Using normalized equations to solve the indetermination problem in the Oaxaca-Blinder Decomposition: an application to the gender wage gap in Brazil

Luiz Guilherme Scorzafave
Elaine Toldo Pazello

1. Introduction

The decomposition proposed by Oaxaca (1973) and Blinder (1973) has been applied in hundreds of studies around the world and also in Brazil[1]. In general, the methodology is applied to separate the wage differentials of distinct groups (men/women, white/non-white) in two components. One related to differences in observable characteristics of the two groups. For example, men could earn more because they have more experience or are more educated than women. Nevertheless this part, called "explained differential", is responsible for only a fraction of the wage gap between the groups. The remaining gap, called "unexplained differential", is attributable to different returns of the characteristics between the two groups. For instance, men and women with the same level of education could receive a different reward.

Some authors attribute this unexplained gap to discrimination, but there is a controversy in the literature if this conclusion can be done. The argument against this idea is that we can only say that a differential is attributable to discrimination if the estimation has considered all variables that affect wages and were different between groups. Obviously, it is hard to believe that any regression specification can assure this.

Apart from this debate, the Oaxaca-Blinder decomposition can be used to assess the contribution of each variable to the explained and unexplained gap. However, Oaxaca and Ramson (1999) show that the differential share attributable to the dummy variables of the model depends on the choice of the reference group. On

[1] For Brazil, see, for instance, Lovell and Wood (1998), Kassouf (1998) and Ometto et alii (1999).

the other hand, the overall gap fraction due to "explained" and "unexplained" is not plagued by this problem.

Additionally, Yun (2005a, 2005b) proposes a method to implement the Oaxaca-Blinder decomposition that solves this indetermination problem. He estimates "normalized" equations, imposing the restriction that the sum of dummies coefficients is zero.

In the Brazilian case, there is not any paper that considers this indetermination related to dummy variables. So, the aim of this paper is to present the indetermination problem and the solution proposed by Yun (2005b) applying this to the Brazilian case. In order to achieve this, in the next section, we show the indetermination problem and the solution proposed by Yun (2005b). Next, we apply this solution to three years in Brazil: 1988, 1996 and 2004. The choice of these years will, particularly, allow a comparison with the results of Giuberti e Menezes-Filho (2005) that uses 1988 and 1996 in their analysis.

2. Identification: problem and solution

2.1. The problem

We present the identification problem in the Oaxaca-Blinder decomposition based on Oaxaca and Ransom (1999). Suppose that you estimate separate wage regressions for males and females. Let V be a variable, defined by a set of dummy variables, denoted by $\{V_{ik}|k=1,...K_1\}$, where $\sum_{k=1}^{K_1} \overline{V}_{ik} = 1$ and i= m,f. For instance, V could be region of residence and, in Brazil, K_1=5 (South, Southeast, North, Northeast, Mid-West). Without loss of generality, $\overline{V}_{i1}$ will be the omitted category. The separately estimated wage equations for males and females at the sample means are given by:

$$\overline{Y}_m = \hat{\beta}_{m0} + \sum_{k=2}^{K_1} \overline{V}_{mk}\hat{\delta}_{mk} + \sum_{j=1}^{N} \overline{X}_m^{(j)\prime}\hat{\beta}_m^{(j)} = \sum_{k=1}^{K_1} \overline{V}_{mk}\hat{\theta}_{mk} + \sum_{j=1}^{N} \overline{X}_m^{(j)\prime}\hat{\beta}_m^{(j)}$$

$$\overline{Y}_f = \hat{\beta}_{f0} + \sum_{k=2}^{K_1} \overline{V}_{fk}\hat{\delta}_{fk} + \sum_{j=1}^{N} \overline{X}_f^{(j)\prime}\hat{\beta}_f^{(j)} = \sum_{k=1}^{K_1} \overline{V}_{fk}\hat{\theta}_{fk} + \sum_{j=1}^{N} \overline{X}_f^{(j)\prime}\hat{\beta}_f^{(j)}$$

