Propensity score reweighted regression in analysis of wage differentials

Philippe Van Kerm
CEPS/INSTEAD
Population & Emploi
3, Avenue de la Fonte
Esch-Sur-Alzette (L-4364),Luxembourg
E-mail: philippe.vankerm@ceps.lu

Michela Bia
CEPS/INSTEAD and University of Florence
Population & Emploi
3, Avenue de la Fonte
Esch-Sur-Alzette (L-4364),Luxembourg
E-mail: michela.bia@ceps.lu

Introduction

The ubiquitous approach in analysis of wage differentials is the Blinder-Oaxaca (BO) decomposition (Blinder A., 1973; Oaxaca R., 1973) Take as raw measure of the difference in pay between two population groups the difference in the average (log) wage, say $\Delta^U = \mu^m - \mu^f$ where f is typically a target group of interest (e.g., women or immigrants) and m is a reference group (e.g., men or natives). If the composition of the two groups differ with respect to observable earnings-related characteristics (e.g., human capital, age, jobs), it is useful to assess how much of Δ^U can be attributed to such endowments difference and what is due to 'pure' differences in pay. The latter is an 'adjusted' difference in average (log) wage that would be observed if the characteristics of the target group were rewarded as those of the reference group, $\Delta^A = \mu^{f|m} - \mu^f$. Δ^A is meant to capture strict differences in compensation cleared of differences in group characteristics and is often interpreted as measuring 'discrimination'. Any remaining difference between Δ^U and Δ^A is reflective of the effect of 'endowment' differences. One important weakness of the regression-based approach however is that it relies on parametric assumptions about the earnings regressions. Misspecification of the regression models may lead to misleading inference about the various components of the BO decomposition. Non-parametric techniques can be used to avoid this problem (Mora R., 2008). However, as emphasized in Fortin et al.'s (2010) survey, these approaches are often much more computationally demanding, face the curse of dimensionality when dealing with a large number of covariates, and/or do not allow straightforward singling out the impact of individual covariates on wage differentials.

This paper considers a middle way that maintains ease of implementation of the regression-based approach and the possibility to make detailed decompositions, but that is more robust in the presence of misspecification than the classic approach. It involves estimating the coefficients from the reference regression m by weighted least squares (rather than OLS) where the weights are function of the relative density of the covariates in the reference and target samples. This procedure, suggested in Fortin et al. (2010), has been demonstrated to lead to improved predictive performance when prediction is made out-of-sample and in the presence of model misspecification (Shimodaira H., 2000; Sugiyama M. and Müller K., 2005). This is directly relevant to the problem at hand since in computing Δ^A the reference (say male) sample regression coefficients are exclusively used to make wage predictions in the target (say female) sample. Specifically, our objective in this paper is to empirically assess the gains of using such a procedure in a case study to the gender pay gap in Luxembourg. We compare estimates

against a simple OLS approach and a fully non-parametric local linear model. We also experiment with alternative ways to specify and estimate the weighting function. While the approach is not as robust to misspecification as a non-parametric model can be, we find that it can offer significant improvement over the standard approach.

Propensity score reweighted pay gap estimation

Let μ^f and μ^m denote average log wage in, respectively, a sample of N^f female workers –the target group– indexed $i = 1, \ldots, N^f$ and a sample of N^m male workers –the reference group– indexed $j = 1, \ldots, N^m$. Also let $D_i \in \{0, 1\}$ denote the group belonging of observation i with $D_i = 1$ identifying female workers and $D_i = 0$ male workers.[1] The raw, unadjusted, wage differential is

$$\Delta^U = \mu^m - \mu^f$$

and the BO decomposition of this wage gap is

$$\Delta^U = \underbrace{(\mu^m - \mu^{f|m})}_{\text{explained/endowment component}} + \underbrace{(\mu^{f|m} - \mu^f)}_{\text{unexplained/discrimination component}}$$

where $\mu^{f|m}$ is the counterfactual average log wage of the reference group of female workers if their characteristics were rewarded as those of the comparison group of men.

