Non-separable Sample Selection Models with Censored Selection Rules

Ivan Fernandez-Val* Aico van Vuuren† Francis Vella‡

February 2017

Abstract

We consider the identification and estimation of non separable sample selection models with censored selection rules. We account for selection via a control function approach and discuss different objects of interest based on; (1) local objects for the selected sample; (2) local objects for the entire population and (3) derived objects obtained from integration over the values of the control variable. We also present results regarding bounds and counterfactual. We derive conditions for identification for these different objects and suggest strategies for estimation. These strategies are illustrated in two examples related to the determinants of females wages and wage growth in the United Kingdom.

Keywords: Sample Selection, Non-separable models, Control function, Quantile and Distribution Regression

JEL-codes: C14, C21, C24

*Boston University.
†University of Gothenburg.
‡Georgetown University. We thank Costas Meghir for the provision of the data. We also thank participants of seminars in Amsterdam, Odense and Gothenburg.

1 Introduction

This paper provides a treatment of the nonseparable sample selection model with partially observed censoring rules. The most common example is a selection rule with censoring at zero, also referred to in the parametric setting as tobit type 3, although other forms of censored selection rules are permissible. A leading empirical example is estimating the determinants of wages when workers report working hours rather than just work/not work decisions. Many other empirical questions are also characterized by this structure. An important benefit of the model, beyond the relaxation of distributional assumptions, is the inherent heterogeneity facilitated through the nonseparability. Our approach to accounting for selection is the use of an appropriately constructed control function. We propose a three step estimation procedure in which we estimate first employ distribution regression to compute the appropriate control function. The second step is estimated either by distribution regression or quantile regression employing the estimated control function. The primary estimands of interest can then be recovered via these second step estimates.

Our paper contributes to the growing literature on nonseparables models with endogeneity (see, for example, Imbens and Newey 2009, Jun 2009 and Masten and Torgovitsky 2013)) and nonseparable sample selection models in the selection literature (for example, Newey 2007)). Our major points of departure is our focus on a censored selection rule and our use of quantile and distributional regression in a selection context. Our paper, however, has three important contributions that distinguish it from the existing literature. First, while Newey (2007) considers the distribution of the outcome variable conditional on selection we provide statements regarding the outcome variable distribution conditional on specific values of the control function. This local identification approach is popular in

many contexts (see, for example, Heckman and Vytlacil, 2005). We also show that when every population observation has a positive probability of being selected, conditional on the outcome of the particular control variable, selection is irrelevant for the distribution of the outcome variable. Hence, we can estimate certain objects of interest conditional on the value of the control variable. We can also estimate global objects by integrating over the entire distribution of our control variable. However we highlight that these global objects require strong assumptions on the identifying explanators which may not always be satisfied in empirical situations. We also consider partial identification for these global objects.

Our paper is also related to the literature on quantile selection models. Arellano and Bonhomme (2016) address selection by modeling the copula of the error terms in the outcome and selection equations. The most important difference to this paper is they consider a binary, rather than a censored, selection equation. Thus we require more information about the selection process. However this has the advantage that one can consider local effects conditional on the control variable which are identified under weaker conditions. In contrast, Arellano and Bonhomme (2016) consider global effects which require stronger support conditions. Moreover, the Arellano and Bonhomme (2016) procedure can be potentially difficult to implement when the model for the copula is highly dimensional. Our method is computationally simple and easily implemented via standard estimation methods found in statistical software packages.

The following section outlines the model and the related literature. Section 3 defines the control variable and provides results regarding the objects of interest in this model. Section 4 provides estimators of these objects. Section 5 provides an empirical example focusing on the determinants of wages for working women in the United Kingdom.

2 Model

The model has the following structure:

$$Y = g(X, \varepsilon) \text{ if } C > 0 \quad (1)$$

$$C = \max\left(h(X, Z, \eta), 0\right) \quad (2)$$

where Y and C are observable random variables with supports $\mathcal{Y}$ and $\mathcal{C}$ respectively, and X and Z are vectors of observable explanatory variables. The functions $g(.)$ and $h(.)$ are unknown and ε and η are potentially mutually dependent unobservables. The primary objective is to estimate functionals related to $g(.)$ noting that Y is only observed when C is above some known threshold. The non observability of Y for specific values of C induces the possibility of selection bias.

The model is a non-parametric representation of the tobit type-3 model and is a variant of the Heckman (1978) selection model. Here the selection equation has a censored rather than a binary outcome. Previous attempts to estimate this model have imposed a variety of restrictions. It was initially examined in a fully parametric setting, imposing additivity and normality, and estimated by maximum likelihood (see Amemiya 1978, 1979). Vella (1993) provided a two-step estimator based on estimating the generalized residual from the C equation and including this as a control function in the outcome equation. Subsequent work by Honoré et al. (1997) and Lee and Vella (2006) relaxed the model's distributional assumptions but imposed an index restriction and separability of the error terms in each equation.

Note that the model can easily be extended to allow for different censoring rules. For example the Y variable could be censored in a number of ways provided there was some region(s) for which it was continuously observed. This allows for top, middle and/or

bottom censoring. Also, although we do not consider it here, our approach also extends to when Y is censored. For example:

$$Y = \max(g(X, \varepsilon), 0) \text{ if } C > 0. \tag{3}$$

Finally,we highlight that in the presence of a Z variable in equation 2 we allow C to be included as a conditioning variable in the outcome equation.

3 Identification

Our approach to estimation is to account for the selection bias through the use of an appropriately constructed control variable. Accordingly, we first establish the existence of such a variable for this model and then define some objects of interest incorporated in (1)-(2).

Assumption 1 (Control Variable) *$(\varepsilon, \eta) \perp\!\!\!\perp X, Z$, η is a continuously distributed scalar with CDF that is strictly increasing on the support of η and $t \longrightarrow h(X, Z, t)$ is strictly increasing a.s.*

This assumption allows for endogeneity between X and ε in the selected population $C > 0$, since in general ε and η are correlated, i.e. $\varepsilon \not\perp\!\!\!\perp X \mid C > 0$. The monotonicity assumption allows a non-monotonic relationship between ε and C because ε and η are allowed to be non-monotonically dependent.

Lemma 1 (Existence of Control Variable) *Define $V = F_{C|X,Z}(C \mid X, Z)$ where F denotes the conditional CDF. Under the model in (1)-(2) and Assumption 1:*

$$\varepsilon \perp\!\!\!\perp X, Z \mid V, C > 0,$$

i.e., V is a control variable for X in the selected population with $C > 0$.

Proof. See appendix A. ■

The intuition behind Lemma 1 is based on three observations. First, it can be shown that there is a one-to-one relationship between V and η implying that the distribution of ε, conditional on η, is identical to the distribution of ε, conditional on V. Moreover, we have $V = F_\eta(\eta)$. The one-to-one relationship follows because V is based on the conditional distribution function of C and since $C > 0$ for the selected sample it is determined by h and, due to assumption 1, h has a one-to-one relationship with η. The second observation is that conditioning on X, Z and η (or equivalently V) makes selection, *i.e.* $C > 0$, deterministic. Therefore, the distribution of ε, conditional on X, Z and η, does not depend on the condition that $C > 0$. The final observation, namely our assumption that $\varepsilon \perp\!\!\!\perp X, Z$, is sufficient to prove the Lemma.

For the model under consideration there are different classes of estimands interesting for econometric inference. These are: (1) local objects; (2) objects based on integration over the control variable; (3) partial identification based on bounds; and (4) counterfactual distributions.

3.1 Local objects

We consider local objects for given values of X conditional on the control variable $V = F_{C|X,Z}(C \mid X, Z)$. Let $\mathcal{X}$ denote the region of the support of X of interest. Before we discuss the definition and identification of these local objects we present the condition characterizing our exclusion restriction.

Assumption 2 (Identification condition of the exclusion restriction) *Define the support of $Z|X = x$ by $\mathcal{Z}(x)$. In addition, if we define the set $\mathcal{Z}_v(x)$ for $(x, v) \in \mathcal{X} \times [0, 1]$*

as:

$$\mathcal{Z}_v(x) = \left\{ z \in \mathcal{Z}(x) | h(x, z, F_\eta^{-1}(v)) > 0 \right\},$$

then $\mathbb{P}(Z \in \mathcal{Z}_v(x) | X = x, V = v) > 0$.

Assumption 2 differs from the standard instrument relevance assumption since it does not imply that Z has an impact on the value of C. Rather, it states that for given values of x and v there must be an outcome of Z such that the observation has probability 1 to be in the selected sample. That is, $C = h(X, Z, F_\eta^{-1}(V))$, must be strictly larger than zero. The strength of this assumption depends not only on the Z but also on the values of x and v.

Definition 2 (Local average structural function) *The local average structural function (ASF) at* $(x, v) \in \mathcal{X} \times [0, 1]$ *is:*

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

$\mu(x, v)$ *is the expected value of the outcome variable for a fixed level of* $X = x$ *conditional on* $V = v$. *It depends on the realization of* V *because* ε *depends on* V *through* η. *The identification result is stated in the following lemma.*

Lemma 3 *Under model (1)-(2) and Assumptions 1 and 2:*

$$\mu(x, v) = \mathbb{E}_{Y|X,V}(Y | X = x, V = v) = \mathbb{E}_{Y|X,V,C>0}(Y | X = x, V = v, C > 0). \qquad (4)$$

Proof. See Appendix B. ■

From Assumption 1 the local ASF is the expected value of the outcome variable conditional on $X = x$ and $V = v$. This also equals the last expression in (4), which is a function

of the data distribution and hence identified. The second equality in (4) follows from $\mathbb{E}_{Y|X,V}(Y|X = x, V = v, C > 0)$ not depending on the outcome of $Z = z$. This follows from Lemma 1 which shows ε does not depend on the outcome of Z. Moreover, although the distribution of Z is different for the selected sample and the entire population, g does not depend on z. Hence, we can condition on any possible level of z including that for which we have that $h(x, z, F_\eta^{-1}(v)) > 0$. This level of z exists due to Assumption 2. The mean outcome of Y conditional on the outcomes of X, Z and V now no longer depends on $C > 0$ which proves the second equality in (4).

