

# Decomposing Wage Distributions: Estimation and Inference

Sergio Firpo  
UBC and PUC-Rio

Nicole Fortin  
UBC

Thomas Lemieux  
UBC

Preliminary and incomplete, July 2005

## Abstract

The paper generalizes the standard Oaxaca-Blinder decomposition (a popular tool of policy analysis) in two important ways: 1) by applying this type of decomposition to any distributional features and not only the mean, and 2) by allowing for a much more flexible wage setting model. In this paper, we focus on wage decompositions but the method also applies to any other outcome variable. We estimate directly the elements of the decomposition instead of first estimating a structural wage-setting model, and indirectly computing the elements of the decomposition using structural parameters. We propose a two-step method to do the decomposition. In Step 1, overall wage gap is divided into a wage structure effect and a composition effect using a (propensity score) reweighting method. In Step 2, these two terms are further divided into the contribution of each individual explanatory variable using a (novel) influence function projection technique. Finally, we illustrate how the method works in practice using three empirical examples.

# 1 Introduction

Oaxaca-Blinder decompositions are a popular tool of policy analysis that have been applied to a wide range of empirical problems. Two key advantages of the approach is that it is simple to implement and easy to interpret. However, simplicity and convenience come at some costs. First, the procedure can only be used to decompose the mean of outcome variables. For example, the initial application of the decomposition was to the case of gender wage gap. More, recently, however, there has been growing interest for how the gender gap varies at different points of the wage distribution (glass ceiling, etc.). More importantly, the large increase in wage an earnings inequality in the United States over the last 25 years have raised a number of equations about how factors like education and experience affects the dispersion, as opposed to the mean, of wages and earnings.

A second, and less appreciated, limitation of the Oaxaca-Blinder decomposition is that using a simple linear specification may bias the decomposition. This important point has been illustrated Barsky et al. (2002) in the case of the black-white wage wealth gap and is illustrated in Figure 1. The problem is that wage equations have to be used to construct a counterfactual mean wage (e.g. the wage women would earn if they were paid like men). Unless the wage equation correctly captures the (potentially complicated) conditional expectation of wages given covariates, the decomposition results will be misleading. As we explain below, this problem is magnified in the case of distributional parameters besides the means where the whole conditional distribution of wages would have to be correctly specified.

In this paper, we propose an alternative two-step procedure to generalize Oaxaca-Blinder decompositions to any distributional parameters of interest. We illustrate the working of our method in the case of the wages where we are interested in separating the effect of the “wage structure” from “composition effects”. An important feature of our approach is that we directly estimate the elements of the decomposition while letting the underlying model for the wage structure as general as possible, instead of attempting to estimate a potentially complicated distributional models of wage setting. Our approach is related to the program evaluation literature where, for example, the average treatment effect is estimated without first estimating an underlying structural model. We clearly state, however, the assumptions required to interpret the "wage structure" effect as purely reflecting changes in the underlying parameters of the wage setting equation.

The first step of our approach consists of estimating the wage structure and composition effects using a reweighting approach (parametric or non-parametric). We show that the key assumption required here is that the error terms in the wage equation are ignorable. Provided that the assumption is satisfied, the underlying wage setting model can be as general as possible. The idea of your first step is very similar to DiNardo, Fortin, Lemieux (1996). Our main contribution here is to clarify the assumptions required for identification by drawing a parallel with the program evaluation literature, and providing analytical formulas for the standard errors.

In the second step, we further decompose the wage structure and composition effects into the contribution of each individual covariate, just like this is usually done with the Oaxaca-Blinder decomposition. To do so, we introduce a novel method based on linear projections of the influence function on the covariates,  $x$ . Intuitively, the influence function represents the contribution of each observation  $i$  to the the distributional statistic of interest (mean, median, etc.). This is simply  $Y_i$  ( $Y_i - \mu$  to be precise) in the case of the mean, but the influence function

can be computed for any other distributional parameter of interest.

For example, the influence function for the median is  $(1/2 - \mathbb{I}\{Y_i \leq me\})/f(me)$ . Projection of this on  $x$  is simply a rescaled linear probability for whether observation  $i$  is above or below the median. This (new) estimation method is examined in more detail in the case of quantiles in a companion paper. The method can be viewed as a way of modelling unconditional quantiles, by contrast with the traditional (conditional) quantile regressions. The usual OLS regression is both a model for conditional mean and the unconditional mean because fitted values average out to the unconditional mean. By contrast, conditional quantile regressions do not share this property, which limits their applicability for policy analysis.

To provide more intuition about how the method works, it is useful to look in more detail at the case of the mean. Consider two groups (or time periods),  $T = 0, 1$ . Imbedded into the Oaxaca-Blinder is a strong assumption about the underlying "structural" model. In particular, it is assumed, namely that for  $T = 0$ ,  $y_{0,i} = x_i\beta_0 + \varepsilon_i$ , and for  $T = 1$ ,  $y_{1,i} = x_i\beta_1 + \varepsilon_i$  where  $E(\varepsilon_i|x_i, T) = 0$ . The zero conditional mean assumption combined with the linear specification means that the conditional expectation of  $y$  given  $x$ , is linear, a strong assumption.

The overall average wage gap can then be written as

$$\begin{aligned} E(y_1|T=1) - E(y_0|T=0) &= E(x|T=1)\beta_1 - E(x|T=0)\beta_0 \\ &= E(x|T=1)[\beta_1 - \beta_0] + [E(x|T=1) - E(x|T=0)]\beta_0 \end{aligned}$$

The first term is the "wage structure effect". The second term is the "composition" effect. The two terms are then subdivided into the contribution of each of the  $k + 1$  explanatory variable (including a constant). The wage structure effect (WS) is:

$$\begin{aligned} WS &= (\beta_1^0 - \beta_0^0) + \sum_{l=1}^k E(x^l|T=1)(\beta_1^l - \beta_0^l) \\ \text{while the composition effect (CE) is given by} \\ CE &= \sum_{l=1}^k [E(x^l|T=1) - E(x^l|T=0)]\beta_0^l \end{aligned}$$

The various elements of the decomposition can be consistently estimated by first estimating the parameters ( $\beta_0$  and  $\beta_1$ ) of the wage equation by OLS and replacing the expectations by sample means. In other words,  $\beta_1$  is estimated by running OLS on the  $T = 1$  sample while  $\beta_0$  is estimated by running OLS on the  $T = 0$  sample.

Our approach does not rely on the linearity of the conditional expectation being linear. What we propose to do instead is to estimate  $\beta_0$  by OLS (linear projection) on the  $T = 0$  sample reweighted to have the same distribution of  $x$  as in period  $T = 1$ . The influence function projections method then allows us to apply the exact same procedure to any other distributional statistic.

## 2 Decomposition of Differences in Wage Distributions: Identification

In what follows we will focus on the differences in wage distributions between two groups, 1 and 0. In order to do so, we introduce in this section some notation that will be used throughout this article. Suppose we observe a random sample of  $N$  individuals after pooling the two populations 1 and 0. Each individual will be indexed by  $i = 1, \dots, N$ . Let  $T_i = 1$  if individual  $i$  is observed in group 1 and  $T_i = 0$ , if observed in group 0. For a worker  $i$ , let  $Y_{1,i}$  be the wage that would be paid in group 1 and  $Y_{0,i}$  be the wage that would be paid in group 0. As of course a given individual  $i$  is only observed in one of the two groups, we either observe  $Y_{1,i}$  or  $Y_{0,i}$  but never both. Therefore for each  $i$  we can define the observed income,  $Y_i$ , as  $Y_i = Y_{1,i} \cdot T_i + Y_{0,i} \cdot (1 - T_i)$ .

Finally, suppose there is a vector of covariates  $X \in \mathcal{X} \subset \mathbb{R}^r$  that we can observe in both groups. Hence, for the individual  $i$  we will observe  $X_i$ .

Wage determination depends on  $X$  and on some unobserved components  $\varepsilon \in \mathbb{R}^m$  through  $Y_{1,i} = g_1(X_i, \varepsilon_i)$  and  $Y_{0,i} = g_0(X_i, \varepsilon_i)$ , where  $g_1(\cdot, \cdot)$  and  $g_0(\cdot, \cdot)$  are unknown real-valued mappings: for  $j = 0, 1$ ,  $g_j : \mathbb{R}^r \times \mathbb{R}^m \rightarrow \mathbb{R}^+ \cup \{0\}$ . We call each one of those  $g_j(\cdot)$  functions the wage structure prevailing at group  $j$ . As we are not imposing any distribution assumption or specific functional form, writing  $Y_1$  and  $Y_0$  in this way does not restrict the analysis in any sense. Note that in fact, it may even be the case that for some  $j$  in  $\{0, 1\}$ ,  $g_j(\cdot, \varepsilon) = h_j(\varepsilon)$ , or in other words,  $X$  does not affect the wage structure process for a given group. We will however assume that  $(T, X, \varepsilon)$  have an unknown joint distribution but that is far from being restrictive.

Given sequence to the definitions, for a given value  $x$  of  $X$ , we define the “propensity-score” as the proportion of people in the combined population of two groups that is in group 1, given that those people have  $X = x$ , that is,  $p(x) = \Pr[T = 1 | X = x]$ . The unconditional proportion of people in group 1 is  $p$  which will be assumed to be positive.

Note that from our sample of size  $N$  of  $(Y, T, X)$  we can estimate the distributions of  $Y_1, X | T = 1$  and  $Y_0, X | T = 0$ , respectively,  $F_{Y_1, X | T}(\cdot | T = 1)$  and  $F_{Y_0, X | T}(\cdot | T = 0)$ , or in a shortened notation simply  $F_{Y_1, X | T=1}$  and  $F_{Y_0, X | T=0}$ . Without further assumptions, we cannot however estimate the distribution of  $Y_0, X | T = 1$ ,  $F_{Y_0, X | T=1}$ , which is the distribution that would have prevailed with the wage structure of group 0 but with individuals from group 1. This distribution will be a counterfactual distribution.