where $\hat{\beta}_{i0}$ is the estimated intercept, $\overline{Y}_i$ is the mean log wage, $\hat{\beta}_i^{(j)}$ is a column vector of estimated slope coefficients for the set of regressors comprising the *jth* variable, $\overline{X}_i^{(j)}$ is a row vector of regressor means for the set of regressors comprising the *jth* variable, and $\hat{\delta}_{ik}$ is the estimated coefficient for the dummy variable V_{ik}, $\hat{\delta}_{ik} = \hat{\theta}_{ik} - \hat{\theta}_{i1}$ and, finally, $\hat{\beta}_{i0} = \hat{\theta}_{i1}$.

Still following Oaxaca and Ramson (1999), we can decompose the mean wage differences as follows:

$$\overline{Y}_m - \overline{Y}_f =$$

$$= (\hat{\beta}_{m0} - \hat{\beta}_{f0}) + \sum_{k=2}^{K_1} \overline{V}_{fk} (\hat{\delta}_{mk} - \hat{\delta}_{fk}) + \sum_{j=1}^{N} \overline{X}_f^{(j)} ' \Delta \hat{\beta}^{(j)} + \sum_{k=2}^{K_1} (\overline{V}_{mk} - \overline{V}_{fk}) \hat{\delta}_{mk} + \sum_{j=1}^{N} \Delta \overline{X}^{(j)} ' \hat{\beta}_m^{(j)} =$$

$$= \sum_{k=2}^{K_1} \overline{V}_{fk} (\hat{\theta}_{mk} - \hat{\theta}_{fk}) + \sum_{j=1}^{N} \overline{X}_f^{(j)} ' \Delta \hat{\beta}^{(j)} + \sum_{k=1}^{K_1} (\overline{V}_{mk} - \overline{V}_{fk}) \hat{\theta}_{mk} + \sum_{j=1}^{N} \Delta \overline{X}^{(j)} ' \hat{\beta}_m^{(j)}$$

where the last two terms in each equality measure the "endowment" effect, while the others capture the "discrimination" effect.

The first thing that Oaxaca and Ramson (1999, p. 156) assign about this decomposition is that "the estimated overall discrimination and the estimated overall endowment effect are invariant to the choice of left-out reference group and the suppression of the constant term in the absence of a left-out reference group." The most important issue, however, is that the contribution of variable V to the discrimination effect is sensitive to the left-out reference group, because the intercept varies with changes in the left-out reference group. To see this, suppose that the last dummy variable has been chosen as left-out reference group. In this case, the "discrimination" effect would be $\sum_{k=1}^{K_1 - 1} \overline{V}_{fk} (\hat{\phi}_{mk} - \hat{\phi}_{mf})$, where $\hat{\phi}_{ik} = \hat{\theta}_{ik} - \hat{\theta}_{iK_1}$.

Oaxaca and Ramson (1999) shows that if there is only one set of dummy variables in the regression, this problem could be solved incorporating $(\hat{\beta}_{m0} - \hat{\beta}_{f0})$ to the contribution of variable *V* to the "discrimination" effect. However, this solution is not valid if there is more than one set of dummy variables in the estimated equation, a very common situation in the context of wage regressions.

2.2. The solution

Yun (2005a) proposed a methodology to disentangle the identification problem, based on normalized regressions. The idea concerning normalized regressions is that if

> "alternative reference groups yield different estimates of the characteristics and coefficients effects for each individual variable, then it is natural to obtain estimates of the two effects for every possible specification of the reference groups and take the average of the estimates of the two effects with various reference groups as the "true" contributions of individual variables to wage differentials" (Yun, 2005a, p. 766)

But Yun (2005b) shows that we do not need to proceed this cumbersome way. It is possible to implement the method estimating only one equation. To illustrate the method, we follow Yun (2005b) and suppose that we have two sets of dummy variables (*d*'s and *q*'s) and also *L* continuous variables (*z*'s) in the model.