Definition 4 (Local average derivative definition) *The local average derivative in the entire population at* $(x, v) \in \mathcal{X} \times [0, 1]$ *is*

$$\delta(x, v) = \mathbb{E}[\partial_x g(x, \varepsilon) \mid V = v]. \tag{5}$$

The local average derivative is the first-order derivative of the local ASF conditional on the assumption that we can interchange differentiation and integration in (5). This is made formal in the next corollary.

Corollary 5 (Local average derivative identification) *Assume that for all* $x \in \mathcal{X}$, $g(x, \varepsilon)$ *is continuously differentiable in* x *a.s.,* $E[|g(x, \varepsilon)|] < \infty$, *and* $E[|\partial_x g(x, \varepsilon)|] < \infty$. *Under model (1)-(2) and Assumptions 1 and 2:*

$$\delta(x, v) = \frac{\partial}{\partial x}\mu(x, v) = \partial_x \mathbb{E}_{Y|X,V,C>0}(Y|X = x, V = v, C > 0).$$

The estimands based on the local average derivative extend in a straightforward manner to quantiles. These are shown in the following two definitions.

Definition 6 (Local structural distribution function) *The local structural distribu-*

tion function (SDF) at $(x, v, \tau) \in \mathcal{X} \times [0,1] \times [0,1]$ *is:*

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

Definition 7 (Local quantile structural function) *The local quantile structural function (QSF) in the entire population at* $(x, v, \tau) \in \mathcal{X} \times [0,1] \times [0,1]$ *is:*

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

$G(y, x, v)$ represents the distribution function for a given level of y and x, conditional on the outcome $V = v$. It can be interpreted as the distribution function evaluated at y when observations with a V equal to v had observed characteristics equal to x. Using Assumption 1, we have:

$$\begin{aligned}\mathbb{E}_{\varepsilon|V=v}[1\{g(x, \varepsilon) \leq y\} | V = v] &= \mathbb{E}_{\varepsilon|V=v}[1\{g(x, \varepsilon) \leq y\} | X = x, V = v] \\ &= F_{Y|X,V}(y|X = x, V = v),\end{aligned}$$

and this implies that $q(\tau, x, v)$ equals the τ-th quantile of Y, conditional on the outcome of X and V, *i.e.* $\mathbb{Q}_\tau[Y|X = x, V = v]$. This equality is also true for its derivatives. Denoting $\delta_\tau(x, v)$ as the derivative of $q(\tau, x, v)$ it is clear that, under the assumptions of Theorem 5, $\delta_\tau(x, v|C > 0)$ equals the derivative of $\mathbb{Q}_\tau[Y|X = x, V = v]$, *i.e.* $\partial_x \mathbb{Q}_\tau[Y|X = x, V = v]$. The identification of the local SDF is discussed in the following lemma.

Lemma 8 (Structural distribution function identification) *Under model (1)-(2) and Assumptions 1 and 2:*

$$G(y, x, v) = F_{Y|X,V}(y|X = x, V = v) = F_{Y|X,V,C>0}(y|X = x, V = v, C > 0). \quad (6)$$

Proof. See Appendix B. ■

The intuition behind the proof of Lemma 8 is similar to the proof of Lemma 3. The last expression in (6) equals the data distribution and is hence identified. This also shows identification of the local QSF.

Even though we have investigated the identification of our objects of interest using the assumption that there is an appropriately excluded variable(s) Z, the existence of such a variable(s) is not a strict requirement for identification of the objects presented in this section. Suppose that, for a particular level of x and v, we can replace Assumption 2, by the assumption that $h(x, F_{\eta}^{-1}(v)) > 0$, then Lemmas 3, 5, and 8 are valid even in the absence of Z. The interpretation of this assumption is that observation with a level of X equal to x and a V equal to v always satisy $C > 0$.

3.2 Objects based on integration over the control variable

We expand the estimands of interest by examining the global counterparts of the local objects outlined above. That is, we can integrate over the distribution function of any random variable $\widetilde{V}$ with a support which is a subset of $[0, 1]$, to obtain:

$$T(x) = \int_{\text{supp}(\widetilde{V})} T(x, v) dF_{\widetilde{V}}(v),$$

where $T(x, v)$ can be any of local objects defined above. $T(x)$ is identified when the distribution of $\widetilde{V}$ is identified and $T(x, v)$ is identified for the support of $\widetilde{V}$. For example, as in Newey (2007) and Imbens and Newey (2009) one can consider the SDF among the selected population, *i.e:*

$$G(y, x|C > 0) = \mathbb{E}_{\varepsilon|C>0}[1\{g(x, \varepsilon) \leq y\} |C > 0],$$

which is the local structural distribution function introduced in Section 4.2 integrated over the distribution of $\widetilde{V} = (V|C > 0)$. It is equal to the distribution function evaluated at y when all observations in the selected sample have observed characteristics equal to x. Due to the endogeneity of $X|C > 0$, this differs from the observed distribution of y, conditional on $C > 0$ and $X = x$. If, for example, the correlation between ε and η is positive and when X has a positive impact on both g and h, then there is an over representation of observations with a low ε and a high X in the selected sample. Hence, for these high levels of X, $G(y, x|C > 0)$ will be lower than the observed distribution. The distribution of $V|C > 0$ is a part of the data distribution and hence identified. The local object $\mathbb{E}_{\varepsilon|V,C>0}[1\{g(x,\varepsilon) \leq y\}|C > 0, V = v]$ is only identified in the joint support of (X, V), conditional on $C > 0$. We make the following assumption that is identical to that made by Imbens and Newey (2009).

Assumption 3 (Common Support) $\mathcal{V}(x) \equiv \text{supp}(V \mid X, C > 0) = \text{supp}(V \mid C > 0) \equiv \mathcal{V}$ *a.s.*

The common support assumption implies that the support of the control variable does not depend on X. We have that

$$G(y, x|C > 0) = \int_{\mathcal{V}} G(y, x, v) dF_V(v|C > 0) = \int_{\mathcal{V}(x)} G(y, x, v) dF_V(v|C > 0), \quad (7)$$

where the last equality is due to Assumption 3. Since we have shown that $G(y, x, v|C > 0)$ is identified for any $v \in \mathcal{V}(x)$, it follows that $G(y, x|C > 0)$ is identified.

We can also estimate the same objects based on the entire population. For example, we can define the structural distribution function among the entire population as

$$G(y, x) = \mathbb{E}_{\varepsilon|C>0}[1\{g(x,\varepsilon) \leq y\}], \quad (8)$$

which is the local structural distribution function introduced in Section 4.2, integrated over the distribution of $\widetilde{V} = V$. The object as defined in (8) is the object of interest in some earlier papers (for example, Buchinsky, 1998). The distribution function of V is uniform by definition and hence it is identified. The object $\mathbb{E}_{\varepsilon|V,C>0}[1\{g(x,\varepsilon) \leq y\}|V = v]$ is only identified whenever there exists a $z \in \mathcal{Z}(x)$ for which $h(x, z, F_\eta^{-1}(v)) > 0$ for every $v \in [0, 1]$. Appendix C proves that the assumption for identification is equal to the assumption that $\mathcal{V}(x)$ equals $[0, 1]$. The intuition behind this is not difficult to understand since:

$$G(y, x) = \int_0^1 G(y, x, v)dv,$$

which is only equal to the left-hand side of (7) when $\mathcal{V}(x) = [0, 1]$. Unfortunately, this assumption is too strong for many applications. For example, in the empirical example below we examine the hours of work decision of females as the selection process. If the data contains some individuals with low levels of education then these individuals should still have a positive probability of selection for any level of unobserved abilities. Otherwise we cannot point identify the objects of interest for the whole population.

3.3 Partial identification

Given the restrictive nature of some of the assumptions required for identification of the objects above we now employ a partial identification approach which relies on less demanding assumptions. We provide results for the bounds for the entire population and apply the same ideas to obtain bounds on $G(y, x|C > 0)$ when Assumption 3 is not satisfied.

Lemma 9 *Define* $P(x) = \int_{[0,1]/\mathcal{V}(x)} dv$. *Then:*

$$G_l(y, x) \leq G(y, x) \leq P(x) + G_l(y, x),$$

with

$$G_l(y,x) = \int_{V(x)} \mathbb{P}(Y \leq y | X = x, V = v) dv.$$

Proof. See appendix D. ■

The proof of Lemma 9 employs the result that we can identify $G(y,x,v)$ for every $v \in \mathcal{V}(x)$ (see Lemma 8), while we construct bounds $0 \leq G(y,x,v) \leq 1$ for every point on the unit interval that does not belong to $\mathcal{V}(x)$. Combining these observations, together with the definition of $G(y,x) = \int_0^1 G(y,x,v)dv$, results in Lemma 9. We can also calculate bounds after integration over the distribution of X. For example:

$$G(y) = \int_{\mathcal{X}} G(y,x) dF_X(x)$$

By integration over the distribution of X and using Lemma 9 we directly obtain that:

$$\int_{\mathcal{X}} G_l(y,x) dF_X(x) \leq G(y) \leq \int_{\mathcal{X}} \{G_l(y,x) + P(x)\} dF_X(x). \tag{9}$$

The distribution of X over which we integrate does not necessarily equal the distribution of X among the entire population (as equation (9)). That is, we can consider counterfactuals such as the distribution of X where one element is fixed to a predetermined level.

3.4 Counterfactual distributions

The use of counterfactual distributions is currently popular in the econometrics literature (see Chernozhoukov, Fernandez-Val and Melly, 2013) although few papers consider their use in a selection context (see Arellano and Bonhomme, 2016, and the references given there). Our counterfactual distributions are based on the integration methods introduced

in Section 3.2. From (1)-(2) and Assumption 1, we have:

$$
\begin{aligned}
&F_Y(y|C>0) \\
&= \int_{\mathcal{X},\mathcal{Z},\mathcal{V}(x,z)} F_Y(y|X=x,V=v)dF_{V,X,Z|C>0}(v,x,z|C>0) \\
&= \int_{\mathcal{X},\mathcal{Z}} \left\{ \int_{\mathcal{V}(x,z)} F_Y(y|X=x,V=v)dF_{V|X,Z,C>0}(v|X=x,Z=z,C>0) \right\} dF_{X,Z|C>0},
\end{aligned}
\tag{10}
$$

where $\mathcal{V}(x,z) = \text{supp}(V|X=x,Z=z,C>0)$ and also the supports $\mathcal{X}$ and $\mathcal{Z}$ are conditional on $C>0$. The final bracketed term represents the distribution of Y for the selected sample conditional on the outcomes of X and Z.