We are interested in analyzing the difference in wage distributions between groups 1 and 0 by looking at some finite parameters of those distributions. Let  $\nu$  be such parameter defined as a functional of wage distributions or, more generally, a functional of the conditional joint distribution of  $(Y_1, Y_0, X) | T$ , that is  $\nu : \mathcal{F}_\nu \rightarrow \mathbb{R}^k$  and  $\mathcal{F}_\nu$  is a class of distribution functions such that  $F \in \mathcal{F}_\nu$  if  $\nu(F) < +\infty$  and  $F(\cdot)$  is continuously differentiable in the support  $(Y_1, Y_0, X) | T$ , with positive first derivative,  $f(\cdot)$ . In order to simplify notation, define  $W_1 = [Y_1, X']'$ ,  $W_0 = [Y_0, X']'$ , and  $W = [Y, X']'$ . The functionals we are interested in are typically finite vectors of parameters. However, in a special case, we will also be interested in the density  $f(y_1, y_0, x | t)$ , which, given the continuity assumption, will be completely determined by  $F$  and, therefore, can be written as  $\nu(F)$  with dimension being  $k = +\infty$ .

The difference in the  $\nu$ 's between the two groups is called here the *overall wage gap* (measured in terms of  $\nu$ ) or alternatively, the  $\nu$ -total wage gap,  $\Delta_O^\nu$ :

$$\Delta_O^\nu = \nu(F_{W_1 | T=1}) - \nu(F_{W_0 | T=0}) \quad (1)$$

Note that Equation 1 is a simple difference in the parameter  $\nu$ . Interesting examples of those simple differences are several. They could be, for example, difference in means ( $E[Y_1 | T = 1] - E[Y_0 | T = 0]$ ), difference in medians, ( $me(Y_1 | T = 1) - me(Y_0 | T = 0)$ ), difference in other quantiles, difference in projection coefficients ( $(E[X \cdot X' | T = 1])^{-1} \cdot E[X \cdot Y_1 | T = 1] - (E[X \cdot X' | T = 0])^{-1} \cdot E[X \cdot Y_0 | T = 0]$ ). We can use the fact that  $X$  is potentially unevenly distributed across groups and try to control for that by decomposing Equation  $\Delta_O^\nu$  in two parts:<sup>1</sup>

$$\Delta_O^\nu = (\nu_{W_1 | T=1} - \nu_{W_0 | T=1}) + (\nu_{W_0 | T=1} - \nu_{W_0 | T=0}) \quad (2)$$

The first term of the sum is the difference in the parameter  $\nu$  between groups 1 and 0 if all individuals under consideration were from group 1. This is the effect of changes in the

<sup>1</sup>We will sometimes refer to the functional  $\nu(F_Z)$  simply as  $\nu_Z$ .

“wage structure”, which is summarized by the functions  $g_1(\cdot, \cdot)$  and  $g_0(\cdot, \cdot)$ . Therefore this first term corresponds to the effect in  $\nu$  of a change from  $g_1(\cdot, \cdot)$  to  $g_0(\cdot, \cdot)$  keeping the distribution  $(X, \varepsilon)|T = 1$ .

The second term corresponds to changes in the distribution of  $(X, \varepsilon)$ , keeping the “wage structure”  $g_0(\cdot, \cdot)$ . In particular, the change is from  $(X, \varepsilon)|T = 1$  to  $(X, \varepsilon)|T = 0$ . This is often called the composition effect.

Note that unless we impose restrictions on the distributions of  $(T, X, \varepsilon)$  we cannot say anything about the first term reflecting differences in the wage structure “controlling for the  $X$ ’s” only. Actually, if we had fixed constant only the distribution of  $X$ ’s, those differences would be confounded by the presence of unobservable components that are part of the wage determination. By the same reason, without further restrictions, we cannot say anything about the second term being the composition effect of changes on  $X$  only. Again, it will be confounded by changes in the distribution  $\varepsilon$  across groups as well.

Under identification assumptions however,  $\Delta_{\mathcal{O}}^{\nu}$  can be written as the sum of two identifiable effects,  $\Delta_{\mathcal{S}}^{\nu}$ , which corresponds to changes in wage structure keeping the distribution of  $X$  fixed as of  $F_{X|T=1}$ ; and  $\Delta_{\mathcal{C}}^{\nu}$ , the composition effect of changing the distribution of  $X$  from  $F_{X|T=1}$  to  $F_{X|T=0}$ . In order to identify those decomposition terms, we need the following condition to hold:

**ASSUMPTION 1** *Let  $(T, X, \varepsilon)$  have a joint distribution. For all  $x$  in  $\mathcal{X}$ :  $\varepsilon$  is independent of  $T$  given  $X = x$ .*

Some comments on Assumption 1 follow. First, this assumption has become popular in empirical research after a series of papers by Rubin and coauthors and by Heckman and coauthors<sup>2</sup>. In the program evaluation literature, this assumption is sometimes called *unconfoundedness* and allows identification of the treatment effect on the treated sub-population. Assumption 1 should be analyzed in a case-by-case situation, as for many exercises it is plausible to hold. In our case, it states that the distribution of the unobserved explaining factors of the wage determination is the same across groups 1 and 0 once we condition on a vector of observed components.

We also make an overlap assumption:

**ASSUMPTION 2** *For all  $x$  in  $\mathcal{X}$ ,  $p(x) < 1$ . Furthermore,  $0 < \Pr[T = 1]$ .*

Under Assumption 2 there is overlap in observable characteristics across groups, in the sense that it does not exist a value of  $x$  in  $\mathcal{X}$  such that it is only observed among people in group 1.

We define three weighing functions that will be very useful in our identification results:

$$\begin{aligned}\omega_1(T) &= \frac{T}{p} \\ \omega_0(T) &= \frac{1-T}{1-p} \\ \omega_{0|1}(T, X) &= \left( \frac{p(X)}{1-p(X)} \right) \cdot \left( \frac{1-p}{p} \right) \cdot \frac{1-T}{1-p}.\end{aligned}$$

We are now able to state the following result on identification of  $\Delta_{\mathcal{S}}^{\nu}$  and  $\Delta_{\mathcal{C}}^{\nu}$ .

---

<sup>2</sup>See, for instance, Rosenbaum and Rubin (1983 and 1984), Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd, (1998).

LEMMA 1 *Under Assumptions 1 and 2,*

(i)

$$\begin{aligned}
F_{W_1|T=1}(w) &= E \left[ \omega_1(T) \cdot \prod_{l=1}^{r+1} \mathbb{I}\{W_l \leq w_l\} \right] \\
F_{W_0|T=0}(w) &= E \left[ \omega_0(T) \cdot \prod_{l=1}^{r+1} \mathbb{I}\{W_l \leq w_l\} \right] \\
F_{W_0|T=1}(w) &= E \left[ \omega_{0|1}(T, X) \cdot \prod_{l=1}^{r+1} \mathbb{I}\{W_l \leq w_l\} \right]
\end{aligned}$$

(ii)  $\Delta_S^\nu = \nu_{W_1|T=1} - \nu_{W_0|T=1}$  and  $\Delta_C^\nu = \nu_{W_0|T=1} - \nu_{W_0|T=0}$  are identifiable.<sup>3</sup>

At this point, is worth having a simple example for illustration purposes. Consider the difference in means  $\mu_{Y_1|T=1} - \mu_{Y_0|T=0}$ :

$$\mu_{Y_1|T=1} - \mu_{Y_0|T=0} = \left( \mu_{Y_1|T=1} - \mu_{Y_0|T=1} \right) + \left( \mu_{Y_0|T=1} - \mu_{Y_0|T=0} \right)$$

and let us focus on each part of the previous sum separately:

$$\begin{aligned}
\mu_{Y_1|T=1} - \mu_{Y_0|T=1} &= \mathbb{E}[g_1(X, \varepsilon) | T = 1] - \mathbb{E}[g_0(X, \varepsilon) | T = 1] \\
&= \mathbb{E}[g_1(X, \varepsilon) | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 1] | T = 1]
\end{aligned}$$

but by Assumption 1:

$$\begin{aligned}
&\mathbb{E}[g_1(X, \varepsilon) | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 1] | T = 1] \\
&= \mathbb{E}[g_1(X, \varepsilon) | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1]
\end{aligned}$$

and expanding it using integrals:

$$\begin{aligned}
&\mathbb{E}[g_1(X, \varepsilon) | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1] \\
&= \int \left( \int (g_1(x, \varepsilon) - g_0(x, \varepsilon)) dF_{\varepsilon|X}(\varepsilon | x) \right) dF_{X|T=1}(x | T = 1)
\end{aligned}$$

so it is clear that such difference is attributable only to differences (in a average or mean sense) in the  $g_j(\cdot)$ 's. The expression  $\int (g_1(x, \varepsilon) - g_0(x, \varepsilon)) dF_{\varepsilon|X}(\varepsilon | x)$  is the so-called ‘‘average partial effect’’ at a given level  $X = x$  of migrating from group 0 to group 1. Its integral using the distribution of covariates at  $T = 1$  yields the average wage gap.

