Inter-American Development Bank
Banco Interamericano de Desarrollo (BID)
Research Department
Departamento de Investigación
Working Paper #612

An Extension of the Blinder-Oaxaca Decomposition to a Continuum of Comparison Groups

by

Hugo Ñopo

Inter-American Development Bank

July 2007

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Ñopo, Hugo.

An extension of the Blinder-Oaxaca Decomposition to a Continuum of Comparison Groups / by Hugo Ñopo.

p. cm. (Research Department Working paper series ; 612)
Includes bibliographical references.

1. Race discrimination—Peru—Economic aspects. 2. Mestizaje—Peru—Wages. 3. Peru—Race relations I. Inter-American Development Bank. Research Dept. II. Title. III. Series.

HD4903 Ñ78 2007
331.133 Ñ78---dc21

Inter-American Development Bank
1300 New York Avenue, N.W.
Washington, DC 20577

The Research Department (RES) produces a quarterly newsletter, *IDEA (Ideas for Development in the Americas)*, as well as working papers and books on diverse economic issues. To obtain a complete list of RES publications, and read or download them please visit our web site at: http://www.iadb.org/res.

Abstract[1]

This paper proposes an extension of the Blinder-Oaxaca decomposition from two to a continuum of comparison groups. The proposed decomposition is then estimated for the case of racial wage differences in urban Peru, exploiting a novel data set that allows the capturing of *mestizaje* (racial mixtures).

Keywords: Blinder-Oaxaca decomposition, Race, Gender, Informality.
JEL Classification Codes: C1, J1, J7, O17.

[1] The findings in this paper do not necessarily represent the views of the Inter-American Development Bank or its Board of Directors. The advice of Chris Taber is especially acknowledged. This paper was motivated by discussions with Jaime Saavedra and Máximo Torero while the author was at the Grupo de Análisis para el Desarrollo (GRADE). Email: hugon@iadb.org

1. Introduction

The Blinder-Oaxaca decomposition has been a valuable tool for the analysis of wage gaps since its conception in the early 1970s (Blinder, 1973 and Oaxaca, 1973). It decomposes the differences in earnings between two groups into two additive elements: one attributed to the existence of differences in observable characteristics between the two groups and the other attributed to differences in the rewards to those characteristics. It has been extensively used during the last three decades and it has been extended in different directions to incorporate quantile analysis (Albrecht et al., 2003), dichotomous outcomes (Fairlie, 2005), censored outcomes (Bauer and Sinning, 2005), count data (Bauer et al., 2006) and non-parametric setups (Ñopo, forthcoming).

So far, however, the decomposition has been used to compare only pairs of groups: females and males, whites and Afro-descendants, rural and urban residents, formal and informal workers, etc. This paper extends the decomposition to a continuum of comparing groups. The applications that can be foreseen for an extension like this are numerous, as a growing literature is challenging dualistic or binary approaches to informality (Maloney, 1999, 2004), racial/ethnic divides (Ñopo et al., 2007), and rural and urban differences (World Bank, 2005), three of the areas in which the traditional Blinder-Oaxaca decomposition has been already applied. Additionally, other areas in which the number of comparing groups can be naturally modeled as a continuum could benefit from this extension of the methodology. This extension to a continuum of groups can easily be combined with the other extensions outlined above that deal with different types of outcomes.

2. The Blinder-Oaxaca Decomposition. Basic Notation and An Alternative Presentation

The basic setup of the Blinder-Oaxaca decomposition works as follows. Let the comparing groups be 1 and 0. The estimator for the outcome gap between these two groups is $\overline{y}^1 - \overline{y}^0$. A departure point for the Blinder-Oaxaca decomposition is the estimation of the regressions $y_i^1 = \beta^1 \cdot x_i^1 + \varepsilon_i^1$ and $y_i^0 = \beta^0 \cdot x_i^0 + \varepsilon_i^0$, where x represents vector of observable characteristics, β their corresponding vectors of rewards and ε the