Assume the outcome model has three different variants which are denoted by subscripts q, r and s. While these model variants are represented by the system (1)-(2), each has it's own distributions of X, Z, ε and η as well as functions g and h. In our empirical application we use the year as a basis for the different variants. Let $F_{Y_{q|r,s}}(\cdot|\cdot)$ be the distribution function of Y when the function g and the distribution of ε are as in model version q, the distribution of X and Z are as in model version r and the function h as well as the distribution of η are as in model version s. Below we provide decompositions of wage growth for the period 1982-99 for women in the United Kingdom. Thus we assume that the wage structure is as in year q, the observed characteristics are as in year r, and participation is as in year s. Based on (10), we have that:

$$
\begin{aligned}
F_{Y_{q|r,s}}(y|C_s>0) = \int_{\mathcal{X}_r\times\mathcal{Z}_r} &\left\{ \int_{\mathcal{V}_s(x,z)} F_{Y_q|X_q,V_q}(y|X_q=x,V_q=v) \right. \\
&\left. dF_{V_s|X_s,Z_s,C_s>0}(v_s|X_s=x,Z_s=z,C_s>0) \right\} dF_{X_r,Z_r|C_r>0}(x,z|C_r>0),
\end{aligned}
\tag{11}
$$

where $\mathcal{X}_t$ and $\mathcal{Z}_t$ are the supports of X_t and Z_t conditional on $C_t>0$, and where the subscripts $t=q,r,s$ indicate the model version. In addition, $\mathcal{V}_t(x,z)$ is the support of V_t

conditional on X_t and Z_t and $C_t > 0$. It can be shown that the lower bound of the support equals $v_0^t(x,z) = F_{C_t|X_t,Z_t}(0|X_t = x, Z_t = z)$. For the object in (**??**) to be identified, the distributions $F_{Y_q|X_q,V_q}(y|X_q = x, V_q = v)$ and $F_{V_s|X_s,Z_s,C_s>0}(v_s|X_s = x, Z_s = z, C_s > 0)$ must be identified on the whole integral. For the distribution $F_{Y_q|X_q,V_q}(y|X_q = x, V_q = v)$, this results in the restrictions that $\mathcal{X}_r \subseteq \mathcal{X}_q$ and $\inf_{z\in\mathcal{Z}_r(x)} v_0^s(x,z) \geq \inf_{z\in\mathcal{Z}_r(x)} v_0^q(x,z)$ for all $x \in \mathcal{X}_r$. Here, $\mathcal{Z}_t(x)$ is the support of Z_t conditional on $X_t = x$. This implies that the observed characteristics in version r should be a subset of the observed characteristics in version q, while the potential outcomes of the control variable V_s should be a subset of the potential outcomes of V_r. For $F_{V_s|X_s,Z_s,C_s>0}(v|X_s = x, Z_s = z_s, C_s > 0)$ to be identified, we require $\mathcal{X}_r \times \mathcal{Z}_r \subseteq \mathcal{X}_s \times \mathcal{Z}_s$. That is, all characteristics (including the outcomes of Z) in version r should be a subset of those in version s.

We do not need an exclusion restriction to identify (**??**). However, if we do not, then the restriction on the support equals $v_0^s(x) \geq v_0^q(x)$. This implies that, conditional on the outcome of X, selection should be at least as restrictive in version s as in version q. In Section 5, we examine female labor supply and define the year s as a year with a low participation level of women, while the year r has a higher level of participation. When using an exclusion restriction, then it is possible that selection in version s is sometimes less restrictive as in version q.

4 Estimation

We propose a multi-step procedure in which we first estimate the control function. We then estimate the local objects and then, if applicable, we estimate the objects of interest that are obtained by integrating over the control function.

4.1 Step 1: Estimation of the control function

The first step is to estimate the control function via distribution regression (see Foresi and Peracchi, 1995 and Chernozhukov, Fernandez-Val and Mely, 2013). Based on the observed value of C_i, $\hat{V}_i = \Lambda(P(X_i, Z_i)^T \widehat{\beta}(C_i))$, where:

$$\hat{\beta}(c) = \arg\max_b \sum_{i=1}^{n} \left[1\{C_i \leq c\} \log \Lambda(P(X_i, Z_i)^T b)\right.$$

$$\left. +1\{C_i > c\} \log \Lambda(P(X_i, Z_i)^T b))\right], \ c \in \mathcal{C}_n,$$

and where Λ, labeled the link function, is typically the normal or logistic distribution function, and $\mathcal{C}_n$ denotes the support of C in the sample. P is a polynomial of X_i and Z_i. The assumption of a specific link function does not impose distributional assumptions on the errors in the true model. The imposition of the single-index structure in this equation is to make estimation feasible in the presence of a high dimensional X.

4.2 Step 2: Estimation of local objects

Following Chernozhukov, Fernandez-Val and Kowalski (2014) we propose a series based estimator. Estimation is conducted over the sample for which $C > 0$. The local ASF, $\mu(x, v)$, can be estimated as the predicted value of a series regression of Y_i on X_i and $\widehat{V}_i$, *i.e.* $\widehat{\mu}(x, v) = P(x, v)^T \widehat{\theta}$, with P represents a polynomial in x and v and

$$\widehat{\theta} = \left[P(x, v)^T P(x, v)\right]^{-1} P(x, v)^T Y.$$

Derivatives of $\widehat{\mu}(x,v)$ yields an estimate of the local average derivative.[1] We can use distribution series regression for the estimation of the SDF. That is, estimate $G(y,x,v)$ using $\widehat{G}(y,x,v) = \Lambda(P(x,v)^T\widehat{\theta})$, where θ is the solution to:

$$\widehat{\theta}(y) = \arg\max_b \sum_{i=1}^{n} \Big[1\{Y_i \leq y\} \log \Lambda(P(X_i, \widehat{V}_i)^T b)) \\ +1\{Y_i > y\} \log \Lambda(P(X_i, \widehat{V}_i)^T b))\Big].$$

We use quantile series regression for the estimation of the QSF and estimate $\widehat{\theta}_\tau$ via a quantile regression of Y_i on $P(X_i, \widehat{V}_i)$. We then estimate $q(\tau, x, v)$ by $P(x,v)^T\widehat{\theta}_\tau$. We can estimate $\delta_\tau(x,v)$ by taking derivatives of this estimator with respect to x, *i.e.* $\nabla_x P(x,v)^T\widehat{\theta}_\tau$.

4.3 Step 3: Estimation of the objects of interest by integration

Estimation of objects which involve integration over the control function can be obtained by summing over the distribution of the control function. We obtain an estimator of the average structural function $\mu(x)$ for the selected sample using:

$$\widehat{\mu}(x|C>0) = \sum_{i=1, C_i>0}^{n} \widehat{\mu}(x, \widehat{V}_i) / \sum_{i=1}^{n} \mathbf{1}(C_i > 0),$$

where the estimator $\widehat{\mu}(x, \widehat{V}_i)$ is from the previous subsection. The estimator of the average derivative can be obtained by taking derivatives of $\widehat{\mu}(X_i, V_i)$ with respect to X_i, *i.e:*

$$\widehat{\delta}(x|C>0) = \sum_{i=1, C_i>0}^{n} \widehat{\delta}(x, \widehat{V}_i) / \sum_{i=1}^{n} \mathbf{1}(C_i > 0) = \sum_{i=1, C_i>0}^{n} \frac{\partial}{\partial x}\widehat{\mu}(x, \widehat{V}_i) / \sum_{i=1}^{n} \mathbf{1}(C_i > 0).$$

[1] An alternative approach is to follow Jun (2009) and Masten and Torgovitsky (2014). These papers acknowledge that with the 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 the logic of their approach can be extended here. That is, while Jun (2009) and Masten and Torgovistsky (2014) account for endogeneity we could employ this approach to control for selection.

An estimator of the structural distribution function, $G(y,x)$, is given by:

$$\widehat{G}(y,x|C>0) = \sum_{i=1}^{n} 1(C_i>0)\widehat{G}(y,x,\widehat{V}_i) / \sum_{i=1}^{n} 1(C_i>0).$$

An estimator of the quantile structural function $q(\tau,x)$ can be obtain by inversion of this estimator. Each of these estimators can be interpreted as estimator for the entire population whenver the assumption discussed at the end of Section 3.2 are satisfied.

4.4 Estimation of bounds

We now present an estimator for the object discussed in Section 3 which provide bounds of partial effects for the entire population. The lower bound of $G(y,x)$ can be estimated using the following algorithm:

Algorithm 10 *1. Estimate the distribution of $V|X, C>0$ using distribution regression and denote it $\widehat{F}_{V|X,C>0}(v \mid X=x, C>0)$.*

2. Estimate the support of $V|X, C>0$ on the selected sample: $\widehat{\mathcal{V}}(x) = \text{supp}(V|X = x, C>0) = \{v \in [0,1] | \widehat{F}_{V|X,C>0}(v \mid X=x, C>0) > h_n\}$, with $h_n \downarrow 0$ for $n \to \infty$.

3. Estimate $G_l(y,x)$ using

$$\widehat{G}_l(y,x) = \frac{1/q_n}{1-2/q_n} \sum_{v=1/q_n, 2/q_n, \ldots, 1-1/q_n; v \in \widehat{\mathcal{V}}(x)} \widehat{G}(y,x,v) \tag{12}$$

where $q_n \downarrow 0$ when $n \to \infty$.