Now, going back to previous expression, note that

$$\begin{aligned}
&\mathbb{E}[g_1(X, \varepsilon) | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1] \\
&= \mathbb{E}[Y_1 | T = 1] - \mathbb{E}[\mathbb{E}[Y_0 | X, T = 0] | T = 1]
\end{aligned}$$

---

<sup>3</sup>Note that even if  $g_1(\cdot, \varepsilon) = h_1(\varepsilon)$  and  $g_0(\cdot, \varepsilon) = h_0(\varepsilon)$  the result from Lemma 1 is unaffected. The intuition is that since  $(X, \varepsilon)$  have a joint distribution, we can use the available information on that distribution to reweigh the effect of the  $\varepsilon$ 's on  $Y$ .

which from definition of  $Y_1$  and  $Y_0$ . Now, using  $Y = Y_1 \cdot T + Y_0 \cdot (1 - T)$ :

$$\begin{aligned} & \mathbb{E}[Y_1 | T = 1] - \mathbb{E}[\mathbb{E}[Y_0 | X, T = 0] | T = 1] \\ &= \mathbb{E}[Y | T = 1] - \mathbb{E}[\mathbb{E}[Y | X, T = 0] | T = 1] \\ &= \mathbb{E}\left[\frac{T}{p} \cdot Y\right] - \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1 - T}{1 - p(X)}\right) \cdot Y | X\right] | T = 1\right] \end{aligned}$$

and the last equality follows by using twice the law of iterated expectations and Assumption 2. Now, using the Bayes' rule, the weights definition, and dividing and multiplying by  $1 - p$ , we have:

$$\begin{aligned} & \mathbb{E}\left[\frac{T}{p} \cdot Y\right] - \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1 - T}{1 - p(X)}\right) \cdot Y | X\right] | T = 1\right] \\ &= \mathbb{E}[\omega_1(T) \cdot Y] - \mathbb{E}\left[\frac{p(X)}{p} \cdot \mathbb{E}\left[\left(\frac{1 - T}{1 - p(X)}\right) \cdot Y | X\right]\right] \\ &= \mathbb{E}[\omega_1(T) \cdot Y] - \mathbb{E}\left[\left(\frac{p(X)}{1 - p(X)}\right) \cdot \left(\frac{1 - p}{p}\right) \cdot \mathbb{E}\left[\left(\frac{1 - T}{1 - p}\right) \cdot Y | X\right]\right] \\ &= \mathbb{E}[\omega_1(T) \cdot Y] - \mathbb{E}[\omega_{01}(T, X) \cdot Y] \end{aligned}$$

which proves that  $\mu_{Y_1|T=1} - \mu_{Y_0|T=1}$ , or simply  $\Delta_S^\nu$  is expressible in terms of observed data  $(Y, T, X)$ .

Now, we want to see that and that  $\mu_{Y_0|T=1} - \mu_{Y_0|T=0}$  can also be identified from the data and that it reflects only differences in the distribution (in this case, in an average sense) of the  $X$ 's. result. Using the law of iterated expectations and the ignorability assumption:

$$\begin{aligned} \mu_{Y_0|T=1} - \mu_{Y_0|T=0} &= \mathbb{E}[g_0(X, \varepsilon) | T = 1] - \mathbb{E}[g_0(X, \varepsilon) | T = 0] \\ &= \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 1] | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1] \\ &\quad + \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 0] \\ &= \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X] | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X] | T = 1] \\ &\quad + \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 0] \\ &= 0 + \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 1] - \mathbb{E}[\mathbb{E}[g_0(X, \varepsilon) | X, T = 0] | T = 0] \end{aligned}$$

and since  $\mathbb{E}[g_0(X, \varepsilon) | X, T = 0]$  is a function of  $X$  only, the last expression captures only differences in the distributions of  $X|T = 1$  and  $X|T = 0$ , fixing the wage structure as the one of group 0 (by using the  $g_0(\cdot)$  function).

A comment on the general result. Non-parametric identification of either the income structure functions  $g_1(\cdot, \cdot)$  and  $g_0(\cdot, \cdot)$ , or the distribution function of  $\varepsilon$  are not necessary for the effects  $\Delta_S^\nu$  and  $\Delta_C^\nu$  to be identified. Therefore, methods based on conditional mean restrictions (the Oaxaca-Blinder extension approach) and methods based on conditional quantile restrictions (the Machado-Mata approach) are based on too strong identification conditions that can be easily relaxed if the final interest is in the terms  $\Delta_S^\nu$  and  $\Delta_C^\nu$ .

We can now turn our attention to estimation and inference. We will consider two separate cases. In the first one, the weighting functions are parametrically estimated. In the second case, we estimate those weighting functions non-parametrically, as they will be functions of observed  $(T, X)$ . In both cases, we derive the distribution theory of final estimators.

### 3 Decomposition of Differences in Wage Distributions: Estimation and Inference

We proceed in this section in the following way. In subsection 3.1, we discuss how one would estimate each of the previous parameters by means of a two-step approach. Then in subsection 3.2 we discuss the asymptotic behavior of their estimators.

#### 3.1 First Step Estimation

We now present the estimation procedure for each one of the following three quantities  $\nu(F_{W_1|T=1})$ ,  $\nu(F_{W_0|T=0})$  and  $\nu(F_{W_0|T=1})$ . Note that for the first two quantities, estimation is very standard as the distributions of  $W_1|T=1$  and  $W_0|T=0$  are identified from data on  $(Y, T, X)$ . The one that requires special attention is estimation of  $\nu(F_{W_0|T=1})$  but that can be dealt with the reweighting scheme under the ignorability assumption. For implementation of the reweighting procedure, we need to compute the weighting function. We now present two usual methods to deal with that problem.

##### 3.1.1 Estimating the Weights

We are interested in estimating weights  $\omega$  that are generally functions of the distribution of  $(T, X)$ . The three weighting functions under consideration are  $\omega_{1|1}(T)$ ,  $\omega_{0|0}(T)$ , and  $\omega_{0|1}(T, X)$ . The first two weights are trivially estimated by:

$$\begin{aligned}\hat{\omega}_{1|1}(T) &= \frac{T}{\hat{p}} \\ \hat{\omega}_{0|0}(T) &= \frac{1-T}{1-\hat{p}}\end{aligned}$$

where  $\hat{p} = N^{-1} \sum_{i=1}^N T_i$ .

The weighting function  $\omega_{0|1}(T, X)$  can be estimated by:

$$\hat{\omega}_{0|1}(T, X) = \frac{1-T}{1-\hat{p}} \cdot \left( \frac{1-\hat{p}}{\hat{p}} \right) \cdot \left( \frac{\hat{p}(X)}{1-\hat{p}(X)} \right)$$

The issue is how to estimate the probability of being in group 1 given  $X$ . Consider first a parametric approach.

**Parametric propensity score estimation** Suppose that  $p(X)$  is correctly specified up to a finite vector of parameters  $\delta_0$ . That is,  $p(X) = p(X; \delta_0)$ . Estimation of  $\delta_0$  follows by maximum likelihood:

$$\hat{\delta}_{MLE} = \arg \max_{\delta} \sum_{i=1}^N T_i \cdot \log(p(X; \delta)) + (1 - T_i) \cdot \log(1 - p(X; \delta))$$

Define, the derivative of  $p(X; \delta)$  with respect to  $\delta$  as  $\dot{p}(X; \delta) = \partial p(X; \delta) / \partial \delta$ . The score function  $s(T, X; \delta)$  will be:

$$s(T, X; \delta) = \dot{p}(X; \delta) \cdot \frac{T - p(X; \delta)}{p(X; \delta) \cdot (1 - p(X; \delta))}$$

following a normalization argument, we suppress the entry for  $\delta$  whenever a function of it is evaluated at the true  $\delta$ . Therefore,

$$s(T, X; \delta_0) = s(T, X) = \dot{p}(X) \cdot \frac{T - p(X)}{p(X) \cdot (1 - p(X))}$$

and finally

$$\hat{\omega}_{0|1}(T, X) = \frac{1 - T}{1 - \hat{p}} \cdot \left( \frac{1 - \hat{p}}{\hat{p}} \right) \cdot \left( \frac{p(X; \hat{\delta}_{MLE})}{1 - p(X; \hat{\delta}_{MLE})} \right)$$

**Nonparametric propensity score estimation** Suppose that  $p(X)$  is completely unknown to the researcher. In that case, following Hirano, Imbens and Ridder (2003), we approximate the log odds ratio by a polynomial series. In practice, this is done by finding a vector  $\hat{\pi}$  that is the solution of the following problem:

$$\hat{\pi} = \arg \max_{\pi} \sum_{i=1}^N T_i \cdot \log(L(H_K(X_i)' \pi)) + (1 - T_i) \cdot \log(1 - L(H_K(X_i)' \pi))$$

where  $L: \mathbb{R} \rightarrow \mathbb{R}$ ,  $L(z) = (1 + \exp(-z))^{-1}$ ; and  $H_K(x) = [H_{K,j}(x)]$  ( $j = 1, \dots, K$ ), a vector of length  $K$  of polynomial functions of  $x \in \mathcal{X}$  satisfying the following properties: (i)  $H_K: \mathcal{X} \rightarrow \mathbb{R}^K$ ; and (ii)  $H_{K,1}(x) = 1$ . If we want  $H_K(x)$  to include polynomials of  $x$  up to the order  $n$ , then it is sufficient to choose  $K$  such that  $K \geq (n + 1)^R$ . The non-parametric estimation comes from the fact that such approximation is refined as the sample size increases, that is,  $K$  will be a function of the sample size  $N$ ,  $K = K(N) \rightarrow +\infty$  as  $N \rightarrow +\infty$ .

In that approach,  $p(x)$  is estimated by  $\hat{p}(x) = L(H_K(x)' \hat{\pi})$ , thus:

$$\hat{\omega}_{0|1}(T, X) = \frac{1 - T}{1 - \hat{p}} \cdot \left( \frac{1 - \hat{p}}{\hat{p}} \right) \cdot \left( \frac{L(H_K(x)' \hat{\pi})}{1 - L(H_K(x)' \hat{\pi})} \right)$$

In Appendix we state a set of assumptions that will guarantee uniform convergence of  $\hat{\omega}(\cdot)$ , which will be used in deriving the asymptotic properties of our final estimators.