residual terms. In this way, the estimators for the expected outcomes can be then estimated as $\overline{y}^1 = \hat{\beta}^1 \cdot \overline{x}^1$ and $\overline{y}^0 = \hat{\beta}^0 \cdot \overline{x}^0$, respectively. The estimator of the gap then becomes $\overline{y}^1 - \overline{y}^0 = \hat{\beta}^1 \cdot \overline{x}^1 - \hat{\beta}^0 \cdot \overline{x}^0$ which, after adding and subtracting $\hat{\beta}^1 \cdot \overline{x}^0$, becomes $\overline{y}^1 - \overline{y}^0 = \hat{\beta}^1 \cdot (\overline{x}^1 - \overline{x}^0) + \left(\hat{\beta}^1 - \hat{\beta}^0\right) \cdot \overline{x}^0$. The component $\hat{\beta}^1 \cdot (\overline{x}^1 - \overline{x}^0)$, denoted by Δ_x, is then interpreted as the part of the gap that is explained by differences in average observable characteristics of the individuals and the other component, $\left(\hat{\beta}^1 - \hat{\beta}^0\right) \cdot \overline{x}^0$, denoted by Δ_0, is interpreted as the part explained by differences in the rewards for those characteristics.

The traditional notation of the Blinder-Oaxaca decomposition outlined above can also be expressed with a single regression. For this purpose, let the dummy variable t_i indicate the group to which individual i belongs ($t_i = 0$ for those individuals i that belong to the base group, $t_i = 1$ for those individuals i that belong to the comparing group). Denoting by y_i the outcome of individual i, the coefficient α_1 on the equation $y_i = \alpha_0 + \alpha_1 t_i + \varepsilon_i$ represents the wage gap between both groups (that is, $\alpha_1 = E[y|t = 1] - E[y|t = 0]$). The Blinder-Oaxaca decomposition is an attempt to explain the α_1 coefficient (the gap) on the basis of observable characteristics. For that purpose, it is necessary to estimate the equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 t_i + \beta_3 x_i t_i + \nu_i \qquad (1)$$

where t is the dummy variable introduced above, x is a n-dimensional vector of observable characteristics, β_1 and β_3 are the corresponding n-dimensional vectors of rewards for those characteristics (β_1 for individuals of group 0 and $\beta_1 + \beta_3$ for individuals of group 1) and β_0 and β_2 are one-dimensional coefficients (intercepts). Then, α_1 can be expressed as

$$\alpha_1 = E[(\beta_0 + \beta_2) + (\beta_1 + \beta_3)x|t = 1] - E[\beta_0 + \beta_1 x|t = 0]$$

which after some re-arrangements becomes

$$\alpha_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)E[x|t = 1] - \beta_0 - \beta_1 E[x|t = 0]$$

$$\alpha_1 = \beta_1(E[x|t = 1] - E[x|t = 0]) + \beta_2 + \beta_3 E[x|t = 1] \quad \alpha_1 = \Delta_x + \Delta_0$$

Now, $\Delta_x = \beta_1(E[x|t = 1] - E[x|t = 0])$ and $\Delta_0 = \beta_2 + \beta_3 E[x|t = 1]$ can be interpreted in the traditional way. Δ_x is the component of the wage gap that is explained by the difference in average characteristics of the individuals, while Δ_0 is the component that remains unexplained and can be attributed to the existence of a combination of discrimination and unobservable characteristics. The extension of this decomposition to a continuum of comparing groups will be introduced in the next section.