$$y = \alpha + \left(\sum_{j=2}^{J} d_j \gamma_j + \sum_{k=2}^{K} q_k \theta_k\right) + \sum_{l=1}^{L} z_l \delta_l + e \qquad (1)$$

This equation is called "usual regression" and he proposes an alternative specification that does not omit the reference group:

$$y = \alpha^* + \left(\sum_{j=1}^{J} d_j \gamma_j^* + \sum_{k=1}^{K} q_k \theta_k^*\right) + \sum_{l=1}^{L} z_l \delta_l + e \qquad (2)$$

Yun (2005b) shows that if we estimate a model omitting, for example, the first category of variable *d*, we can obtain the estimates for γ_j that would prevailed if the group *r* is omitted, simply doing $\gamma_j - \gamma_r$, and the intercept changes from $\alpha + \gamma_1$ to $\alpha + \gamma_r$. But taking on this averaging approach implies to impose that $\sum_{j=1}^{J} \gamma_j^* = 0$ and $\sum_{k=1}^{K} \theta_k^* = 0$, as Suits (1984) states. Particularly, "since these restrictions do not have unique solutions, he specifies the coefficients of the normalized regression as $\gamma_j^* = \gamma_j + m_\gamma$ and $\theta_j^* = \theta_j + m_\theta$, and refines the problem of deriving the normalized regressions as finding values of m_γ and

m_θ. It turns out that their values are $m_\gamma = -\sum_{j=1}^{J} \gamma_j / J$ and $m_\theta = -\sum_{k=1}^{K} \theta_k / K$, where $\gamma_1 = \theta_1 = 0$" (Yun, 2005b, p. 3). Considering this, Yun (2005b) proposes the "normalized equation":

$$y = (\alpha + \bar{\gamma} + \bar{\theta}) + \left(\sum_{j=1}^{J} d_j (\gamma_j - \bar{\gamma}) + \sum_{k=1}^{K} q_k (\theta_k - \bar{\theta})\right) + \sum_{l=1}^{L} z_l \delta_l + e \qquad (3),$$

where $\bar{\gamma} = \sum_{j=1}^{J} \gamma_j / J$, $\bar{\theta} = \sum_{k=1}^{K} \theta_k / K$ and $\gamma_1 = \theta_1 = 0$.

If we estimate the equation (2) for men and women separately, we can implement the Oaxaca decomposition of the wage equation that is invariant to the choice of the omitted category in the dummy variables.

3. An application to the Brazilian case

In this section we apply the solution provided by Yun (2005b) to solve the identification problem. Initially, we obtain the normalized regressions and after that we apply the Oaxaca decomposition. We use data from a Brazilian National Household Survey (*Pesquisa Nacional de Amostra por Domicilios – PNAD*), conducted annually by *Instituto Brasileiro de Geografia e Est*atística (*IBGE*), for three years (1988, 1996 and 2004). The table 1 presents the description of the data.

Table 1 — Proportion of workers in each group and wage gap

	1988		1996		2004	
Variables	**Men**	**Women**	**Men**	**Women**	**Men**	**Women**
Wage gap	0.486		0.279		0.216	
25-29 years old	0.232	0.243	0.203	0.201	0.203	0.198
30-34 years old	0.215	0.225	0.214	0.221	0.193	0.196
35-39 years old	0.187	0.192	0.194	0.205	0.184	0.191
40-44 years old	0.152	0.153	0.168	0.174	0.172	0.177
45-49 years old	0.118	0.110	0.131	0.125	0.142	0.140
50-54 years old	0.096	0.077	0.090	0.075	0.106	0.097
0-3 years of schooling	0.274	0.247	0.222	0.181	0.164	0.120
4 years of schooling	0.239	0.199	0.155	0.137	0.116	0.095
5-7 years of schooling	0.101	0.084	0.164	0.140	0.163	0.135
8 years of schooling	0.093	0.081	0.120	0.104	0.118	0.102
9-10 years of schooling	0.036	0.038	0.047	0.045	0.055	0.051
11 years of schooling	0.125	0.165	0.162	0.197	0.238	0.273
12 + years of schooling	0.128	0.177	0.130	0.196	0.147	0.223
White	0.632	0.616	0.608	0.618	0.550	0.579
North	0.038	0.039	0.047	0.044	0.062	0.055
Mid-West	0.070	0.070	0.075	0.074	0.083	0.083
Northeast	0.182	0.199	0.196	0.208	0.211	0.206
Southeast	0.551	0.536	0.519	0.512	0.484	0.491
South	0.150	0.146	0.163	0.162	0.160	0.165
Part-time	0.017	0.156	0.031	0.151	0.034	0.145
Metropolitan Area	0.448	0.482	0.388	0.413	0.363	0.387
Observations	34570	21897	42041	28910	55159	42434

