More on identification in detailed wage decompositions

Javier Gardeazabal and Arantza Ugidos *†
University of the Basque Country

11th April 2002

Abstract

A simple solution to the identification problem in detailed wage decompositions is proposed and illustrated with an empirical application.

Keywords: wage discrimination, wage decomposition, identification.
JEL: J31, J71.

*We gratefully acknowledge financial support from the Spanish Ministerio de Ciencia y Tecnología (BEC2000-1394) and Ministerio de Trabajo y Asuntos Sociales (Instituto de la Mujer 33/00) .

†Mailing address: Javier Gardeazabal, Dpto. Fundamentos del Análisis Económico, Universidad del País Vasco, 48015 Bilbao, Spain. E-mail: jepgamaj@bs.ehu.es

1 Introduction

Wage decompositions of the sort proposed by Oaxaca (1973) and Blinder (1973) are often used by researchers to decompose wage differentials of two demographic groups into differences in characteristics and differences in returns to those characteristics. The latter is used as an estimate of the degree of discrimination in the labor market. Researchers are often interested in dividing differences in returns to characteristics into the separate contributions of each individual variable. In empirical applications, most explanatory variables are categorical. All explanatory variables except experience and tenure are measured with dummies. In a recent paper Oaxaca and Ransom (*The Review of Economics and Statistics*, 1999) point out that "...conventional decomposition methodology cannot identify the separate contributions of dummy variables to the wage decomposition, because it is only possible to estimate the relative effects of a dummy variable. So the discrimination component is not invariant to the choice of the "left out" reference group." They show that standard estimates of the contributions of individual dummy variables to the wage decomposition are not identified but the overall measure of wage discrimination is. More recently, Horrace and Oaxaca (2001) prove that the "intuitively appealing" method for estimating gender wage gaps by industry proposed by Fields and Wolff (1995) also suffers from an identification problem, as its results vary according to the choice of the left out reference group. They propose another measure of the overall wage gap by industry invariant to the choice of left out reference group. Nevertheless, identification of the contribution of individual dummy variables to the wage decomposition remains an issue. The failure to identify the contribution of individual dummy variables to the wage discrimination raises one additional problem: It is not possible to compare the results of different studies as they may use different left out groups.

In this note we propose a method for solving this identification problem. Identification can be attained by means of a normalization restriction on the coefficients of each set of dummy variables. This normalization restriction allows us to estimate the contribution of all individual dummy variables, including the typical left out reference groups. This way of proceeding is very well known in econometrics, but is not usually employed because the use of a left out reference group poses no identification problems for most econometric applications.

We first illustrate the identification problem in the simplest of all possible scenarios in section 2. Section 3 describes the proposed solution to the identification problem. Section 4 illustrates the magnitude of the problem and the solution with an empirical application to the Spanish labor market.

Section 5 concludes.

2 The identification problem

To illustrate this potential identification problem, let us consider the following example. Suppose the only explanatory variable is education and there are J categories of studies. The equation considered is a linear regression of the form

$$w_g = \beta_{0g} + \sum_{j=1}^{J} \beta_{jg} D_{jg} + u_g \tag{1}$$

where w_g is the (log) wage of a person belonging to demographic group g, β_{0g} and β_{jg} are parameters, D_{jg} is a dummy variable that takes the value of one when the individual has studies in category j and zero otherwise and u_g is a zero mean disturbance term. This model cannot be estimated, since there is exact multicolinearity (the constant term is the sum of the J dummies). Typically, one of the dummies, let us say the first one, is excluded from the regression to attain identification. By adding and subtracting β_{1g} in (1) and taking into account that $\sum_{j=1}^{J} D_{jg} = 1$, we obtain

$$w_g = \gamma_{0g} + \sum_{j=2}^{J} \gamma_{jg} D_{jg} + u_g, \tag{2}$$

where $\gamma_{0g} = \beta_{0g} + \beta_{1g}$ and $\gamma_{jg} = \beta_{jg} - \beta_{1g}$ and the regression equation includes all dummies but the first one. As long as the interpretation of the transformed coefficients is taken into account, this specification poses no problem for most econometric applications. However, for discrimination studies this specification may result in erroneous inference.

For the case of wage discrimination between males (m) and females (f), assuming that the estimated male wage structure is nondiscriminatory, the wage decomposition is[1]

[1] Oaxaca and Ransom (1995) examine four different methods of estimating wage discrimination. These methods differ with respect to the implicitly assumed nondiscriminatory wage structure. The problem of identifying the contribution of a dummy variable to discrimination arises in all these methods. The solution proposed in this paper, illustrated for the case when the estimated male wage structure is nondiscriminatory, can also be used with the other methods used for estimating discrimination.

