continuous and $x \mapsto g(x, \cdot)$ is differentiable. An example is the average derivative conditional on $X = x$ in the selected population:

$$\delta_S(x) = \mathbb{E}[\delta(x, V) \mid X = x, C > 0],$$

which is (4.5) with $\theta(x, v) = \delta(x, v)$ and $x_0 = x$. This object is point identified in the selected population under Assumption 1 because the integral is over $\mathcal{V}(x)$, the support of V conditional on $X = x$ in the selected population. Another example is the average derivative in the selected population:

$$\delta_S = \mathbb{E}[\delta(X, V) \mid C > 0],$$

which is point identified under Assumption 1 because the integral is over $\mathcal{XV}$, the support of (X, V) in the selected population. This is a special case of the generic global effect:

$$\theta_S = \int \theta(x, v) dF_{XV|C>0}(x, v). \tag{4.6}$$

4.4 Counterfactual distributions

We consider linear functionals of the global effects including counterfactual distributions constructed by integration of the DSF with respect to different distributions of the explanatory variables and control function. These counterfactual distributions enable wage decompositions and other counterfactual analyses (e.g., DiNardo et al., 1996, Chernozhukov et al., 2013, Firpo et al., 2011, and Arellano and Bonhomme, 2017).

We focus on functionals for the selected population. To simplify notation, we use a superscript s to denote these functionals, instead of explicitly conditioning on $C > 0$. The decompositions are based on the following representation of the observed distribution of Y:

$$G_Y^s(y) = \int F_{Y|Z,V}^s(y \mid z, v) dF_{Z,V}^s(z, v) = \frac{\int G(y, x, v) 1(h(z, v) > 0) dF_{Z,V}(z, v)}{\int 1(h(z, v) > 0) dF_{Z,V}(z, v)}. \tag{4.7}$$

This representation follows from $F_{Y|Z,V}^s(y \mid z, v) = G(y, x, v)$ by Lemma 1, and:

$$dF_{V,Z}^s(v, z) = \frac{\mathbb{P}(C > 0 \mid Z = z, V = v) dF_{V,Z}(v, z)}{\mathbb{P}(C > 0)} = \frac{1(h(z, v) > 0) dF_{V,Z}(v, z)}{\int 1(h(z, v) > 0) dF_{V,Z}(v, z)},$$

by Bayes' rule and monotonicity of $v \mapsto h(z, v)$.

We construct counterfactual distributions by combining the component distributions G and $F_{Z,V}$ with the selection rule h from different populations corresponding to different time periods or demographic groups. Let G_t and F_{Z_k,V_k} denote the distributions in groups t and k, and h_r denote the selection rule in group r. The counterfactual distribution of Y when G is as in group t, $F_{Z,V}$ is as in group k, while the selection rule is identical to group r, is:

$$G_{Y_{\langle t|k,r \rangle}}^s(y) := \frac{\int G_t(y, x, v) 1(h_r(z, v) > 0) dF_{Z_k,V_k}(z, v)}{\int 1(h_r(z, v) > 0) dF_{Z_k,V_k}(z, v)}. \tag{4.8}$$

Note that under this definition the observed distribution in group t is $G_{Y_{\langle t|t,t \rangle}}^s$. The integral in (4.8) is non-parametrically identified if $\mathcal{Z}_k \subseteq \mathcal{Z}_r$ and $(\mathcal{XV}_k \cap \mathcal{XV}_r) \subseteq \mathcal{XV}_t$. These conditions are weaker than $\mathcal{ZV}_k \subseteq \mathcal{ZV}_r$ and $\mathcal{XV}_k \subseteq \mathcal{XV}_t$, which guarantee that h_r and G_t are identified for all combinations of z and v over which we integrate.[6] By monotonicity of $v \mapsto h(z, v)$, the condition $h_r(z, v) > 0$ is equivalent to:

$$v > F_{C_r|Z}(0 \mid z), \tag{4.9}$$

[6] We can weaken the condition $\mathcal{ZV}_k \subseteq \mathcal{ZV}_r$ to $\mathcal{Z}_k \subseteq \mathcal{Z}_r$ because $v \mapsto h_r(z, v)$ is monotone and we only need to know if $h_r(z, v) > 0$. We weaken the condition $\mathcal{XV}_k \subseteq \mathcal{XV}_t$ to $[\mathcal{XV}_k \cap \mathcal{XV}_r] \subseteq \mathcal{XV}_t$ by adopting the convention that G_t is equal to some arbitrary constant whenever it is not identified at $(x, v) \in \mathcal{XV}_k \setminus \mathcal{XV}_r$ because $h_r(z, v) \leq 0$ if $(x, v) \notin \mathcal{XV}_r$ and $G_t(y, x, v) 1(h_r(z, v) > 0) = 0$.

where $F_{C_r|Z}$ is the distribution of C conditional on Z in group r.

We can decompose the difference in the observed distribution between group 1 and group 0 using counterfactual distributions:

$$G^s_{Y_{\langle 1|1,1\rangle}} - G^s_{Y_{\langle 0|0,0\rangle}} = \underbrace{[G^s_{Y_{\langle 1|1,1\rangle}} - G^s_{Y_{\langle 1|1,0\rangle}}]}_{(1)} + \underbrace{[G^s_{Y_{\langle 1|1,0\rangle}} - G^s_{Y_{\langle 1|0,0\rangle}}]}_{(2)} + \underbrace{[G^s_{Y_{\langle 1|0,0\rangle}} - G^s_{Y_{\langle 0|0,0\rangle}}]}_{(3)},$$

where (1) is a selection effect due to the change in the selection rule given the distribution of the explanatory variables and the control function, (2) is a composition effect due to the change in the distribution of the explanatory variables and the control function, and (3) is a structural effect from the change in the conditional distribution of the outcome given the explanatory variables and control function.

4.5 Comparison with the Heckman Selection Model

Our decomposition of the distribution yields effects that differ from those derived for the Heckman selection model (HSM) by Mulligan and Rubinstein (2008). Relative to our selection effect, we show that the selection effect in the HSM excludes a component, includes another we attribute to the composition effect, and contains one that cannot be separately identified from the structural effect in nonseparable models. We illustrate this via a comparison of the decomposition of the observed mean of Y, which in our notation for counterfactual distributions, has the form:

$$\mu^s_{Y_{\langle t|k,r\rangle}} := \frac{\int \mu^t(x,v)1(h_r(z,v)>0)dF_{Z_k,V_k}(z,v)}{\int 1(h_r(z,v)>0)dF_{Z_k,V_k}(z,v)}. \tag{4.10}$$

Suppose that (3.1) and (3.2) follow the HSM:[7]

$$Y_t = \alpha_t' X_t + \epsilon_t \qquad \text{if } C_t > 0,$$
$$C_t = \max\{\gamma_t' Z_t + \eta_t, 0\},$$

where the first element of the vector X_t is the constant term, and ϵ_t and η_t follow a standard bivariate normal distribution with correlation ρ_t independently of Z_t. The mean of Y_t in the selected population (after integration over the distribution of Z_t)

[7]For simplicity we suppress the i subscript and the t subscript denotes period t.

equals:

$$\mu^s_{Y_{\langle t|t,t\rangle}} = \alpha_t' \, \mathbb{E}[X_t \mid C_t > 0] + \rho_t \, \mathbb{E}[\lambda(\gamma_t' Z_t) \mid C_t > 0],$$

where $\lambda(\cdot)$ denotes the inverse Mills ratio. We decompose the difference in $\mu^s_{Y_{\langle t|t,t\rangle}}$ between $t = 0$ and $t = 1$ into selection, composition and structural effects.

Mulligan and Rubinstein (2008) define the selection effect as:

$$\rho_1 \, \mathbb{E}[\lambda(\gamma_1' Z_1) \mid C_1 > 0] - \rho_0 \, \mathbb{E}[\lambda(\gamma_0' Z_0) \mid C_0 > 0].$$

This effect consists of three components:

$$\begin{aligned}\rho_1 \int & [\lambda(\gamma_1' z)\Pi_1(\gamma_1' z) - \lambda(\gamma_0' z)\Pi_1(\gamma_0' z)] dF_{Z_1}(z) \\ & + \rho_1 \left\{ \int \lambda(\gamma_0' z)\Pi_1(\gamma_0' z) dF_{Z_1}(z) - \int \lambda(\gamma_0' z)\Pi_0(\gamma_0' z) dF_{Z_0}(z) \right\} \\ & \qquad + (\rho_1 - \rho_0)' \, \mathbb{E}\left[\lambda\left(\gamma_0' Z_0\right) \middle| \, C_0 > 0\right], \quad (4.11)\end{aligned}$$

where $\Pi_k(\gamma_r' z) = \Phi(\gamma_r' z) / \int \Phi(\gamma_r z) dF_{Z_k}(z)$ is the counterfactual probability of selection in group k when the selection equation is as in group r. The term $\Pi_k(\gamma_r' z) F_{Z_k}(z)$ can therefore be interpreted as the counterfactual distribution of Z in the population k when the selection equation is as in group r. The first two terms capture effects from changes in the composition of the selected population in terms of observable characteristics. The first term results from applying the selection equation from group 0 to group 1 holding the composition of group 1 fixed, whereas the second term results from changing the distribution of characteristics from group 1 to group 0. The third term captures the effect from a change in the composition of the selected population in terms of unobservables through the correlation coefficient.[8] In our view, the first and third terms are rightful components of the selection effect, whereas the second term belongs to the composition effect as it is driven by changes in the distribution of Z.

We now obtain the effects of our decomposition and compare them with the

[8] Mulligan and Rubinstein (2008) assume that the covariates are homogeneous across groups, $F_{Z_0} = F_{Z_1}$, such that the second component drops out.

selection effect of Mulligan and Rubinstein (2008). The LASF in the HSM is:

$$\mu(x, v) = \alpha_t' x + \rho_t \Phi^{-1}(v). \tag{4.12}$$

Plugging this expression into (4.10) and some straightforward calculations given in Appendix B.3 yield:

$$\mu^s_{Y_{\langle t|k,r\rangle}} = \int \left[\alpha_t' x + \rho_t \lambda(\gamma_r' z)\right] \Pi_k(\gamma_r' z) dF_{Z_k}(z). \tag{4.13}$$

Our selection effect is:

$$\begin{aligned}\mu^s_{Y_{\langle 1|1,1\rangle}} - \mu^s_{Y_{\langle 1|1,0\rangle}} = \alpha_1' &\int x[\Pi_1(\gamma_1' z) - \Pi_1(\gamma_0' z)] dF_{Z_1}(z) \\ &+ \rho_1 \int [\lambda(\gamma_1' z)\Pi_1(\gamma_1' z) - \lambda(\gamma_0' z)\Pi_1(\gamma_0' z)] dF_{Z_1}(z).\end{aligned} \tag{4.14}$$

The first term is the effect on the average wage due to changes in the composition of the selected population in terms of observable characteristics resulting from applying the selection equation from group 0 to group 1. It is positive when the selected population contains relatively more individuals with characteristics that are associated with higher average wages. This term is missing in the selection effect in equation (4.11). The second term is the corresponding effect for the unobserved characteristics and corresponds to the first term in (4.11). Our composition effect is:

$$\begin{aligned}\mu^s_{Y_{\langle 1|1,0\rangle}} - \mu^s_{Y_{\langle 1|0,0\rangle}} = \alpha_1' &\left\{ \int x\Pi_1(\gamma_0' z) dF_{Z_1}(z) - \int x\Pi_0(\gamma_0' z) dF_{Z_0}(z) \right\} + \\ &\rho_1 \left\{ \int \lambda(\gamma_0' z)\Pi_1(\gamma_0' z) dF_{Z_1}(z) - \int \lambda(\gamma_0' z)\Pi_0(\gamma_0' z) dF_{Z_0}(z) \right\}.\end{aligned}$$

The first term is the change in the average wage resulting directly from differences in the distribution of the observed characteristics. The second term is the same as the second term in (4.11). Finally, our structural effect is:

$$\mu^s_{Y_{\langle 1|0,0\rangle}} - \mu^s_{Y_{\langle 0|0,0\rangle}} = (\alpha_1 - \alpha_0)' \, \mathbb{E}\left[X_0 | \, C_0 > 0\right] + (\rho_1 - \rho_0) \, \mathbb{E}\left[\lambda\left(\gamma_0' Z_0\right) | \, C_0 > 0\right].$$

The first term reflects the impact of the change in the returns to observed characteristics. The second term captures the type and degree of selection and is the same as the third term in equation (4.11). As the expectation involving the inverse Mills ratio is positive, the contribution of this term is positive whenever $\rho_1 > \rho_0$.

Finally, we show, via an example, that the two terms of the structural effect cannot be identified separately in general when the model is nonseparable. Consider the multiplicative version of the HSM:

$$Y_t = X_t'\alpha_t\epsilon_t, \qquad \text{if } C_t > 0, \tag{4.15}$$

$$C_t = \max\{\gamma_t'Z_t + \eta_t, 0\}, \tag{4.16}$$

where we weaken the parametric assumption on the joint distribution of ϵ_t and η_t to $\eta_t \sim N(0,1)$ and:

$$\mathbb{E}[\epsilon_t \mid Z_t, C_t > 0] = \rho_t\lambda(\gamma_t'Z_t).$$

In this case α_t and ρ_t are not identified separately from the moment condition:

$$\mathbb{E}[Y_t \mid Z_t, C_t > 0] = X_t'\alpha_t\rho_t\lambda(\gamma_t'Z_t),$$

which is the only information of the model about α_t and ρ_t.

While this illustrative discussion focuses on the decomposition of the mean, others have also investigated the decomposition of the distribution. Arellano and Bonhomme (2017) and Maasoumi and Wang (2018) use the Machado and Mata (2005) method to derive selection effects at different points of the distribution. That is, they simulate the distribution of wages that would result if everybody in the population worked based on estimates of the model's parameter obtained via the correction method of Arellano and Bonhomme (2017). They interpret the difference between this distribution and the uncorrected distribution of wages as a selection effect. This differs from our approach which investigates the counterfactual distribution for a more restrictive selection regime.

5 Empirical Results

5.1 Hours Equation

Figure 3 plots the employment rates of males and females, defined as the percentage of the sample reporting positive annual hours. The male employment rate fluctuates cyclically around a downward trend starting at 90.0 percent, reaching a minimum of 82.1 percent in 2012, and increasing to 83.1 percent in 2016. The female rate increases from a low of 56.5 percent in 1976 to a high of 75.3 percent in 2001. It is 70.0 percent in 2016. Figure 3 also plots average annual hours worked for wage earners. For males they vary cyclically around a slightly upward trend. They decrease from 2032 in 1976 to a sample minimum of 1966 hours in 1983. They then increase reaching a maximum of 2160 hours in 2001 before ending at 2103 hours. For females, average annual hours increase by nearly 20 percent, from 1515 to 1792 hours, between 1976 and 2000. The increase continues more slowly after 2000 reaching 1838 hours in 2016. These patterns suggest important movements along both the intensive and extensive margins of labor supply.

We estimate the control function, defined as the conditional distribution function of hours, via a distribution regression of annual hours of work on the conditioning variables for the whole sample separately by gender for each year, as outlined in the Appendix C. This first step includes those reporting zero hours of work. The conditioning variables include age and age squared, dummy variables for the highest educational attainment reported (less than high school, high school, some college, or college or more), a dummy variable for marital status (married or not), a dummy variable for being nonwhite, a set of dummy variables for the region of residence (North-East, South, Central and West), the number of children in the household, the number of household members, a dummy variable for the presence of children aged less than 5 years and a dummy variable for the presence of unrelated individuals in the household. We also include the interaction of the educational dummies with age and age squared. Each of these variables appears in the hourly wage models except those capturing family composition.

As our focus is on the wage equation, we do not discuss the results for the hours equation although we highlight the role of the exclusion restrictions. Given our selection rule the assumption that annual hours do not affect the hourly wage rate means that the variation in hours across individuals is a source of identification. That is, the variation in hours induces movement in the control function for the sample of workers. We also use measures of family household composition as explanators of hours which do not directly affect wages. While one can argue that household composition may affect hourly wage rates, we regard these restrictions as reasonable. Note that similar restrictions have been previously employed (see, for example, Mulligan and Rubinstein, 2008). Given the potentially contentious use of these exclusion restrictions, we explore the impact of not using them below.

5.2 Wage Equation and the Impact of Education

We now focus on the determinants of individual hourly wage rates by estimating wage equations, separately for males and females, for each of the cross sections by distribution regression over the subsample reporting a positive wage. The conditioning variables are those in the hours equation except for the household composition variables. We also include the control function and its square, and the other conditioning variables are all interacted with the control function.

Substantial empirical evidence suggests that the increasing wage dispersion partially reflects changes in the impact of education (see for example, Autor et al., 2008, and Murphy and Topel, 2016). We explore this by estimating the impact of education on wages. We treat an individual's education level as exogenous and the "endogeneity/selection" here reflects that from the hours of work decision.

Education is represented by three dummy variables indicating that an individual has acquired, as her/his highest level of schooling, "high school", "some college" or "college or more". The excluded category is "less than high school". Since 1992, the CPS measures educational attainment by the highest year of school or degree completed rather than the previously employed "highest year of school attendance". Although the educational recode by the IPUM-CPS aims at maximizing comparability over time, there is a discontinuity between 1991 and 1992 in how those with a high school degree and some college are classified. Note that the educational attainment of the workforce increases dramatically over our sample period. Male workers with at most a high-school degree fell from 64.1 percent in 1976 to 38.4 percent in 2016, while those with at least a college degree rose from 20.4 percent in 1976 to 35.2 percent in 2016. The trends for females are even more striking as those with at most a high-school degree fell from 69.1 percent in 1976 to 29.7 percent in 2016, while those with at least a college degree rose from 16.1 percent in 1976 to 40.1 percent in 2016.

We estimate the various treatment effects introduced above. For example, the average treatment effects are estimated as:

$$\frac{1}{n_e}\sum_{i=1,E_i=e}\widehat{\mu}(X_i,e',\widehat{V}_i)-\frac{1}{n_e}\sum_{i=1,E_i=e}\widehat{\mu}(X_i,e_i,\widehat{V}_i),$$

where E_i is education level of individual i, e and e' are the actual and counterfactual education levels respectively, n_e is the number of observations that have an education level equal to e, X_i denotes other observed characteristics, and $\widehat{V}_i$ denotes the estimated control function for individual i. The function $\widehat{\mu}$ is the estimated LASF based on a flexibly specified regression of log wages on (X_i, E_i, V_i). The quantile treatment effects can be obtained via inversion of the estimated LSDF's. As we use log wages these effects can be interpreted as percentage changes. To satisfy the identification restriction we calculate the wage increase that individuals with education, e, would receive if they achieved a higher level of education, e'.