We use the same algorithm for the upper bound with the third step replaced by:

$$\widehat{G}_l(y,x) + \widehat{P}(x) =$$
$$\frac{1/q_n}{1-2/q_n} \sum_{v=1/q_n,2/q_n,\ldots,1-1/q_n} \left\{ 1(v \in (\widehat{\mathcal{V}}|X))\widehat{F}_{Y|X,V=v,C>0}(y \mid X = X_i, V = \widehat{V}_i) + 1(v \notin (\widehat{\mathcal{V}}(x)) \right\}. \tag{13}$$

Using equation (9) we can also estimate bounds of the SDF where we integrate over the distribution of X. These represents bounds that are unconditional on the outcomes of X, *i.e.* $G(y)$, can be obtained by averaging over the empirical distribution of X. Hence, the lower bound becomes:

$$\widehat{G}_l(y) = \frac{1}{n} \sum_{i=1}^{n} \widehat{G}_l(y, X_i), \tag{14}$$

while the upper bound is given by $\widehat{G}_l(y) + \sum_{i=1}^{n} P(X_i)$.

4.5 Estimation of counterfactual distributions

The object defined in (**??**) can be obtained by estimating $F_{Y_q|X_q,V_q}(y|X_q = x, V_q = v)$ for every x and v and then weighting them on the basis of estimated densities that appear at the left-hand side of (**??**). This, however, would result in an extremely complex estimator. Accordingly. we make the following rank invariance assumption of η to derive a simpler estimator of (**??**).

Assumption 4 (Rank invariance) *We assume that* $V_q = V_r = V_s$.

Rank invariance does not make assumptions regarding the distributions of η_t or the function h_t for $t = q, r, s$. It assumes that an observation that is ranked in one of the model versions has the same rank in any other model version. Using assumption 4, the

sample analog of object (**??**) equals

$$\widehat{F}_{Y_{q|r,s}}(y|C_s > 0) = \frac{\sum_{i=1}^{n_r} \mathbf{1}(\widehat{V}_{r,i} > \widehat{v}_0^s(x_{r,i}, z_{r,i}))\widehat{F}_{Y_q|X_q}(y|X_q = x_{r,i}, V_q = \widehat{V}_{r,i})}{\sum_{i=1}^{n_r} \mathbf{1}(\widehat{V}_{r,i} > \widehat{v}_0^s(x_{r,i}, z_{r,i}))}, \tag{15}$$

where the indicator function implies that we take an average over the subsample of model version r for which we have that the estimated level of V_r is above the censoring point in model version s. This estimate has many components estimated in earlier steps. Therefore, we have the following three-step procedure.

Algorithm 11 *1. Estimate both $\widehat{v}_0^s(x_{r,i}, z_{r,i})$ and $\widehat{V}_{r,i}$ for $i = 1, \ldots, n_r$ using distribution regression.*

2. Estimate $\widehat{F}_{Y_q|X_q,Z_q}(y|X_q = x_{r,i}, Z_q = z_{r,i}, V_q = \widehat{V}_{r,i})$ for $i = 1, \ldots, n_r$ using distribution regression.

3. Estimate $\widehat{F}_{Y_{q|r,s}}(y|C_s > 0)$ using equation (15).

5 Application: United Kingdom wage regressions

We illustrate our approach through two examples related to the earnings of females in the United Kingdom. First, we examine the role of selection bias from the hours decision in estimating the returns to schooling for female workers. Second, we provide a decomposition of earnings growth for the same sample in the presence of selection bias. We use data from the United Kingdom Family Expediture Survey (FES) for the years 1978 to 1999. Using the same data source Blundell, Reed and Stoker (2003) study wage growth for males while Blundell et al. (2007) examine wage inequality for both males and females. As we use their data selection rules we refer to these papers for further details. The FES is a repeated cross section of households and contains detailed information on the number of weekly

hours worked and the hourly wage of the individual. We restrict the sample to females that report an education level and only include working women who report working weekly hours of at most 70 and an hourly wage of at least 0.01 pounds. This reduces the total number of observations from 96,402 to 94,985. This produces a data set of over 4,100 observations per year and with approximately 2,600 working females. Unless otherwise stated our analysis is performed on a year by year basis.

The outcome variable is the log-hourly wage defined as the nominal weekly earnings divided by the number of hours worked and deflated by the quarterly UK retail price index. Following Blundell, Reed and Stoker (2003) we use the simulated out-of-work benefits income as an exclusion restriction in the hours equation. We refer to their paper for details and note that the UK benefits system makes the use of out-of-work benefits as an exclusion restriction appropriate since, in contrast to other European countries, unemployment benefits are not related to income prior to the period out of work. Still, Blundell et al. (2007) argue that the system of housing benefits can have a positive relationship with in-work potential. We do not consider these additional issues and refer to Blundell et al. (2007) for a potential solution using a monotonicity restriction in place of an exclusion restriction in the hours equation. Note equations (*i.e* (1) and (2)) characterize the model of Blundell, Reed and Stoker (2003) when $g(\cdot)$ and $h(\cdot)$ are linear and when ε and η are normally distributed.

5.1 Returns to education

We first estimate selection adjusted returns to education via the use of quantile regressions. The model contains two dummy variables variables capturing school leaving age (ie. the individual left school at the age of 17-18 years of age and 19 years or older with the excluded group those who left an earlier age), a quartic in age, and dummy variables

	Q1	Q2	Q3
Leaving school at the age of 17-18 years			
1979	0.136	0.178	0.235
	(0.021)	(0.024)	(0.032)
1989	0.173	0.274	0.303
	(0.022)	(0.021)	(0.032)
1999	0.211	0.303	0.292
	(0.027)	(0.031)	(0.025)
Leaving school at the age of 19 years or older			
1979	0.443	0.565	0.645
	(0.025)	(0.024)	(0.037)
1989	0.457	0.609	0.612
	(0.026)	(0.025)	(0.038)
1999	0.518	0.653	0.642
	(0.028)	(0.032)	(0.028)

Table 1: Regression results for the returns to education

denoting the individual has a partner and the region in which she lives. We report how the returns have changed over time by estimating the model for the 1979, 1989 and 1999 cross sections. The unadjusted estimates are shown in Table 1 and indicate that there is an increase in the returns to education at higher quantiles. There is also some indication that at some quantiles there is an increase in the return to education over time.

We now turn to the implementation of our procedure. We first discuss the first step estimator of the control variable. We estimate this by distribution regression and include the variables from the wage regression plus the exclusion restriction discussed above. Figure 1 reports kernel density estimates for the control variable for the different education levels and 3 different years: 1979, 1989 and 1999. Note that we estimate each year separately and plot the control variable for the appropriate subsample in each cross section. These kernel density estimates are corrected to ensure V cannot be lower than 0 (see also Guerre, Perigne and Vuong, 2000). The distribution of V is skewed to the right implying that there is the possibility of sample selection. In addition, the lower bound of the support of V for the lowest level of education is around 0.2-0.3. Thus from Lemma 3 this implies that any local estimate for this education level at or below the first quartile of the

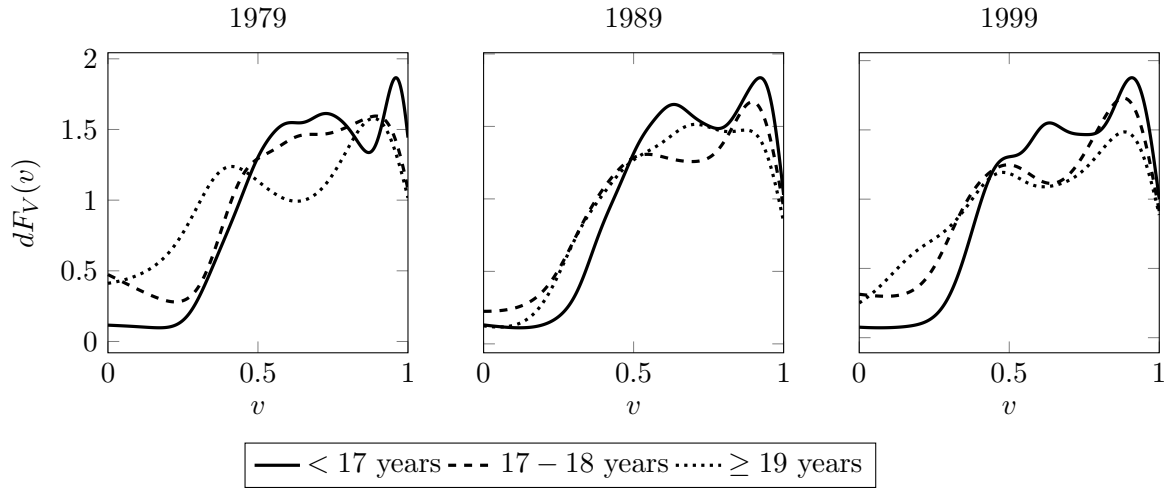

Figure 1: Kernel density estimates for the estimated control variable V for different levels of education.

control variable should not be interpreted as an object for the entire population. Such an interpretation seems less problematic for the other education levels and appears appropriate for all education levels when we consider median values or higher of the control variable.

The results for the wage equation are listed in Table 2. We capture selection by including and interact V and V^2 with the education dummies. We evaluate the results by assessing the estimate at specific values of the control function, *i.e.* the impact at $V = 0.25$ equals:

$$\beta_{17-18\text{yrs.},\tau} + \beta_{17-18\text{yrs.},V,\tau} \times 0.25 + \beta_{17-18\text{yrs.},V^2,\tau} \times 0.25^2$$

with $\beta_{17-18\text{yrs.},\tau}$ the quantile coefficient at τ for 17-18 years dummy and $\beta_{17-18\text{yrs.},V,\tau}$ and $\beta_{17-18\text{yrs.},V^2,\tau}$ are the coefficients of interaction terms of this variable with the control variable.

We find a large increase in the returns to education for all levels of V among all quantiles and for each year. Lowly educated women have a lower likelihood to participate on the

basis of their observables. Thus, among these women only those with a high draw from V participate. This is not true for the more highly educated women. If we make the plausible assumption that η and ε have a positive correlation then the selected sample contains a group of relatively high earning low educated women. Therefore, a simple comparison of the conditional quantiles between the low educated (and hence high ability) and the high educated (and hence average ability) women, as is done in standard quantile regression, results in an underestimation of the returns to education.