$$\overline{w}_m - \overline{w}_f = \underbrace{\widehat{\gamma}_{0m} - \widehat{\gamma}_{0f} + \sum_{j=2}^{J} \left(\widehat{\gamma}_{jm} - \widehat{\gamma}_{jf}\right) \overline{D}_{jf}}_{discrimination} + \underbrace{\sum_{j=2}^{J} \widehat{\gamma}_{jm} \left(\overline{D}_{jm} - \overline{D}_{jf}\right)}_{characteristics}$$

where $\overline{w}_g$, $g = m, f$ are the sample averages of (log) wages, $\widehat{\gamma}_{jg}$ are OLS estimates and $\overline{D}_{jg}$, are the average value of the dummy variables (the proportion of individuals in group g with education level j). According to this decomposition, the contribution of dummy variable j to discrimination is $\left(\widehat{\gamma}_{jm} - \widehat{\gamma}_{jf}\right) \overline{D}_{jf}$. This is the product of two terms: the difference in returns to dummy variable j, $\left(\widehat{\gamma}_{jm} - \widehat{\gamma}_{jf}\right)$ and the proportion of females with studies in category j, $\overline{D}_{jf}$. The contribution of each variable to discrimination is not invariant to the left out reference group. Changing the left out reference group always generates a change in the quantitative contribution of a dummy variable to discrimination and sometimes even changes the qualitative result from discrimination against one group to discrimination against the other group.

The following example illustrates the case of a qualitative change in results. Suppose that we find that $\widehat{\gamma}_{2m} < \widehat{\gamma}_{2f}$. Can we say that the return to education level $j = 2$ is greater for women than for men? The answer is no: suppose that $\beta_{1m} > \beta_{1f}$ and $\beta_{2m} > \beta_{2f}$, that is, returns to levels 1 and 2 of education are greater for men than for women. However, if $(\beta_{1m} - \beta_{1f}) > (\beta_{2m} - \beta_{2f}) > 0$, then $\gamma_{2m} < \gamma_{2f}$. Provided our econometric estimates are good enough we would get the result $\widehat{\gamma}_{2m} < \widehat{\gamma}_{2f}$, apparently indicating discrimination against men when, in fact, there is discrimination against women. Of course, the change in the contribution of a dummy to discrimination may be only quantitative, depending on the left out group. Therefore, it is very important to take into account the return of the omitted category in evaluating the difference in returns between men and women.

3 An identification restriction

The contribution to discrimination of each individual dummy variable can be easily identified through the introduction of an identification restriction. We estimate equation (1) subject to

$$\sum_{j=1}^{J} \beta_{jg} = 0. \tag{3}$$

This restriction can be interpreted as a normalizing restriction on the coefficients of the dummy variables. This sort of restriction is typically introduced in ANOVA analysis. Formally, by imposing restriction (3) we are restricting the feasible linear combinations of the set of dummies to be orthogonal to the constant term. Real applications include several sets of dummy variables, thus requiring one additional restriction such as (3) for each set of dummy variables, but for the sake of simplicity, we will continue our exposition with the example of just one set of dummy variables.

Introducing the normalizing restriction (3) slightly complicates the estimation problem as the OLS estimator cannot now be used directly. Estimation of equation (1) subject to (3) can be pursued by means of restricted least squares, but a much simpler way to proceed is as follows. Solving for β_{1g} in equation (3) and substituting the result in (1) we get

$$w_g = \beta_{0g} + \sum_{j=2}^{J} \beta_{jg}(D_{jg} - D_{1g}) + u_g, \tag{4}$$

where the dummy variables are expressed in differences with respect to the dummy of the left out reference group. Therefore, the parameters can be easily estimated by OLS on the transformed regression (4). Provided no additional econometric problems are present, β_{0g}, β_{jg}, $j = 2, ..., J$, can be consistently estimated by OLS, and hence are identified. In addition, a consistent estimate of the coefficient of the omitted category is given by $\widehat{\beta}_{1g} = -\sum_{j=2}^{J} \widehat{\beta}_{jg}$, where $\widehat{\beta}_{jg}$, $j = 2, ..., J$ are OLS estimates. Thus, β_{1g} is also identified. Furthermore, a consistent estimate of the standard error of $\widehat{\beta}_{1g}$ is $\sqrt{\mathbf{1}'\widehat{V}\mathbf{1}}$, where $\mathbf{1}$ is a $(J-1)$ vector of ones and $\widehat{V}$ is a consistent estimate of the covariance matrix of $(\widehat{\beta}_{2g}, \widehat{\beta}_{3g}, ..., \widehat{\beta}_{Jg})'$.

Finally, the estimated wage decomposition is

$$\overline{w}_m - \overline{w}_f = \widehat{\beta}_{0m} - \widehat{\beta}_{0f} + \sum_{j=1}^{J} \left(\widehat{\beta}_{jm} - \widehat{\beta}_{jf}\right) \overline{D}_{jf} + \sum_{j=1}^{J} \widehat{\beta}_{jm} \left(\overline{D}_{jm} - \overline{D}_{jf}\right), \tag{5}$$

where now $\left(\widehat{\beta}_{jm} - \widehat{\beta}_{jf}\right) \overline{D}_{jf}$ is an estimate of the "true" contribution of the $j-th$ dummy variable to the wage gap, usually attributed to discrimination. Notice that the dummy variables used in decomposition (5) are not the transformed dummies but the originals.