The first thing that is interesting to say is that gender wage gap is narrowing in Brazil since 1988, from 0.487 to 0.216 in 2004. In terms of education, it has occurred a substantial improvement in the Brazilian situation since 1988. For instance, the proportion of men with 11 years of schooling has increased 11.3 percentage points between 1988 and 2004. Despite this fact, women continue to be more educated than men. In turn, there is a larger fraction of women working among 30 and 44 years old, while men are the majority

among older and younger workers. With regard to the region of residence, the data shows a concentration of Brazilian workers in Southeast, Northeast and South regions and, remarkably, a falling proportion of workers living in metropolitan areas. However, the most interesting fact is the increasing proportion of men working in part-time activities along with a decrease in this number among women. Despite this, 14.5% of the female workers were in part time activities while there were only 3.4% of men in this situation in 2004.

After this short discussion of the descriptive statistics, we are able to evaluate the results of the methodology adopted by the paper. First we will analyze the results of the regression estimates for 2004, which are in the Appendix[2]. The dependent variable is the logarithm of hourly wage and the independent variables are dummies for age (25-29,30-34, 35-39, 40-44, 45-49, 50-54), schooling (0-3, 4, 5-7, 8, 9-10, 11, 12 or more), race (white, non-white), region (North, Northeast, Southeast, South, Mid-West), part-time (working less than 20 hours per week) and metropolitan area.

The coefficients signals are aligned with the expected. In terms of age, the coefficients show a positive relationship between wage and age. The results also indicate that the higher the educational level, the greater the wage. It is interesting to observe the sheepskin effect in the educational estimates. The white coefficient has the signal normally obtained, independent of the gender, that is, white workers earn more than comparable non-white workers. The regional dummies are different for women and men. For women, the results show that the wage is higher in Mid-West and Southeast compared to North and smaller in Northeast compared to North; there is not difference between South and North. For men, the wages are larger in Mid-West, Southeast and South compared to North and smaller in Northeast compared to North. The part-time coefficient has a positive signal for both women and men. Finally, the metropolitan coefficient indicates that individuals living in metropolitan areas earn higher wages than others.

The Oaxaca decomposition technique permits identifying the factors that explain the wage differential between men and women, dividing them into two types: a part attributed to the observables characteristics and another part attributed to the “market return” to these characteristics. The second part could be attributed to “discrimination”, because men and

[2] For others years, the signal and magnitude of the variables are very similar. The results can be obtained directly with the authors.

women receive different prices for their characteristics. The table 2 shows the evolution of gender wage gap and the decomposition analysis.

Table 2: Evolution of wage gender gap and of decomposition analysis

	1988	1996	2004
Variables (age, schooling, race, region, part-time, metropolitan area)	-0.1443	-0.1796	-0.1849
Coefficients	0.6307	0.4588	0.4007
Δln(wage per hour)	0.4863	0.2791	0.2158

As argued above, the data shows a significant decrease in the gender wage gap between 1988 and 2004. In 1988, men's wage was 63% higher than women's one, but in 2004 this advantage dropped to 24%. The characteristics contribute to reduce the gap and the coefficients to raise the gap. If we were sure that the model was including all characteristics that explain the gap, evidence would be indicating the existence of discrimination favoring men. However, is important to say, the main source of the fall in the gap between the years was the decline of the 'discrimination' term.