Bounds for the returns to education are in Table 3. These are estimated using algorithm 10 where the results for all education levels, *i.e.* the first column of Table 3, is estimated using (14). The estimated lower bound of the median, $\underline{\widehat{Q}}_{0.5}(y)$, then solves $\widehat{G}_l(\underline{\widehat{Q}}_{0.5}(y)) = 0.5$. For the other columns we calculate the same statistic as in (14) but set the education level of every individual equal to that of the corresponding column. Upper bounds are calculated in a similar manner. We compare our bounds with the worst-case-Manski bounds. It appears the Manski bounds are not informative about the impact of education in these data. For example, for 1989 the lower bound of the highest education level is even below the upper bound of the lowest education level. The same result arises for all years when comparing the individuals who left education at 17-18 years of age with those with the lowest education level. Our bounds provide a more precise description of the returns to education. Using the estimated bounds we do find that the returns to education, measured by the difference in hourly wages between any education level and the lowest level of education, was positive for all years and for all education levels. Moreover, can conclude that the returns to education increased as the lower bound for the returns to education for the highest education level for 1999 is higher than the upper bound for 1979 (*i.e.* 0.635 versus 0.624).

	Q1	Q2	Q3
Leaving school at the age of 17-18			
	$V = 0.25$		
1979	0.266	0.206	0.284
	(0.022,0.510)	(-0.095,0.507)	(-0.045,0.613)
1989	0.256	0.095	0.516
	(0.095,0.416)	(-0.149,0.339)	(0.283,0.749)
1999	0.273	0.085	0.451
	(0.076,0.469)	(-0.081,0.251)	(0.251,0.651)
	$V = 0.5$		
1979	0.233	0.196	0.307
	(0.145,0.321)	(0.091,0.300)	(0.199,0.414)
1989	0.319	0.236	0.411
	(0.246,0.393)	(0.144,0.328)	(0.321,0.502
1999	0.350	0.244	0.412
	(0.255,0.445)	(0.159,0.330)	(0.315,0.508
	$V = 0.75$		
1979	0.191	0.145	0.267
	(0.123,0.260)	(0.050,0.240)	(0.169,0.365)
1989	0.305	0.265	0.307
	(0.240,0.370)	(0.182,0.348)	(0.219,0.394)
1999	0.345	0.282	0.318
	(0.256,0.433)	(0.197,0.367)	(0.239,0.397)
Leaving school at the age of 19 or older			
	$V = 0.25$		
1979	1.119	0.991	0.993
	(0.766,1.471)	(0.681,1.301)	(0.782,1.203)
1989	0.950	0.568	1.015
	(0.658,1.243)	(0.304,0.832)	(0.698,1.332)
1999	0.850	0.476	1.058
	(0.570,1.130)	(0.226,0.727)	(0.802,1.315)
	$V = 0.5$		
1979	0.866	0.597	0.884
	(0.695,1.037)	(0.455,0.738)	(0.792,0.976)
1989	0.723	0.526	0.788
	(0.620,0.827)	(0.388,0.664)	(0.677,0.900)
1999	0.753	0.577	0.822
	(0.645,0.861)	(0.480,0.674)	(0.742,0.902)
	$V = 0.75$		
1979	0.610	0.392	0.631
	(0.474,0.746)	(0.271,0.514)	(0.542,0.720)
1989	0.535	0.474	0.562
	(0.430,0.641)	(0.359,0.589)	(0.475,0.650)
1999	0.633	0.544	0.594
	(0.526,0.740)	(0.454,0.635)	(0.502,0.685)

Table 2: Estimates of the returns to education, our method using a control function. Bootstrapped confidence intervals are in between parentheses.

	All education levels	$\leq$ 16 years	17-18 years	$\geq$ 19 years
1979				
Uncorrected median	1.414	1.351	1.592	1.993
Manski bounds at median	(1.142, 1.781)	(1.107, 1.688)	(1.271, 1.956)	(1.763, 2.199)
Our bounds	(1.255, 1.313)	(1.196, 1.265)	(1.402, 1.509)	(1.948, 2.026)
Difference lowest			(0.137, 0.313)	(0.439, 0.624)
1989				
Uncorrected median	1.645	1.520	1.800	2.171
Manski bounds at median	(1.311, 2.070)	(1.222, 1.929)	(1.548, 2.090)	(1.884, 2.400)
Our bounds	(1.411, 1.606)	(1.323, 1.479)	(1.665, 1.870)	(1.958, 2.222)
Difference lowest			(0.186, 0.547)	(0.479, 0.899)
1999				
Uncorrected median	1.851	1.639	1.949	2.343
Manski bounds	(1.494, 2.260)	(1.331, 2.109)	(1.692, 2.221)	(2.121, 2.496)
Our bounds	(1.636, 1.792)	(1.479, 1.587)	(1.791, 1.948)	(2.222, 2.349)
Difference lowest			(0.204, 0.469)	(0.635, 0.870)

Table 3: Bounds for the global median of wages for different levels of education.

5.2 Decomposition of the wage increase

We now employ our procedure to extend the analysis of Blundell, Reed and Stoker (2003) and decompose the growth of female wages while incorporating a role for selection and changes in the distribution of the wages.[2] Figure 2 reports the participation rate of women over these years. Participation was around 65 percent in the years before the recession in the beginning 1980's. The participation rate of women then dropped considerably to a low of 58 percent in 1982 and stayed low until 1985. Subsequently female participation rates have been increasing almost monotonically over time reaching a participation rate just under 70 percent at the end of the century.

Using algorithm 11 we decompose the observed changes in the distribution of wages

[2]Blundell, Reed and Stoker (2003) provide the decomposition of male wages in the same period while employing a parametric approach to account for selection.

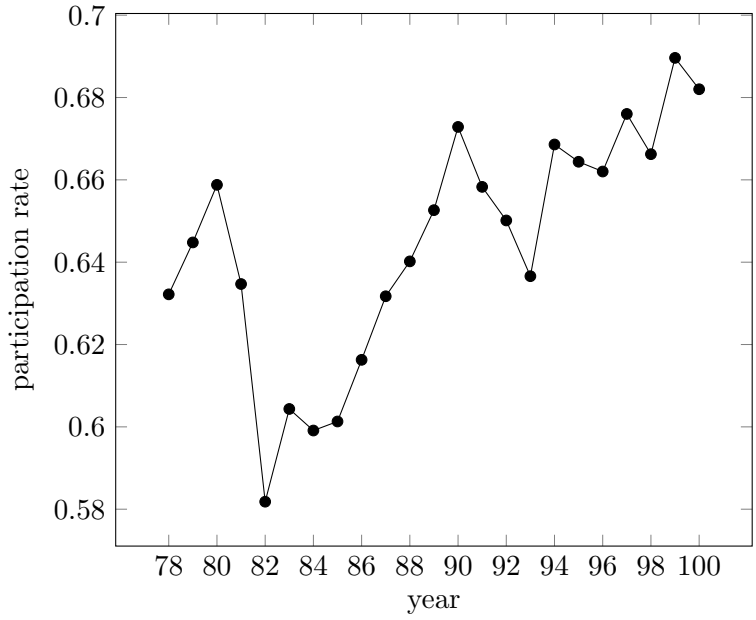

Figure 2: British females: labor market participation rate

between 1982 (year 0) and any subsequent year (t) into the following components:

$$F_{Y_{t|t,t}}(y|C_t>0) - F_{Y_{0|0,0}}(y|C_0>0) = \underbrace{F_{Y_{t|t,t}}(y|C_t>0) - F_{Y_{t|t,0}}(y|C_0>0)}_{(1)}$$

$$+ \underbrace{F_{Y_{t|t,0}}(y|C_0>0) - F_{Y_{t|0,0}}(y|C_0>0)}_{(2)} + \underbrace{F_{Y_{t|0,0}}(y|C_0>0) - F_{Y_{0|0,0}}(y|C_0>0)}_{(3)},$$

where (1) captures the difference in the wage distribution between the specified year and 1982 due to selection (the selection component); (2) reflects the difference in the wage distribution due to the composition of workers (the composition component); and (3) represents the difference in the wage distribution due to the changing wage structure (the wage structure component). We do not focus on these components but report the differences in the quantiles. For example, the selection component equals:

$$\Delta^1_\tau = \mathbb{Q}_\tau(Y_{(t|t,t)}|C_t>0) - \mathbb{Q}_\tau(Y_{(t|t,0)}|C_0>0)$$

and similarly, we introduce Δ^2_τ and Δ^3_τ.

Figure 3 shows the time series of the different components and the total difference in wages from 1982 to 1999. Similar to Blundell, Reed and Stoker (2003), who find a large change in the wage dispersion for males in this period, we find the total increase to be much larger for Q2 and Q3 than for D1 and Q1. This is especially true since 1991, where we do not find any improvements at the bottom of the distribution while the improvements at the top remain. In addition, we find the increase in the wages is primarily due to the wage structure component and this is especially true at the bottom of the distribution where the wage structure component is almost as high as the total increase. Only at the upper part of the distribution can the composition component explain a substantial part of the wage growth. Here, even though women at the upper part of the distribution did increase their wages in the 90s, this wage increase was almost completely based on their improvements of the composition of the work force and not on improvements in the wage structure. The selection component is small in absolute values and negative. The negative effect is expected as comparing all later years with 1982 we make the sample more selective. Thus, since it is likely the "best" women were working in 1982 the wages will increase by dropping the less able women in the later years from the sample. We also find that the selection effect is largest at D1 and Q1, while it is almost non-existent at Q3. This is also expected. That is, women at the top of the distribution worked both in 1982 as in any other year and therefore we may not change the sample at all by imposing the high level of selection in 1982 on the later years.

