DISCUSSION PAPER SERIES

IZA DP No. 8855

The Wage Return to Education: What Hides Behind the Least Squares Bias?

Corrado Andini

February 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Wage Return to Education: What Hides Behind the Least Squares Bias?

Corrado Andini
Universidade da Madeira, CEEAplA and IZA

Discussion Paper No. 8855
February 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8855
February 2015

ABSTRACT

The Wage Return to Education: What Hides Behind the Least Squares Bias?*

This paper combines the approach by Guimarães and Portugal (2010) with the methodology of Gelbach (2015) to investigate the determinants of the least squares bias of the wage return to education. We find that disregarding individual fixed effects is highly problematic, accounting for 95% of the bias. In contrast, disregarding firm fixed effects has marginal consequences.

JEL Classification: I21, J31

Keywords: wages, education, least squares

Corresponding author:

Corrado Andini
Universidade da Madeira
Campus da Penteada
9000-390 Funchal
Portugal
E-mail: andini@uma.pt

* The views expressed in this paper are those of the author and do not necessarily reflect those of the institutions he is affiliated with. Part of this paper has been written while the author was visiting the Banco de Portugal.

1. Introduction

Individuals invest a lot of resources in education. The latter is seen to have positive effects on human development. However, the estimation of the magnitude of these effects is surrounded by technical difficulties. Though economists started to be interested in the link between individual wages and schooling a long time ago, the first agent-based models only appeared in the 1960s and 1970s. After the seminal contributions by Becker (1964), Ben-Porath (1967) and Mincer (1974), a lot of progress has been made. Nevertheless, a number of questions remain open. One of these, perhaps not the most important but for sure the most studied, is about the magnitude of the wage return to schooling.

The vast literature on schooling returns has stressed that the least squares estimation of the wage return to education is, in general, biased. This point was originally made by Griliches (1977) who discussed the existence of two types of biases. The first, known as the "ability bias", is due to the correlation between individual unobserved ability and schooling. The second, known as the "attenuation bias", is due to measurement errors in the schooling variable.

In general, the least squares estimator of the schooling coefficient is biased if an omitted variable in the wage-schooling model is correlated with schooling. Besides the case of individual unobserved heterogeneity, which is often discussed by researchers, other factors may play a role. An example is firm unobserved heterogeneity: employers with higher unobserved managerial skills may provide incentives for their high-school workers to complete college education.

The most straightforward way to solve the omitted-variable problem is to estimate a wage-schooling model without omitted variables. In many cases, this approach is not feasible due to data constraints. In this paper, we take advantage of a rich matched employer-employee dataset, known as *Quadros de Pessoal* (QP). We work with over 24 million observations (24344086), tracking the whole private-sector population of Portuguese wage earners and related employers from 2002 to 2012. Thus, inference is unlikely to be an issue here.

In particular, using the QP dataset and the approach proposed by Guimarães and Portugal (2010), we are able to estimate two high-dimensional vectors of fixed effects, one for employees and one for employers, in a wage-schooling model which also controls for observed individual and firm characteristics as well as year fixed effects.

Then, using the methodology proposed by Gelbach (2015), we look at the difference between the schooling coefficient in a restricted model - where wages are explained by education attainment only - and the schooling coefficient in an unrestricted model. We decompose the difference into four groups of determinants: observed covariates, year fixed effects, individual fixed effects, and firm fixed effects.

The paper focuses on the wage return to graduate education in Portugal, but our methodology can be applied to other levels of education attainment and other countries.

2. Data

QP data are gathered annually by the Portuguese Ministry of Employment.

All firms with wage earners are required to complete the survey. The mandatory nature of the survey implies extremely high rates of response. Public-sector employees and household workers are excluded from coverage.

The data refer to the firm situation during a reference week in October. Workers on short-term leave (sickness, vacation, strikes, and maternity leave) and those on leave for compulsory military service are also reported.

The data include both firm characteristics (such as location, industry, employment, sales, ownership, capital) and worker characteristics (such as gender, age, occupation, schooling, date of hire, hours of work, and earnings).

Schooling data report the highest completed level of education, which we have divided in five broad categories: first cycle (4 years) or less; second cycle (6 years); third cycle (9 years); high school (12 years); graduate (bachelor degree or more).

Descriptive statistics are provided in Table 1.