The next three tables show the gender wage gap decomposition using the traditional and the normalized equation, respectively in 1988, 1996 and 2004.

Table 3 – Gender wage gap decomposition – 1988

Independent variables	Variable	% of Δ ln(wage)	Traditional equation		Normalized equation	
			Coefficient	% of Δ ln(wage)	Coefficient	% of Δ ln(wage)
Age	0.0039	0.8	-0.0910	-18.7	-0.0177	-3.7
Schooling	-0.1185	-24.4	-0.0157	-3.2	0.0097	2.0
Race	0.0041	0.8	0.0149	3.1	0.0031	0.6
Region	0.0053	1.1	-0.0074	-1.5	0.0102	2.1
Part-time	-0.0270	-5.6	0.0025	0.5	-0.0700	-14.5
Metropolitan Area	-0.0121	-2.5	-0.0303	-6.2	0.0035	0.7
Constant			0.7576	155.8	0.6915	143.5
Total	-0.1443	-29.7	0.6304	129.7	0.6304	130.9

Remark: Δ ln(wage) = 0.4863

Table 4 – Gender wage gap decomposition – 1996

Independent variables	Variable	% of Δ ln(wage)	Traditional equation		Normalized equation	
			Coefficient	% of Δ ln(wage)	Coefficient	% of Δ ln(wage)
Age	0.0010	0.3	-0.0618	-22.1	-0.0097	-3.5
Schooling	-0.1181	-42.3	0.0747	26.7	-0.0174	-6.2
Race	-0.0021	-0.8	0.0232	8.3	0.0041	1.5
Region	0.0048	1.7	0.0313	11.2	0.0024	0.9
Part-time	-0.0592	-21.2	0.0092	3.3	-0.1401	-50.2
Metropolitan Area	-0.0060	-2.1	-0.0241	-8.6	0.0069	2.5
Constant		0.0	0.4063	145.6	0.6126	219.4
Total	-0.1796	-64.4	0.4588	164.4	0.4588	164.3
Remark: Δ ln(wage) = 0.2791						

Table 5 – Gender wage gap decomposition – 2004

Independent variables	Variable	% of Δ ln(wage)	Traditional equation		Normalized equation	
			Coefficient	% of Δ ln(wage)	Coefficient	% of Δ ln(wage)
Age	0.0001	0.1	-0.0602	-27.9	-0.0067	-3.1
Schooling	-0.1267	-58.7	0.0633	29.3	-0.0052	-2.4
Race	-0.0056	-2.6	-0.0020	-0.9	-0.0002	-0.1
Region	-0.0016	-0.7	0.0474	22.0	0.0102	4.7
Part-time	-0.0466	-21.6	0.0090	4.2	-0.1225	-56.8
Metropolitan Area	-0.0046	-2.1	-0.0376	-17.4	0.0142	6.6
Constant			0.3808	176.5	0.5109	236.7
Total	-0.1849	-85.7	0.4007	185.7	0.4007	185.7
Remark: Δ ln(wage) = 0.2158						

Columns 1 and 2 show the variables contribution to the wage gap (that is the same in the traditional and normalized regressions). The most important variables are schooling and the part-time dummy. These variables contribute to diminish the wage differential. The

estimates (in Appendix) show a positive relation between these variables and wage. So, as women are in average more educated than men and they are the majority in part-time occupations, these variables contribute to reduce the gender wage gap. The contribution of other variables is really very small.