Figures 4 to 6 present the confidence intervals for the different components. These are produced via 1000 bootstrapped samples and are uniform for our three components. This implies that the upper bound of the quantiles equal $\widehat{\mathbb{Q}}(Y_{q|t,s}) + c(p)s_k$, where:

$$s_k^2 = \frac{\mathbb{Q}_{0.75}(\widehat{\Delta}^k)^* - \mathbb{Q}_{0.25}(\widehat{\Delta}^k)^*}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)},$$

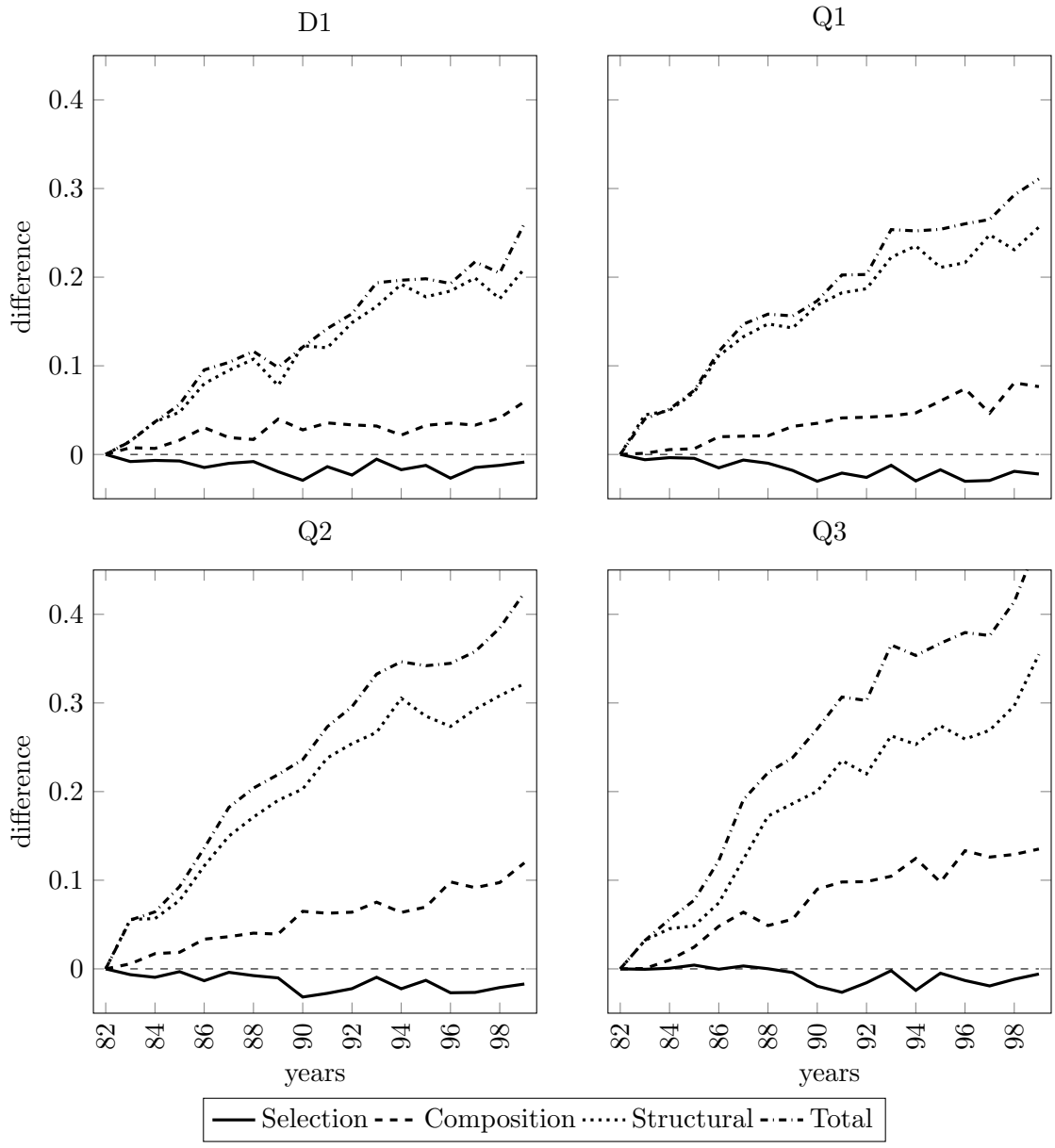

Figure 3: Contributions of the different components of the decomposition of the wage increase since 1982 at different quantiles.

with Φ the standard normal distribution and an asterisk indicates that these are bootstrapped values. The value of $c(p)$ equals:

$$c(p) = \text{p quantile of} \left(\max_{k=1,2,3} \{ |\widehat{\Delta}^{k*(j)} - \widehat{\Delta}^{k}| / s_k \} \right),$$

where the superscript (j) implies the j-th bootstrap sample. We find the confidence intervals to be wide for the selection component. It implies that only for Q1 the confidence interval lies completely below the x-axis in 1990. At all other points of the distribution, and in all other years, we cannot conclude there is a statisically significant effect from the change in selection component. This is partially due to the low level of the selection component but also because of the low number of observations. Figure 7 reproduces the selection components and its confidence intervals when we pool four years of data. We obtain more stable results and the confidence intervals are much smaller. However, we cannot reject the null hypothesis that changes in the selection component did not affect the distribution at any of the quantiles apart from Q1. We can conclude from this that the selection effect did not have a big impact on the wage distribution in the years of our analysis. Arellano and Bonhomme (2016) make similar conclusions using a model with a binary selection equation and looking at the global impact for the entire population. Even at the beginning of the 90s, where the selection component is the highest, we find that the selection component would have increased the first quartile by only a little over 2 percent based on a total increase of almost 20 percent. There are two reasons for this result. The first is that the change in labor force participation was not that drastic for women in the UK for this period of time. Second, even if we take account of the small increase in labor force participation, then the selection effect is smaller than what we would have expected in the case that only less able women would have entered the labor market. For example,

if we look at the 8 percent increase in labor force participation from 1982 to 1990, then if only less able women would have entered the labor market in that period of time, it would imply a shift of the quantile function by about 12 percent (*i.e.* 0.08 / 0.68). At the first quartile, this is a difference of 0.15 in log wage points for the raw data and this is far above the 0.02 that we predict from our model. This implies that, even though we do predict that on average the women who entered the labor market in the period 1982 to 1990 were less able than the ones that already entered the labor market before 1982, the differences in their abilities were small on average.

References

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

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

[3] Arellano, M. and S. Bonhomme (2016), "Quantile selection models', *Econometrica*, **84**, forthcoming.

[4] Blundell, R., A. Gosling, H. Ichimura, C. Meghir (2007), "Changes in the distribution of male and female wages accounting for employment composition using bounds", *Econometrica*, **75**, 323-63.

[5] Blundell, R., H. Reed and T.M. Stoker (2003), "Interpreting aggregate wage growth: the role of labor market participation", *American Economic Review*, **93**, 1114-31.

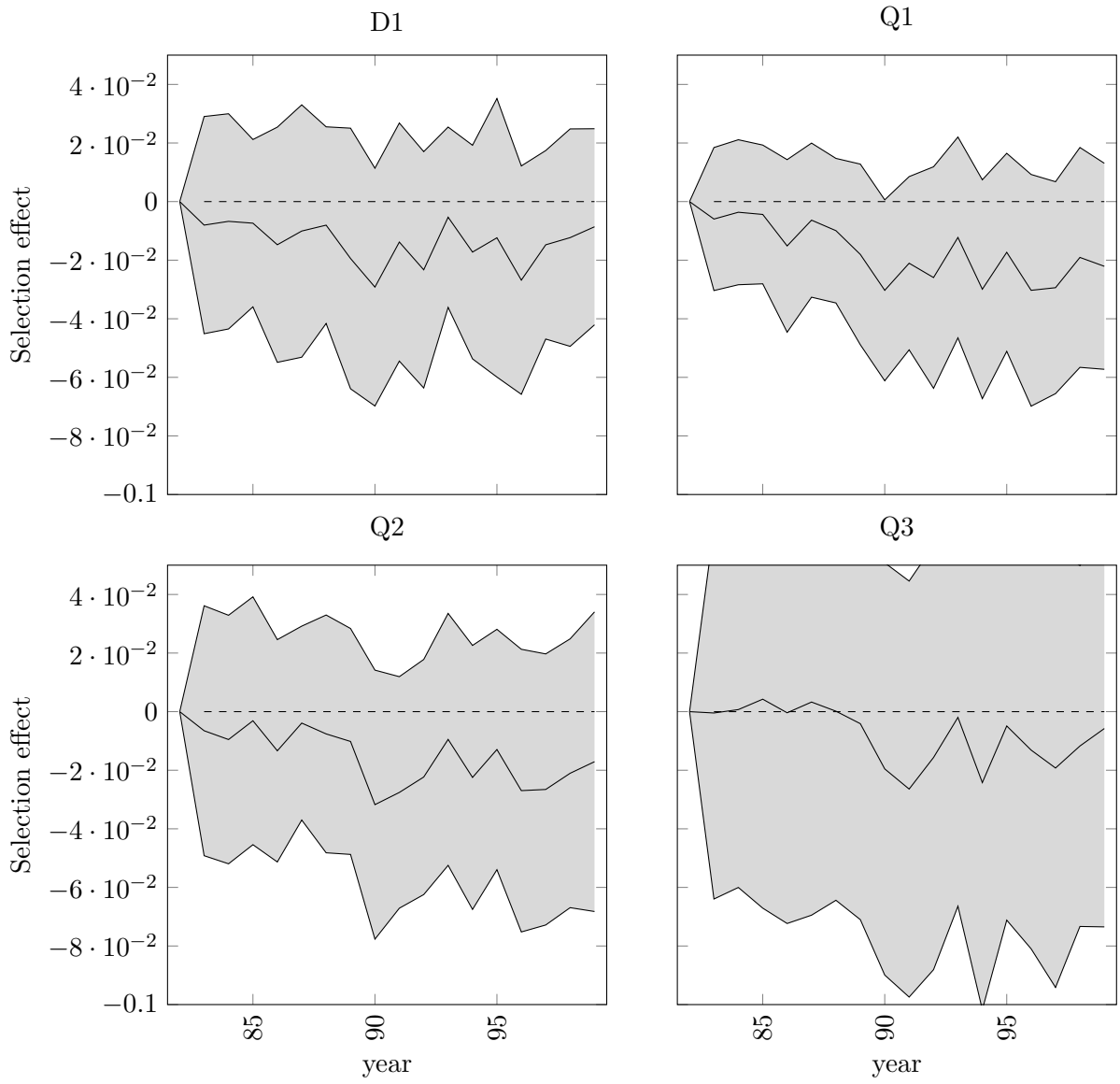

Figure 4: Upper- and lower bound of a bootstrap 95 percent confidence interval for the selection component.