3. Empirical strategy

To apply the Gelbach (2015) methodology, we define two models: the restricted model and an unrestricted one. The restricted model is as follows:

(1) $$lnw_{ijt} = \alpha + \beta graduate_{ijt} + e_{ijt}$$

where lnw_{ijt} is the logarithm of the gross hourly wage and $graduate_{ijt}$ is a variable equal to one if the individual i employed in firm j at time t holds at least a bachelor degree (zero otherwise). If e_{ijt} is uncorrelated with $graduate_{ijt}$, then the least squares estimate of β is unbiased.

The unrestricted model is as follows:

(2) $$lnw_{ijt} = \alpha + \beta graduate_{ijt} + \psi_1 X_{ijt} + \psi_2 f_t + \psi_3 \hat{f}_i + \psi_4 \hat{f}_j + u_{ijt}$$

where X_{ijt} is a matrix of observed covariates including both individual and firm characteristics (individual age, individual age squared, individual tenure, firm size, firm age, and firm capital), f_t is a vector of year fixed effects accounting for the business

cycle, $\hat{f}_i$ is a vector of estimated individual fixed effects, and $\hat{f}_j$ is a vector of estimated firm fixed effects. The coefficients ψ_3 and ψ_4 are expected to be close to one.

Both $\hat{f}_i$ and $\hat{f}_j$ are obtained by applying the estimation approach proposed by Guimarães and Portugal (2010) to a slightly modified version of model (2) where individual and firm fixed effects are vectors to be estimated rather than explanatory variables.

The education variable is time-varying, allowing us to estimate both the coefficient β and the individual fixed-effect vector, even when the latter is correlated with education.

If model (2) holds, then e_{ijt} is correlated with $graduate_{ijt}$, and the least squares estimate of β in model (1) is biased. To the best of our knowledge, we are the first to estimate to what extent this bias is due to missing observed covariates, missing year effects, missing individual effects, and missing firm effects. This is the key contribution of this paper.

4. Results

The least squares estimates for model (1) and model (2) are provided in Table 2 and Table 3. The unrestricted model (2) explains 99.82% of the wage variability.

The baseline group ($graduate_{ijt} = 0$) is formed by individuals with secondary, primary and no education at all. Thus, the estimate of β in the restricted model tends to be larger than in the case where only high-school workers are used as baseline[1]. Obviously, the choice of the baseline group is not the key point of this paper. The same holds for the choice of the group of interest (the chosen group is quite heterogeneous, including even individuals with a doctoral degree).

The coefficient of graduate education in the restricted model is 0.814. The coefficient of graduate education in the unrestricted model is 0.105. This means that, controlling for all the covariates in (2), the causal effect of graduate education on hourly wages is 10.5%. The difference between the two estimates is equal to 0.709.

The Gelbach decomposition in Table 4 suggests that the difference 0.709 ≅ –0.030+0.005+0.732+0.002. The value –0.030 is due to the observed covariates. The value 0.005 is due to the year fixed effects. The value 0.732 is due to individual fixed effects and the value 0.002 is due to firm fixed effects.

The difference (0.709) is the sum of a large upward bias (0.005+0.732+0.002) and a small downward bias (–0.030). The whole bias in absolute value is 0.769.

Disregarding our observed covariates implies a small downward bias. Disregarding individual fixed effects explains 95% of the whole bias in absolute value and 99% of the upward bias. Intuitively, when a wage-schooling model does not control for individual

[1] See the online appendix at http://www3.uma.pt/andini/Documentos/online_appendix.pdf

fixed effects, since the education variable is highly persistent over time, its coefficient mainly captures the effect of individual fixed characteristics. Thus, the effect of education on wages is overstated.

Individual fixed effects include both observed time-invariant characteristics, such as gender (information on race is missing in the data), and unobserved time-invariant characteristics, such as genetic ability. To disentangle, we perform the above exercise for males and females, separately. Since the results for males and females in Table 5 and Table 6 are very similar to those reported for all individuals, gender is likely to play a minor role.

5. Conclusion

This paper has combined the approach by Guimarães and Portugal (2010) with the methodology of Gelbach (2015) to investigate the determinants of the least squares bias of the wage return to education. We have used matched employer-employee data for Portugal, over the 2002-2012 period. We have found that disregarding individual fixed effects is highly problematic, accounting for 95% of the whole least squares bias. In contrast, disregarding firm fixed effects has marginal consequences.

References

Becker, G. (1964) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: Columbia University Press.

Ben-Porath, J. (1967) "The production of human capital over the life cycle". *Journal of Political Economy*, 75(4): 352-365.