Figures 6 to 8 present the average treatment effects, and their 95 percent confidence level bands, for various contrasts in education levels.[9] The average treatment effects for the "high school" to "less than high school" comparison, shown in Figure 6, reveal that the effect is increasing but reasonably flat. Figure 7 provides the estimated treatment effect for the contrast between "some college" and "less than high school". For males it is 33 percent in 1976 and 45 percent in 2016. It peaks at 50 percent in 2008. Although there are episodic declines, there is a sizable increase over the sample period. For females there is a notable increase during the 1980's although generally the pattern and magnitude of the changes are similar to those of males.

Figure 8 provides the treatment effect for the difference between "college" and "less than high school", and it is generally large. This premium increases from 50

[9]The asymptotic theory and the associated proofs for our estimates are provided in Appendices D and F. Our inference procedures are provided in Appendix E.

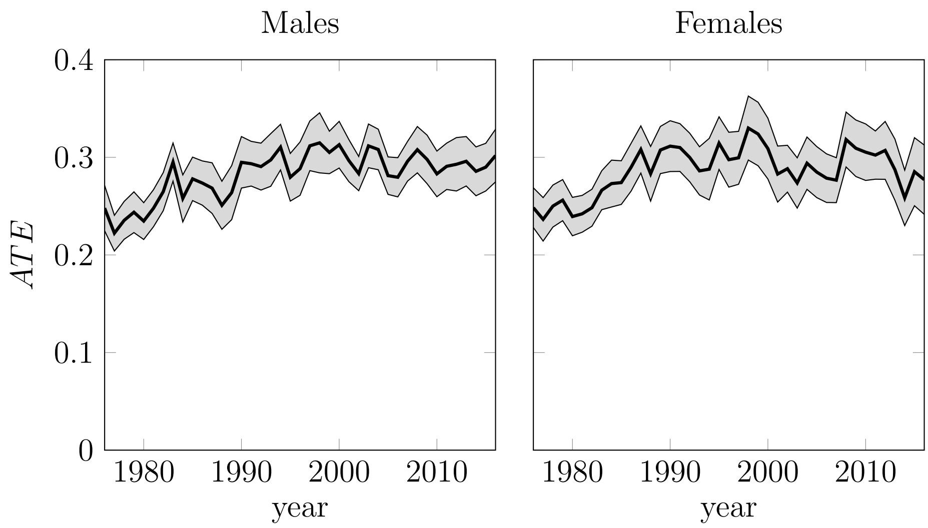

Figure 6: Average treatment effect of education, high school versus less than high school with correction.

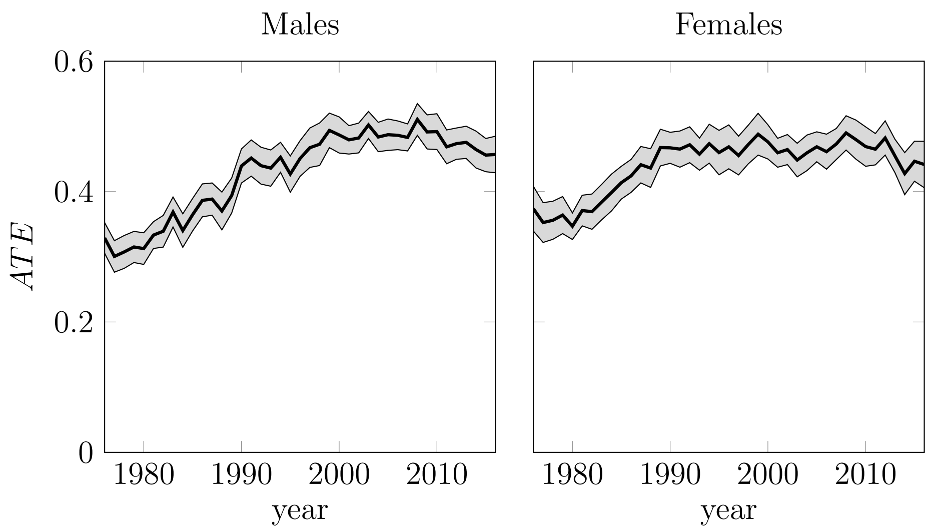

Figure 7: Average treatment effect of education, some college versus less than high school with correction.

to 89 percent for males and from 69 to 88 percent for females. For males there is a steady increase with several large jumps. The occasional decreases appear to either reflect sampling issues or cyclical influences. The evidence for females is similar although there is a large decline in the late 1970's before large increases in both the 1980's and 1990's. The growth in the final decades of the previous millennium is not observed in the new millennium.

The average treatment effects fail to reflect the heterogeneity in the specific

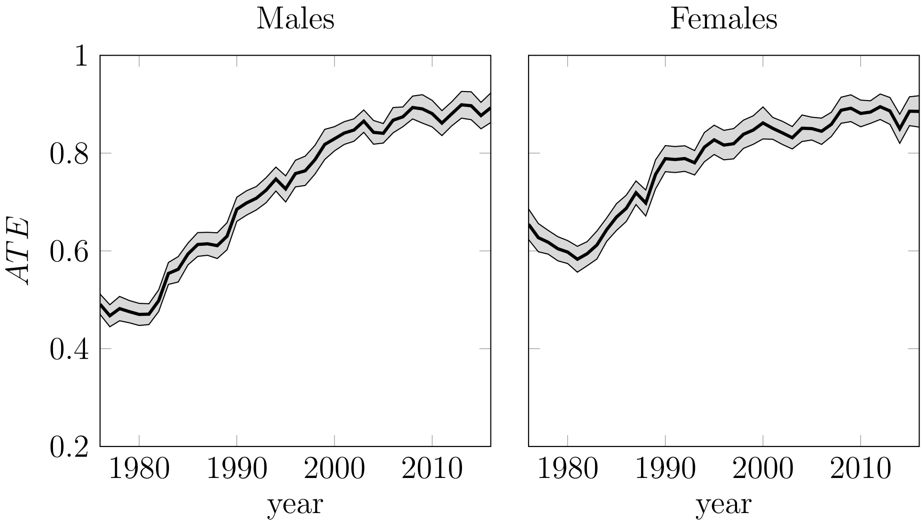

Figure 8: Average treatment effect of education, college versus less than high school with correction.

educational treatments. Accordingly, Figures 9 to 11 provide the quantile treatment effects at the 25th, 50th and 75th percentiles.[10] Figure 9 presents the "high school"/"less than high school" comparison. Several features are notable. First, despite the growth in this premium from the period 1976 to 2000, the profiles are reasonably flat over large parts of the sample period. Second, for males there are some differences across quantiles with those at the 75th percentile showing the strongest evidence of growth. Third, females have relatively similar profiles at the three quantiles we examine. Figure 10 contrasts "some college" to "less than high school" and provides a clearer indication that this premium has increased with particularly strong evidence of large increases in the 1980's and part of the 1990's for females. The variability across the quantiles is small. Figure 11 presents the quantile treatment effects for the college premium. There is a notable increase in the college effect for males at the first quartile until around 2000 before it flattens for the remainder of the sample period. There is a more prolonged increase at the median. This reflects a more heterogeneous effect with the quantile effect at the first quartile starting at around 52 percent before ending at 81 percent, while the median increases to 93 percent in 2013 before ending at 90 percent. The increase is even more evident at the third quartile. Similar patterns appear for females although with some additional

[10]The quantile treatment effects are based on location in the wage distribution. Thus the treatment effects are not necessarily increasing as the quantile at which they are evaluated increases.

features. There is dramatic growth at the median and below in the 1980's and parts of the 1990's and some leveling out over the 2000's. There is a steady increase at the third quartile over the 1980's and 1990's and periods of equally fast growth in some post 2000 periods. For females, there is significant heterogeneity in the effects across quantiles.

The following conclusions can be drawn regarding these estimated education treatment effects. First, the returns to education have increased over our sample period and the premia for the higher education levels have increased remarkably. Second, the treatment effects, particularly for the college premium, display heterogeneity with the quantile treatment effects showing that the treatment effect for college education is very high for both males and females at the upper quantiles. Finally, the general trends and level of the treatment effects are similar across gender although the trends in wages differ by gender. This suggests that the education premia is not the only contributing factor to the different patterns of within gender wage inequality. While we do not attempt to isolate the role of the other conditioning variables, they are included in the decomposition exercise. Note that these results are for the selectivity adjusted estimates. Although they are not reported here, the unadjusted estimates are similar. We also estimated local effects corresponding to different values of the control function. Those estimated effects, available from the authors, did not reveal anything remarkable.

5.3 Decompositions

For the decompositions the base years are set to 1976 for females and 2010 for males. The increasing female participation rate makes the choice of 1976 seem reasonable as it assumes that those individuals with a certain combination of x and v working in 1976 have a positive probability of working in any other year. A sensible choice of base year for males is 2010, the bottom of the financial crisis, as it has the lowest level of participation over our sample period. Nevertheless, this is somewhat harder to defend. The different base years for the genders means that we can compare trends but not wage levels across gender. The decompositions are presented in Figures 12 to 16. They capture the impact of changes in the specified components on the change of the wage distribution. For example, the selection effect captures

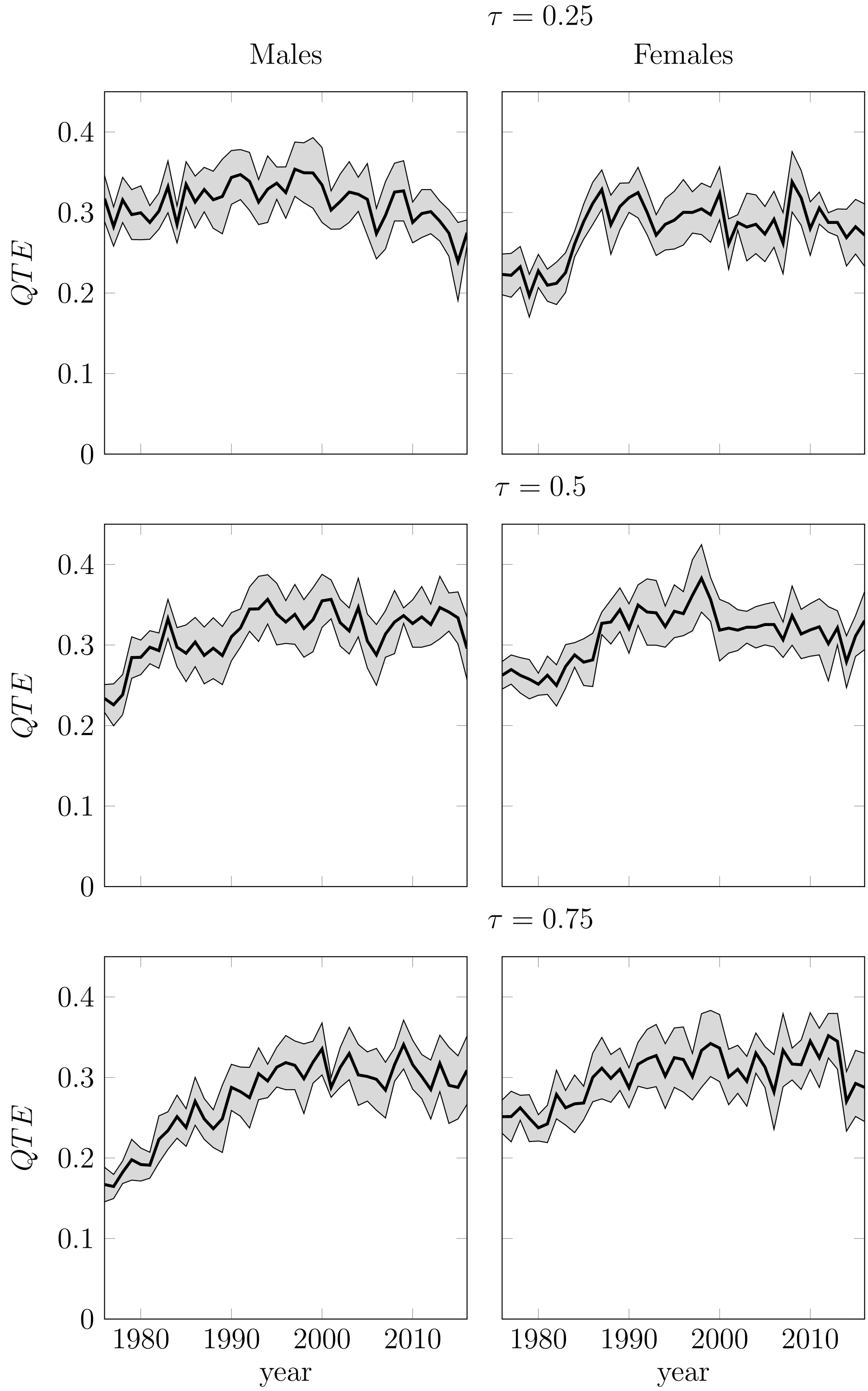

Figure 9: Quantile treatment effect of education, high school versus less than high school with correction.

the change in the wage distribution due to a change in the selection process.

We commence with the median presented in Figure 14. During our sample

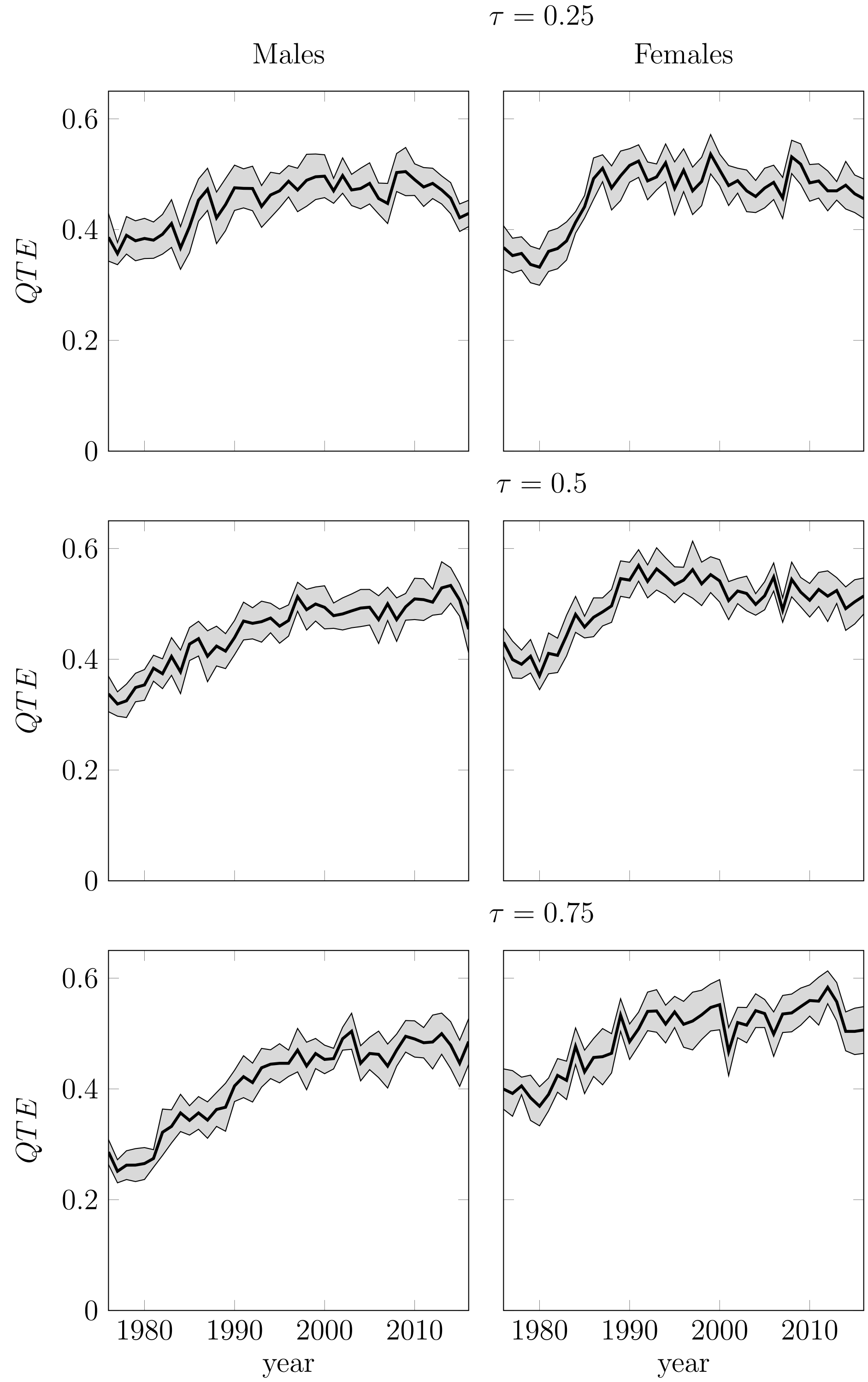

Figure 10: Quantile treatment effect of education, some college versus less than high school with correction.

period the male median wage decreases by 13 percent while that of females increased by 20 percent. The total and structural effects are very similar for males with