### 3.1.2 Reweighting Estimates of Distributions

We are interested in the estimation and inference of  $\nu_{W_1|T=1}$ ,  $\nu_{W_0|T=0}$ ,  $\nu_{W_0|T=1}$  and their differences.  $\Delta_O^\nu$ ,  $\Delta_S^\nu$  and  $\Delta_C^\nu$ . It can be shown that under certain regularity conditions, estimators of those objects will be distributed asymptotically normally. We now show how to estimate those quantities and derive their asymptotic distributions for two major cases: finite vector of parameters, just called  $\nu$  and the density  $f(\cdot)$ .

Start with the finite vector of parameters. Estimation follows by a plug-in approach. The quantity  $\nu_Z = \nu(F_Z)$  is just the functional  $\nu$  evaluated at the distribution  $F_Z$ . Replacing the c.d.f. by the empirical distribution function produces the estimators of interest:  $\hat{\nu}_{Y_1|T=1} =$

$\nu\left(\widehat{F}_{Y_1|T=1}\right); \widehat{\nu}_{Y_0|T=0} = \nu\left(\widehat{F}_{Y_0|T=0}\right); \widehat{\nu}_{Y_0|T=1} = \nu\left(\widehat{F}_{Y_0|T=1}\right)$  where

$$\begin{aligned}\widehat{F}_{W_1|T=1}(w) &= N^{-1} \sum_{i=1}^N \widehat{\omega}_1(T_i) \cdot \prod_{l=1}^{r+1} \mathbb{I}\{W_{i,l} \leq w_l\} \\ \widehat{F}_{W_0|T=0}(w) &= N^{-1} \sum_{i=1}^N \widehat{\omega}_0(T_i) \cdot \prod_{l=1}^{r+1} \mathbb{I}\{W_{i,l} \leq w_l\} \\ \widehat{F}_{W_0|T=1}(w) &= N^{-1} \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i, X_i) \cdot \prod_{l=1}^{r+1} \mathbb{I}\{W_{i,l} \leq w_l\}\end{aligned}$$

To be concrete, consider three examples: the mean  $\mu$ , the median  $me$ , and the variance  $\sigma^2$  of the conditional distributions  $Y_1|T=1$ ,  $Y_0|T=0$  and  $Y_0|T=1$ . Starting with the mean:  $\widehat{\mu}_{Y_1|T=1} = N^{-1} \sum_{i=1}^N \widehat{\omega}_{1|1}(T_i) \cdot Y_i$ ;  $\widehat{\mu}_{Y_0|T=0} = N^{-1} \sum_{i=1}^N \widehat{\omega}_{0|0}(T_i) \cdot Y_i$ ; and  $\widehat{\mu}_{Y_0|T=1} = N^{-1} \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i) \cdot Y_i$ . The median will be estimated by:  $\widehat{me}_{Y_1|T=1} = \arg \min_q \sum_{i=1}^N \widehat{\omega}_{1|1}(T_i) \cdot |Y_i - q|$ ;  $\widehat{me}_{Y_0|T=0} = \arg \min_q \sum_{i=1}^N \widehat{\omega}_{0|0}(T_i) \cdot |Y_i - q|$ ; and  $\widehat{me}_{Y_0|T=1} = \arg \min_q \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i) \cdot |Y_i - q|$ . Finally, the variance will be:  $\widehat{\sigma}_{Y_1|T=1}^2 = N^{-1} \sum_{i=1}^N \widehat{\omega}_{1|1}(T_i) \cdot \left(Y_i - \widehat{\mu}_{Y_1|T=1}\right)^2$ ;  $\widehat{\sigma}_{Y_0|T=0}^2 = N^{-1} \sum_{i=1}^N \widehat{\omega}_{0|0}(T_i) \cdot \left(Y_i - \widehat{\mu}_{Y_0|T=0}\right)^2$ ; and  $\widehat{\sigma}_{Y_0|T=1}^2 = N^{-1} \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i) \cdot \left(Y_i - \widehat{\mu}_{Y_0|T=1}\right)^2$ .

The densities  $f_{W_1|T}(w|t)$  and  $f_{W_0|T}(w|t)$  cannot be estimated by the plug-in method using  $\nu\left(\widehat{F}\right)$  since  $\widehat{F}$  does not have a density and, therefore,  $\nu\left(\widehat{F}\right)$  is not defined. However, estimation follows by a weighted kernel density estimation. Note that three densities of our interest can be defined as the following limiting functionals  $f_{W_j|T}(w|t=s) = \lim_{h \rightarrow 0} \left\{ h^{-1} \cdot E \left[ K_h(W_j - w) | T = s \right] \right\}$ , where  $j, s = 0, 1$ ,  $K_h(\cdot)$  is a  $\mathbb{R}^{r+1} \rightarrow \mathbb{R}$  kernel function such that  $K_h(W_j - w) = h^{-1} \cdot K\left(\frac{W_{j,1}-w_1}{h_1}, \frac{W_{j,2}-w_2}{h_2}, \dots, \frac{W_{j,r+1}-w_{r+1}}{h_{r+1}}\right)$ ,  $h = \prod_{l=1}^{r+1} h_l$  and  $[h_1, \dots, h_{r+1}]'$  is an  $r+1$  vector of bandwidths. The expectations inside the limits can be identified as they are just particular cases of Theorem 1:  $f_{W_j|T}(w|t=s) = \lim_{h \rightarrow 0} \left\{ h^{-1} \cdot E \left[ \omega_{j|s}(T_i, X_i) \cdot K_h(W - w) \right] \right\}$ , for  $j, s = 0, 1$ . The associated estimators are, therefore:

$$\begin{aligned}\widehat{f}_{W_1|T}(w|t=1) &= N^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_{1|1}(T_i) \cdot K_h(W_i - w) \\ \widehat{f}_{W_0|T}(w|t=0) &= N^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_{0|0}(T_i) \cdot K_h(W_i - w) \\ \widehat{f}_{W_0|T}(w|t=1) &= N^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i, X_i) \cdot K_h(W_i - w)\end{aligned}$$

Such approximation of  $f(\cdot)$  by  $\widehat{f}(\cdot)$  is refined as the sample size increases, that is,  $h$  will be a function of the sample size  $N$ ,  $h = h(N) \rightarrow 0$  as  $N \rightarrow +\infty$ .

### 3.1.3 Asymptotic Distribution

We now show that the plug-in estimators  $\widehat{\nu}$  are asymptotically normal and compute their asymptotic variances. We start assuming that the estimators  $\widehat{\nu}$  are asymptotically linear in the following sense:

ASSUMPTION 3 (ASYMPTOTIC LINEARITY) *If  $\{Z_1, Z_1, \dots, Z_N\}$ , a random sample of size  $N$  from  $F_Z$  were available then the plug-in estimator  $\nu(\widehat{F}_Z)$  of  $\nu(F_Z)$  could be expressed as  $\nu(\widehat{F}_Z) = \nu(F_Z) + N^{-1} \sum_{i=1}^N \psi^\nu(Z_i; F_Z) + o_p(1/\sqrt{N})$ , where  $\widehat{F}_Z(z) = N^{-1} \sum_{i=1}^N \mathbb{I}\{Z_i \leq z\}$ .*

There are estimators of functionals of the data distribution that are exactly linear, as those that are based on sample moments. There are others that can be linearized and the remainder term will approach zero as the sample size increases. One example of an estimator in that class is the sample median. The influence function of the sample median,  $\psi^{me}$  is known to be  $\psi^{me}(z; F_Z) = \left(\frac{dF_Z(z)}{dz}\bigg|_{z=me}\right)^{-1} \cdot (1/2 - \mathbb{I}\{z \leq me\})$ . For the other examples already discussed, the respective influence functions are  $\psi^\mu(z; F_Z) = z - \mu$ , for the mean, and  $\psi^{\sigma^2}(z; F_Z) = (z - \int \alpha \cdot dF_Z(\alpha))^2 - \sigma^2$  for the variance.

Under ignorability the estimators  $\widehat{\nu}_{W_1|T=1}$ ,  $\widehat{\nu}_{W_0|T=0}$ , and  $\widehat{\nu}_{W_0|T=1}$ , proposed in a previous section will remain asymptotically linear. We consider two cases: Parametric and non-parametric first step:

THEOREM 1 (a) *Under the above assumptions:*

- (i)  $\sqrt{N} \cdot (\widehat{\nu}_{W_1|T=1} - \nu_{W_1|T=1}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \omega_{1|1}(T_i) \cdot \psi^\nu(Y_i, X_i; F_{W_1|T=1}) + o_p(1) \xrightarrow{D} N(0, V_{1|1})$ ,
- (ii)  $\sqrt{N} \cdot (\widehat{\nu}_{W_0|T=0} - \nu_{W_0|T=0}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \omega_{0|0}(T_i) \cdot \psi^\nu(Y_i, X_i; F_{W_0|T=0}) + o_p(1) \xrightarrow{D} N(0, V_{0|0})$
- (iii) (a) *if in addition we assume [parametric], then:*

$$\begin{aligned} \sqrt{N} \cdot (\widehat{\nu}_{W_0|T=1} - \nu_{W_0|T=1}) &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \omega_{0|1}(T_i, X_i) \cdot \psi^\nu(Y_i, X_i; F_{W_0|T=1}) \\ &+ (\omega_{1|1}(T_i) - \omega_{0|1}(T_i, X_i)) \cdot \frac{\dot{p}(X_i)'}{p(X_i)} \cdot (E[s(T, X) \cdot s(T, X)'])^{-1} \\ &\cdot E \left[ \frac{\dot{p}(X)}{1-p(X)} \cdot E[\psi^\nu(Y, X; F_{W_0|T=1}) \mid X, T=0] \right] + o_p(1) \xrightarrow{D} N(0, V_{0|1,P}) \end{aligned}$$