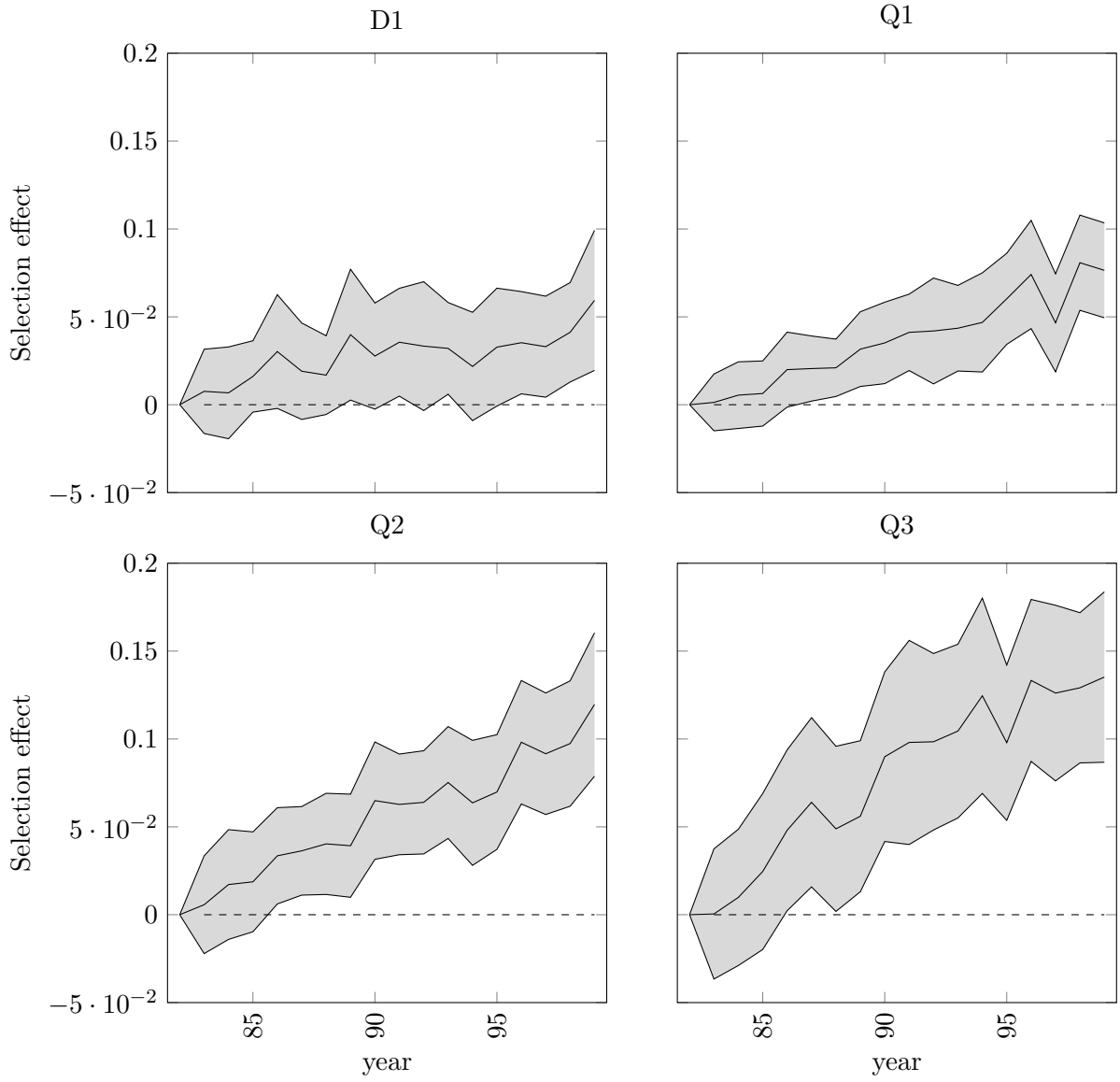

Figure 5: Upper- and lower bound of a bootstrap 95 percent confidence interval for the composition component.

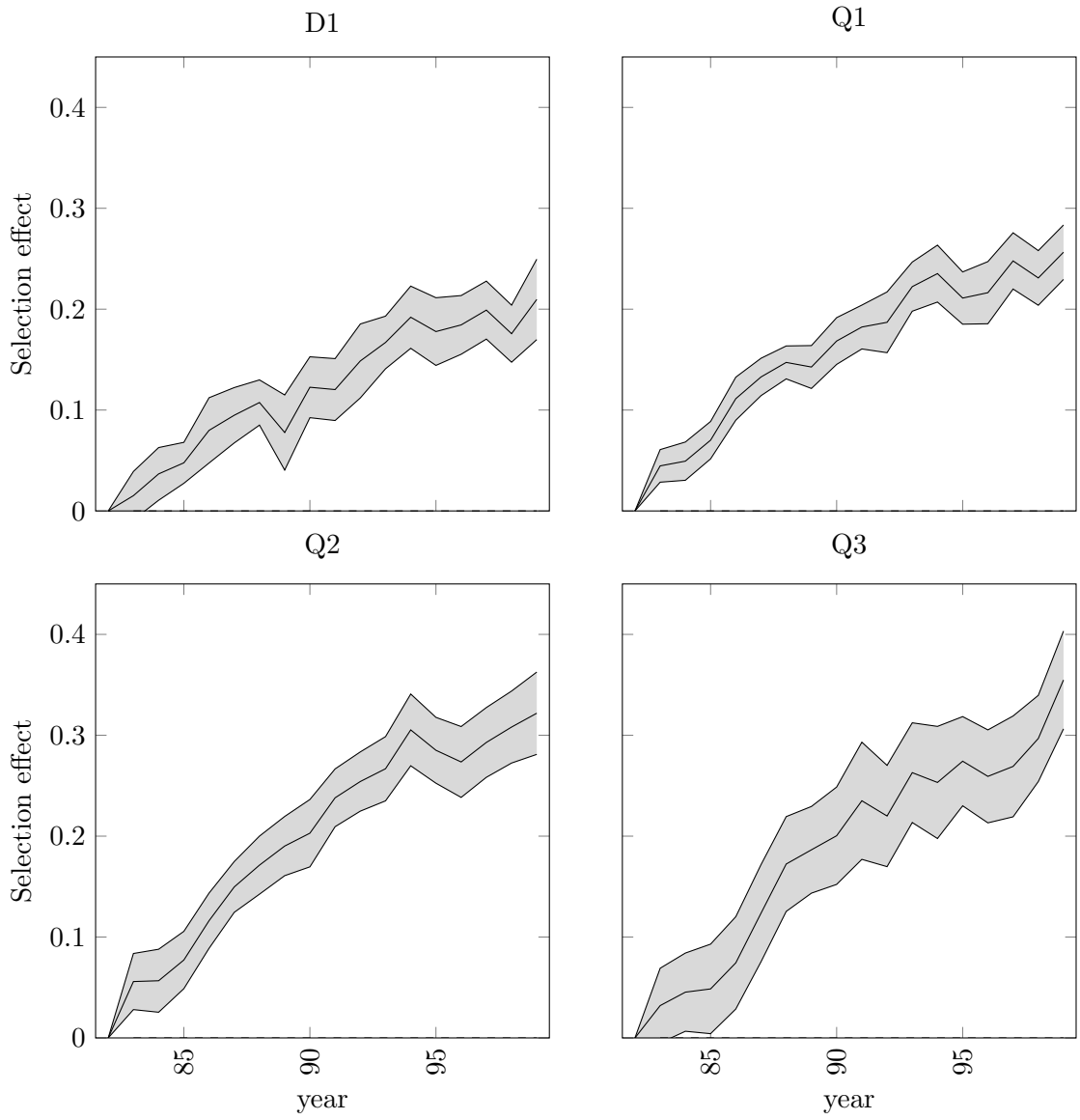

Figure 6: Upper- and lower bound of a bootstrap 95 percent confidence interval for the wage structure effect