Gelbach, J. (2015) "When do covariates matter? And which ones, and how much?". *Journal of Labor Economics*, forthcoming.

Guimarães, P., and Portugal, P. (2010) "A simple feasible alternative procedure to estimate models with high-dimensional fixed effects". *Stata Journal*, 10(4): 628-649.

Griliches, Z. (1977) "Estimating the returns to schooling: Some econometric problems". *Econometrica*, 45(1): 1-22.

Mincer, J. (1974) *Schooling, Experience, and Earnings*. New York: Columbia University Press.

Table 1. Descriptive statistics

	mean	sd	min	max
lnw	1.36	0.53	-2.9	8
age	37.99	10.89	15.0	65
1st cycle or less	0.24	0.43	0.0	1
2nd cycle	0.21	0.41	0.0	1
3rd cycle	0.23	0.42	0.0	1
high school	0.19	0.40	0.0	1
graduate	0.13	0.33	0.0	1
female	0.44	0.50	0.0	1
lntenure	3.80	1.42	0.0	6
lnsize	3.99	2.32	0.0	10
lnfage	3.39	1.93	0.0	8
lncap	12.59	3.16	-4.6	23

Table 2. Restricted model, all individuals

lnw	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
graduate	.8145058	.0030897	263.62	0.000	.8084501	.8205614
cons	1.267337	.0010912	1161.40	0.000	1.265198	1.269475

Table 3. Unrestricted model, all individuals

lnw	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
graduate	.1052288	.0001581	665.48	0.000	.1049189	.1055387
age	.044351	.0000295	1503.82	0.000	.0442932	.0444088
age2	-.000284	3.64e-07	-780.87	0.000	-.0002847	-.0002833
lntenure	.0249025	.0000372	668.86	0.000	.0248295	.0249754
lnsize	.0319752	.0000312	1026.33	0.000	.0319141	.0320362
lnfage	.0085287	.0000261	327.27	0.000	.0084777	.0085798
lncap	.0021394	.0000237	90.37	0.000	.002093	.0021858
2003	-.0010705	.0002203	-4.86	0.000	-.0015022	-.0006388
2004	.0296107	.0002176	136.06	0.000	.0291842	.0300373
2005	.0479057	.0002143	223.53	0.000	.0474856	.0483257
2006	.0469013	.0002133	219.90	0.000	.0464833	.0473193
2007	.0576	.0002114	272.46	0.000	.0571856	.0580143
2008	.0563281	.0002103	267.86	0.000	.055916	.0567403
2009	.0815485	.0002122	384.29	0.000	.0811326	.0819645
2010	.0808263	.0002206	366.45	0.000	.080394	.0812586
2011	.0621374	.0002229	278.77	0.000	.0617005	.0625743
2012	.0518804	.0002437	212.84	0.000	.0514027	.0523581
indiv effects	1.000008	.0001124	8893.28	0.000	.9997874	1.000228
firm effects	1.00291	.0009521	1053.40	0.000	1.001044	1.004776
cons	.0000354	.0005872	0.06	0.952	-.0011156	.0011863

Table 4. Gelbach decomposition, all individuals

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
covariates	-.0306352	.0018488	-16.57	0.000	-.0342588	-.0270116
year effects	.0050678	.0001753	28.91	0.000	.0047243	.0054113
indiv effects	.7323195	.0029542	247.89	0.000	.7265295	.7381096
firm effects	.0025248	.0003226	7.83	0.000	.0018925	.0031571
difference	.709277	.0030878	229.70	0.000	.7032249	.715329

Table 5. Gelbach decomposition, males

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
covariates	-.0002639	.0026167	-0.10	0.920	-.0053926	.0048648
year effects	.004272	.0002456	17.40	0.000	.0037907	.0047532
indiv effects	.7664909	.0041445	184.94	0.000	.7583678	.774614
firm effects	.0020843	.0004616	4.52	0.000	.0011796	.0029891
difference	.7725833	.0044141	175.03	0.000	.7639319	.7812347

Table 6. Gelbach decomposition, females

	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
covariates	-.0577261	.002594	-22.25	0.000	-.0628102	-.0526419
year effects	.0058011	.0002507	23.14	0.000	.0053096	.0062925
indiv effects	.7312546	.0040009	182.77	0.000	.7234129	.7390962
firm effects	.0031039	.0004469	6.95	0.000	.0022281	.0039798
difference	.6824335	.0039916	170.97	0.000	.6746102	.6902569