- (iii) (b) *otherwise, if in addition we assume [non-parametric], then:*

$$\begin{aligned} \sqrt{N} \cdot (\widehat{\nu}_{W_0|T=1} - \nu_{W_0|T=1}) &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \omega_{0|1}(T_i, X_i) \cdot \psi^\nu(Y_i, X_i; F_{W_0|T=1}) \\ &+ (\omega_{1|1}(T_i) - \omega_{0|1}(T_i, X_i)) \cdot E[\psi^\nu(Y, X; F_{W_0|T=1}) \mid X_i, T=0] + o_p(1) \xrightarrow{D} N(0, V_{0|1,NP}) \end{aligned}$$

where

$$\begin{aligned} V_{1|1} &= E \left[ (\omega_{1|1}(T) \cdot \psi^\nu(Y, X; F_{W_1|T=1}))^2 \right] \\ V_{0|0} &= E \left[ (\omega_{0|0}(T) \cdot \psi^\nu(Y, X; F_{W_0, X|T=0}))^2 \right] \end{aligned}$$

$$V_{0|1,P} = E \left[ \left( \omega_{0|1}(T, X) \cdot \psi^\nu(Y, X; F_{W_0|T=1}) \right. \right. \\ \left. \left. + (\omega_{1|1}(T) - \omega_{0|1}(T, X)) \cdot \frac{\dot{p}(X)'}{p(X)} \cdot (E[s(T, X) \cdot s(T, X)'])^{-1} \right. \right. \\ \left. \left. \cdot E \left[ \frac{\dot{p}(X)}{1-p(X)} \cdot E[\psi^\nu(Y, X; F_{W_0|T=1}) \mid X, T=0] \right] \right)^2 \right]$$

$$V_{0|1,NP} = E \left[ \left( \omega_{0|1}(T, X) \cdot \psi^\nu(Y, X; F_{W_0|T=1}) \right. \right. \\ \left. \left. + (\omega_{1|1}(T) - \omega_{0|1}(T, X)) \cdot E[\psi^\nu(Y, X; F_{W_0|T=1}) \mid X, T=0] \right)^2 \right]$$

The density estimators  $\hat{f}_{W_j|T}(w|t=s)$ ,  $j, s = 0, 1$  are, for a fixed bandwidth  $h$ , exactly linear. The derivation of the asymptotic distribution of those estimators follows the same structure of plug-in estimators, but a stronger smoothness condition has to be imposed, since now two simultaneous approximations are taking place: the p-score estimation (and the associated smoothness parameter  $K(N)$ ) and the density estimation (and the associated smoothness parameter  $h(N)$ ).

**THEOREM 2** (a) *Under the above assumptions:*

(i)  $\sqrt{N \cdot h} \cdot (\hat{f}_{W_1|T=1}(w) - f_{W_1|T=1}(w)) \xrightarrow{D} N(\lambda_{1|1}(w), V_{1|1}(w))$ ,

(ii)  $\sqrt{N \cdot h} \cdot (\hat{f}_{W_0|T=0}(w) - f_{W_0|T=0}(w)) \xrightarrow{D} N(\lambda_{0|0}(w), V_{0|0}(w))$

(iii) (a) *if in addition we assume [parametric], then:*

$\sqrt{N \cdot h} \cdot (\hat{f}_{W_0|T=1}(w) - f_{W_0|T=1}(w)) \xrightarrow{D} N(\lambda_{0|1}(w), V_{0|1,P}(w))$

(iii) (b) *otherwise, if in addition we assume [non-parametric], then:*

$\sqrt{N \cdot h} \cdot (\hat{f}_{W_0|T=1}(w) - f_{W_0|T=1}(w)) \xrightarrow{D} N(\lambda_{0|1}(w), V_{0|1,NP}(w))$

where

$$V_{1|1}(w) = f_{W_1|T=1}(w) \cdot \int (K(u))^2 du$$

$$V_{0|0}(w) = f_{W_0|T=0}(w) \cdot \int (K(u))^2 du$$

$$V_{0|1,P}(w) = E \left[ \left( \omega_{0|1}(T, X) \cdot K_h(W - w) \right. \right. \\ \left. \left. + (\omega_{1|1}(T) - \omega_{0|1}(T, X)) \cdot \frac{\dot{p}(X)'}{p(X)} \cdot (E[s(T, X) \cdot s(T, X)'])^{-1} \right. \right. \\ \left. \left. \cdot E \left[ \frac{\dot{p}(X)}{1-p(X)} \cdot E[K_h(W - w) \mid X, T=0] \right] \right)^2 \right]$$

$$V_{0|1,NP}(w) = E \left[ \left( \omega_{0|1}(T, X) \cdot K_h(W - w) + (\omega_{1|1}(T) - \omega_{0|1}(T, X)) \cdot E[K_h(W - w) | X, T = 0] \right)^2 \right]$$

and

$$\begin{aligned} \lambda_{1|1}(w) &= \frac{h^2}{2} \text{tr} \left( \left( \int uu' K(u) du \right) \cdot \frac{\partial^2 f_{W_1|T=1}(w)}{\partial w \partial w'} \right) \\ \lambda_{0|0}(w) &= \frac{h^2}{2} \text{tr} \left( \left( \int uu' K(u) du \right) \cdot \frac{\partial^2 f_{W_0|T=0}(w)}{\partial w \partial w'} \right) \\ \lambda_{0|1}(w) &= \frac{h^2}{2} \text{tr} \left( \left( \int uu' K(u) du \right) \cdot \frac{\partial^2 \left( f_{W_1|T=1}(w) \cdot \frac{p(x)}{p \cdot (1-p(x))} \right)}{\partial w \partial w'} \right) \end{aligned}$$

### 3.1.4 Variance Estimation

We now show how to estimate the asymptotic variance of estimators that we have presented so far. We then state some sufficient conditions for that estimator to be consistent. The four objects that we need to estimate are  $V_{1|1}$ ,  $V_{0|0}$ ,  $V_{0|1,P}$  and  $V_{0|1,NP}$ . Respective estimation of  $V_{1|1}$  and  $V_{0|0}$  follows trivially by:

$$\begin{aligned} \widehat{V}_{1|1} &= N^{-1} \sum_{i=1}^N \left( \widehat{\omega}_{1|1}(T_i) \cdot \widehat{\psi}^\nu \left( Y_i, X_i; \widehat{F}_{W_1|T=1} \right) \right)^2 \\ \widehat{V}_{0|0} &= N^{-1} \sum_{i=1}^N \left( \widehat{\omega}_{0|0}(T_i) \cdot \widehat{\psi}^\nu \left( Y_i, X_i; \widehat{F}_{W_0|T=0} \right) \right)^2 \end{aligned}$$

where  $\widehat{\psi}^\nu(\cdot; \cdot)$  is a consistent estimator of  $\psi^\nu(\cdot; \cdot)$ . For example, for the median of  $Y_1|T=1$ , we would use  $\widehat{\psi}^{me} \left( y; \widehat{F}_{Y_1|T=1} \right) = \left( \widehat{f}_{Y_1|T=1} \left( \widehat{me}_{Y_1|T=1} \right) \right)^{-1} \cdot (1/2 - \mathbb{I}\{y \leq \widehat{me}_{Y_1|T=1}\})$  where  $\widehat{f}_{Y_1|T=1}(\cdot)$  is a consistent estimator for the density of  $Y_1|T=1$ ,  $f_{Y_1|T=1}(\cdot)$ .

Estimation of both  $V_{0|1,P}$  and  $V_{0|1,NP}$  are more complex as they involve in both cases estimation of  $E[\psi^\nu(Y, X; F_{W_0|T=1}) | X, T = 0]$ . We propose estimating that by:

$$\widehat{E}[\psi^\nu(Y, X; F_{W_0|T=1}) | X, T = 0] = H_K(X)' \widehat{\gamma}$$

where

$$\widehat{\gamma} = \left( \sum_{i=1}^N \widehat{\omega}_{0|0}(T_i) \cdot H_K(X_i) \cdot H_K(X_i)' \right)^{-1} \cdot \sum_{j=1}^N \widehat{\omega}_{0|0}(T_j) \cdot H_K(X_j) \cdot \widehat{\psi}^\nu \left( Y_j, X_j; \widehat{F}_{W_0|T=1} \right)$$

Thus:

$$\begin{aligned} \widehat{V}_{0|1,P} = N^{-1} \sum_{i=1}^N & \left( \widehat{\omega}_{0|1}(T_i, X_i) \cdot \widehat{\psi}^\nu \left( Y_i, X_i; \widehat{F}_{W_0, X|T=1} \right) + (\widehat{\omega}_{1|1}(T_i) - \widehat{\omega}_{0|1}(T_i, X_i)) \cdot \frac{\dot{p}(X_i; \widehat{\delta})'}{p(X_i; \widehat{\delta})} \right. \\ & \left. \cdot \left( \sum_{j=1}^N s(T_j, X_j; \widehat{\delta}) \cdot s(T_j, X_j; \widehat{\delta})' \right)^{-1} \cdot \sum_{l=1}^N \frac{\dot{p}(X_l; \widehat{\delta})'}{1 - p(X_l; \widehat{\delta})} \cdot H_K(X_l)' \widehat{\gamma} \right)^2 \end{aligned}$$

and

$$\widehat{V}_{0|1,NP} = N^{-1} \sum_{i=1}^N \left( \widehat{\omega}_{0|1}(T_i, X_i) \cdot \widehat{\psi}^\nu \left( Y_i, X_i; \widehat{F}_{W_0, X|T=1} \right) + (\widehat{\omega}_{1|1}(T_i) - \widehat{\omega}_{0|1}(T_i, X_i)) \cdot H_K(X_i)' \widehat{\gamma} \right)^2$$

**THEOREM 3 : (Consistent Estimation of the Asymptotic Variance):** Under the above assumptions,  $\widehat{V}_{\Delta_\tau} - V_{\Delta_\tau} = o_p(1)$ .