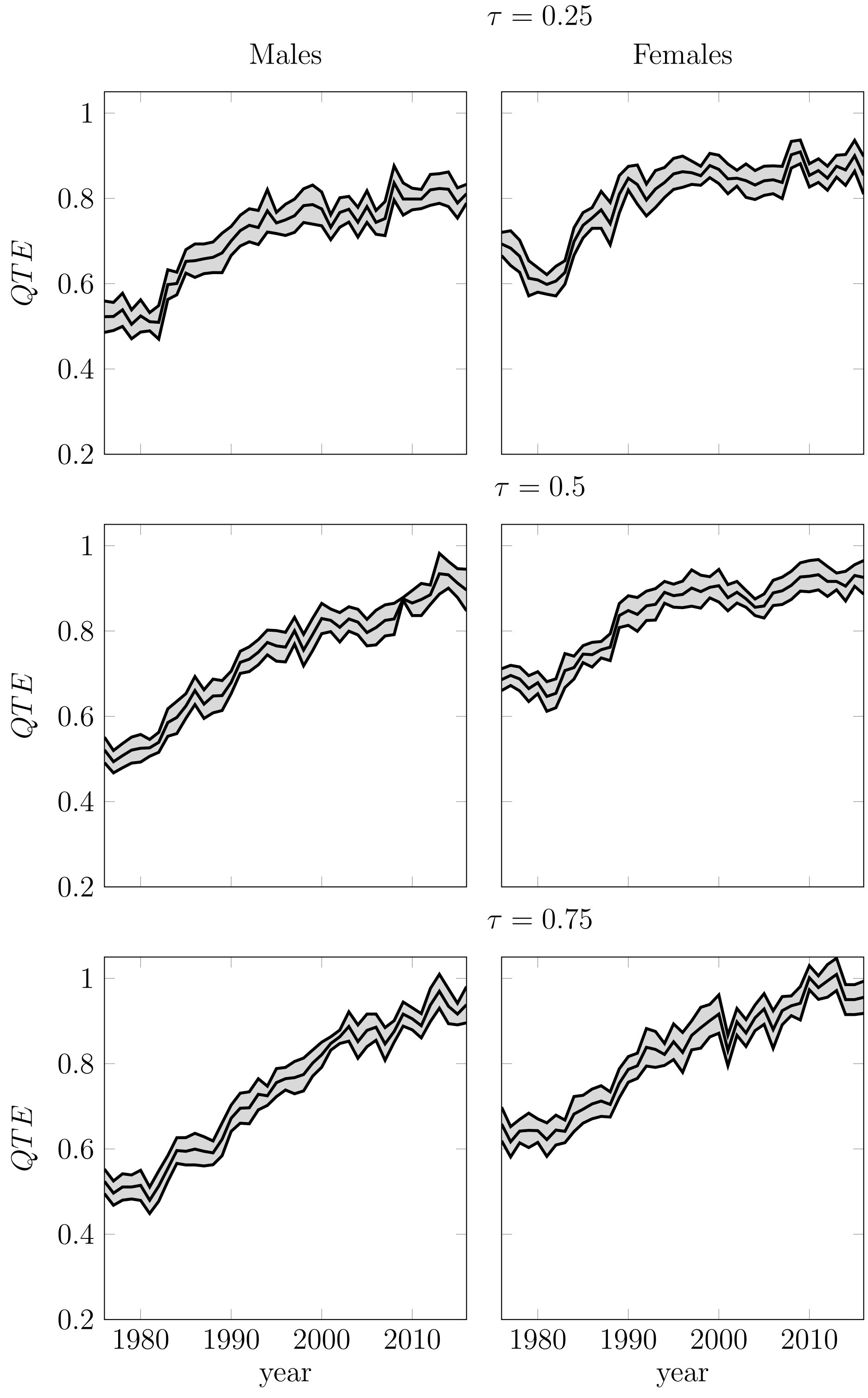

Figure 11: Quantile treatment effect of education, college versus less than high school with correction.

the small difference due to the composition effect. There is no evidence of any change in the selection effects. The composition effect increases the median wage by

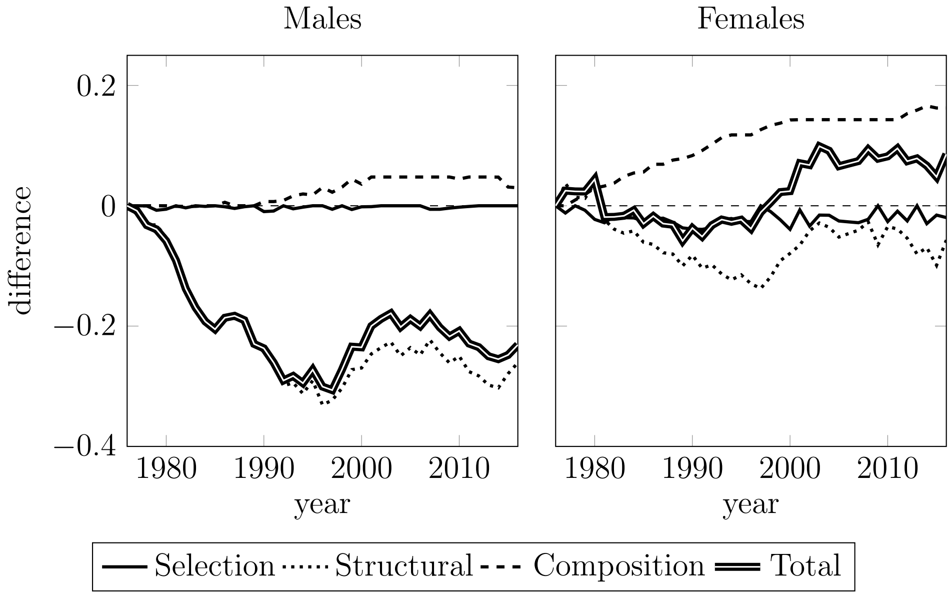

Figure 12: Decompositions at D1.

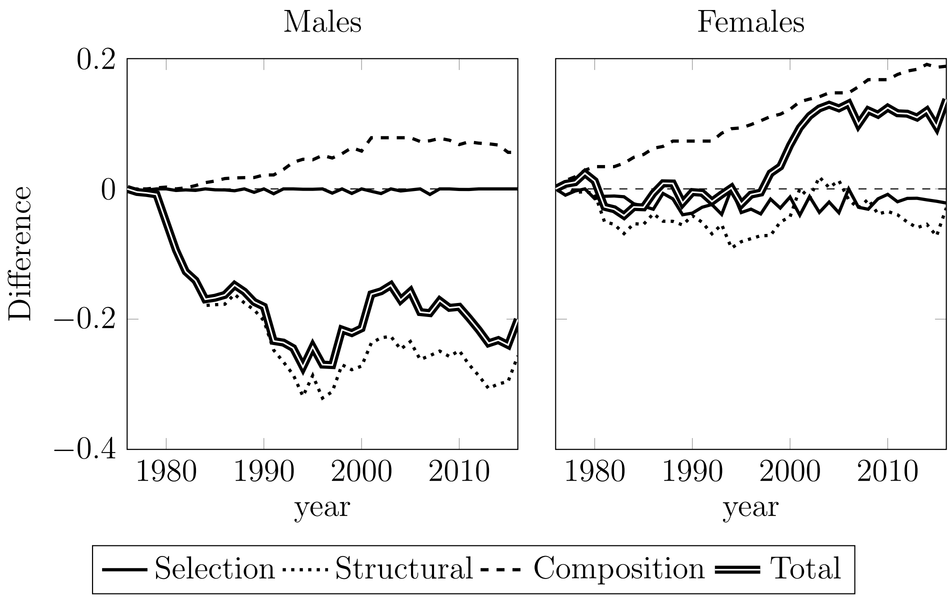

Figure 13: Decompositions at Q1.

around 2 percent and most likely reflects the increasing educational attainment of the workforce. The structural component, which appears to be strongly procylical, produces large negative effects. While there are some upturns, coinciding with periods of improving economic conditions, the structural effect is negative for the entire period. This is a striking result noting that it is consistent with Chernozhukov

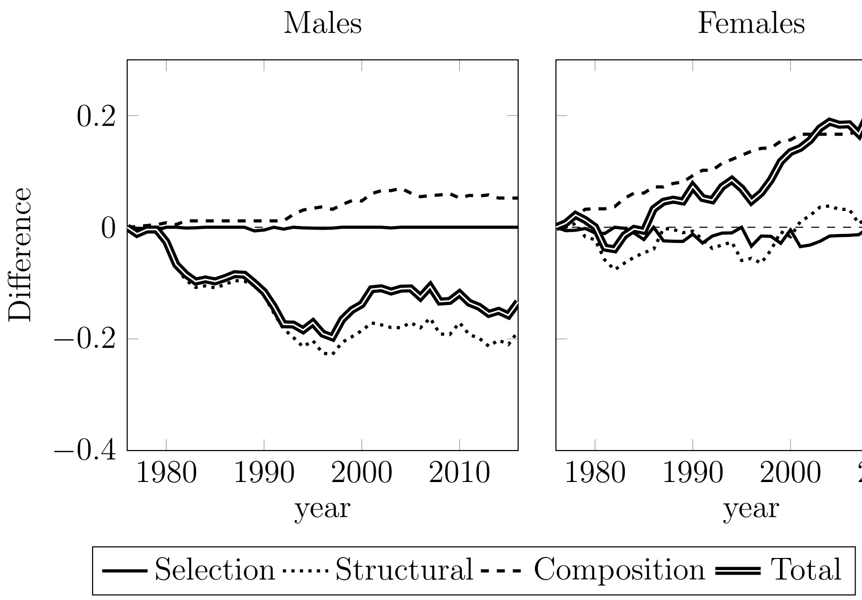

Figure 14: Decompositions at Q2.

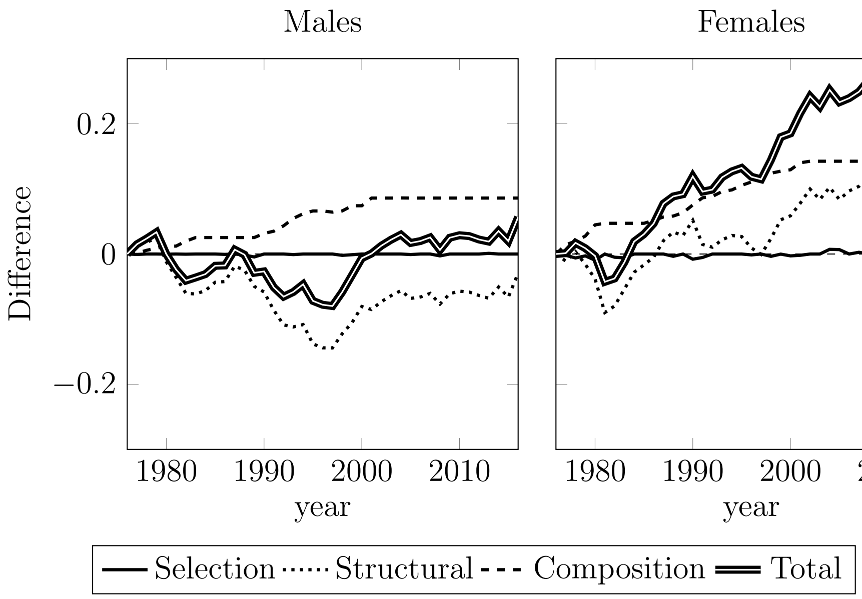

Figure 15: Decompositions at Q3.

et al. (2013) who perform a similar exercise for the period 1979 to 1988 without correcting for sample selection.

Figure 14 reveals that the increase in the female median wage is entirely driven by the composition effect. The structural effect is negative for the whole period, but is less substantial than that for males. The turning points appear to coincide across

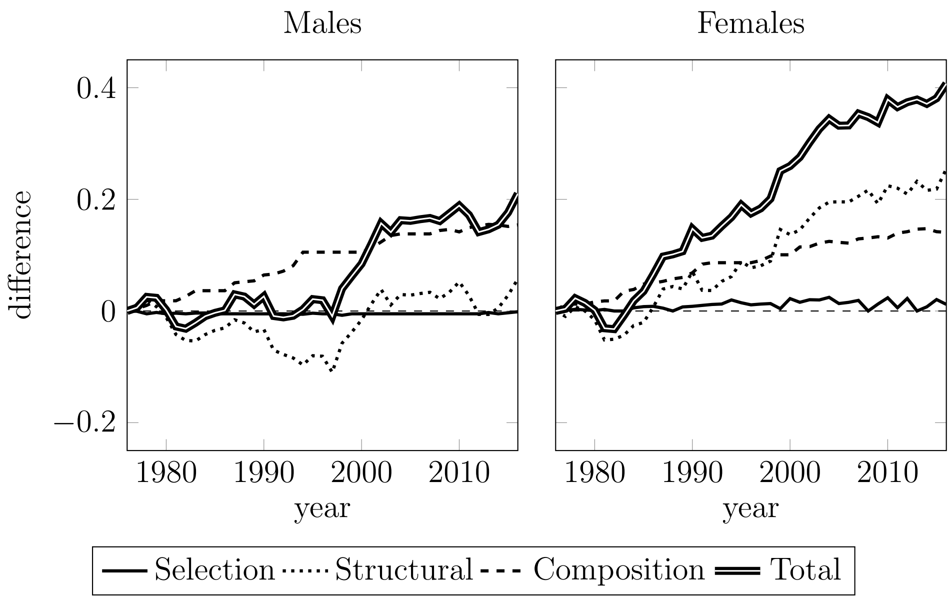

Figure 16: Decompositions at D9.

genders. The contribution of the selection component is negative but negligible. This contrasts with Mulligan and Rubinstein (2008) who find a large change in the selection effect at the mean over time. Specifically, they find that the selection effect changes from negative and large to positive and large. We do not find evidence for such a drastic effect, but as highlighted above, our selection effect differs from that employed by Mulligan and Rubinstein (2008). If their assertion that the correlation between the error terms increased over time is correct, implying that the selected sample of females became increasingly more productive relative to the total population, then our selection effect might underestimate the total change in the wage distribution due to changes in the selected sample over time.

We do not, however, find support for the conclusion of Mulligan and Rubinstein that the least productive females, in terms of unobservables, were working in the 1970's and 1980's. Our selection effect measures the difference between the observed distribution in any given year and the resulting distribution if the "least likely to participate" females among the selected sample would not have participated. If the "least likely females" to participate would have been the most productive, this would reduce the wage in the year of evaluation and reflect a positive change in the selection effect in our figures for the 1970's and the 1980's. We do not find any support for a positive change in the selection effect.

The decompositions for the median suggests that the decline in the median male wage is due to the prices associated with the male skill characteristics. However, the evidence above established that the returns to schooling have generally increased and the male labor force has become increasingly more educated. This suggests that the wages of those individuals with the lowest levels of education must have markedly decreased. A similar pattern is observed for the female median wage although the less substantial negative structural impact is offset by the composition effects.

Figures 12 and 13 report the decompositions at the 10th and 25th percentiles and they are remarkably similar for males. They suggest the reductions in the male wage rate capture the prices associated with the human capital of individuals located at these lower quantiles. There is evidence of a negative structural component of around 25 to 30 percent at each of these quantiles in both the late 1990's and the late 2010's. While these effects are somewhat offset by the composition effects, the overall effect on wages is negative. The wage reduction appears to be driven by how worker's characteristics in this part of the wage distribution are valued. There are no signs of selection effects for males. Our results at the first decile for the period 1979 to 1988 appear to differ to those in Chernozhukov et al. (2013). However, that study breaks the structural effect into separate components due to changes in the mandatory minimum wage, unionization, and the returns to any other characteristic. Not surprisingly, changes in the mandatory minimum wage have a large impact on the 10th percentile. Our estimate of the structural effect combines these various components and also includes the variation in the prices of unobservables. Our results are consistent with their study at any other quantile.

The evidence at the 10th and 25th percentiles for females is very different to that for males. First, there are greater differences between the 10th and 25th percentiles. The negative structural effects for females are more evident at the 10th percentile. The structural effects are small at the 25th percentile and offset by the composition effects. For both the 10th and 25th percentiles the overall wage changes become positive in the late 1990's and generally increase over the remainder of the sample.

Figures 15 and 16 present the decompositions at the 75th and 90th percentiles. The male wage at the third quartile shows a small increase. The structural component displays a similar pattern to that for females at the lower quantiles discussed above. That is, initially there is a large decrease before rebounding and remaining relatively flat from the early 2000's. Unlike the lower quantiles the negative changes resulting from the structural component are not sufficiently large to dominate the positive composition effects so the overall wage growth at the 75th percentile is positive from the early 2000's onwards.

The changes in the female hourly wage rate at the 75th percentile highlights that the larger movements have occurred at the upper quantiles of the wage distribution. The changes in the structural component are initially negative before turning positive around the mid 1980's. From the beginning of the 1980's the positively trending structural component combines with the composition effect to produce a steadily increasing wage. However, while the decomposition at the 75th pecentile suggests that the structural components are an important contribution to inequality at higher quantiles of the distribution, the results at the 90th percentile are even more supportive of this perspective. For males the structural component is less negative than at lower quantiles and this combined with the positive composition effect produces a wage gain for the whole period. For females the structural component is positive from the middle of the 1980's and has a larger positive effect than the composition effect. The two effects combine to produce a remarkable 41 percent growth in the wage.

Consider now the selection effects recalling they capture changes in the selection rule while assuming that the same unobservables have an impact on the hours of work decision as those in the year being evaluated. Thus, there can only be a change in the selection effect if the selection rule, and hence the employment rate, has changed over time. Figures 14 to 16 suggest that selection effects cannot explain the observed changes in the males' wage distribution. At the 50th, 75th and the 90th percentiles the selection effect is essentially zero. This result is not surprising as males in this area of the wage distribution have a strong commitment to the labor force. Moreover, there is likely to be relatively little movement on either the extensive or intensive margin given males' level of commitment to full-time employment. One might suspect that it would be more likely to uncover changes in the selection effects at the lower parts of the wage distribution as these individuals are likely to have a weaker commitment to full-time employment. However, the evidence does not support this.

Now focus on the selection effects for females. There is little evidence of changes in selection at higher quantiles. Similar to males, females located in this part of the wage distribution are likely to have had a relatively strong commitment to employment in 1976 and thus there were no substantial moves in their hours distribution. However, at the 10th and 25th percentiles the selection effects seem economically important.[11] At the 10th percentile in 2016 the selection effect contribution is 2.2 percent, while the total wage change is between 8 and 9 percent. Thus the female wage was lowered by 2 percent due to the increased participation of females. This is consistent with the "positive selection" finding of Mulligan and Rubinstein (2008) for the 1990's. We also find a similar relationship for the late 1970's and 1980's. Our results generally suggest a positive relationship between the control variable V and wages at the bottom of the distribution. This implies that those with the highest number of working hours, after conditioning on their observed characteristics, had the highest wages. The trends implied by our results suggest that selection becomes more important as we move further down the female wage distribution. This is similar to the findings of Arellano and Bonhomme (2017) who study the evolution of female wages in the British labor market for the period 1978 to 2000.

Before we further discuss our results we focus on two important issues. Namely, the validity of our exclusion restrictions and the composition of our wage sample. We noted above that the use of household composition variables as exclusion restrictions can be seen as controversial in this setting for both males and females. Accordingly we reproduced the decompositions, excluding first the household variables from both the hours and wage equations and then including them in both equations. For each of these two new specifications the model is now identified by the variation in the number of hours worked. Although we do not present the results here, neither model produces any remarkable changes with respect to the presence or magnitude of selection effects. The only notable difference with the specification used above is the presence of occasionally larger negative selection effects at the bottom decile for females for the specification which excludes the family composition variables from both equations.