Regression-based estimation of $\mu^{f|m}$ is done by modelling the relationship between observed covariates X and log wage (Y) with a separate linear regression model in each of the two groups:

$$Y_i = X_i\beta^d + e_i \qquad i = 1, \ldots, N^d \qquad d \in \{m, f\}$$

where $\mathrm{E}[e_i|X_i, D_i] = 0$. Under these assumptions, μ^f, μ^m and $\mu^{f|m}$ can be expressed as

$$\mu^f = \frac{1}{N^f}\sum_{i=1}^{N^f} X_i\beta^f, \qquad \mu^m = \frac{1}{N^m}\sum_{j=1}^{N^m} X_j\beta^m, \qquad \mu^{f|m} = \frac{1}{N^f}\sum_{i=1}^{N^f} X_i\beta^m.$$

and the adjusted wage differential is

$$\Delta^A = \mu^{f|m} - \mu^f = \frac{1}{N^f}\sum_{i=1}^{N^f} X_i(\beta^m - \beta^f) = \bar{X}^f(\beta^m - \beta^f)$$

where $\bar{X}^f$ is a row vector of covariate means in the sample of interest.

Application of reweighted regression in this context involves estimating the regression coefficients β_f by ordinary least squares and β_m by *weighted* least squares:

$$\hat{\beta}^f = \left(\sum_{i=1}^{N^f} X_i'X_i\right)^{-1}\sum_{i=1}^{N^f} X_i'Y_i, \qquad \hat{\beta}^m = \left(\sum_{j=1}^{N^m} X_j'\omega(X_j)X_j\right)^{-1}\sum_{j=1}^{N^m} X_j'\omega(X_j)Y_j$$

where the reweighting factor $\omega(X_j)$ is a function of the ratio of the probability density of covariates at X_j in groups m (f^m) and f (f^f). We restrict attention here to functions of the form

$$\omega(X_j;\lambda) = \left(\frac{f^f(X_j)}{f^m(X_j)}\right)^\lambda \qquad \lambda \in [0, 1]$$

[1] For clarity of exposition, this article is framed in terms of wage differentials across gender, but the arguments and techniques apply to any continuous outcome variable that can be modelled in a regression setting and any two populations of interest, such as immigrants and natives, etc.

This specification leads to the classic unweighted OLS case with $\lambda = 0$, and gives the baseline fully reweighted case discussed in Fortin et al. (2010, pp. 45–49) with $\lambda = 1$.

The intuition is that when model parameters are used for predictions out-of-sample one should give more weight at the estimation stage to observations close to the points that will be used for predictions. Observations in the reference sample that have no close counterparts in the target sample are discarded –or more precisely down-weighted. This reweighting is however irrelevant when the parametric model is correctly specified and therefore generates valid prediction anywhere on the support of definition of covariates.

Superior predictive performance in the target sample of the reweighted regression based on the density ratio in the presence of misspecification is formally demonstrated in Shimodaira (2000). The reweighted model with $\lambda = 1$ is shown to lead to asymptotically optimal predictions in the prediction sample. Intermediate values of $\lambda < 1$ may be preferred in finite samples. Optimal choice for λ in finite sample, in terms of minimizing expected prediction error, depends on unknown population parameters but Shimodaira (2000) suggests that estimate by minimizing the following information criteria:

$$\hat{\lambda} = \min_{\lambda}\left\{\left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}\left(\frac{\hat{\epsilon}_j^2}{\hat{\sigma}^2}+\log(2\pi\hat{\sigma}^2)\right)\right)+2\left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}\left(\frac{\hat{\epsilon}_j^2}{\hat{\sigma}^2}\hat{h}_j+\frac{\omega(x_j;\lambda)}{2\hat{c}_\omega}\left(\frac{\hat{\epsilon}_j^2}{\hat{\sigma}^2}-1\right)^2\right)\right)\right\}$$

where $\hat{\epsilon}_j = (Y_j - X_j\hat{\beta}^m)$ is the WLS regression residual for observation j, $\hat{c}_\omega = \sum_{j=1}^{N_m}\omega(X_j;\lambda)$ is the sum of weights, $\hat{\sigma}^2 = \sum_{j=1}^{N_m}\omega(X_j;\lambda)\hat{\epsilon}_j^2/\hat{c}_\omega$ is an estimate of the residual variance from the WLS regression, and $\hat{h}_j$ is the j^{th} element of the diagonal of the hat matrix of the WLS model.[2, 3]

Note however that Shimodaira's IC does not capture the variance inflation associated with estimation of the probability density functions ratio which is not normally known in applications (see *supra*). This potentially limits its useability in practice.