### 3.2 Second Step: Influence Function Projections

Let's first illustrate how the second step of our estimation procedure works in the case of the mean, where we have

$$\nu_{W_1|T=1} = E(Y_1|X, T=1), \nu_{W_0|T=0} = E(Y_0|X, T=0), \text{ and } \nu_{W_0|T=1} = E(Y_0|X, T=1).$$

Estimates  $\widehat{\nu}$  of the  $\nu$ 's are obtained by computing sample means of  $Y$  using the three estimated weights  $\widehat{\omega}_1$ ,  $\widehat{\omega}_0$  and  $\widehat{\omega}_{0|1}$ . For example,

$$\widehat{E}(Y_1|X, T=1) = \sum_{i=1}^N \widehat{\omega}_1(T_i) Y_i.$$

Now consider the three sets of projection coefficients:

$$\begin{aligned} \widehat{\gamma}_1 &= \left( \sum_{i=1}^N \widehat{\omega}_1(T_i) X_i X_i' \right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_1(T_i) X_i Y_i \\ \widehat{\gamma}_0 &= \left( \sum_{i=1}^N \widehat{\omega}_0(T_i) X_i X_i' \right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_0(T_i) X_i Y_i \\ \widehat{\gamma}_{0|1} &= \left( \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i, X_i) X_i X_i' \right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i, X_i) X_i Y_i \end{aligned}$$

By definition, it follows that

$$\widehat{\nu}_{W_1|T=1} = \widehat{E}[X, T=1] \widehat{\gamma}_1, \widehat{\nu}_{W_0|T=0} = \widehat{E}[X, T=0] \widehat{\gamma}_0, \text{ and}$$

$$\widehat{\nu}_{W_0|T=1} = \widehat{E}[X, T=1] \widehat{\gamma}_{0|1}.$$

We can thus rewrite the decomposition for the mean as:

$$\begin{aligned} \widehat{\Delta}_O^\mu &= \left( \widehat{E}[X, T=1] \widehat{\gamma}_1 - \widehat{E}[X, T=1] \widehat{\gamma}_{0|1} \right) \\ &+ \left( \widehat{E}[X, T=1] \widehat{\gamma}_{0|1} - \widehat{E}[X, T=0] \widehat{\gamma}_0 \right) \end{aligned}$$

This is very similar to the standard Oaxaca-Blinder decomposition except that the counterfactual mean wage is  $\widehat{E}[X, T = 1] \widehat{\gamma}_{0|1}$  instead of  $\widehat{E}[X, T = 1] \widehat{\gamma}_0$  in the standard decomposition. This difference between  $\widehat{\gamma}_{0|1}$  and  $\widehat{\gamma}_0$  is linked to the **Shortcoming 2** of the standard decomposition mentioned above. We later show an empirical example where this difference turns out to be quite important

Let  $\psi^\mu(Y)$  be the influence function for the mean, where  $\psi^\mu(Y) = Y - \mu$ . Consider the rescaled influence function

$$\phi^\mu(Y) = \mu + \psi^\mu(Y) = Y.$$

In the case of the median, the rescaled influence function is

$$\phi^{me}(Y) = me + (1/2 - \mathbb{I}\{Y \leq me\})/f(me).$$

More generally, the rescaled influence function of the  $q^{th}$  quantile  $y^q$  is

$$\phi^q(Y) = y^q + (q - \mathbb{I}\{Y \leq y^q\})/f(y^q).$$

The rescaled influence function can be computed for each observation by plugging the sample estimate of the quantile,  $\widehat{y}^q$ , and estimating the density at the sample quantile,  $\widehat{f}(\widehat{y}^q)$ .

Since the rescaled influence function is equal to  $Y$  in the case of the mean, the projections used above turn out to be projection of the rescaled influence functions on  $X$ . By analogy, rescaled influence function for other distributional parameters such as quantiles can also be projected on  $X$ . For example, for the  $q^{th}$  quantile we have:

$$\begin{aligned} \widehat{\gamma}_1^q &= \left( \sum_{i=1}^N \widehat{\omega}_1(T_i) X_i X_i' \right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_1(T_i) X_i \widehat{\phi}_1^q(Y_i) \\ \widehat{\gamma}_0^q &= \left( \sum_{i=1}^N \widehat{\omega}_0(T_i) X_i X_i' \right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_1(T_i) X_i \widehat{\phi}_0^q(Y_i) \\ \widehat{\gamma}_{0|1}^q &= \left( \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i, X_i) X_i X_i' \right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_{0|1}(T_i, X_i) X_i \widehat{\phi}_{1|0}^q(Y_i) \end{aligned}$$

We can thus decompose this quantile as:

$$\begin{aligned} \widehat{\Delta}_O^q &= \left( \widehat{E}[X, T = 1] \widehat{\gamma}_1^q - \widehat{E}[X, T = 1] \widehat{\gamma}_{0|1}^q \right) \\ &\quad + \left( \widehat{E}[X, T = 1] \widehat{\gamma}_{0|1}^q - \widehat{E}[X, T = 0] \widehat{\gamma}_0^q \right) \end{aligned}$$

This generalizes the Oaxaca-Blinder decomposition to any quantile. The same technique can also be used to decomposition standard distributional measures like the variance or the Gini coefficient.

## 4 Empirical Applications

We use different empirical applications to illustrate various features of our decomposition method. We focus on three particular issues The first issue we look at the issues we look at in this section are:

1. Illustrate specification problems in the standard decomposition by running Mincer wage equations for men in 1973 and 2003

2. Show how influence function projection work by looking at the effect of unions on wages by quantiles of the wage distribution (and contrast with traditional quantile regressions)
3. Decompose 1973-2003 changes in wage inequality (variance and 90-10 gap).

#### 4.1 Misspecification of a Mincer-type wage equation

Table 2 shows standard estimates of a Mincer wage equation for men in 1973 and 2003. The 1973 data are from the May Supplement of the Current Population Survey (CPS). The 2003 data are from the Outgoing Rotation Groups (ORG) Supplement of the CPS. More detail on the these data is provided in Lemieux (2005). We both report the most standard version of the Mincer equation where log wages are regressed on a linear function of years of education and a quadratic function of potential experience, as well as a extended specification where a quartic specification in experience is used (as in Murphy and Welch, 1990).

The Mincer equation is an interesting case to explore because there is growing evidence that the effect of education on wages has become more and more convex over time (Deschenes, 2002, Mincer 1997). This is thus a prime case where the simple linear specification is likely incorrect, and where the estimates from a linear specification may well depend on the distribution of characteristics. For example, since the level of education has increased substantially over time, estimating the Mincer equation by OLS in 2003 should yield a larger effect of education on wages when the 2003 distribution of characteristics is used than when the 2003 data is reweighted to get the same distribution of characteristics as in 1973.

Table 1 shows that this is indeed what happens. For example, in the models with quadratic experience (columns 1 to 3), the return to education increase from 0.068 in 1973 to 0.111 in 2003, a 0.043 increase. Column 3 shows estimates from a model in which the 2003 data has been reweighted to the 1973 distribution of education and experience.<sup>4</sup> In terms of the previous notation, the coefficients in columns 3 are  $\hat{\gamma}_{0|1}$  (T=0 for 2003, T=1 for 1973) compared to  $\hat{\gamma}_0$  for column 2. The returns to education decreases from 0.111 to 0.101, which represents about a quarter of the increase between 1973 and 2003. This means that using a linear specification for education overstates the increase in the "price" of education because the true conditional expectation function is convex and the level of education is higher in 2003 than in 1973. In terms of a Oaxaca-Blinder decomposition, this means that the decomposition would attribute to changes in the regression coefficients (i.e. to "price effects") a component linked to misspecification of the conditional expectation function. Our alternative decomposition "corrects" for this problem by using  $\hat{\gamma}_{0|1}$  instead of  $\hat{\gamma}_0$ . The results with the quartic specification for experience (columns 4 to 6) are very similar.

There is a much smaller change in the return to experience than in the return to education. For example, the coefficient on the linear term increases from 0.042 in 1973 to 0.046 in 2003. Reweighting actually reduces the coefficient instead to 0.038. Figure 2 shows that the same pattern of results holds for the whole experience profile. This means that the whole increase in the experience profile (and even more) is a spurious consequence of the fact that 1) the simple Mincer equation is misspecified in 2003, and 2) the distribution of characteristics change in such a way that the linear approximation becomes steeper than in 1973 (just like fitting the mode in Figure 1 for higher values of X yields a steeper regression curve).

---

<sup>4</sup>Reweighting is performed by estimating a logit with a full set of experience and education dummies, plus a set of interactions between education dummies and a quartic in experience.

We conclude from this first example that misspecification problems can substantially bias the standard Oaxaca-Blinder decomposition.

## 4.2 Influence Function Projections

To illustrate how the influence function projections work in practice, we focus on the impact of union on wages (for men) which is well known to have a differential impacts at different points in the wage distribution (Card (1996)). There are several reasons for this. First, union both increase the conditional mean of wages (the “between” effect) and decrease the conditional distribution of wages. This means that unions tend to increase wages in low wage quantiles where both the between and within group effects go in the same direction, but can decrease wages in high wage quantiles where the between and within group effects go in opposite directions. This is compounded by the fact that the union wage gap generally declines for higher than lower skills levels.