Now focus on the issues related to our wage sample composition. As we construct our control function using the hours of annual work variable we are imposing, for

[11] The confidence intervals for these selection effects are presented in Figure 19 in Appendix A. They indicate that for many time periods they are statistically significantly different from zero.

example, that an individual who works 2 hours a week for 50 weeks of the year is the same, in terms of unobservables, as one who works 50 hours a week for 2 weeks of the year. Given the interpretation of the control function this may be unreasonable. We check for the sensitivity of our results to this assumption by censoring the data at various points of the hours distribution. Our empirical work above employs wages for all those working positive hours so we reproduce the decomposition exercises after censoring the data at the 5th, 10th and 25th points of the hours distribution. That is, we exclude the individuals working lower numbers of hours from our wage sample while retaining them in the estimation of the hours equation. We also repeat the decomposition exercise using the full time full year selection rule employed by Mulligan and Rubinstein (2008). The increased censoring has the impact of making the wage sample more homogenous. Note, however, that as this exercise is changing the sample size of the wage sample it is not reasonable to compare results at the same quantiles across samples. However, with the exception of the structural effects for females at the third quartile the results for the structural, composition and total effects at different quantiles are very similar for the different censoring rules and the FTFY samples. The only notable difference is the absence of the selection effects except for the larger samples. That is, once we censor the bottom of the hours distribution the selection effects even disappear for the lower quantiles for females. This is consistent with our earlier discussion that the selection effects, as measured by our definition, captures the impact of the inclusion of those with a weaker commitment to market employment. We highlight that we find no evidence of a change in selection effects using the FTFY sample corresponding to the Mulligan and Rubinstein (2008) selection rule. However, as we discussed in detail above, this reflects the different definition of selection effects and the inability to disentangle components of their selection effect from the structural effect in our setting.

6 Discussion

A number of our empirical results are notable. The impact of the educational treatment varies drastically by level of attainment. While the return to completing high school and obtaining some college, relative to not completing high school, clearly increases wages, they do not appear to be the major driving forces of increasing inequality. In contrast the college premium for both genders has increased dramatically over the sample period and has important implications for inequality. This is consistent with several other studies dating back to Murphy and Welch (1992).

The decompositions reveal a number of findings. The mechanisms driving wages differ by location in the wage distribution and across gender. For males, the fall in wages at the median and below appears partially due to the penalty associated with lack of education and other forces which are negatively affecting the lower skilled. This is reflected by the large negative structural effects for this area of the wage distribution. As more workers are receiving higher education the composition effects are positive and somewhat offset the negative structural effects. However, the large, and increasing educational premium, signals that the "penalty" to not being educated has increased. The negative structural effects for males are not restricted to the lower part of the wage distribution. For females, there are wage increases at each quantile we examined and these reflect, in part, positive and increasing composition effects. Moreover, while the structural effects are generally negative for the whole period at the 10th and 25th percentiles, they are typically positive at the quantiles we examined above the median. Most notably, the structural effects for females are not dampening wage growth to the same extent as for males and at some quantiles these structural effects are even substantial contributors to wage growth. At the lower parts of the female wage distribution the impact of selection is negative and can be substantial. Selection effects become more important as we move down the wage distribution. The patterns at the 10th and 25th percentiles are suggestive of possible larger selection effects at lower percentiles.

Our investigation does not provide direct insight into the macro factors generating these wages profiles. Nor does it provide evidence on the role of institutional factors which disproportionately influence certain sectors of the work force. However it does seem that the mechanisms affecting the wages at the bottom are very different than those influencing the top. The evidence at the top is supportive of an increasing skill premium. At the bottom it appears that the prevailing considerations are those associated with the lack of protection of lower wage workers. These include the decreases in the real value of the minimum wage, the reduction in unionization and the union premium, and increases in employer bargaining power.

We now directly examine the issue of inequality. The 90/10th percentile ratios for

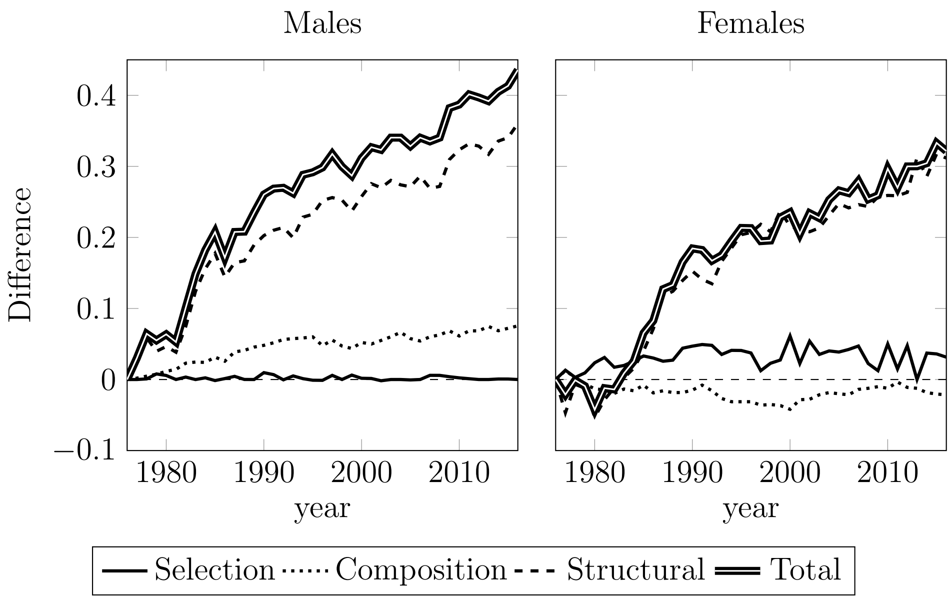

Figure 17: Decompositions of D9-D1 ratio.

hourly wages for males and females have increased by 55 and 38 percent respectively. Our evidence illustrates that hourly wage growth at different locations in the wage distribution is affected differently by the relevant factors. However, we are unable to directly infer from that evidence the respective contribution of these factors to the changes in inequality. For males recall that wage changes at the 10th percentile were due to the large negative structural effects. There was no evidence of selection effects and a small positive composition effect offsets the decreases from around the early 1990's. The large gains at the 90th percentile reflected a steadily increasing composition effect and a structural effect which contributed both negatively and positively over the sample period. There are no signs of changes in selection effects. Figure 17 provides a decomposition of the changes in the 90/10 ratio for males. Unsurprisingly, the large increase in the 90/10 ratio for males is almost entirely due to structural effects. These reflect the negative structural effects at the first decile and not the positive effects at the ninth decile. The composition effects increase inequality via their large positive contribution at the ninth decile.

The evidence is more difficult to interpret for females. There is an increasingly positive composition effect at the first decile but the negative structural effect produces a decline in wages. At the ninth decile, there is a steadily increasing composition effect and a structural effect which is generally increasing wages. The large

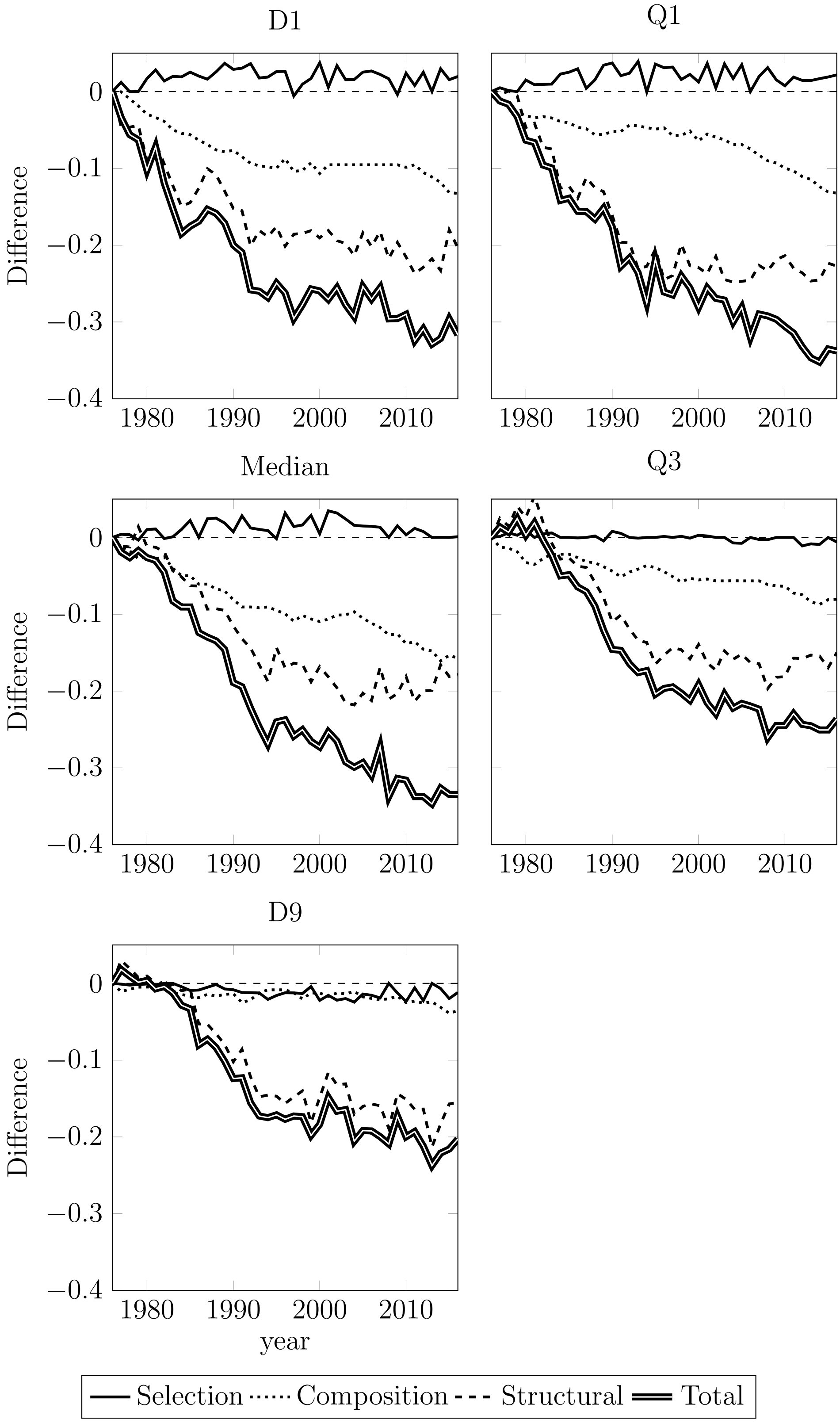

Figure 18: Decompositions of the changes of the wages of males divided by the wages of females.

increase in the total effect for the majority of the sample period reflects the sum of these two positive effects. At the first decile, the evidence suggests that the changes in the selection effects are negatively affecting wage growth, while there is no sign of selection at the ninth decile. Figure 17 presents the decomposition of the change in the 90/10 ratio for females. Given the various issues related to wage growth, it is not surprising that the change in the 90/10 ratio appears to reflect almost entirely a structural effect. The composition effect has slightly decreased this measure of inequality due to the large composition effects at the lower part of the wage distribution. The change in the selection effects has increased the 90/10 ratio. Moreover, its contribution in some periods is a relatively large fraction of the total change.

Although our primary focus is not gender inequality, we examine the trends in the male/female hourly wage ratio at different points of the wage distribution. These are reported in Figure 18. First, at all locations of the wage distribution females appear to be catching up. Moreover, the largest gains appear at the median and below. This result should be treated with caution as it appears largely due to the reduction in the male wage and not large increases in the female wage. This is confirmed by a re-examination of Figure 1. Second, the improvement in the relative performance of females is almost entirely due to structural effects at all of the quantiles reported in Figure 18. As the earlier evidence suggested the sign and the size of the structural effects on the individual gender specific wages varied by sample period and location in the wage distribution, it is surprising that the impact on the gender wage ratios is so clear. The evidence suggests a relative improvement in the value of female labor at all points of the wage distribution. Third, the composition effects also have steadily increased the relative performance of females at all quantiles noting that the size of the effect diminishes as we move up the wage distribution. The effect is small at the ninth decile. Finally, the selection effects are increasing gender inequality although they only appear at the lower parts of the wage distribution.

7 Conclusions

This paper documents the changes in female and male wages over the period 1976 to 2016. We decompose these changes into structural, composition and selection components by proposing and implementing an estimation procedure for nonseparable models with selection. We find that male real wages at the median and below have decreased over our sample period despite an increasing skill premium and an increase in educational attainment. The reduction is primarily due to large decreases of the wages of the individuals with a low level of education. Wages at the upper quantiles of the distribution have increased drastically due to a large and increasing skill premium. Combined with the decreases at the lower quantiles, this increase in the upper quantiles has substantially increased wage inequality. Female wage growth at lower quantiles is modest although the median wage has grown steadily. The increases at the upper quantiles for females are substantial and reflect increasing skill premia. These changes have resulted in a substantial increase in female wage inequality. As our sample period is associated with large changes in the participation rates and the hours of work of females we explore the role of changes in "selection" in wage movements. We find that the impact of these changes in selection is to decrease the wage growth of those at the lower quantiles with very little evidence of selection effects at other locations in the female wage distribution. The selection effects appear to increase wage inequality.

References

[1] Acemoglu D., and Autor D. (2011), "Skills, tasks and technology: Implications for employment and earnings" In D. Card and O. Ashenfelter (eds.), *Handbook of Labor Economics*, Vol. 4B: 1043–1171. Amsterdam: Elsevier Science, North-Holland.

[2] Amemiya, T. (1978), "The estimation of a simultaneous equation generalized probit model", *Econometrica*, **46**, 1193–1205.

[3] Amemiya, T. (1979), "The estimation of a simultaneous tobit model", *International Economic Review*, **20**, 169–81.

[4] Angrist J., Chernozhukov V., and Fernández-Val I. (2006), "Quantile regression under misspecification, with an application to the U.S. wage structure", *Econometrica*, **74**, 539–63.

[5] Arellano M., and Bonhomme S. (2017), "Quantile selection models with an application to understanding changes in wage inequality", *Econometrica*, **85**, 1–28.

[6] Autor D.H., Katz L.F., and Kearney M.S. (2008), "Trends in U.S. wage inequality: Revising the revisionists", *Review of Economics and Statistics*, **90**, 300–23.

[7] Chen S., (1997) "Semiparametric estimation of the Type-3 Tobit model", *Journal of Econometrics*, **80**, 1–34.

[8] Chernozhukov, V., Fernández-Val I. and Kowalski A. (2015), "Quantile regression with censoring and endogeneity", *Journal of Econometrics*, **186**, 201–21.

[9] Chernozhukov, V., Fernández-Val I., and Melly B. (2013), "Inference on counterfactual distributions", *Econometrica*, **81**, 2205–68.

[10] Chernozhukov, V., Fernández-Val I., Newey, W., Stouli S., and Vella F. (2017), "Semiparametric estimation of structural functions: methods and inference", working paper, University of Bristol.

[11] Chesher, A. (2003), "Identification in nonseparable models", *Econometrica*, **71**, 1401–44.

[12] DiNardo J., Fortin N.M., and Lemieux T. (1996), "Labor market institutions and the distribution of wages , 1973–1992: A semiparametric approach", *Econometrica*, **64**, 1001–44.

[13] Fernández-Val I., Van Vuuren A., and F. Vella (2017), "Nonseparable sample selection models with censored selection rules", working paper.

[14] Flood S., King M., Ruggles S., and Warren J. R. (2015), "Integrated public use microdata series, Current Population Survey: Version 4.0 [Machine-readable database]", University of Minnesota, mimeo.

[15] Florens, J., J.J. Heckman, C. Meghir and E. Vytlacil (2008), "Identification of treatment effects using control functions in models with continuous,

endogenous treatment and heterogeneous effects", *Econometrica*, **76**, 1191–1206.

[16] Foresi S., and Peracchi F. (1995), "The conditional distribution of excess returns: An empirical analysis", *Journal of the American Statistical Association*, **90**, 451–66.

[17] Fortin N.M., Lemieux T., and Firpo S. (2011), "Decomposition methods in economics." In D. Card and O. Ashenfelter (eds.), *Handbook of Labor Economics*, Vol. 4A: 1–102. Amsterdam: Elsevier Science, North-Holland.

[18] Hahn, J. (1995), "Bootstrapping quantile regression estimators", *Econometric Theory*, **11**, 105–21.

[19] Heckman J.J. (1974), "Shadow prices, market wages and labor supply", *Econometrica*, **42**, 679–94.

[20] Heckman J.J. (1979), "Sample selection bias as a specification error", *Econometrica*, **47**, 153–61.

[21] Heckman, J.J. and E. Vytlacil (2005), "Structural equations, treatment effects, and econometric policy evaluation", *Econometrica*, **73**, 669–738.

[22] Hoderlein, S. and E. Mammen (2007), "Identification of marginal effects in nonseparable models without monotonicity", *Econometrica*, **75**, 1513–18.

[23] Honoré B., Kyriazidou E., and Udry C. (1997), "Estimation of type 3 tobit models using symmetric trimming and pairwise comparison", *Journal of Econometrics*, **76**, 107–28.

[24] Imbens G. W., and Newey W.K. (2009), "Identification and estimation of triangular simultaneous equations models without additivity", *Econometrica*, **77**, 1481–1512.

[25] Juhn C., Murphy K.M., and Pierce B. (1993), "Wage inequality and the rise in returns to skill", *Journal of Political Economy*, **101**, 410–42.

[26] Jun, S.J. (2009), "Local structural quantile effects in a model with a nonseparable control variable", *Journal of Econometrics*, **151**, 82–97.

[27] Katz L. F., and Autor D. H. (1999), “Changes in the wage structure and earnings inequality.” In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Vol. 3A: 1463–1555. Amsterdam: Elsevier Science, North-Holland.

[28] Katz L.F., and Murphy K.M. (1992), “Changes in relative wages, 1963–87: Supply and demand factors”, *Quarterly Journal of Economics*, **107**, 35–78.

[29] Koenker R., and Bassett G. (1978), “Regression quantiles”, *Econometrica*, **46**, 33–50.

[30] Krueger A., and Posner E. (2018), “A proposal for protecting low-income workers from monopsony and corruption.” In J. Shambaugh and R. Nunn (eds.), *Revitalizing Wage Growth. Policies to Get American Workers a Raise*, 139–156. Washington, DC: Brookings.

[31] Lee D.S. (1999), “Wage inequality in the United States during the 1980s: Rising dispersion or falling minimum wage?”, *Quarterly Journal of Economics*, **114**, 977–1023.

[32] Lee M.J., and Vella F. (2006), ”A semiparametric estimator for censored selection models with endogeneity”, *Journal of Econometrics*, **130**, 235–52.

[33] Lemieux T. (2006), “Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?” *American Economic Review*, **96**, 461–98.

[34] Ma S. and M. Kosorok (2005), “Robust semiparametric M-estimation and the weighted bootstrap”, *Journal of Multivariate Analysis*, **96**, 190-217.

[35] Maasoumi E., and Wang, L. (2018) ”The Gender Gap in the Earnings Distribution,” *Journal of Political Economy*, forthcoming.

[36] Macahdo J., and Mata J. (2005), ”Counterfactual decomposition of changes in wage distributions using quantile regression”, *Journal of Applied Econometrics*, **20**, 445–465.