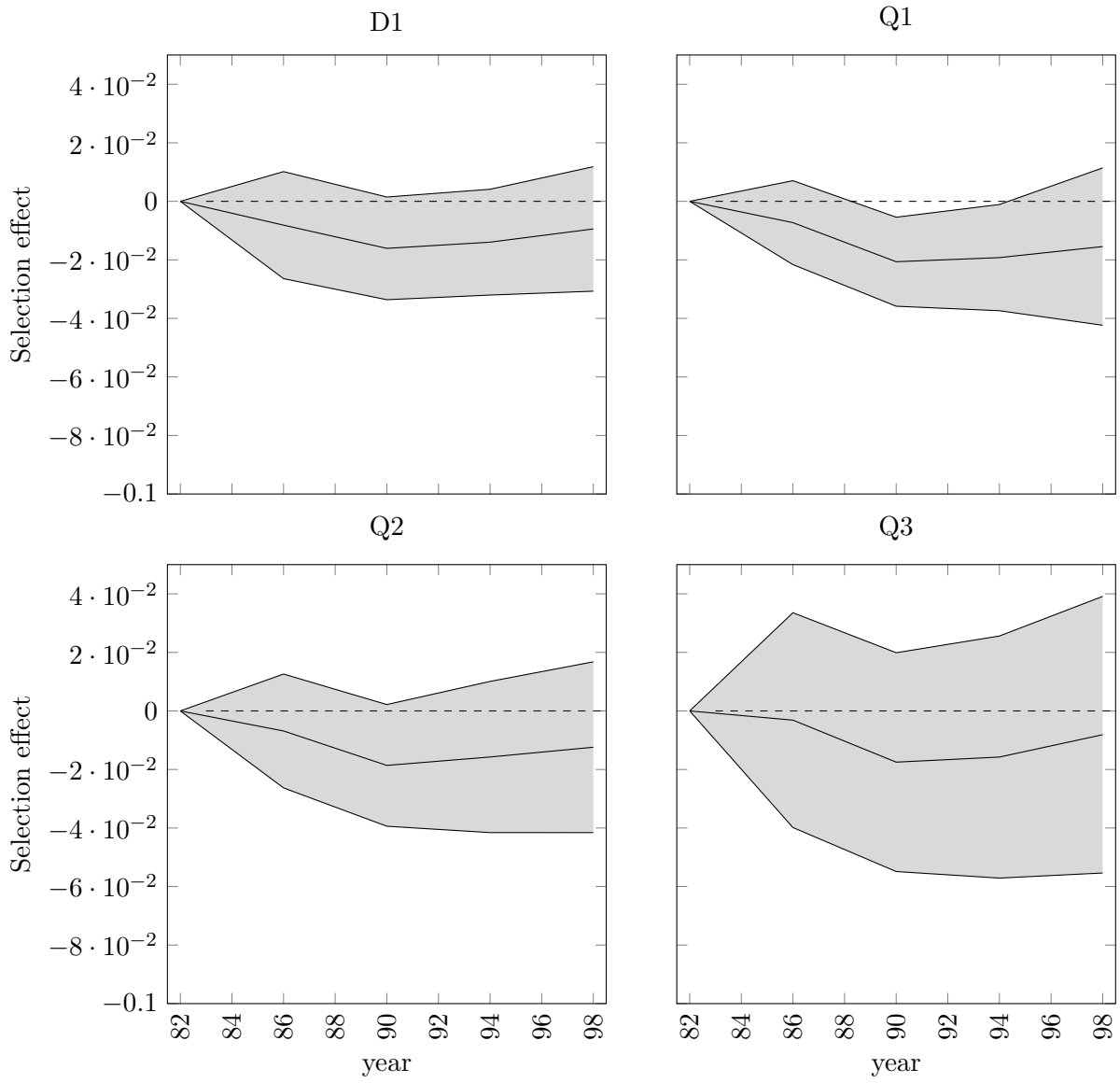

Figure 7: Upper- and lower bound of a bootstrap 95 percent confidence interval for the selection component. The results are based on pooled observations from 1982-1985, 1986-1989, ...

[6] Buchinsky, M. (1998), "The dynamics of changes in the female wage distribution in the USA: a quantile regression approach", *Journal of Applied Econometrics*, **13**, 1–30.

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

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

[9] Das. M., W. K. Newey and F. Vella (2003), "Nonparametric estimation of sample selection models", *Review of Economic Studies*, **70**, 33-58.

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

[11] Guerre, E. I. Perrigne and Q. Vuong (2000), "Optimal nonparametric estimation of first-price auctions", *Econometrica*, **68**, 525–74.

[12] Harmon, C. and I. Walker (1995), "Estimates of the economic return to education for the United Kingdom", *American Economic Review*, **85**, 1278–86.

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

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

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

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

[17] Lee, M.J, and F. Vella (2006), "A semi-parametric estimator for censored selection models with endogeneity", *Journal of Econometrics*, **130**, 235–52.

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

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

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

Appendix

A Proof of 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 $h(X, Z, \eta) > 0$), by Assumption 1,

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

Since this holds for any function $a(\varepsilon)$ and any $h(X, Z, \eta) > 0$, X and ε are independent conditional on η and $C > 0$. The result follows because η is a one-to-one function of $F_{C|X,Z}(C \mid X, Z)$ when $C > 0$ since $\eta = h^{-1}(X, Z, C)$ if $C > 0$ by Assumption 1, and for $c > 0$

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

where F_η is the cdf of η. ■

B Proof of Lemma 3

Proof. We start with the first equality in (4). Using Lemma 1, we have

$$\begin{aligned}\mathbb{E}[Y \mid X = x, V = v] &= \int g(x,e) dF_{\varepsilon|X}(e \mid X = x, V = v) \\ &= \int g(x,e) dF_{\varepsilon|V}(e \mid V = v).\end{aligned} \tag{16}$$

Next, we proof the second equality. Because of Assumption 1 and $\eta = F_\eta^{-1}(V)$, we have that

$$\begin{aligned}\delta(x,v) &= \int g(x,\varepsilon) dF_\varepsilon(e|V = v, X = x) \\ &= \int g(x,\varepsilon) dF_\varepsilon(e|\eta = F_\eta^{-1}(v), X = x) \\ &= \int g(x,\varepsilon) dF_\varepsilon(e|\eta = F_\eta^{-1}(v), X = x, Z = z),\end{aligned}$$

for every z. More in particular we have this for any z that has the restriction $h(x,z,F_\eta^{-1}(v)) > 0$. Note that the set of outcomes of Z for which this is true, is not an empty set by assumption. This implies that if this is true, we can condition on this outcome, *i.e.*

$$\delta(x,v) = \int g(x,\varepsilon) dF_\varepsilon \left(e|\eta = F_\eta^{-1}(v), X = x, Z = z, h(x,z,F_\eta^{-1}(v)) > 0\right).$$

Using the definition of C, we obtain

$$\delta(x,v) = \int g(x,\varepsilon) dF_\varepsilon(e|\eta = F_\eta^{-1}(v), X = x, Z = z, C > 0).$$

Again, using Assumption 1 and the identity $\eta = F_\eta^{-1}(V)$, we obtain

$$\delta(x,v) = \int g(x,\varepsilon) dF_\varepsilon(e|V = v, X = x, C > 0) = \delta(x,v|C > 0)$$

Proof of Lemma 8
The first equality is proven in the text. For the second equality, we have

$$\begin{aligned}F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|V = v) &= F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|V = v; X = x; Z = z) \\ &= F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|\eta = F_\eta^{-1}(v); X = x; Z = z)\end{aligned}$$

for every z. More in particular we have this for any z that has the restriction $h(x,z,F_\eta^{-1}(v)) > 0$. Note that this is not an empty set by assumption. This implies that

$$F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|V = v) = F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|\eta = F_\eta^{-1}(v); X = x; Z = z; h(x,z,F_\eta^{-1}(v)) > 0)$$

Using the definition of C, we obtain

$$F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|V = v) = F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|\eta = F_\eta^{-1}(v); X = x; Z = z; C > 0)$$

Again, using Assumption 1, we obtain

$$\begin{aligned}F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|V = v) &= F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|\eta = F_\eta^{-1}(v); C > 0) \\ &= F_{g(x,\varepsilon)|V}(g(x,\varepsilon)|V = v; C > 0)\end{aligned}$$

■

C Proof of the identity in assumptions

The distribution of $V|X=x,Z=z,C>0$ equals

$$\begin{aligned}F_{V|X,Z,C>0}(v|C>0,X=x,Z=z) &= \mathbb{P}(F_{h(x,z,\eta)}(h(x,z,\eta)|X=x,Z=z)\leq v|h(x,z,\eta)>0,X=x,Z=z)\\ &= \mathbb{P}(F_\eta(\eta)\leq v|\eta>h^{-1}(x,z,0),X=x,Z=z)\\ &= \mathbb{P}(\eta\leq F_\eta^{-1}(v)|\eta>h^{-1}(x,z,0),X=x,Z=z)\\ &= \frac{v-F_\eta(h^{-1}(x,z,0))}{1-F_\eta(h^{-1}(x,z,0))}\end{aligned} \tag{17}$$

where the first equality uses definitions, the second equality uses the fact that $F_{h(x,z,\eta)}(h(x,z,q)|X=x,Z=z)=F_\eta(q)$, the third equality is rewriting and the last equality is based on conditional probability. Equation (17) implies that the support of $V|X=x,Z=z,C>0$ equals $[F_\eta(h^{-1}(x,z,0)),1]$ and hence the support of $V|Z=z,C>0$ equals $[\inf_{z\in\mathcal{Z}(x)}F_W(h^{-1}(x,z,0)),1]$. This implies that $v\in\mathcal{V}(x)$ implies that

$$v\geq\inf_{z\in\mathcal{Z}(x)}F_\eta(h^{-1}(x,z,0))\Leftrightarrow F_\eta^{-1}(v)\geq\inf_{z\in\mathcal{Z}(x)}h^{-1}(x,z,0)\Leftrightarrow\sup_{z\in\mathcal{Z}(x)}h(x,z,F_\eta^{-1}(v))\geq 0$$

■

D Proof of Lemma 9

Since for every x and e, we have that $0\leq\mathbf{1}(g(x,e)\leq y)\leq 1$,

$$0\leq\mathbf{1}(g(x,e)\leq y)dF_{\varepsilon|V}(e|V=v)=G(y,x,v)\leq 1$$

and since $P(x)=\int_{[0,1]/\mathcal{V}(x)}dv$, we also have that

$$0\leq\int_{[0,1]/\mathcal{V}(x)}G(y,x,v)dv\leq P(x)$$

while

$$G_l(y,x)=\int_{\mathcal{V}(x)}G(y,x,v)dv$$

Summing up these two equations, we obtain

$$G_l(y,x)\leq\int_0^1 G(y,x,v)dv=G(y,x)\leq P(x)+G_l(y,x)$$

■

* There are many economic examples that justify the system of (1) and (2). For example, suppose that $g(X, \varepsilon)$ equals (log) hourly wages of an individual. As in Blundell, Reed and Stoker (2003), we can assume that this hourly wage only depends on the level of human capital possessed by the worker and where X contains observed (age, education, region) and ε contains unobserved human capital characteristics. Suppose that the worker's utility U only depends on current spending (S) and leisure (L), *i.e.* $U = U(S, L)$. Moreover, suppose that the worker is endowed with a total of T hours per period, to be divided into leisure and working time, C. Finally, we assume that the worker has non-labor income equal to S^* and obtains a level of income equal to S_0 when out of work in top of this non-labor income. With the restriction that current spending equals current income, we obtain that

$$C = \arg\max_{c} U(Yc, T - c),$$

with the restriction that $C > C_0$ and $U(YC + S^*, T - C) > U(S^* + S_0, T)$ and where C_0 is the minimum hours of work that can be provided by the worker. Standard economic analysis shows that C is a function of X, ε, S^* and the parameters of the utility function, while the participation restriction, *i.e.* $C > C_0$ is also affected by income when out of work, S_0. In addition, even if $g(X, \varepsilon)$ is a separable function, then C is non-separable unless we make very specific assumptions about the utility function. Note that our system of equations allows for the fact that both S^* and S_0 are correlated with the unobserved human capital variables contained in ε.