4 Application to the Spanish gender wage gap

The data come from the Spanish sample of the Survey of Wage Structure carried out in the European Union in October 1995. In Spain the survey

was conducted by the *Instituto Nacional de Estadística* (INE) at establishment level. This survey covers information on individuals working for firms with ten or more employees from all sectors and provinces. To give an idea of how representative the sample is, workers at firms with ten or more employees accounted for 70.75% (72.95% of men and 66.74% of women) of the total working population in Spain in October 1995. The SWE contains very detailed information about each worker's wage, individual and job characteristics.[2] Following the usual practice in the field, the factors controlled for in wage equations are education, experience (proxied by age) and tenure. To consider the demand side of the labor market, sector and regional dummies are also included in the wage equations. We also control for firm size, the type of labor agreement that determines wages at firms, whether firms are publicly or privately owned, and the occupation and type of contract of each individual. Except for age and tenure, all explanatory variables are categorical.

To illustrate the identification problem and the solution method proposed, we report the results for education dummies, leaving out the results of the other variables. We group individuals into five education groups: EDU1, EDU2,...,EDU5.[3] Table 1 reports the contribution to discrimination of each education dummy using EDU1 as the left out reference group in column 1, EDU2 in column 2, ... , EDU5 in column 5, and working according to our proposed method in column 6.[4] Entries in columns 1 to 5 are calculated as $100\left(\widehat{\gamma}_{jm}-\widehat{\gamma}_{jf}\right)\overline{D}_{jf}/(\overline{w}_m-\overline{w}_f)$, and entries in column 6 are calculated as $100\left(\widehat{\beta}_{jm}-\widehat{\beta}_{jf}\right)\overline{D}_{jf}/(\overline{w}_m-\overline{w}_f)$. For instance, the percentage contribution of EDU1 to discrimination is 0.142 when the left out reference group is EDU2 and 0.222 when the left out reference group is EDU3. It is clear that the quantitative contribution of one individual dummy changes with the left out reference group. Sometimes, the choice of left out group can change the contribution of one dummy variable from negative to positive. For instance, the contribution of EDU2 to discrimination varies from -10.867 (ref: EDU4) to 3.218 (ref: EDU3). The last column of Table 1 reports the results obtained

[2]From an original sample size of 177,114 we removed all those observations corresponding to trainees (1,170), those who did not work the entire month of October (5,192), those who worked part time (6,306), those who did not report their wages (25) and those whose reported wage was less than 100 Spanish pesetas per hour (151). The final sample size is 164,270: 129,061 men and 35,209 women.

[3]Respectively, less than primary studies, primary studies, secondary studies (including high school and three-year vocational studies), three-year college education (also including five-year vocational studies) and five-year college education (including Master's diplomas and Ph.D.'s).

[4]The set of excluded dummies for all other variables is the same for all columns.

imposing the normalization restriction on the coefficients. Using this normalization restriction we are able to identify the "true" contribution of each individual dummy variable invariant with the left out reference group.

The example used in section 2 can be illustrated using the returns to EDU4 and EDU5 reported in column 6. Men earn a higher return for education levels EDU4 and EDU5 than women. The contribution of EDU4 to discrimination is 1.678 percentage points and the contribution of EDU5 is 0.266. Therefore, EDU4 and EDU5 contribute positively to discrimination against women. However, inference based on the regression omitting EDU4 shows the contribution of EDU5 to discrimination to be -0.589, indicating discrimination against men.

5 Conclusions

In this note we provide a simple way of identifying the contribution of individual dummy variables to wage discrimination. Identification is attained by means of a normalizing restriction on the coefficients of each set of dummy variables. The introduction of these restrictions allows us to identify the contribution to wage discrimination of all categories, including those typically left out as reference groups. One advantage of this method is to provide a unified way of dealing with dummy explanatory variables, facilitating the comparison of results obtained by different researchers.

References

[1] Blinder, A. S. (1973) "Wage discrimination: reduced form and structural estimates," *Journal of Human Resources* 8, 436-455.

[2] Fields J. and E. N. Wolff (1995) "Interindustry wage differentials and the gender wage gap," *Industrial and Labor Relations Review*, 49 (1), 105-20.

[3] Horrace, W.C. and R. L. Oaxaca (2001) "Inter-industry wage differentials and the gender wage gap: an identification problem," *Industrial and Labor Relations Review*, 54 (3), 611-18.

[4] Oaxaca, R. (1973), "Male-female wage differentials in urban labor markets." *International Economic Review*, 14(3), 693-709.

[5] Oaxaca, R. and M. R. Ransom (1994), "On discrimination and the decomposition of wage differentials," *Journal of Econometrics*, 61, 5-21.

[6] Oaxaca, R. and M. R. Ransom (1999), "Identification in detailed wage decompositions," *The Review of Economics and Statistics* 81(1), 154-7.

Table 1: Contribution to discrimination

	1	2	3	4	5	6
EDU1		0.142	0.222	-0.131	-0.010	0.045
EDU2	-5.645		3.218	-10.867	-6.045	-3.868
EDU3	-3.432	-1.246		-5.454	-3.587	-2.744
EDU4	1.252	2.606	3.377		1.156	1.678
EDU5	0.049	0.739	1.132	-0.589		0.266