[37] Masten, M. and Torgovitsky A. (2014), “Instrumental variables estimation of a generalized correlated random coefficients model”, working paper, Duke University, Durham.

[38] Matzkin, R. (2003), "Nonparametric estimation of nonadditive random functions", *Econometrica*, **71**, 1339–75.

[39] Meyer B.D., Mok W.K.C., and Sullivan J.X. (2015), "Household surveys in crisis", *Journal of Economic Perspectives*, **29**, 199–226.

[40] Mulligan C., and Rubinstein Y. (2008), "Selection, investment, and women's relative wages over time", *Quarterly Journal of Economics*, **123**, 1061–1110.

[41] Murphy K.M., and Topel R.H. (2016), "Human capital investment, inequality, and economic growth", *Journal of Labor Economics*, **34**, S99–S127.

[42] Murphy K.M., and Welch F. (1992), "The structure of wages", *Quarterly Journal of Economics*, **57**, 285–326.

[43] Newey, W.K. (2007), "Nonparametric continuous/discrete choice models", *International Economic Review*, **48**, 1429–39.

[44] Praestgaard, J. and Wellner J. (1993), "Exchangeably weighted bootstraps of the general empirical process", *Annals of Probability*, **21**, 2053–86.

[45] Vella, F. (1993), "A simple estimator for models with censored endogenous regressors", *International Economic Review*, **34**, 441–57.

[46] Welch F. (2000), "Growth in women's relative wages and inequality among men: One phenomenon or two?", *American Economic Review Papers & Proceedings*, **90**, 444–49.

A Figures

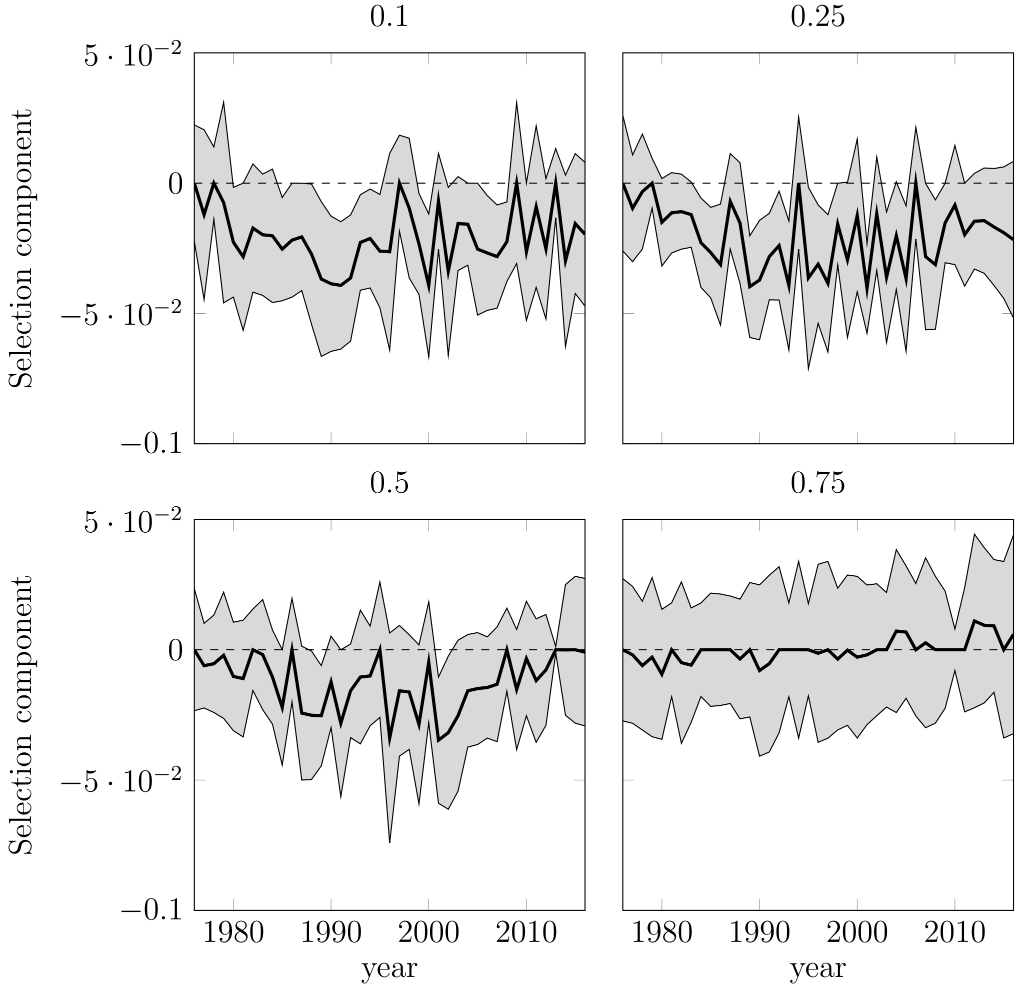

Figure 19: Selection component and 95% confidence intervals for females.

B Proofs of Section 4

B.1 Lemma 1

Proof. The proof is similar to the proof of Theorem 1 in Newey (2007). For any bounded function $a(\varepsilon)$ and $C > 0$ (and hence $h(Z, \eta) > 0$), by Assumption 1,

$$\mathbb{E}\left[a(\varepsilon) \mid Z = z, \eta = q, C > 0\right] = \mathbb{E}\left[a(\varepsilon) \mid \eta = q, C > 0\right]$$

Since this holds for any function $a(\varepsilon)$ and any $h(Z, \eta) > 0$, Z and ε are independent conditional on η and $C > 0$. The result follows because η is a one-to-one function of

$F_{C|Z}(C \mid Z)$ when $C > 0$ since $\eta = h^{-1}(Z, C)$ if $C > 0$ by Assumption 1, and for $c > 0$

$$\begin{aligned} F_{C|Z}(c \mid Z = z) &= \mathbb{P}(\max(h(Z,\eta),0) \leq c \mid Z = z) \\ &= \mathbb{P}(h(Z,\eta) \leq c \mid Z = z) = \mathbb{P}(\eta \leq h^{-1}(Z,c) \mid Z) = h^{-1}(Z,c), \end{aligned}$$

where we use the normalization $\eta \sim U(0,1)$. ■

B.2 Theorem 1

Proof. Define the generic local object

$$\theta(x,v) = \mathbb{E}_\varepsilon[\Gamma(x,\varepsilon) \mid V = v]$$

for some function $\Gamma(x,e) : \mathcal{X} \times \mathcal{E} \to \mathbb{R}^k; k \in \mathbb{N}^+$, where $\mathcal{E}$ is the support of ε. Using Assumption 1, this equals

$$\theta(x,v) = \mathbb{E}_\varepsilon[\Gamma(x,\varepsilon) \mid Z = z, V = v].$$

Since conditional on $Z = z$ and $V = v$, we have that $C = \max\{h(z,v), 0\}$ and since $(x,v) \in \mathcal{XV}$, there is a $z \in Z$ such that $C = h(z,v) > 0$ and hence

$$\begin{aligned} \theta(x,v) &= \mathbb{E}_\varepsilon[\Gamma(x,\varepsilon) \mid Z = z, V = v, h(z,v) > 0] \\ &= \mathbb{E}_\varepsilon[\Gamma(x,\varepsilon) \mid Z = z, V = v, C > 0] \\ &= \mathbb{E}_\varepsilon[\Gamma(X,\varepsilon) \mid Z = z, V = v, C > 0], \end{aligned}$$

where the third line is due to $X \subseteq Z$. This third line is identical to the right-hand side of (4.1) when $\Gamma(x,e) = g(x,e)$. Along the same lines, this also proves Corollary 1, with $\Gamma(x,e) = \partial_x g(x,e)$. Note that this also proves identification of the LDSF, since

$$G(y,x,v) = \mathbb{E}_\varepsilon[\mathbf{1}\{g(x,\varepsilon) \leq y\} \mid V = v]$$

and hence the proof is completed by using $\Gamma(x,e) = \mathbf{1}\{g(x,e) \leq y\}$. ■

B.3 Proof of equation (4.13)

Note that $h_r(z,v)$ implies in this case that $v \geq \Phi(-\gamma_r' z)$. Hence, we can rewrite (4.10) after substitution of (4.12) as

$$\mu^s_{Y_{\langle t|k,r\rangle}} := \frac{\int \int_{\Phi(-\gamma_r' z)}^1 \left[\alpha_t' x + \rho \Phi^{-1}(v)\right] dv dF_{Z_k}(z)}{\int \int_{\Phi(-\gamma_r' z)}^1 dv dF_{Z_k}(z)}.$$

The result follows from the change of variables $w = \Phi^{-1}(v)$ and calculation of the integrals.

C Estimation

The effects of interest are all identified by functionals of the distribution of the observed variables and the control function in the selected population. The control function is the distribution of the censoring variable C conditional on all the explanatory variables Z. We propose a multistep semiparametric method based on least squares, distribution and quantile regressions to estimate the effects. The reduced form specifications used in each step can be motivated by parametric restrictions on the model (3.1)–(3.2). We refer to Chernozhukov et al. (2017) for examples of such restrictions.

Throughout this section, we assume that we have a random sample of size n, $\{(Y_i * 1(C_i > 0), C_i, Z_i)\}_{i=1}^n$, of the random variables $(Y * 1(C > 0), C, Z)$, where $Y * 1(C > 0)$ indicates that Y is observed only when $C > 0$.

C.1 Estimation of the control function

We estimate the control function using logistic distribution regression (Foresi and Peracchi, 1995, and Chernozhukov et al., 2013). More precisely, for every observation in the selected sample, we set:

$$\widehat{V}_i = \Lambda(R_i^{\mathrm{T}} \widehat{\pi}(C_i)), \quad R_i := r(Z_i), \quad i = 1, \dots, n, C_i > 0,$$

where, for $c \in \mathcal{C}_n$, the empirical support of C,

$$\widehat{\pi}(c) = \arg\max_{\pi \in \mathbb{R}^{d_r}} \sum_{i=1}^{n} \left[1\{C_i \leq c\} \log \Lambda(R_i^{\mathrm{T}} \pi)) + 1\{C_i > c\} \log \Lambda(-R_i^{\mathrm{T}} \pi) \right],$$

Λ is the logistic distribution, and $r(z)$ is a d_r-dimensional vector of transformations of z with good approximating properties such as polynomials, B-splines and interactions.

C.2 Estimation of local objects

We can estimate the local average, distribution and quantile structural functions using flexibly parametrized least squares, distribution and quantile regressions, where we replace the control function by its estimator from the previous step.

For reasons explained in Section D, our estimation method is based on a trimmed sample with respect to the censoring variable C. Therefore, we introduce the following trimming indicator among the selected sample

$$T = 1(C \in \overline{\mathcal{C}})$$

where $\overline{\mathcal{C}} = (0, \bar{c}]$ for some $0 < \bar{c} < \infty$, such that $P(T = 1) > 0$.

The estimator of the LASF is $\widehat{\mu}(x, v) = w(x, v)^{\mathrm{T}} \widehat{\beta}$, where $w(x, v)$ is a d_w-dimensional vector of transformations of (x, v) with good approximating properties, and $\widehat{\beta}$ is the ordinary least squares estimator:[12]

$$\widehat{\beta} = \left[\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i^{\mathrm{T}} T_i \right]^{-1} \sum_{i=1}^{n} \widehat{W}_i Y_i T_i, \quad \widehat{W}_i := w(X_i, \widehat{V}_i).$$

[12]An alternative approach is to follow Jun (2009) and Masten and Torgovitsky (2014). These papers acknowledge that with an index restriction the parameters of interest can be estimated in the presence of a control function by estimation over subsamples for which the control function has a *similar* value. While each of these papers considers a random coefficients model with endogeneity their approach is applicable here.

The estimator of the LDSF is $\widehat{G}(y, x, v) = \Lambda(w(x, v)^{\mathrm{T}}\widehat{\beta}(y))$, where $\widehat{\beta}(y)$ is the logistic distribution regression estimator:

$$\widehat{\beta}(y) = \arg\max_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^{n} \left[1\{Y_i \leq y\} \log \Lambda(\widehat{W}_i^{\mathrm{T}} b)) + 1\{Y_i > y\} \log \Lambda(-\widehat{W}_i^{\mathrm{T}} b))\right] T_i.$$

Similarly, the estimator of the LQSF is $\widehat{q}(\tau, x, v) = w(x, v)^{\mathrm{T}}\widehat{\beta}(\tau)$, where $\widehat{\beta}(\tau)$ is the Koenker and Bassett (1978) quantile regression estimator

$$\widehat{\beta}(\tau) = \arg\min_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^{n} \rho_\tau(Y_i - \widehat{W}_i^{\mathrm{T}} b) T_i.$$

Estimators of the local derivatives are obtained by taking derivatives of the estimators of the local structural functions. Thus, the estimator of the LADF is:

$$\widehat{\delta}(x, v) = \partial_x w(x, v)^{\mathrm{T}} \widehat{\beta},$$

and the estimator of the LQDF is:

$$\widehat{\delta}_\tau(x, v) = \partial_x w(x, v)^{\mathrm{T}} \widehat{\beta}(\tau).$$

C.3 Step 3: Estimation of global effects

We obtain estimators of the generic global effects by approximating the integrals over the control function by averages of the estimated local effects evaluated at the estimated control function. The estimator of the effect (4.3) is

$$\widehat{\theta}_S(x) = \sum_{i=1}^{n} T_i \widehat{\theta}(x, \widehat{V}_i) / \sum_{i=1}^{n} T_i.$$

This yields the estimators of the ASF for $\widehat{\theta}(x, v) = \widehat{\mu}(x, v)$ and DSF at y for $\widehat{\theta}(x, v) = \widehat{G}(y, x, v)$. The estimator of the QSF is then obtained by inversion of the estimator of the DSF.[13] We form an estimator of the effect (4.5) as

$$\widehat{\theta}_S(x \mid x_0) = \sum_{i=1}^{n} T_i K_i(x_0) \widehat{\theta}(x, \widehat{V}_i) / \sum_{i=1}^{n} T_i K_i(x_0),$$

for $K_i(x_0) = 1(X_i = x_0)$ when X is discrete or $K_i(x_0) = k_h(X_i - x_0)$ when X is continuous, where $k_h(u) = k(u/h)/h$, k is a kernel, and h is a bandwidth such as $h \to 0$ as $n \to 0$. Finally, the estimator of the effect (4.6) is

$$\widehat{\theta}_S = \sum_{i=1}^{n} T_i \widehat{\theta}(X_i, \widehat{V}_i) / \sum_{i=1}^{n} T_i.$$

[13] We can use the generalized inverse

$$\widehat{q}_S(\tau, x) = \int_0^\infty 1(\widehat{G}_S(y, x) \leq \tau) dy - \int_{-\infty}^0 1(\widehat{G}_S(y, x) > \tau) dy,$$

which does not require that the estimator of the DSF $y \mapsto \widehat{G}_S(y, x)$ be monotone.

C.4 Step 4: Estimation of counterfactual distributions

Based on equations (4.8) and (4.9), the estimator (or sample analog) of the counterfactual distribution is:

$$\widehat{G}^s_{Y_{\langle t|k,r\rangle}}(y) = \sum_{i=1}^{n} \Lambda(\widehat{W}_i^{\mathrm{T}}\widehat{\beta}_t(y))1[\widehat{V}_i > \Lambda(R_i^{\mathrm{T}}\widehat{\pi}_r(0))]/n^s_{kr},$$

where the average is taken over the sample values of $\widehat{V}_i$ and Z_i in group k, $n^s_{kr} = \sum_{i=1}^n 1[\widehat{V}_i > \Lambda(R_i^{\mathrm{T}}\widehat{\pi}_r(0))]$, $\widehat{\beta}_t(y)$ is the distribution regression estimator of step 2 in group t, and $\widehat{\pi}_r(0)$ is the distribution regression estimator of step 1 in group r. Here we are estimating the components $F^s_{Y_t}$ by logistic distribution regression in group t and the component $F^s_{Z_k}$ by the empirical distribution in group k.

D Asymptotic theory

We derive large sample theory for some of the local and global effects. We focus on average effects for the sake of brevity. The theory for distribution and quantile effects can be derived using similar arguments, see, for example, Chernozhukov et al. (2015) and Chernozhukov et al. (2017). Throughout the analysis we treat the dimensions of the flexible specifications used in all the steps as fixed, so that the model parameters are estimable at a $\sqrt{n}$ rate. The model remains semiparametric because some of the parameters are function-valued such as the parameters of the control variable.[14]

In what follows, we shall use the following notation. We let the random vector $A = (Y * 1(C > 0), C, Z, V)$ live on some probability space $(\Omega_0, \mathcal{F}_0, P)$. Thus, the probability measure P determines the law of A or any of its elements. We also let $A_1, ..., A_n$, i.i.d. copies of A, live on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which contains the infinite product of $(\Omega_0, \mathcal{F}_0, P)$. Moreover, this probability space can be suitably enriched to carry also the random weights that appear in the weighted bootstrap. The distinction between the two laws P and $\mathbb{P}$ is helpful to simplify the notation in the proofs and in the analysis. Calligraphic letters such as $\mathcal{Y}$ and $\mathcal{X}$ denote the supports of $Y * 1(C > 0)$ and X; and $\mathcal{YX}$ denotes the joint support of (Y, X). Unless explicitly mentioned, all functions appearing in the statements are assumed to be measurable.

We now state formally the assumptions. The first assumption is about sampling and the bootstrap weights.

Condition 1 (Sampling and Bootstrap Weights) *(a) Sampling: the data* $\{Y_i * 1(C_i > 0), C_i, Z_i\}_{i=1}^n$ *are a sample of size* n *of independent and identically distributed observations from the random vector* $(Y * 1(C > 0), C, Z)$*. (b) Bootstrap weights:* $(\omega_1, ..., \omega_n)$ *are i.i.d. draws from a random variable* $\omega \geq 0$*, with* $\mathbb{E}_P[\omega] = 1$*,* $\mathrm{Var}_P[\omega] = 1$*, and* $\mathbb{E}_P|\omega|^{2+\delta} < \infty$ *for some* $\delta > 0$*; live on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*; and are independent of the data* $\{Y_i * 1(C_i > 0), C_i, Z_i\}_{i=1}^n$ *for all* n.

The second assumption is about the first stage where we estimate the control function

$$\vartheta_0(c, z) := F_C(c \mid z).$$

[14] Chernozhukov at al. (2017) discuss the trade-offs between imposing parametric restrictions in the model and the support conditions required for nonparametric identification of the effects of interest.