Table 2 shows detailed estimates of influence function projections at the 10<sup>th</sup>, 50<sup>th</sup> and 90<sup>th</sup> quantiles using 1984 ORG CPS data for men. The results are also compared with the OLS benchmark. Interestingly, the effect of unions first increase from 0.203 at the 10th quantile to 0.339 at the median before turning negative (-0.124) at the 90<sup>th</sup> quantile. Notice also how the effect of education changes at the different quantiles. For example, the effect of post-graduate education (relative to high school) increases from 0.143 at the 10<sup>th</sup> quantile, to 0.512 at the median and 0.880 at the 90<sup>th</sup> quantile. Since both high school graduates and college post-graduates are relatively unlikely to be in the lowest ten percent of the wage distribution, post-graduate education has relatively little effect. By contrast, post-graduate education greatly increases the probability of earning above the 90<sup>th</sup> quantile, thus explaining the much larger coefficient at this quantile.

Table 3 contrast the effect of unions in our influence function projections to standard estimates based on (conditional) quantile regressions. Remember the key difference between the two methods. The influence function projections indicate the impact of unions (in this case) on the unconditional quantiles of wages, while quantile regressions indicate the impact of unions on conditional quantiles. The quantile regression estimates in the second of Table 3 show, as expected, that unions increase the location of the conditional wage distribution (i.e. positive effect on the median) but also reduce conditional wage dispersion. This explains why the effect of unions monotonically declines as quantiles increase. One cannot infer from the quantile regressions, however, what is the overall effect of unions on the unconditional wage distribution. The key problem is that, unlike conditional means, conditional quantiles do not aggregate up to unconditional quantiles. For example, the fact that unions increase the conditional median by 0.196 does not say anything about the effect of unions on the unconditional median.

In fact, the first row of Table 3 show that the effect of unions on unconditional quantiles estimated using our influence function projections are quite different from the conditional quantile estimates. For instance, the effect on the median (0.339) largely exceed the conditional quantile regression estimate of 0.196. Furthermore, the effect is a non-monotonic function of quantiles. The effect first increases from 0.054 at the 5<sup>th</sup> quantile to 0.374 at the 25<sup>th</sup> quantile before declining to eventually reach a large negative effect of -0.229 at the 95<sup>th</sup> quantile. The large effect at the top end reflects the fact that compression effects dominate everything else at the top end. Union wages are more compressed and do not exhibit the “fat” upper tail of the non-union wage distribution. As a result, unions have large and negative impact on the probability of earning less more than the 95<sup>th</sup> quantile.

The story for the lowest quantiles is a bit different. In 1984, the 5th quantile turns out to exactly be the minimum wage of \$3.35. The wage density is quite high because of the bunching at the minimum wage, which explains why estimated coefficients are relatively small (the inverse of the density appears in the equation for the influence function). Intuitively, the idea is that unions and other wage determination variables have little impact at the bottom of the wage distribution because everybody is bunched up at the minimum wage anyway.

### 4.3 Decomposing 1973-2003 changes in the wage distribution

We finally show in Table 4 how our two step procedure can be used to decompose four parameters, the mean, the median, the variance and the 90-10 gap, of the wage distribution between 1973 and 2003. The date used are once again the May 1973 and the 2003 ORG CPS for men. The first step estimates are obtained by fitting the logit model discussed above to reweight the 2003 wage distribution. The second step estimates are obtained by running the influence function projections on a set of education (six) and experience dummies (nine). Instead of reporting the elements of the decomposition for each dummy, we sum up the effects for education and experience.

In the case of the mean, remember that the (rescaled) influence function is simply  $\phi^\mu(Y) = Y$ . In the case of the variance we have  $\phi^{\sigma^2}(Y) = (Y - \mu)^2$ . The influence function for the median is defined earlier while the one of the 90-10 gap is just the difference between the two corresponding influence functions:

$$\phi^{90-10}(Y) = \phi^9(Y) - \phi^1(Y) = (y^9 - y^1) + (.9 - \mathbb{I}\{Y \leq y^9\})/f(y^9) - (.1 - \mathbb{I}\{Y \leq y^1\})/f(y^1).$$

The results in column 1 and 2 show that the small decline in the mean and the larger decline in the median are due to the offsetting impact of wage structure and composition effects. For example, the large secular increase in educational achievement should have resulted in a 0.174 increase in the mean and 0.248 increase in the median. This was more than offset by the general decline in wages (wage structure effects).

To the best of your knowledge, no Oaxaca-Blinder type decomposition for the median (in column 2) have ever been reported in the literature. This illustrates the usefulness of our proposed method since the median is a fairly standard distributional parameters. The median is also of particular policy interest since policy measures may or may not be supported by a majority of the population depending on the impact on the median.

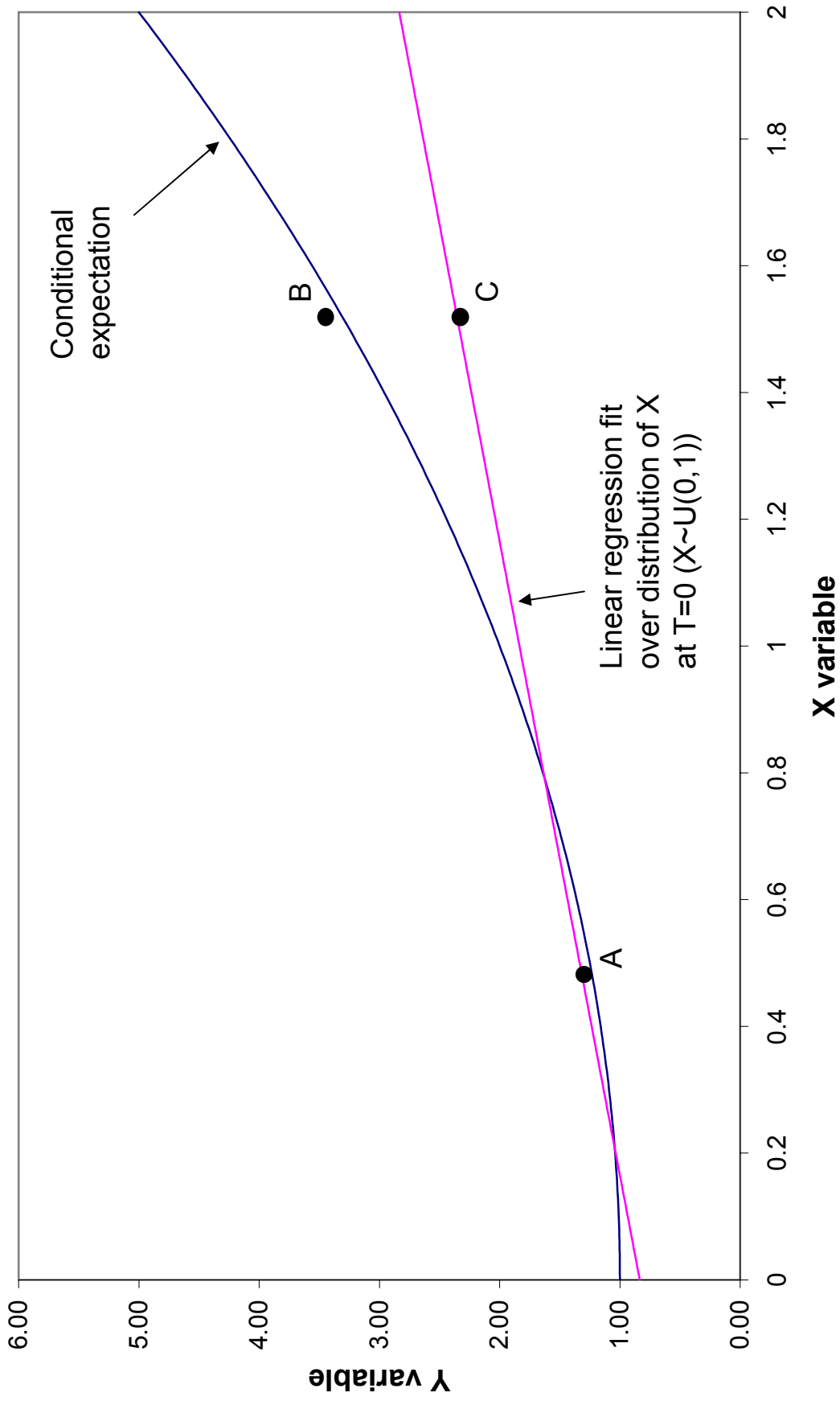
Finally, the results of columns 3 and 4 confirm existing findings about the sources of changes in wage inequality. Both wage structure and composition effects (Lemieux, 2005) are important. Changes in returns to education are also a more important component of changes in the wage structure than changes in returns to experience.

## REFERENCES

- ABADIE, A., (2005), "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*.
- BISHOP, J., J. FORMBY, AND P. THISTLE, (1992) "Convergence of the South and Non-South Income Distributions, 1969-1979," *The American Economic Review*, Vol. 82, No. 1., pp. 262-272.
- BLINDER, A., (1973), "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436-455.
- BOURGUIGNON, F., F. FERREIRA, P. LEITE, (2002) "Beyond Oaxaca-Blinder: accounting for differences in household income distributions across countries" PUC-Rio, TD #452.
- CRAMÉR, (1946)
- COWELL, F., (1995), *Measuring Inequality*. (2nd edition). Harvester Wheatsheaf, Hemel Hempstead.
- DINARDO, J., (2002), "Propensity Score Reweighting and Changes in Wage Distributions", Unpublished paper, University of Michigan, Department of Economics.
- DINARDO, J., N. FORTIN, AND T. LEMIEUX, (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64, 1001-1044.
- FERGUSON, T., (1996), *A Course in Large Sample Theory*,
- FIRPO, S. (2004), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *UBC Department of Economics Discussion Paper*, No. 04-01.
- HECKMAN, J., H. ICHIMURA, AND P. TODD, (1997), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261-294.
- HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD, (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.
- HIRANO, K., G. IMBENS, AND G. RIDDER, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," forthcoming, *Econometrica*.
- HOEFFDING, W., (1948)
- JUHN, C., K. MURPHY, AND B. PIERCE, (1993), "Wage Inequality and the Rise in Returns to Skill," *The Journal of Political Economy*, 101, 410-442.
- KOENKER, R., AND G. BASSETT, (1978), "Regression Quantiles," *Econometrica*, 46, 33-50.
- LEHMANN, E., (1998), *Elements of Large Sample Theory*.
- LEMIEUX, T., (2002), "Decomposing Changes in Wage Distributions: a Unified Approach," *The Canadian Journal of Economics*, 35, 646-688.