The coefficients effects are very different when we use the traditional compared to the normalized equation. This highlights the importance of the methodology applied here. For instance, the results using the traditional equation indicate that while age contributes to diminish the gap, schooling contributes to raise it. On the other hand, when the normalized regression is utilized, the coefficients effects of the schooling and age turn on to act in the same direction and moreover lose importance. However, the main change is that part-time dummy gains relevance in this decomposition. In 1988, the return of this characteristic contributes to reduce the total differential in 14.5% and in 2004 in 56.8%, that is, the women's comparative advantage in these occupations could possibly explain the falling wage gap between 1988 and 2004.

4. Conclusion

There are hundreds of works all over the world that implement the Oaxaca-Blinder decomposition. However, most of these works are plagued by the identification problem when a set of dummy variables is used, as Oaxaca and Ramson (1999) shows.

In this paper, we apply the solution proposed by Yun (2005a, 2005b) to the Brazilian gender wage gap estimation. Our first finding is that gender gap is narrowing in Brazil since 1988. The results also show that as women are more educated and more engaged in part time activities than men, these factors contribute to reduce the gender gap. On the other hand, the difference in the constant term between men and women explain the entire wage differential. However, the increasing difference in part time coefficients between men and women is contributing to alleviate this situation and it can be pointed as responsible for the narrowing gender wage gap in Brazil since 1988.

Giuberti and Menezes-Filho (2005) that does not use the any correction to identification problem conclude that different returns related to age are important to explain

the wage gap and part-time dummy does not have importance. But, as showed here, solving the indetermination problem, these results are inverted.

References

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. **The Journal of Human Resources,** v. 8, n.7, p. 436-455.

Giuberti, A. Menezes-Filho, N. (2005). Discriminação dos rendimentos por gênero: uma comparação entre o Brasil e os Estados Unidos. **Economia Aplicada**, v.9, n.3, p.369-383.

Kassouf, A. L. (1998). Wage gender discrimination and segmentation in the brazilian labor market. Brazilian Journal of Applied Economics, v.2, n.2, p. 243-269.

Lovell, P. A. & Wood, C. H. (1998). Skin color, racial identity and life chances in Brazil. **Latin American Perspectives**, v.25, n.3, p. 90-109.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. **International Economic Review**, v.14, p.693-709.

Oaxaca, R. e Ramson, M. (1999). Identification in detailed wage decompositions. **Review of Economics and Statistics**, v. 81, n.1, p. 154-157.

Ometto, A. M., Hoffman, R., & Alves, M. C. (1999). Participação da mulher no mercado de trabalho: Discriminação em Pernambuco e São Paulo. **Revista Brasileira de Economia**, v. 53, n.3, p. 287-322.

Suits, D. (1984). Dummy variables: Mechanics v. Interpretation. **Review of Economics and Statistics**, v. 66, n. 1, p. 177-180.

Yun, M. (2005a). A simple solution to the identification problem in detailed wage decompositions. **Economic Inquiry**, v. 43, n.4, p. 766-772.

Yun, M. (2005b). Normalizad Equation and Decomposition Analysis: computation and inference. **IZA Discussion Paper**, Institute for the Study of Labor, n. 1822.

Appendix

Table A1 – Estimated Coefficients - Wage Equation – Men and Women -2004

	Men	Women
25-29 years old	-0.319	-0.437
30-34 years old	-0.176	-0.263
35-39 years old	-0.105	-0.149
40-44 years old	-0.045	-0.081
45-49 years old	0.013**	-0.022*
4 years of schooling	0.150	0.233
5-7 years of schooling	0.253	0.338
8 years of schooling	0.414	0.486
9-10 years of schooling	0.465	0.586
11 years of schooling	0.781	0.861
12 + years of schooling	1.543	1.581
White	0.191	0.187
Mid-West	0.096	0.141
Northeast	-0.327	-0.302
Southeast	0.036	0.097
South	-0.002**	0.053
Part-time	0.419	0.682
Metropolitan Area	0.191	0.088
Constant	0.317	0.698
Adjusted- R^2	0.455	0.432
F-Test p-value	0.000	0.000

Remark: ** not significant; *significant at 10% level; the remaining coefficients are significant at 5%.