We assume a logistic distribution regression model for the conditional distribution of C in the trimmed support, $\overline{\mathcal{C}}$, that excludes censored and extreme values of C. The purpose of the upper trimming is to avoid the upper tail in the modeling and estimation of the control variable, and to make the eigenvalue assumption in Condition 2(b) more plausible. We consider a fixed trimming rule, which greatly simplifies the derivation of the asymptotic properties. Throughout this section, we use bars to denote trimmed supports with respect to C, e.g., $\overline{\mathcal{CZ}} = \{(c,z) \in \mathcal{CZ} : c \in \overline{\mathcal{C}}\}$, and $\overline{\mathcal{V}} = \{\vartheta_0(c,z) : (c,z) \in \overline{\mathcal{CZ}}\}$.

Condition 2 (First Stage) *(a) Trimming: we consider the trimming rule as defined by the indicator* $T = \mathbf{1}(C \in \overline{\mathcal{C}})$*. (b) Model: the distribution of* C *conditional on* Z *follows the distribution regression model in the trimmed support* $\overline{\mathcal{C}}$, i.e.,

$$F_C(c \mid Z) = F_C(c \mid R) = \Lambda(R^{\mathrm{T}}\pi_0(c)), \quad R = r(Z),$$

for all $c \in \overline{\mathcal{C}}$*, where* Λ *is the logit link function; the coefficients* $c \mapsto \pi_0(c)$ *are three times continuously differentiable with uniformly bounded derivatives;* $\overline{\mathcal{R}}$ *is compact; and the minimum eigenvalue of* $\mathbb{E}_P\left[\Lambda(R^{\mathrm{T}}\pi_0(c))[1-\Lambda(R^{\mathrm{T}}\pi_0(c))]RR^{\mathrm{T}}\right]$ *is bounded away from zero uniformly over* $c \in \overline{\mathcal{C}}$.

For $c \in \overline{\mathcal{C}}$, let

$$\widehat{\pi}^b(c) \in \arg\min_{\pi \in \mathbb{R}^{\dim(R)}} \sum_{i=1}^n \omega_i \{1(C_i \le c)\log\Lambda(R_i^{\mathrm{T}}\pi) + 1(C_i > c)\log\Lambda(-R_i^{\mathrm{T}}\pi)\},$$

where either $\omega_i = 1$ for the unweighted sample, to obtain the estimator; or ω_i are the bootstrap weights to obtain bootstrap draws of the estimator. Then set

$$\vartheta_0(c,r) = \Lambda(r^{\mathrm{T}}\pi_0(c)); \; \widehat{\vartheta}^b(c,r) = \Lambda(r^{\mathrm{T}}\widehat{\pi}^b(c)),$$

if $(c,r) \in \overline{\mathcal{CR}}$, and $\vartheta_0(c,r) = \widehat{\vartheta}^b(c,r) = 0$ otherwise.

Theorem 4 of Chernozhukov et al. (2015) established the asymptotic properties of the DR estimator of the control function. We repeat the result here as a lemma for completeness and to introduce notation that will be used in the results below. Let $\|f\|_{T,\infty} := \sup_{a\in\mathcal{A}} |T(c)f(a)|$ for any function $f : \mathcal{A} \mapsto \mathbb{R}$, and $\lambda = \Lambda(1-\Lambda)$, the density of the logistic distribution.

Lemma 2 (First Stage) *Suppose that Conditions 1 and 2 hold. Then, (1)*

$$\begin{aligned}
\sqrt{n}(\widehat{\vartheta}^b(c,r) - \vartheta_0(c,r)) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n e_i \ell(A_i,c,r) + o_{\mathbb{P}}(1) \rightsquigarrow \Delta^b(c,r) \text{ in } \ell^\infty(\overline{\mathcal{CR}}),\\
\ell(A,c,r) &:= \lambda(r^{\mathrm{T}}\pi_0(c))[1\{C \le c\} - \Lambda(R^{\mathrm{T}}\pi_0(c))] \times\\
&\quad \times r^{\mathrm{T}}\mathbb{E}_P\left\{\Lambda(R^{\mathrm{T}}\pi_0(c))[1-\Lambda(R^{\mathrm{T}}\pi_0(c))]RR^{\mathrm{T}}\right\}^{-1} R,\\
\mathbb{E}_P[\ell(A,c,r)] &= 0, \mathbb{E}_P[T\ell(A,C,R)^2] < \infty,
\end{aligned}$$

where $(c,r) \mapsto \Delta^b(c,r)$ *is a Gaussian process with uniformly continuous sample paths and covariance function given by* $\mathbb{E}_P[\ell(A,c,r)\ell(A,\tilde{c},\tilde{r})^{\mathrm{T}}]$*. (2) There exists* $\widetilde{\vartheta}^b : \overline{\mathcal{CR}} \mapsto [0,1]$ *that obeys the same first order representation uniformly over* $\overline{\mathcal{CR}}$*, is close to* $\widehat{\vartheta}^b$ *in the sense that* $\|\widetilde{\vartheta}^b - \widehat{\vartheta}^b\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$ *and, with probability approaching one, belongs to a* *bounded function class Υ such that the covering entropy satisfies:*[15]

$$\log N(\epsilon, \Upsilon, \|\cdot\|_{T,\infty}) \lesssim \epsilon^{-1/2}, \quad 0 < \epsilon < 1.$$

The next assumptions are about the second stage. We assume a flexible linear model for the conditional distribution of Y given (X, V) in the trimmed support $C \in \overline{\mathcal{C}}$, impose compactness conditions, and provide sufficient conditions for identification of the parameters. Compactness is imposed over the trimmed support and can be relaxed at the cost of more complicated and cumbersome proofs.

Condition 3 (Second Stage) *(a) Model: the expectation of Y conditional on (X, V) in the trimmed support $C \in \overline{\mathcal{C}}$ is*

$$\mathbb{E}(Y \mid X, V, C \in \overline{\mathcal{C}}) = W^{\mathrm{T}}\beta_0, \quad V = F_{C|Z}(C \mid Z), \quad W = w(X, V).$$

(b) Compactness and moments: the set $\overline{\mathcal{W}}$ is compact; the derivative vector $\partial_v w(x, v)$ exists and its components are uniformly continuous in $v \in \overline{\mathcal{V}}$, uniformly in $x \in \overline{\mathcal{X}}$, and are bounded in absolute value by a constant, uniformly in $(x, v) \in \overline{\mathcal{X}\mathcal{V}}$; $\mathbb{E}(Y^2 \mid C \in \overline{\mathcal{C}}) < \infty$; and $\beta_0 \in \mathcal{B}$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_w}$. (c) Identification and nondegeneracy: the matrix $J := \mathbb{E}_P[WW^{\mathrm{T}}\, T]$ is of full rank; and the matrix $\Omega := \mathrm{Var}_P[f_1(A) + f_2(A)]$ is finite and is of full rank, where

$$f_1(A) := \{W^{\mathrm{T}}\beta_0 - Y\}WT,$$

and, for $\dot{W} = \partial_v w(X, v)|_{v=V}$,

$$f_2(A) := \mathbb{E}_P[\{[W^{\mathrm{T}}\beta_0 - Y]\dot{W} + W^{\mathrm{T}}\beta_0 W\}T\ell(a, C, Z)]\big|_{a=A}.$$

Let

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^{\dim(W)}} \sum_{i=1}^{n} T_i(Y_i - \beta^{\mathrm{T}}\widehat{W}_i)^2, \quad \widehat{W}_i = w(X_i, \widehat{V}_i), \quad \widehat{V}_i = \widehat{\vartheta}(C_i, R_i),$$

where $\widehat{\vartheta}$ is the estimator of the control function in the unweighted sample; and

$$\widehat{\beta}^b = \arg\min_{\beta \in \mathbb{R}^{\dim(W)}} \sum_{i=1}^{n} \omega_i T_i(Y_i - \beta^{\mathrm{T}}\widehat{W}_i^b)^2, \quad \widehat{W}_i^b = w(X_i, \widehat{V}_i^b), \quad \widehat{V}_i^b = \widehat{\vartheta}^b(C_i, R_i),$$

where $\widehat{\vartheta}^b$ is the estimator of the control function in the weighted sample. The following lemma establishes a central limit theorem and a central limit theorem for the bootstrap for the estimator of the coefficients in the second stage.

Let $\rightsquigarrow_{\mathbb{P}}$ denote bootstrap consistency, *i.e.* weak convergence conditional on the data in probability as defined in Appendix F.1.

Lemma 3 (CLT and Bootstrap FCLT for $\widehat{\beta}$) *Under Conditions 1–3, in $\mathbb{R}^{d_w}$,*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow J^{-1}G, \quad \textit{and} \quad \sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) \rightsquigarrow_{\mathbb{P}} J^{-1}G,$$

where $G \sim N(0, \Omega)$ and J and Ω are defined in Assumption 3(c).

[15] See Appendix F for a definition of the covering entropy.

The properties of the estimator of the LASF, $\widehat{\mu}(x,v) = w(x,v)^{\mathrm{T}}\widehat{\beta}$, and its bootstrap version, $\widehat{\mu}^b(x,v) = w(x,v)^{\mathrm{T}}\widehat{\beta}^b$, are a corollary of Lemma 3.

Corollary 2 (FCLT and Bootstrap FCLT for LASF) *Under Assumptions 1–3, in* $\ell(\overline{\mathcal{XV}})$,

$$\sqrt{n}(\widehat{\mu}(x,v) - \mu(x,v)) \rightsquigarrow Z(x,v) \text{ and } \sqrt{n}(\widehat{\mu}^b(x,v) - \widehat{\mu}(x,v)) \rightsquigarrow_{\mathbb{P}} Z(x,v),$$

where $(x,v) \mapsto Z(x,v) := w(x,v)^{\mathrm{T}}J^{-1}G$ *is a zero-mean Gaussian process with covariance function*

$$\mathrm{Cov}_P[Z(x_0,v_0), Z(x_1,v_1)] = w(x_0,v_0)^{\mathrm{T}}J^{-1}\Omega J^{-1}w(x_1,v_1).$$

To obtain the properties of the estimator of the ASFs, we define $W_x := w(x,V)$, $\widehat{W}_x := w(x,\widehat{V})$, and $\widehat{W}_x^b := w(x,\widehat{V}^b)$. The estimator and its bootstrap draw of the ASF in the trimmed support, $\mu_S(x) = \mathbb{E}_P\{\beta_0^{\mathrm{T}}W_x \mid T=1\}$, are $\widehat{\mu}_S(x) = \sum_{i=1}^n \widehat{\beta}^{\mathrm{T}}\widehat{W}_{xi}T_i/n_T$, and $\widehat{\mu}_S^b(x) = \sum_{i=1}^n e_i\widehat{\beta}^{bT}\widehat{W}_{xi}^bT_i/n_T^b$, where $n_T = \sum_{i=1}^n T_i$ and $n_T^b = \sum_{i=1}^n e_iT_i$. The estimator and its bootstrap draw of the ASF on the treated in the trimmed support, $\mu_S(x \mid x_0) = \mathbb{E}_P\{\beta_0^{\mathrm{T}}W_x \mid T=1, X=x_0\}$, are $\widehat{\mu}_S(x \mid x_0) = \sum_{i=1}^n \widehat{\beta}^{\mathrm{T}}\widehat{W}_{xi}K_i(x_0)T_i/n_T(x_0)$, and $\widehat{\mu}_S^b(x) = \sum_{i=1}^n e_i\widehat{\beta}^{bT}\widehat{W}_{xi}^b\, K_i(x_0)T_i/n_T^b(x_0)$, where $n_T(x_0) = \sum_{i=1}^n K_i(x_0)T_i$ and $n_T^b(x_0) = \sum_{i=1}^n e_iK_i(x_0)T_i$. Let $p_T := P(T=1)$ and $p_T(x) := P(T=1, X=x)$. The next result gives large sample theory for these estimators. The theory for the ASF on the treated is derived for X discrete, which is the relevant case in our empirical application.

Theorem 4 (FCLT and Bootstrap FCLT for ASF) *Under Assumptions 1–3, in* $\ell(\overline{\mathcal{X}})$,

$$\sqrt{np_T}(\widehat{\mu}_S(x) - \mu_S(x)) \rightsquigarrow Z(x) \text{ and } \sqrt{np_T}(\widehat{\mu}_S^b(x) - \widehat{\mu}_S(x)) \rightsquigarrow_{\mathbb{P}} Z(x),$$

where $x \mapsto Z(x)$ *is a zero-mean Gaussian process with covariance function*

$$\mathrm{Cov}_P[Z(x_0), Z(x_1)] = \mathrm{Cov}_P[W_{x_0}^{\mathrm{T}}\beta_0 + \sigma_{x_0}(A), W_{x_1}^{\mathrm{T}}\beta_0(v) + \sigma_{x_1}(A) \mid T=1],$$

with

$$\sigma_x(A) = \mathbb{E}_P\{W_x^{\mathrm{T}}T\}J^{-1}[f_1(A) + f_2(A)] + \mathbb{E}_P\{\dot{W}_x^{\mathrm{T}}\beta_0T\ell(a,X,R)\}\big|_{a=A}.$$

Also, if $p_T(x_0) > 0$, *in* $\ell(\overline{\mathcal{X}})$,

$$\sqrt{np_T(x_0)}(\widehat{\mu}_S(x \mid x_0) - \mu_S(x \mid x_0)) \rightsquigarrow Z(x \mid x_0) \text{ and }$$
$$\sqrt{np_T(x_0)}(\widehat{\mu}_S^b(x \mid x_0) - \widehat{\mu}_S(x \mid x_0)) \rightsquigarrow_{\mathbb{P}} Z(x \mid x_0),$$

where $x \mapsto Z(x \mid x_0)$ *is a zero-mean Gaussian process with covariance function*

$$\mathrm{Cov}_P[Z(x \mid x_0), Z(\tilde{x} \mid x_0)] = \mathrm{Cov}_P[W_x^{\mathrm{T}}\beta_0 + \psi_x(A), W_{\tilde{x}}^{\mathrm{T}}\beta_0 + \psi_{\tilde{x}}(A) \mid T=1, X=x_0],$$

Theorem 4 can be used to construct confidence bands for the ASFs, $x \mapsto \mu_S(x)$ and $x \mapsto \mu_S(x \mid x_0)$, over regions of values of x via Kolmogorov-Smirnov type statistics and weighted bootstrap, and to construct confidence intervals for average treatment effects, $\mu(x_1) - \mu(x_0)$ and $\mu(x_1 \mid x_0) - \mu(x_0 \mid x_0)$, via t-statistics and weighted bootstrap.

E Inference

We use weighted bootstrap to make inference on all the objects of interest (Praestgaard and Wellner, 1993; Hahn, 1995). This method obtains the bootstrap version of the estimator of interest by repeating all the estimation steps including random draws from a distribution as sampling weights. The weights should be positive and come from a distribution with unit mean and variance such as the standard exponential. Weighted bootstrap has some theoretical and practical advantages over empirical bootstrap. Thus, it is appealing that the consistency can be proven following the strategy set forth by Ma and Kosorok (2005), and the smoothness induced by the weights helps dealing with discrete covariates with small cell sizes. The implementation of the bootstrap for the local and global effects is summarized in the following algorithm:

Algorithm 5 (Weighted Bootstrap) *For $b = 1, \ldots, B$, repeat the following steps: (1) Draw a set of weights $(\omega_1^b, \ldots, \omega_n^b)$ i.i.d. from a distribution that satisfies Condition 1(b) such as the standard exponential distribution. (2) Obtain the bootstrap draws of the control function, $\widehat{V}_i^b = \Lambda(R_i^{\mathrm{T}} \widehat{\pi}^b(C_i))$, $i = 1, \ldots, n$, where for $c \in \mathcal{C}_n$,*

$$\widehat{\pi}^b(c) = \arg \max_{\pi \in \mathbb{R}^{d_r}} \sum_{i=1}^{n} \omega_i^b \left[1\{C_i \leq c\} \log \Lambda(R_i^{\mathrm{T}} \pi)) + 1\{C_i > c\} \log \Lambda(-R_i^{\mathrm{T}} \pi) \right].$$

(3) Obtain the bootstrap draw of the local effect, $\widehat{\theta}^b(x, v)$. For the LASF, $\widehat{\theta}^b(x, v) = \widehat{\mu}^b(x, v) = w(x, v)^{\mathrm{T}} \widehat{\beta}^b$, where

$$\widehat{\beta}^b = \left[\sum_{i=1}^{n} \omega_i^b \widehat{W}_i^b (\widehat{W}_i^b)^{\mathrm{T}} T_i \right]^{-1} \sum_{i=1}^{n} \omega_i^b \widehat{W}_i^b Y_i T_i, \quad \widehat{W}_i^b := w(X_i, \widehat{V}_i^b).$$

For the LDSF, $\widehat{\theta}^b(x, v) = \widehat{G}^b(y, x, v) = \Lambda(w(x, v)^{\mathrm{T}} \widehat{\beta}^b(y))$, where

$$\widehat{\beta}^b(y) = \arg \max_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^{n} \omega_i^b \left[1\{Y_i \leq y\} \log \Lambda(b^{\mathrm{T}} \widehat{W}_i^b) + 1\{Y_i > y\} \log \Lambda(-b^{\mathrm{T}} \widehat{W}_i^b) \right] T_i.$$

For the LQSF, $\widehat{\theta}^b(x, v) = \widehat{q}^b(\tau, x, v) = w(x, v)^{\mathrm{T}} \widehat{\beta}^b(\tau)$, where

$$\widehat{\beta}^b(\tau) = \arg \min_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^{n} \omega_i^b \rho_\tau (Y_i - b^{\mathrm{T}} \widehat{W}_i^b) T_i.$$

(4) Obtain the bootstrap draw of the global effects as

$$\widehat{\theta}_S^b(x) = \sum_{i=1}^{n} \omega_i^b T_i \widehat{\theta}^b(x, \widehat{V}_i^b) / \sum_{i=1}^{n} \omega_i^b T_i,$$

$$\widehat{\theta}_S^b(x \mid x_0) = \sum_{i=1}^{n} \omega_i^b T_i K_i(x_0) \widehat{\theta}^b(x, \widehat{V}_i^b) / \sum_{i=1}^{n} \omega_i^b T_i K_i(x_0),$$

or

$$\widehat{\theta}_S^b = \sum_{i=1}^{n} \omega_i^b T_i \widehat{\theta}^b(X_i, \widehat{V}_i^b) / \sum_{i=1}^{n} \omega_i^b T_i.$$

F Proofs of Appendix D

F.1 Notation

In what follows ϑ denotes a generic value of the control function. It is convenient also to introduce some additional notation, which will be extensively used in the proofs. Let $V_i(\vartheta) := \vartheta(Z_i)$, $W_i(\vartheta) := w(X_i, V_i(\vartheta))$, and $\dot{W}_i(\vartheta) := \partial_v w(X_i, v)|_{v=V_i(\vartheta)}$. When the previous functions are evaluated at the true values we use $V_i = V_i(\vartheta_0)$, $W_i = W_i(\vartheta_0)$, and $\dot{W}_i = \dot{W}_i(\vartheta_0)$. Recall that $A := (Y * 1(C > 0), C, Z, V)$, $T(c) = 1(c \in \overline{\mathcal{C}})$, and $T = T(C)$. For a function $f : \mathcal{A} \mapsto \mathbb{R}$, we use $\|f\|_{T,\infty} = \sup_{a \in \mathcal{A}} |T(c)f(a)|$; for a K-vector of functions $f : \mathcal{A} \mapsto \mathbb{R}^K$, we use $\|f\|_{T,\infty} = \sup_{a \in \mathcal{A}} \|T(c)f(a)\|_2$. We make functions in Υ as well as estimators $\widehat{\vartheta}$ to take values in $[0, 1]$. This allows us to simplify notation in what follows.