- MACHADO, J. AND J. MATA, (2004), "Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression", *Journal of Applied Econometrics*, (forthcoming).
- NEWBY, W., (1995), "Convergence Rates for Series Estimators," in *Advances in Econometrics and Qualitative Economics: Essays in Honor of C.R. Rao*, G. Maddal, P.C. Phillips, and T.N. Srinivasan, eds., Cambridge US, Basil-Blackwell.
- NEWBY, W., (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147-168.
- OAXACA, R., (1973), "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693-709.
- ROSENBAUM, P., AND D. RUBIN, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.
- VAN DER VAART, A., (1998), *Asymptotic Statistics*.
- WAND, M. AND M. JONES, (1994), *Kernel Smoothing*. Chapman & Hall/CRC.

**Figure 1: Misspecified Conditional Expectation Function**



**Figure 2: Experience profiles estimated in the 1973 and 2003 CPS**

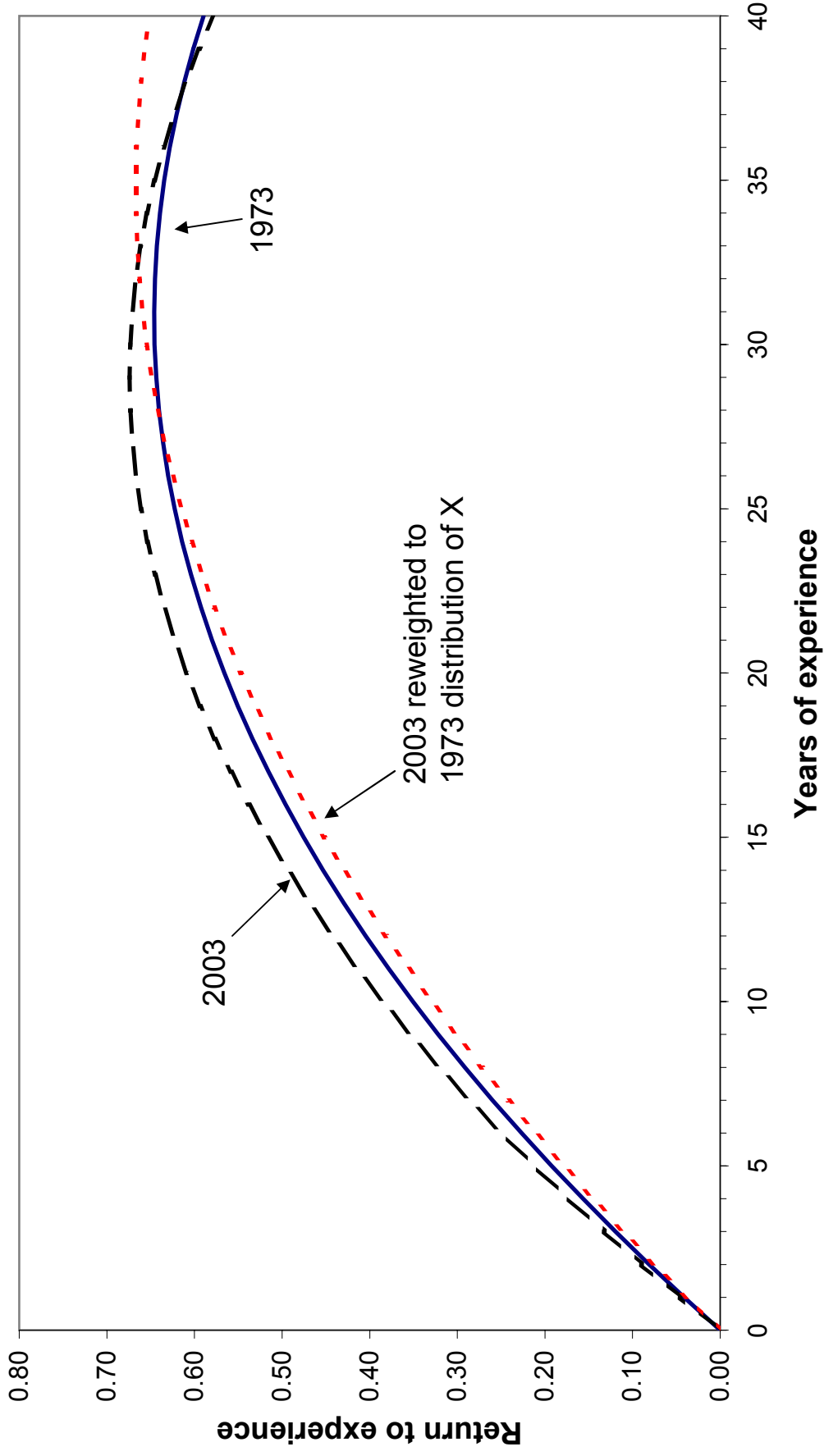


Table 1: Mincer equations for men, 1973 and 2003

	1973	2003	2003	1973	2003	2003
			reweighted			reweighted
Education	0.068 (0.001)	0.111 (0.001)	0.101 (0.001)	0.069 (0.001)	0.113 (0.001)	0.103 (0.001)
Experience	0.042 (0.001)	0.046 (0.001)	0.038 (0.000)	0.087 (0.003)	0.080 (0.003)	0.073 (0.002)
Exper <sup>2</sup> (/100)	-0.068 (0.002)	-0.080 (0.001)	-0.055 (0.001)	-0.402 (0.026)	-0.293 (0.023)	-0.248 (0.017)
Exper <sup>3</sup> (/1000)				0.082 (0.008)	0.041 (0.008)	0.028 (0.005)
Exper <sup>4</sup> (/10000)				-0.006 (0.001)	-0.002 (0.001)	0.000 (0.001)

Table 2: OLS vs Influence function projections, 1984 CPS

	OLS	Influence function projections		
		10th	50th	90th
	(1)	(2)	(3)	(4)
Union	0.184 (0.003)	0.203 (0.006)	0.339 (0.005)	-0.124 (0.007)
Nonwhite	-0.132 (0.005)	-0.126 (0.008)	-0.154 (0.007)	-0.103 (0.010)
Married	0.137 (0.004)	0.201 (0.006)	0.150 (0.005)	0.042 (0.008)
Educ 0-8	-0.353 (0.007)	-0.303 (0.011)	-0.469 (0.010)	-0.261 (0.014)
Educ 9-11	-0.188 (0.005)	-0.355 (0.008)	-0.194 (0.007)	-0.074 (0.010)
Educ 13-15	0.131 (0.004)	0.052 (0.007)	0.173 (0.006)	0.161 (0.009)
Educ 16	0.409 (0.005)	0.201 (0.008)	0.472 (0.007)	0.599 (0.010)
Educ 17+	0.476 (0.005)	0.143 (0.009)	0.512 (0.008)	0.880 (0.011)
Exper 0-4	-0.547 (0.006)	-0.599 (0.011)	-0.639 (0.010)	-0.455 (0.014)
Exper 5-9	-0.263 -0.006	-0.078 -0.010	-0.356 -0.009	-0.375 -0.013
Exper 10-14	-0.146 (0.006)	-0.037 (0.010)	-0.186 (0.009)	-0.250 (0.013)
Exper 15-19	-0.056 (0.006)	-0.026 (0.011)	-0.074 (0.010)	-0.081 (0.014)
Exper 25-29	0.036 (0.007)	0.003 (0.012)	0.038 (0.011)	0.078 (0.015)
Exper 30-34	0.039 (0.007)	0.000 (0.012)	0.034 (0.011)	0.082 (0.016)
Exper 35-39	0.041 (0.008)	0.021 (0.013)	0.029 (0.012)	0.073 (0.016)
Exper 40+	0.010 (0.008)	0.041 (0.013)	0.006 (0.012)	-0.030 (0.016)

Table 3: Comparing union effect in influence function and quantile regression model

	5th	10th	25th	50th	75th	90th	95th
Influence function regressions	0.054 (0.003)	0.203 (0.006)	0.374 (0.006)	0.339 (0.005)	0.074 (0.005)	-0.124 (0.007)	-0.229 (0.011)
Quantile regressions	0.309 (0.007)	0.295 (0.003)	0.261 (0.001)	0.196 (0.001)	0.140 (0.004)	0.095 (0.006)	0.069 (0.005)

Note: Estimated using 1984 CPS data. Other regressors are the same as in Table 2.

Table 4: Decomposition of 1973-2003 changes in wage distribution

	Mean	Median	Variance	90-10
Overall 1973-2003 change	-0.030	-0.067	0.109	0.252
Wage structure effect	-0.241	-0.314	0.064	0.109
Education	0.042	0.012	0.033	0.065
Experience	0.012	-0.011	-0.011	0.007
Constant	-0.295	-0.315	0.041	0.037
Composition effect	0.211	0.247	0.045	0.143
Education	0.174	0.248	0.006	0.018
Experience	0.027	0.033	0.032	0.078
Constant	0.009	-0.034	0.007	0.047

Note: First-step model estimated using logit model with full set of experience and education dummies, plus interaction between quartic in experience and education dummies. Second step model estimated using dummies for 6 education categories and 9 experience categories shown in Table 2. Omitted group is high school graduates with 20-24 years of experience. The "experience" and "education" effects in the table are the sum of the effects for education and experience dummies.