Along the same lines as Shimodaira's information criteria, 'reweighted R^2' measures can be useful to assess improvement in fit in the support of the data covered by the target sample when using WLS:

$$\text{RW}^2 = 1 - \left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}\hat{\epsilon}_j^2\right)\left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}(Y_j-\tilde{\mu}^m)^2\right)^{-1}$$

where $\tilde{\mu}^m = \left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}\right)^{-1}\left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}Y_j\right)$ is the reweighted mean log wage in the reference sample. RW^2 captures the fit of the WLS regression model in the reweighted male sample which, by construction, has the same covariate distribution as the female sample over which predictions will be made. This can be compared to the equivalently reweighted R^2 obtained when the OLS coefficients are used:

$$\text{RW}_0^2 = 1 - \left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}(Y_j-X_j\hat{\beta}_0^m)^2\right)\left(\sum_{j=1}^{N^m}\frac{f^f(X_j)}{f^m(X_j)}(Y_j-\tilde{\mu}^m)^2\right)^{-1}$$

where $\hat{\beta}_0^m$ denotes OLS coefficient estimates. The degree to which RW^2 exceeds RW_0^2 can be taken as an indication of the improvement in predictive accuracy of the WLS compared to the OLS in the target sample.

Direct estimation of the density ratio can however be replaced by estimation of the propensity score for which simple estimators are available even with large K. As is well-known (see e.g.

[2]The hat matrix is $H = X(X'WX)^{-1}X'W$ where W is the vector of weights $\omega(X_j;\lambda)$.

[3]See Sugiyama M. and Müller K. (2005) for an alternative approach to selecting λ.

Rosenbaum and Rubin, 1983), simple application of Bayes' rule yields

$$\frac{f^f(X_j)}{f^m(X_j)} = \frac{p(X_j)}{1-p(X_j)} \frac{1-\pi}{\pi}$$

where $\pi = \Pr(D_i = 1)$ is the (unconditional) probability of belonging to the target group and $p(x) = \Pr(D_i = 1|x)$ is the propensity score, that is, the conditional probability of belonging to the target group for a covariate vector x.

Illustrative example and Monte Carlo simulations

Before assessing the gain to using propensity score reweighting in a full-fledge empirical analysis, we first illustrate the technique on a simulated example.

We replicate Shimodaira's (2002) model with observations on a single covariate x drawn from two population groups distributed $x \sim \mathrm{N}\left(\mu^d, \tau^d\right)$, with $\mu^f = 0$, $\tau^f = 0.3^2$, $\mu^m = 0.5$, $\tau^m = 0.5^2$. The resulting density ratio is

$$\frac{f^f(x)}{f^m(x)} \propto \exp\left(-\frac{(x-\bar{\mu})^2}{2\bar{\tau}}\right)$$

where $\bar{\tau} = ((\tau^f)^{-1} - (\tau^m)^{-1})^{-1} = 0.38^2$ and $\bar{\mu} = ((\tau^f)^{-1}\mu^f - (\tau^m)^{-1}\mu^m) = -0.28$. (See Figure 1 for a graphical illustration of resulting samples). Log wages are given by

$$y = -x + x^3 + \epsilon \qquad \epsilon \sim \mathrm{N}\left(0, 0.3^2\right)$$

for both groups. The adjusted wage differential is therefore nil.