We adopt the standard notation in the empirical process literature (see, e.g., Van der Vaart, 1998),

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

and

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

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

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

We also use the concepts of covering entropy and bracketing entropy in the proofs. The covering entropy $\log N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the logarithm of the minimal number of $\|\cdot\|$-balls of radius ϵ needed to cover the set of functions $\mathcal{F}$. The bracketing entropy $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the logarithm of the minimal number of ϵ-brackets in $\|\cdot\|$ needed to cover the set of functions $\mathcal{F}$. An ϵ-bracket $[\ell, u]$ in $\|\cdot\|$ is the set of functions f with $\ell \leq f \leq u$ and $\|u - \ell\| < \epsilon$.

We 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 $Z_n^b \rightsquigarrow_{\mathbb{P}} Z$ in $\mathbb{Z}$ if

$$\sup_{h \in \mathrm{BL}_1(\mathbb{Z})} \left|\mathbb{E}_P^b h\left(Z_n^b\right) - \mathbb{E}_P h(Z)\right| \to_{\mathbb{P}^b} 0, \tag{F.1}$$

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

F.2 Proof of Lemma 3

The proof strategy follows closely the argument put forth in Chernozhukov et al. (2015) to deal with the dimensionality and entropy properties of the first step distribution regression estimators.

F.2.1 Auxiliary Lemmas

We start with 2 results on stochastic equicontinuity and a local expansion for the second stage estimators that will be used in the proof of Lemma 3.

Lemma 4 (Stochastic equicontinuity) *Let $\omega \geq 0$ be a positive random variable with $\mathrm{E}_P[\omega] = 1$, $\mathrm{Var}_P[\omega] = 1$, and $\mathbb{E}_P|\omega|^{2+\delta} < \infty$ for some $\delta > 0$, that is independent of $(Y * 1(C > 0), Z, C, V)$, including as a special case $\omega = 1$, and set, for $A = (\omega, Y * 1(C > 0), Z, C, V)$,*

$$f_1(A, \vartheta, \beta) := \omega \cdot [W(\vartheta)^{\mathrm{T}}\beta - Y] \cdot W(\vartheta) \cdot T.$$

Under Assumptions 1–3 the following relations are true

(a) Consider the set of functions

$$\mathcal{F} = \{f_1(A, \vartheta, \beta)^{\mathrm{T}}\alpha : (\vartheta, \beta) \in \Upsilon_0 \times \mathcal{B}, \alpha \in \mathbb{R}^{\dim(W)}, \|\alpha\|_2 \leq 1\},$$

where $\mathcal{B}$ is a compact set under the $\|\cdot\|_2$ metric containing β_0, Υ_0 is the intersection of Υ, defined in Lemma 2, with a neighborhood of ϑ_0 under the $\|\cdot\|_{T,\infty}$ metric. This class is P-Donsker with a square integrable envelope of the form ω times a constant.

(b) Moreover, if $(\vartheta, \beta) \to (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric, then

$$\|f_1(A, \vartheta, \beta) - f_1(A, \vartheta_0, \beta_0)\|_{P,2} \to 0.$$

(c) Hence for any $(\widetilde{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric such that $\widetilde{\vartheta} \in \Upsilon_0$,

$$\|\mathbb{G}_n f_1(A, \widetilde{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n f_1(A, \vartheta_0, \beta_0)\|_2 \xrightarrow{\mathbb{P}} 0.$$

(d) For for any $(\widehat{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric so that

$$\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n}), \text{ where } \widetilde{\vartheta} \in \Upsilon_0,$$

we have that

$$\|\mathbb{G}_n f_1(A, \widehat{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n f_1(A, \vartheta_0, \beta_0)\|_2 \xrightarrow{\mathbb{P}} 0.$$

Proof of Lemma 4. The proof is divided in subproofs of each of the claims.
Proof of Claim (a). The proof proceeds in several steps.
Step 1. Here we bound the bracketing entropy for

$$\mathcal{I}_1 = \{[W(\vartheta)^{\mathrm{T}}\beta - Y]T : \beta \in \mathcal{B}, \vartheta \in \Upsilon_0\}.$$

For this purpose consider a mesh $\{\vartheta_k\}$ over Υ_0 of $\|\cdot\|_{T,\infty}$ width δ, and a mesh $\{\beta_l\}$ over $\mathcal{B}$ of $\|\cdot\|_2$ width δ. A generic bracket over $\mathcal{I}_1$ takes the form

$$[i_1^0, i_1^1] = [\{W(\vartheta_k)^{\mathrm{T}}\beta_l - \kappa\delta - Y\}T, \{W(\vartheta_k)^{\mathrm{T}}\beta_l + \kappa\delta - Y\}T],$$

where $\kappa = L_W \max_{\beta \in \mathcal{B}} \|\beta\|_2 + L_W$, and $L_W := \|\partial_v w\|_{T,\infty} \vee \|w\|_{T,\infty}$.

Note that this is a valid bracket for all elements of $\mathcal{I}_1$ because for any ϑ located within δ from ϑ_k and any β located within δ from β_l,

$$
\begin{aligned}
|W(\vartheta)^{\mathrm{T}}\beta - W(\vartheta_k)^{\mathrm{T}}\beta_l|T &\leq |(W(\vartheta) - W(\vartheta_k))^{\mathrm{T}}\beta|T + |W(\vartheta_k)^{\mathrm{T}}(\beta - \beta_l)|T \\
&\leq L_W\delta \max_{\beta\in\mathcal{B}} \|\beta\|_2 + L_W\delta \leq \kappa\delta, \qquad \text{(F.2)}
\end{aligned}
$$

and the $\|\cdot\|_{P,2}$-size of this bracket is given by

$$\|i_1^0 - i_1^1\|_{P,2} \leq \sqrt{2}\kappa\delta.$$

Hence, counting the number of brackets induced by the mesh created above, we arrive at the following relationship between the bracketing entropy of $\mathcal{I}_1$ and the covering entropies of Υ_0, and $\mathcal{B}$,

$$
\begin{aligned}
\log N_{[]}(\epsilon, \mathcal{I}_1, \|\cdot\|_{P,2}) &\lesssim \log N(\epsilon^2, \Upsilon_0, \|\cdot\|_{T,\infty}) + \log N(\epsilon^2, \mathcal{B}, \|\cdot\|_2) \\
&\lesssim 1/(\epsilon^2 \log^4 \epsilon) + \log(1/\epsilon),
\end{aligned}
$$

and so $\mathcal{I}_1$ is P-Donsker with a constant envelope.

Step 2. Similarly to Step 1, it follows that

$$\mathcal{I}_2 = \{W(\vartheta)^{\mathrm{T}}\alpha T : \vartheta \in \Upsilon_0, \alpha \in \mathbb{R}^{\dim(W)}, \|\alpha\|_2 \leq 1\}$$

also obeys a similar bracketing entropy bound

$$\log N_{[]}(\epsilon, \|\cdot\|_{P,2}) \lesssim 1/(\epsilon^2 \log^4 \epsilon) + \log(1/\epsilon)$$

with a generic bracket taking the form $[i_2^0, i_2^1] = [\{W(\vartheta_k)^{\mathrm{T}}\alpha - \kappa\delta\}T, \{W(\vartheta_k)^{\mathrm{T}}\alpha + \kappa\delta\}T]$. Hence, this class is also P-Donsker with a constant envelope.

Step 3. In this step we verify the claim (a). Note that $\mathcal{F} = \omega \cdot \mathcal{I}_1 \cdot \mathcal{I}_2$. This class has a square-integrable envelope under P. The class $\mathcal{F}$ is P-Donsker by the following argument. Note that the product $\mathcal{I}_1 \cdot \mathcal{I}_2$ of uniformly bounded classes is P-Donsker, e.g., by Theorem 2.10.6 of Van der Vaart and Wellner (1996). Under the stated assumption the final product of the random variable ω with the P-Donsker class remains to be P-Donsker by the Multiplier Donsker Theorem, namely Theorem 2.9.2 in Van der Vaart and Wellner (1996).

Proof of Claim (b). The claim follows by the Dominated Convergence Theorem, since any $f_1 \in \mathcal{F}$ is dominated by a square-integrable envelope under P, and, $W(\vartheta)^{\mathrm{T}}\beta T \to W^{\mathrm{T}}\beta_0 T$ and $|W(\vartheta)^{\mathrm{T}}\alpha T - W^{\mathrm{T}}\alpha T| \to 0$ in view of the relation such as (F.2).

Proof of Claim (c). This claim follows from the asymptotic equicontinuity of the empirical process $(\mathbb{G}_n[f_1], f_1 \in \mathcal{F})$ under the $L_2(P)$ metric, and hence also with respect to the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric in view of Claim (b).

Proof of Claim (d). It is convenient to set $\widehat{f}_1 := f_1(A, \widehat{\vartheta}, \widehat{\beta})$ and $\widetilde{f}_1 := f_1(A, \widetilde{\vartheta}, \widetilde{\beta})$. Note that

$$
\begin{aligned}
\max_{1\leq j\leq \dim W} |\mathbb{G}_n[\widehat{f}_1 - \widetilde{f}_1]|_j &\leq \max_{1\leq j\leq \dim W} |\sqrt{n}\mathbb{E}_n[\widehat{f}_1 - \widetilde{f}_1]|_j + \max_{1\leq j\leq \dim W} |\sqrt{n}\mathbb{E}_P(\widehat{f}_1 - \widetilde{f}_1)|_j \\
&\lesssim \sqrt{n}\mathbb{E}_n[\widehat{\zeta}\,] + \sqrt{n}\ \mathrm{E}_P[\widehat{\zeta}\,] \\
&\lesssim \mathbb{G}_n[\widehat{\zeta}\,] + 2\sqrt{n}\mathbb{E}_P[\widehat{\zeta}\,],
\end{aligned}
$$

where $|f|_j$ denotes the jth element of the application of absolute value to each element of the vector f_1, and $\widehat{\zeta}$ is defined by the following relationship, which holds with probability approaching one,

$$\max_{1\leq j\leq \dim W} |\widehat{f}_1 - \widetilde{f}_1|_j \lesssim \omega|W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y|\|W(\widehat{\vartheta}) - W(\widetilde{\vartheta})\|_2 + \omega|W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - W(\widetilde{\vartheta})^{\mathrm{T}}\widetilde{\beta}| \lesssim \omega(k+|Y|)\Delta_n =: \widehat{\zeta},$$

where k is a constant such that $k \geq L_W \max_{\beta\in\mathcal{B}} \|\beta\|_2$ with $L_W = \|\partial_v w\|_{T,\infty} \vee \|w\|_{T,\infty}$, and $\Delta_n = o(1/\sqrt{n})$ is a deterministic sequence such that

$$\Delta_n \geq \|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty}.$$

The second inequality result follows from

$$|W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y|\|W(\widehat{\vartheta}) - W(\widetilde{\vartheta})\|_2 \lesssim (k+|Y|)\Delta_n, \text{ and } |W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - W(\widetilde{\vartheta})^{\mathrm{T}}\widetilde{\beta}| \lesssim k\Delta_n.$$

Then, by part (c) the result follows from

$$\mathbb{G}_n[\widehat{\zeta}\,] = o_{\mathbb{P}}(1), \qquad \sqrt{n}\mathbb{E}_P[\widehat{\zeta}\,] = o_{\mathbb{P}}(1).$$

Indeed,

$$\|\widehat{\zeta}\|_{P,2} \lesssim \sqrt{\mathbb{E}_P\omega^2(k^2 + \mathrm{E}_P(Y^2 \mid C \in \overline{\mathcal{C}}))\Delta_n^2} = o(1) \Rightarrow \mathbb{G}_n[\widehat{\zeta}\,] = o_{\mathbb{P}}(1),$$

and

$$\|\widehat{\zeta}\|_{P,1} \leq \mathbb{E}_P|\omega|\cdot(k + \mathbb{E}_P(|Y| \mid C \in \overline{\mathcal{C}})\Delta_n = o(1/\sqrt{n}) \Rightarrow \mathbb{E}_P|\widehat{\zeta}| = o_{\mathbb{P}}(1/\sqrt{n}).$$

Lemma 5 (Local expansion) *Under Assumptions 1–3, for*

$$\begin{aligned}
&\widehat{\delta} = \sqrt{n}(\widetilde{\beta} - \beta_0) = O_{\mathbb{P}}(1);\\
&\widehat{\Delta}(c,r) = \sqrt{n}(\widehat{\vartheta}(c,r) - \vartheta_0(c,r)) = \sqrt{n}\ \mathbb{E}_n[\ell(A,c,r)] + o_{\mathbb{P}}(1) \text{ in } \ell^\infty(\overline{\mathcal{CR}}),\\
&\|\sqrt{n}\ \mathbb{E}_n[\ell(A,\cdot)]\|_{T,\infty} = O_{\mathbb{P}}(1),
\end{aligned}$$

we have that

$$\sqrt{n}\ \mathbb{E}_P[\{W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y\}W(\widehat{\vartheta})T] = J\widehat{\delta} + \sqrt{n}\ \mathbb{E}_n\left[f_2(A)\right] + o_{\mathbb{P}}(1),$$

where

$$f_2(a) = \mathbb{E}_P\{[W^{\mathrm{T}}\beta_0 - Y]\dot{W} + W\dot{W}^{\mathrm{T}}\beta_0\}T\ell(a,C,R).$$

Proof of Lemma 5.
Uniformly in $Z \in \overline{\mathcal{Z}}$,

$$\begin{aligned}
&\sqrt{n}\mathbb{E}_P\{W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y \mid Z\}T\\
&= \sqrt{n}\mathbb{E}_P\{W^{\mathrm{T}}\beta_0 - Y \mid Z\}T + \{W(\bar{\vartheta}_\xi)^{\mathrm{T}}\widehat{\delta} + \dot{W}(\bar{\vartheta}_\xi)^{\mathrm{T}}\bar{\beta}_\xi\widehat{\Delta}(C,R)\}T\\
&= \sqrt{n}\mathbb{E}_P\{W^{\mathrm{T}}\beta_0 - Y \mid Z\}T + \{W^{\mathrm{T}}\widehat{\delta} + \dot{W}^{\mathrm{T}}\beta_0\widehat{\Delta}(C,R)\}T + \rho_Z,\\
&\bar{\rho} = \sup_{\{Z\in\overline{\mathcal{Z}}\}} |\rho_Z| = o_{\mathbb{P}}(1),
\end{aligned}$$

where $\bar{\vartheta}_\xi$ is on the line connecting ϑ_0 and $\widehat{\vartheta}$ and $\bar{\beta}_\xi$ is on the line connecting β_0 and $\widetilde{\beta}$.

The first equality follows by the mean value expansion. The second equality follows by uniform continuity of $W(\cdot)$ and $\dot{W}(\cdot)$, $\|\widehat{\vartheta}-\vartheta_0\|_{T,\infty} \xrightarrow{\mathbb{P}} 0$ and $\|\widetilde{\beta}-\beta_0\|_2 \xrightarrow{\mathbb{P}} 0$.

Since the entries of W and $\dot{W}$ are bounded, $\widehat{\delta} = O_{\mathbb{P}}(1)$, and $\|\widehat{\Delta}\|_{T,\infty} = O_{\mathbb{P}}(1)$, with probability approaching one,

$$\begin{aligned}\sqrt{n}\mathbb{E}_P\{W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta}-Y\}W(\widehat{\vartheta})T &= \mathbb{E}_P\{W^{\mathrm{T}}\beta_0-Y\}\dot{W}T\widehat{\Delta}(C,R)\\ &\quad+\mathbb{E}_P\{WW^{\mathrm{T}}T\}\widehat{\delta}+\mathbb{E}_P\{W\dot{W}^{\mathrm{T}}\beta_0T\widehat{\Delta}(C,R)\}+O_{\mathbb{P}}(\bar{\rho})\\ &= J\widehat{\delta}+\mathbb{E}_P[\{W^{\mathrm{T}}\beta_0-Y\}\dot{W}+W\dot{W}^{\mathrm{T}}\beta_0]T\widehat{\Delta}(C,R)+o_{\mathbb{P}}(1).\end{aligned}$$

Substituting in $\widehat{\Delta}(x,r) = \sqrt{n}\ \mathbb{E}_n[\ell(A,x,r)] + o_{\mathbb{P}}(1)$ and interchanging $\mathbb{E}_P$ and $\mathbb{E}_n$,

$$\mathbb{E}_P[\{W^{\mathrm{T}}\beta_0-Y\}\dot{W}+W\dot{W}^{\mathrm{T}}\beta_0]T\widehat{\Delta}(C,R) = \sqrt{n}\ \mathbb{E}_n[g(A)]+o_{\mathbb{P}}(1),$$

since $\|[\{W^{\mathrm{T}}\beta_0-Y\}\dot{W}+\ W\dot{W}^{\mathrm{T}}\beta_0]T\|_{P,2}$ is bounded . The claim of the lemma follows.

F.2.2 Proof of Lemma 3.

The proof is divided in two parts corresponding to the CLT and bootstrap CLT.