3. Decomposing the Gap in a Continuous Setup

The setup outlined above in which t is a binary variable can be extended into another in which t is continuous. Provided that $y_i = \alpha_0 + \alpha_1 t_i + \varepsilon_i$; the slope coeffcient α_1 can be rewritten as $\alpha_1 = \frac{cov(y,t)}{var(t)}$ Just for presentation purposes, without loss of generality, let us assume $E(t) = 0$ and $E(t^2) = 0$. Then $\alpha_1 = E[yt]$ or $\alpha_1 = \int_D tE[y|t]dF(t)$, where D represents the domain and F represents the cumulative density function for t. Next, using the Mincerian equation (Mincer) in the new expression for α_1 will lead to

$$\alpha_1 = \int_D t(\beta_0 + \beta_1 E[x|t] + \beta_2 t + \beta_3 tE[x|t])dF(t)$$

Using the properties assumed on the first two moments of t we have

$$\alpha_1 = \int_D (\beta_1 + \beta_3 t)E[x|t]tdF(t) + \beta_2 \qquad (2)$$

Noting that $\beta_3 E[x]$ is equivalent to $\int_D (\beta_1 + \beta_3 t) E[x] t dF(t),$ we can add the former and subsract the latter from the right-hand side of (alpha-1-old) and after some re-arrangements, obtain

$$\alpha_1 = \left[\int_D (\beta_1 + \beta_3 t)(E[x|t] - E[x]) t dF(t) \right] + [\beta_2 + \beta_3 E[x]] \qquad \alpha_1 = \widetilde{\Delta_x} + \widetilde{\Delta_0}. \tag{3}$$

which now has an analogous interpretation to the Blinder-Oaxaca decomposition.

$\widetilde{\Delta_x} = \int_D (\beta_1 + \beta_3 t)(E[x|t] - E[x]) t dF(t)$ is the component of α_1 that can be explained by the aggregated differences in average characteristics between individuals of type t and the average individual. $\widetilde{\Delta_0} = \beta_2 + \beta_3 E[x]$ is the component of α_1 that remains unexplained.

The two moment conditions about t were assumed only for expositional purposes. Without imposing any of those, the decomposition would be expressed as:

$$\widetilde{\Delta_x} = \frac{\int_D (\beta_1 + \beta_3 t)(E[x|t] - E[x])(t - E[t]) dF(t)}{var(t)} \qquad \text{and} \qquad \widetilde{\Delta_0} = \beta_2 + \beta_3 E[x]$$

Note that the expression for the unexplained component, $\widetilde{\Delta_0}$, as a function of β_2, β_3 and $E[x]$, does not change.

Furthermore, instead of adding $\beta_3 E[x]$ and subtracting its equivalent form in (alpha-1-old) one can add the conditional versions of them $\beta_3 E[x|t = 0]$ and $\int_D (\beta_1 + \beta_3 t) E[x|t = 0] t dF(t).$ In such way the decomposition becomes:

$$\widetilde{\Delta_x} = \frac{\int_D (\beta_1 + \beta_3 t)(E[x|t] - E[x|t = 0])(t - E[t]) dF(t)}{var(t)} \qquad \text{and} \qquad \widetilde{\Delta_0} = \beta_2 + \beta_3 E[x|t = 0]$$

which could be interpreted in light of a slightly different counterfactual situation. The base comparing group would be formed by those with $t = 0$ instead of everybody. As a

final comment, it is straightforward to verify that the restriction of *t* to a binary case (0 and 1) under this latter counterfactual delivers the usual Blinder-Oaxaca decomposition.

4. An Application: Racial Wage Gaps in Urban Peru

The data for this application are taken from Ñopo et al. (2007). This is a nationally representative data set on individuals in urban areas that, in addition to capturing information on human capital characteristics and labor markets outcomes, captures racial information in a novel way. Individuals' race is characterized in this data set by a vector of intensities, as observed by trained pollsters, along the White and Indigenous dimensions. Such intensities are determined on the basis of individuals' observable characteristics (skin color, hair color, hair color, shape of eyes and shape of lips, among other traits). In this way, an individual's racial characteristics are represented by a vector $(I_W, I_I) \in \mathbb{N}_{10} x \mathbb{N}_{10}$ where a higher intensity in a particular dimension denotes more observable characteristics of the individual that make her/him resemble a typical White or Indigenous individual, respectively.