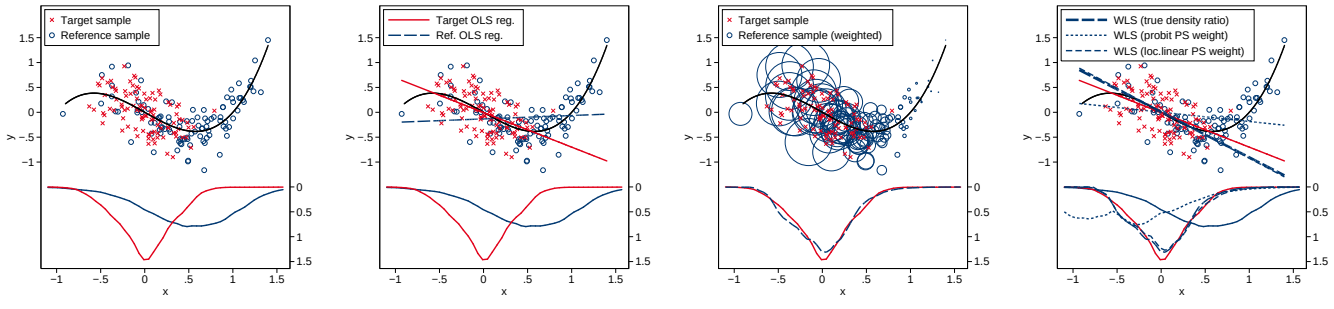
The true relationship between log wage and x is cubic but we consider a misspecified model where the relationship between x and y is assumed linear:

$$y = \alpha + \beta x + e \qquad \mathrm{E}(e|x) = 0.$$

Figure 1 illustrates the model and the reweighting principle on one simulated sample of 100 'males' (blue markers and lines) and 100 'females'. Panel (a) shows a scatter plot of the data along with kernel estimates of the covariate x density distributions in the two groups. The black line traces $\mathrm{E}(y|x)$ from the true model. With the chosen model, the reference sample (males) is spread over a broader (and on average higher) range of values for x than the target sample (females). Crucially, $\mathrm{E}(y|x)$ is approximately linear over the range of the target sample but the relationship is non-linear over the full range of the reference sample. As a consequence the fitted linear regression in the female sample (shown as a solid red line in Panel (b)) approximates $\mathrm{E}(y|x)$ closely, but the linear regression in the male sample (shown as a dashed blue line in Panel (b)) is off-target and very different from the line of the female sample (despite both groups having in fact identical non-linear conditional means). Using coefficients from these misspecified linear regressions leads to an estimate of the adjusted wage gap of –0.126 to the advantage of women (where it is truly nil).

The reweighting strategy involves assigning differential weights to observations from the reference, male sample such that the probability density function of covariates in the reweighted sample is as close as possible to the density in the target, female sample. This is illustrated in Panel (c) in which male observations are plotted with markers proportional to their weight (weights in this plot are defined as the ratio of the true population density functions $f^f(x)$ and $f^m(x)$). Reweighted regression lines are shown in Panel (d). The reweighted estimates are much closer to the target regression estimates at least where the weights are based on the true density ratio (which is not normally known in real data applications) and on the propensity score estimated with a local linear smoother. In this particular example, failure of the parametric probit propensity score model is attributable to

Figure 1: Random samples of 100 reference observations (blue dots) and 100 target observations (red crosses) with regression lines and density estimates



(a) Samples and true reg.line (b) Unw.lin.reg. (c) Rew.samp.(true.ratio) (d) Weight.lin.reg.

incapacity of the probit model to estimate a valid weight for the smallest x value in the reference sample. Linearity in x of the single index component of the probit model implies that this particular observation receives a large weight where it should in fact be down-weighted. This is a reminder that a naive linear index propensity score model is likely to fail whenever the target sample is 'sandwiched' within the reference sample (with observations both below and above the target sample observations).