CLT: In this part we show $\sqrt{n}(\widehat{\beta}-\beta_0) \rightsquigarrow J^{-1}G$ in $\mathbb{R}^{d_w}$.
Step 1. This step shows that $\sqrt{n}(\widehat{\beta}-\beta_0) = O_{\mathbb{P}}(1)$. Recall that

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^{d_w}} \mathbb{E}_n[(Y-W(\widehat{\vartheta})^{\mathrm{T}}\beta)^2T].$$

Due to convexity of the objective function, it suffices to show that for any $\epsilon > 0$ there exists a finite positive constant B_ϵ such that ,

$$\liminf_{n\to\infty}\mathbb{P}\left(\inf_{\|\eta\|_2=1}\sqrt{n}\eta^{\ \mathrm{T}}\mathbb{E}_n\left[\widehat{f}_{1,\eta,B_\epsilon}\right]>0\right)\geq 1-\epsilon, \tag{F.3}$$

where

$$\widehat{f}_{1,\eta,B_\epsilon}(A) := \left\{W(\widehat{\vartheta})^{\mathrm{T}}(\beta_0+B_\epsilon\eta/\sqrt{n})-Y\right\}W(\widehat{\vartheta})T.$$

Let

$$f_1(A) := \left\{W^{\mathrm{T}}\beta_0-Y\right\}WT.$$

Then uniformly in $\|\eta\|_2 = 1$,

$$\begin{aligned}\sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_n[\widehat{f}_{1,\eta,B_\epsilon}] &= \eta^{\mathrm{T}}\mathbb{G}_n[\widehat{f}_{1,\eta,B_\epsilon}]+\sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_P[\widehat{f}_{1,\eta,B_\epsilon}]\\ &=_{(1)} \eta^{\mathrm{T}}\mathbb{G}_n[f_1]+o_{\mathbb{P}}(1)+\eta^{\mathrm{T}}\sqrt{n}\mathbb{E}_P[\widehat{f}_{1,\eta,B_\epsilon}]\\ &=_{(2)} \eta^{\mathrm{T}}\mathbb{G}_n[f_1]+o_{\mathbb{P}}(1)+\eta^{\mathrm{T}}J\eta B_\epsilon+\eta^{\mathrm{T}}\mathbb{G}_n[f_2]+o_{\mathbb{P}}(1)\\ &=_{(3)} O_{\mathbb{P}}(1)+o_{\mathbb{P}}(1)+\eta^{\mathrm{T}}J\eta B_\epsilon+O_{\mathbb{P}}(1)+o_{\mathbb{P}}(1),\end{aligned}$$

where relations (1) and (2) follow by Lemma 4 and Lemma 5 with $\widetilde{\beta} = \beta_0 + B_\epsilon\eta/\sqrt{n}$, respectively, using that $\|\widehat{\vartheta}-\widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta}\in\Upsilon$, $\|\widetilde{\vartheta}-\vartheta_0\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ and $\|\beta_0+B_\epsilon\eta/\sqrt{n}-\beta_0\|_2 = O(1/\sqrt{n})$; relation (3) holds because f_1 and f_2 are P-Donsker by step-2 below. Since J is positive definite, with minimal eigenvalue bounded away from zero, the inequality (F.3) follows by choosing B_ϵ as a sufficiently large constant.

Step 2. In this step we show the main result. Let

$$\widehat{f}_1(A) := \left\{W(\widehat{\vartheta})^{\mathrm{T}}\widehat{\beta} - Y\right\} W(\widehat{\vartheta})T.$$

From the first order conditions of the least squares problem,

$$\begin{aligned} 0 = \sqrt{n}\mathbb{E}_n\left[\widehat{f}_1\right] &= \mathbb{G}_n\left[\widehat{f}_1\right] + \sqrt{n}\mathbb{E}_P\left[\widehat{f}_1\right] \\ &=_{(1)} \mathbb{G}_n[f_1] + o_{\mathbb{P}}(1) + \sqrt{n}\mathbb{E}_P\left[\widehat{f}_1\right] \\ &=_{(2)} \mathbb{G}_n[f_1] + o_{\mathbb{P}}(1) + J\sqrt{n}(\widehat{\beta} - \beta_0) + \mathbb{G}_n[f_2] + o_{\mathbb{P}}(1), \end{aligned}$$

where relations (1) and (2) follow by Lemma 4 and Lemma 5 with $\widetilde{\beta} = \widehat{\beta}$, respectively, using that $\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta} \in \Upsilon$, and $\|\widetilde{\vartheta} - \vartheta\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2, and $\|\widehat{\beta} - \beta_0\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$.

Therefore by invertibility of J,

$$\sqrt{n}(\widehat{\beta} - \beta_0) = -J^{-1}\mathbb{G}_n(f_1 + f_2) + o_{\mathbb{P}}(1).$$

By the Central Limit Theorem

$$\mathbb{G}_n(f_1 + f_2) \rightsquigarrow G \text{ in } \mathbb{R}^{d_w}, \quad G \sim N(0, \Omega), \quad \Omega = \mathbb{E}_P[(f_1 + f_2)(f_1 + f_2)^{\mathrm{T}}],$$

where Ω is specified in the lemma. Conclude that

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow J^{-1}G \text{ in } \mathbb{R}^{d_w}.$$

Bootstrap CLT: In this part we show $\sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) \rightsquigarrow_{\mathbb{P}} J^{-1}G$ in $\mathbb{R}^{d_w}$.
Step 1. This step shows that $\sqrt{n}(\widehat{\beta}^b - \beta_0) = O_{\mathbb{P}}(1)$ under the unconditional probability $\mathbb{P}$. Recall that

$$\widehat{\beta}^b = \arg\min_{\beta \in \mathbb{R}^{\dim(W)}} \mathbb{E}_n[\omega(Y - W(\widehat{\vartheta}^b)^{\mathrm{T}}\beta)^2 T],$$

where ω is the random variable used in the weighted bootstrap. Due to convexity of the objective function, it suffices to show that for any $\epsilon > 0$ there exists a finite positive constant B_ϵ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\inf_{\|\eta\|_2=1} \sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_n\left[\widehat{f}^b_{1,\eta,B_\epsilon}\right] > 0\right) \geq 1 - \epsilon, \tag{F.4}$$

where

$$\widehat{f}^b_{1,\eta,B_\epsilon}(A) := \omega \cdot \left\{[W(\widehat{\vartheta}^b)^{\mathrm{T}}(\beta_0 + B_\epsilon\eta/\sqrt{n})] - Y\right\} W(\widehat{\vartheta}^b)T.$$

The result then follows by an analogous argument to step 1 in the proof of the CLT, which we do not repeat here.
Step 2. In this step we show that $\sqrt{n}(\widehat{\beta}^b - \beta_0) = -J^{-1}\mathbb{G}_n(f_1^b + f_2^b) + o_{\mathbb{P}}(1)$ under the unconditional probability $\mathbb{P}$. Let

$$\widehat{f}_1^b(A) := \omega \cdot \{W(\widehat{\vartheta}^b)^{\mathrm{T}}\widehat{\beta}^b - Y\}W(\widehat{\vartheta}^b)T.$$

From the first order conditions of the least squares problem in the weighted sample,

$$
\begin{aligned}
0 = \sqrt{n}\mathbb{E}_n\left[\widehat{f}_1^b\right] &= \mathbb{G}_n\left[\widehat{f}_1^b\right] + \sqrt{n}\mathbb{E}_P\left[\widehat{f}_1^b\right] \\
&=_{(1)} \mathbb{G}_n[f_1^b] + o_{\mathbb{P}}(1) + \sqrt{n}\mathbb{E}_P\left[\widehat{f}_1^b\right] \\
&=_{(2)} \mathbb{G}_n[f_1^b] + o_{\mathbb{P}}(1) + J\sqrt{n}(\widehat{\beta}^b - \beta_0) + \mathbb{G}_n[f_2^b] + o_{\mathbb{P}}(1),
\end{aligned}
$$

where relations (1) and (2) follow by Lemma 4 and Lemma 5 with $\widetilde{\beta} = \widehat{\beta}^b$, respectively, using that $\|\widehat{\vartheta}^b - \widetilde{\vartheta}^b\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta}^b \in \Upsilon$ and $\|\widetilde{\vartheta}^b - \vartheta_0\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2, and $\|\widehat{\beta}^b - \beta_0\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$. Therefore by invertibility of J,

$$\sqrt{n}(\widehat{\beta}^b - \beta_0) = -J^{-1}\mathbb{G}_n(f_1^b + f_2^b) + o_{\mathbb{P}}(1).$$

Step 3. In this final step we establish the behavior of $\sqrt{n}(\widehat{\beta}^b - \widehat{\beta})$ under $\mathbb{P}^b$. Note that $\mathbb{P}^b$ denotes the conditional probability measure, namely the probability measure induced by draws of $\omega_1, \ldots, \omega_n$ conditional on the data $A_1, ..., A_n$. By Step 2 of the proof of the CLT and Step 2 of the proof of the bootstrap CLT, we have that under $\mathbb{P}$:

$$\sqrt{n}(\widehat{\beta}^b - \beta_0) = -J^{-1}\mathbb{G}_n(f_1^b + f_2^b) + o_{\mathbb{P}}(1), \ \sqrt{n}(\widehat{\beta} - \beta_0) = -J^{-1}\mathbb{G}_n(f_1 + f_2) + o_{\mathbb{P}}(1).$$

Hence, under $\mathbb{P}$

$$\sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) = -J^{-1}\mathbb{G}_n(f_1^b - f_1 + f_2^b - f_2) + r_n = -J^{-1}\mathbb{G}_n((\omega - 1)(f_1 + f_2)) + r_n,$$

where $r_n = o_{\mathbb{P}}(1)$. Note that it is also true that

$$r_n = o_{\mathbb{P}^b}(1) \text{ in } \mathbb{P}\text{-probability},$$

where the latter statement means that for every $\epsilon > 0$, $\mathbb{P}^b(\|r_n\|_2 > \epsilon) = o_{\mathbb{P}}(1)$. Indeed, this follows from Markov inequality and by

$$\mathbb{E}_{\mathbb{P}}[\mathbb{P}^b(\|r_n\|_2 > \epsilon)] = \mathbb{P}(\|r_n\|_2 > \epsilon) = o(1),$$

where the latter holds by the Law of Iterated Expectations and $r_n = o_{\mathbb{P}}(1)$.

Note that $f_1^b = \omega \cdot f_1$ and $f_2^b = \omega \cdot f_2$, where f_1 and f_2 are P-Donsker by step-2 of the proof of the first part and $\mathbb{E}_P\omega^2 < \infty$. Then, by the Conditional Multiplier Central Limit Theorem, e.g., Lemma 2.9.5 in Van der Vaart and Wellner (1996),

$$G_n^b := \mathbb{G}_n((\omega - 1)(f_1 + f_2)) \rightsquigarrow_{\mathbb{P}} G \text{ in } \mathbb{R}^{d_w}.$$

Conclude that

$$\sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) \rightsquigarrow_{\mathbb{P}} J^{-1}G \text{ in } \mathbb{R}^{d_w}.$$

F.3 Proof of Theorem 4

In this section we use the notation $W_x(\vartheta) = w(x, V(\vartheta))$ such that $W_x = w(x, V(\vartheta_0))$.

We focus on the proof for the estimator of the ASF, because the proof for the estimator of the ASF on the treated can be obtained by analogous arguments. The results for the estimator of the ASF follow by a similar argument to the proof of Lemma 3 using Lemmas

6 and 7 in place of Lemmas 4 and 5, and the delta method. For the sake of brevity, here we just outline the proof of the FCLT.

Let $\psi_x(A,\vartheta,\beta) := W_x(\vartheta)^{\mathrm{T}}\beta T$ such that $\mu_S(x) = \mathbb{E}_P\psi_x(A,\vartheta_0,\beta_0)/\ \mathbb{E}_P T$ and $\widehat{\mu}_S(x) = \mathbb{E}_n\psi_x(A,\widehat{\vartheta},\widehat{\beta})/\mathbb{E}_n T$. Then, for $\widehat{\psi}_x := \psi_x(A,\widehat{\vartheta},\widehat{\beta})$ and $\psi_x := \psi_x(A,\vartheta_0,\beta_0)$,

$$\begin{aligned}\sqrt{n}\left[\mathbb{E}_n\psi_x(A,\widehat{\vartheta},\widehat{\beta}) -\ \mathbb{E}_P\psi_x(A,\vartheta_0,\beta_0)\right] \quad &= \quad \mathbb{G}_n\left[\widehat{\psi}_x\right] + \sqrt{n}\mathbb{E}_P\left[\widehat{\psi}_x - \psi_x\right] \\ &=_{(1)} \quad \mathbb{G}_n[\psi_x] + o_{\mathbb{P}}(1) + \sqrt{n}\mathbb{E}_P\left[\widehat{\psi}_x - \psi_x\right] \\ &=_{(2)} \quad \mathbb{G}_n[\psi_x] + o_{\mathbb{P}}(1) + \mathbb{G}_n[\sigma_x] + o_{\mathbb{P}}(1),\end{aligned}$$

where relations (1) and (2) follow by Lemma 6 and Lemma 7 with $\widetilde{\beta} = \widehat{\beta}$, respectively, using that $\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta} \in \Upsilon$, and $\|\widetilde{\vartheta} - \vartheta\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2, and $\sqrt{n}(\widehat{\beta} - \beta_0) = -J^{-1}\mathbb{G}_n(f_1 + f_2) + o_{\mathbb{P}}(1)$ from step 2 of the proof of Lemma 3.

The functions $x \mapsto \psi_x$ and $x \mapsto \sigma_x$ are P-Donsker by Example 19.7 in Van der Vaart (1998) because they are Lipschitz continuous on $\overline{\mathcal{X}}$. Hence, by the Functional Central Limit Theorem

$$\mathbb{G}_n(\psi_x + \sigma_x) \rightsquigarrow Z(x) \text{ in } \ell^\infty(\overline{\mathcal{X}}),$$

where $x \mapsto Z(x)$ is a zero mean Gaussian process with uniformly continuous sample paths and covariance function

$$\mathrm{Cov}_P[\psi_{x_0} + \sigma_{x_0}, \psi_{x_1} + \sigma_{x_1}], \quad x_0, x_1 \in \overline{\mathcal{X}}.$$

The result follows by the functional delta method applied to the ratio of $\mathbb{E}_n\psi_x(A,\widehat{\vartheta},\widehat{\beta})$ and $\mathbb{E}_n T$ using that

$$\begin{pmatrix} \mathbb{G}_n\psi_x(A,\widehat{\vartheta},\widehat{\beta}) \\ \mathbb{G}_n T \end{pmatrix} \rightsquigarrow \begin{pmatrix} Z(x) \\ Z_T \end{pmatrix},$$

where $Z_T \sim N(0, p_T(1-p_T))$,

$$\mathrm{Cov}_P(Z(x), Z_T) = G_T(x)p_T(1-p_T),$$

and

$$\begin{aligned}&\mathrm{Cov}_P[\psi_{x_0} + h_{x_0}, \psi_{x_1} + \sigma_{x_1} \mid T = 1] \\ &\qquad = \frac{\mathrm{Cov}_P[\psi_{x_0} + \sigma_{x_0}, \psi_{x_1} + \sigma_{x_1}] - \mu_T(x_0)\mu_T(x_1)p_T(1-p_T)}{p_T}.\end{aligned}$$

Lemma 6 (Stochastic equicontinuity) *Let $\omega \geq 0$ be a positive random variable with $\mathbb{E}_P[\omega] = 1$, $\mathrm{Var}_P[\omega] = 1$, and $\mathbb{E}_P|\omega|^{2+\delta} < \infty$ for some $\delta > 0$, that is independent of $(Y1(C > 0), C, Z, V)$, including as a special case $\omega = 1$, and set, for $A = (\omega, Y1(C > 0), C, Z, V)$,*

$$\psi_x(A,\vartheta,\beta) := \omega \cdot W_x(\vartheta)^{\mathrm{T}}\beta \cdot T.$$

Under Assumptions 1–3, the following relations are true.

(a) Consider the set of functions

$$\mathcal{F} := \{\psi_x(A,\vartheta,\beta) : (\vartheta,\beta,x) \in \Upsilon_0 \times \mathcal{B} \times \overline{\mathcal{X}}\},$$

where $\overline{\mathcal{X}}$ is a compact subset of $\mathbb{R}$, $\mathcal{B}$ is a compact set under the $\|\cdot\|_2$ metric containing β_0, Υ_0 is the intersection of Υ, defined in Lemma 2, with a neighborhood *of ϑ_0 under the $\|\cdot\|_{T,\infty}$ metric. This class is P-Donsker with a square integrable envelope of the form ω times a constant.*

(b) Moreover, if $(\vartheta, \beta) \to (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric , then

$$\sup_{x \in \overline{\mathcal{X}}} \|\psi_x(A, \vartheta, \beta) - \psi_x(A, \vartheta_0, \beta_0)\|_{P,2} \to 0.$$

(c) Hence for any $(\widetilde{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric such that $\widetilde{\vartheta} \in \Upsilon_0$,

$$\sup_{x \in \overline{\mathcal{X}}} \|\mathbb{G}_n \psi_x(A, \widetilde{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n \psi_x(A, \vartheta_0, \beta_0)\|_2 \to_{\mathbb{P}} 0.$$

(d) For for any $(\widehat{\vartheta}, \widetilde{\beta}) \xrightarrow{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric , so that

$$\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n}), \text{ where } \widetilde{\vartheta} \in \Upsilon_0,$$

we have that

$$\sup_{x \in \overline{\mathcal{X}}} \|\mathbb{G}_n \psi_x(A, \widehat{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n \psi_x(A, \vartheta_0, \beta_0)\|_2 \xrightarrow{\mathbb{P}} 0.$$

Proof of Lemma 6. The proof is omitted since it is similar to the proof of Lemma 4.

Lemma 7 (Local expansion) *Under Assumptions 1–3 , for*

$$\begin{aligned} &\widehat{\delta} = \sqrt{n}(\widetilde{\beta} - \beta_0) = O_{\mathbb{P}}(1); \\ &\widehat{\Delta}(x, r) = \sqrt{n}(\widehat{\vartheta}(x, r) - \vartheta_0(x, r)) = \sqrt{n}\ \mathbb{E}_n[\ell(A, x, r)] + o_{\mathbb{P}}(1) \text{ in } \ell^\infty(\overline{\mathcal{X}\mathcal{R}}), \\ &\|\sqrt{n}\ \mathbb{E}_n[\ell(A, \cdot)]\|_{T,\infty} = O_{\mathbb{P}}(1), \end{aligned}$$

we have that

$$\sqrt{n}\ \left\{\mathbb{E}_P W_x(\widehat{\vartheta})^{\mathrm{T}} \widetilde{\beta} T - \mathbb{E}_P W_x^{\mathrm{T}} \beta_0 T\right\} = \ \mathbb{E}_P\{W_x T\}^{\mathrm{T}} \widehat{\delta} + \mathbb{E}_P\{\dot{W}_x^{\mathrm{T}} \beta_0 T \ell(a, X, R)\}\big|_{a=A} + \bar{o}_{\mathbb{P}}(1),$$

where $\bar{o}_{\mathbb{P}}(1)$ denotes order in probability uniform in $x \in \overline{\mathcal{X}}$.

Proof of Lemma 7. The proof is omitted because it is similar to the proof of Lemma 5.