Then, the ratio $t = \frac{I_W}{I_W + I_I}$ can be interpreted as an indicator of "Whiteness" of an individual. By construction, that ratio lies within the interval [0,1] and for the practical purposes of this estimation can be assumed to be a continuous indicator. Then, an earnings equation can be estimated using this indicator as a regressor, as well as sex, age (and its square), years of schooling, marital status, years of occupational experience (and its square), occupation (eight dummies for nine occupational groups), firm size (four dummies), city (a dummy that distinguishes Lima from the rest of the nation), mother's educational level (two dummies), number of sick days during the last year, migratory condition (one dummy) and hours worked per week. The results obtained for an estimation with the whole national sample are $\alpha_1 = 0.516$ and $\widetilde{\Delta_0} = 0.076$. According to these figures, the average individual with the highest white intensity (t=1) earns approximately 68% more than the average individual in per hour terms. This difference is partially explained by the fact that the average individual with the highest white intensity has individual characteristics that exceeds those of the average individual. After accounting for those differences in observable characteristics, there is still a difference of

8 percent in favor of the white individuals. Analogous regressions and decompositions were performed for different subsets of the data. The results are reported in Table 1. The first column corresponds to the estimators of the slope (α_1), and the second column corresponds to the slope's unexplained component $(\tilde{\Delta}_0)$; each row corresponds to a particular partition of the data, as labeled. The numbers in parenthesis are standard errors, computed by bootstrap (10,000 iterations).

Table 1. Racial Coefficient Decomposition by Different Partition Criteria

	Coefficient Alpha 1	Unexplanied Component
All	0.516	0.076
	(0.179)	(0.113)
Geographic		
Lima	0.524	0.228
	(0.226)	(0.175)
Rest of the Nation	0.089	-0.039
	(0.275)	(0.172)
Gender		
Females	0.547	0.184
	(0.264)	(0.202)
Males	0.512	-0.032
	(0.248)	(0.151)
Type of employment		
Private Wage Earners	0.661	0.134
	(0.249)	(0.167)
Public Wage Earners	0.247	0.244
	(0.358)	(0.290)
Self-Employed	0.516	0.056
	(0.325)	(0.222)

The results suggest that there are more evidences of differences in pay that cannot be explained by differences in observable characteristics among those who live in Lima, females and those who work as salaried employees (either in the private sector or the public sector). However, the standard errors are such that none of these are statistically different than zero.

References

Albrecht, J., A. Björklund and S. Vroman. 2003. "Is There a Glass Ceiling In Sweden?" *Journal of Labor Economics* 21: 145-177.

Bauer, T., and M. Sinning. 2005. "Blinder-Oaxaca Decomposition for Tobit Models." IZA Discussion Paper 1795. Bonn, Germany: Institute for the Study of Labor (IZA).

Bauer, T., S. Göhlmann and M. 2006. "Gender Differences in Smoking Behavior." IZA Discussion Paper 2259. Bonn, Germany: Institute for the Study of Labor (IZA).

Blinder, A. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources* 7(4): 436-55.

Fairlie, R. 2005. "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models." *Journal of Economic and Social Measurement* 30: 305-316.

Maloney, W. 1999. "Does Informality Imply Segmentation in Urban Labor Markets? Evidence from Sectoral Transitions in Mexico." World Bank Economic Review 13: 275-302.

----. 2004. "Informality Revisited." *World Development* 32: 1159-1178.

Ñopo, H. Forthcoming. "Matching as a Tool to Decompose Wage Gaps." Forthcoming in *Review of Economics and Statistics*.

Ñopo, H, J. Saavedra and M. Torero. 2007. "Ethnicity and Earnings in a Mixed Race Labor Market." *Economic Development and Cultural Change* 55(4).

Oaxaca, R. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14(3): 693-709.

Perry, G. et al. editors. 2005. *Beyond the City: The Rural Contribution to Development*. Washington, DC, United States: World Bank.