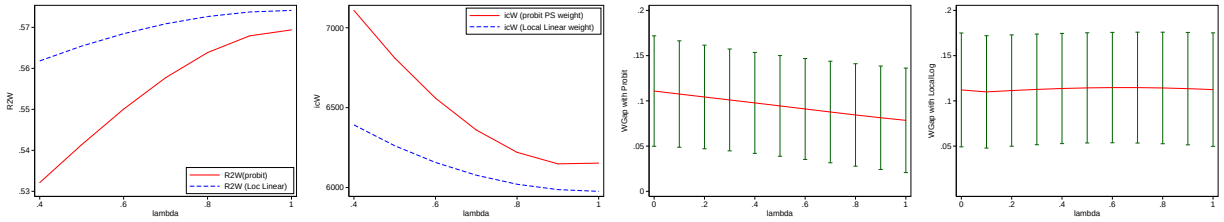
Gender wage differentials in Luxembourg

To assess the impact of reweighting in a real data empirical application, we analyze gender wage differentials using data from the *Panel Socio-Economique Liewen zu Lëtzebuerg*, a longitudinal survey on income and living conditions representative of the Luxembourgish population. Recent estimates from STATEC, the national statistical office of Luxembourg, suggest that women are paid on average 20 percent less than men in Luxembourg, and that approximately half of this gap can be accounted for by differences in human capital and job characteristics (STATEC, 2007). Estimates of comparable magnitude based on PSELL-3/EU-SILC data are reported in Van Kerm (2009). This is similar to estimates found in other European countries (see e.g. Gregory M, 2009). In our analysis of wage differentials, we consider the period 2003-2008 and we extract a sample of 25- to 55-year-old male and female workers. We keep both private and public sector employees (with the exception of civil servant from international institutions) but drop self-employed workers (for whom gross hourly wage is ill-defined). For similar reasons, we also drop employees recorded to work in agriculture. We pool all samples from 2003 to 2008. We focus on gender differences in gross hourly wage which is computed as gross monthly salary in current job (including paid overtime) divided by 4.32 times work hours in a normal week on the job.[4] Wages are expressed in constant January 2007 prices. In the present application we aim to compute adjusted wage gap estimates after controlling for both human capital characteristics (including age, nationality, education and actual years of work experience) and employment and contract types (sector of activity, firm size, supervisory activity, fixed-term contract, and part-time indicator). We take this particular position here since controlling for employment and contract types tends to lead to covariate imbalance by gender. A sample of results are reported in Figure 2: estimates of RW^2, Shimodaira's IC and adjusted wage gap.[5]

[4]To avoid results being driven by a small number of possibly mis-measured wages we excluded observations with hourly wages below 3 or above 60.

[5]We also extended the application of propensity score reweighting to estimation of quantile differences. Additional results are omitted, but available upon request.

Figure 2: R2, Information Criterion and Total Wage Gap estimates



(a) R2 weighted (b) IC weighted (c) Tot.WGap (ProbitW) (d) Tot.WGap (Loc.W)

These results suggest a choice of $\lambda = 1$ among the 10 values compared here for all models. This suggests that intermediate choices of λ for large samples may not necessarily pay off. Bear in mind however that Shimodaira's IC does not take into account estimation of the propensity scores. The total wage gap estimates we get by applying the probit weighted regressions, compared to those from the no-weighted model, are noteworthy: about 0.06 in the weighted regression model against 0.11 in the standard OLS model. Surprisingly, we find no significant differences between the weighted and unweighted models if we consider a local linear smoother to estimate reweighting factors at about 0.10 with the unweighted model.

REFERENCES

Blinder A. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, **8**, 436-455, (1973).

Fortin N., Lemieux T. and Firpo S. Decomposition methods in economics *NBER, wp 16045, National Bureau of Economic Research, Cambridge MA, USA*, (2010).

Gregory M. Gender and Economic inequality. *Oxford Handbook of Economic Inequality, Oxford University Press, Oxford, UK*, 284-312, (2009).

Mora R. A nonparametric decomposition of the Mexican American average wage differences. *Journal of Applied Econometrics*, **23**(4), 463-485, (2008).

Oaxaca R. L. Male-Female wage differentials in urban labour markets. *International Economic Review*, **14**, 673-709, (1973).

Rosenbaum P. and Rubin D. The central rolel of propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41-55, (1983).

Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood *Journal of Statistical Planning and Inference*, **90**, 227-244, (2000).

Sugiyama M. and Müller K. Input-dependent estimation of generalization error under covariate shift *Statistics and Decisions*, **23**(4), 249-279, (2005).

STATEC Egalité hommes-femmes, mythe ou réalité? *Cahier économique 105, Luxembourg*, (2007)

Van Kerm P. Generalized measures of wage differentials. *IRISS wp 2009-08, CEPS/INSTEAD, Differdange, Luxembourg*, (2009).

Acknowledgements

This research is part of the MeDIM project (*Advances in the Measurement of Discrimination, Inequality and Mobility*) supported by the Luxembourg 'Fonds National de la Recherche' (contract FNR/06/15/08) and by core funding for CEPS/INSTEAD from the Ministry of Higher Education and Research of Luxembourg.