ECONOMICS BULLETIN

A computational trick for calculating the Blinder-Oaxaca decomposition and its standard error

Jutta Heinrichs
Associated Economic Consultants

Peter Kennedy
Simon Fraser University

Abstract

To compute the Blinder-Oaxaca decomposition and associated standard errors a practitioner needs to be comfortable using vector and matrix software manipulations. This paper proposes a computational trick for producing these computations by running an artificial regression.

Citation: Heinrichs, Jutta and Peter Kennedy, (2007) "A computational trick for calculating the Blinder-Oaxaca decomposition and its standard error." *Economics Bulletin*, Vol. 3, No. 66 pp. 1-7
Submitted: December 3, 2007. **Accepted:** December 5, 2007.
URL: http://economicsbulletin.vanderbilt.edu/2007/volume3/EB-07C00003A.pdf

1. Introduction

A popular method of analyzing wage discrimination is the Blinder-Oaxaca decomposition, introduced in Blinder (1973) and Oaxaca (1973). In this methodology the sample average wage difference between, say, males and females, is broken into two parts. One part measures the impact of differences in the male/female parameters of the wage determination equation, and the other part measures the impact of male/female endowment differences. The former part is referred to as that part due to discrimination. Calculation of this measure is undertaken in three steps: 1) run a regression on the male data, 2) run a regression on the female data, and 3) using these regression results, compute the two parts described above. A fourth step is required to calculate standard errors for these estimates. Murray (2006, pp.293-5) has a good textbook exposition. Steps three and four can be awkward unless a practitioner is comfortable using software to manipulate vectors and matrices; the purpose of this paper is to suggest a means of simplifying these calculations.

2. The Blinder–Oaxaca Decomposition

Suppose the wage determination function for the males is given as

$$\mathrm{w}_{\mathrm{m}} = \mathrm{X}_{\mathrm{m}}\beta_m + \varepsilon_m$$

where w is an N_m×1 vector of observations on wages of N_m individuals, X is an N×K matrix of observations on K explanatory variables, β is a K×1 vector of parameters, and ε is an N×1 vector of errors. The m subscript denotes males; the wage determination function for females is written by replacing the m subscript with an f subscript. Because the regression line passes through the average of the observations,

$$\overline{\mathrm{w}}_{\mathrm{m}} = \overline{\mathrm{X}}_{\mathrm{m}}\hat{\beta}_m$$

where the bar denotes average and the hat denotes the ordinary least squares estimate. The error term disappears because the sum of the ordinary least squares errors is zero. From this the difference between the male and female average wages in the sample can be written as

$$\overline{\mathrm{w}}_{\mathrm{m}} - \overline{\mathrm{w}}_{\mathrm{f}} = \overline{\mathrm{X}}_{\mathrm{m}}\hat{\beta}_m - \overline{\mathrm{X}}_{\mathrm{f}}\hat{\beta}_f$$

Subtracting and adding $\overline{\mathrm{X}}_{\mathrm{m}}\hat{\beta}_f$ we get

$$\overline{\mathrm{w}}_{\mathrm{m}} - \overline{\mathrm{w}}_{\mathrm{f}} = \overline{\mathrm{X}}_{\mathrm{m}}\left(\hat{\beta}_m - \hat{\beta}_f\right) + \left(\overline{\mathrm{X}}_{\mathrm{m}} - \overline{\mathrm{X}}_{\mathrm{f}}\right)\hat{\beta}_f \qquad (1)$$

This decomposes the sample male/female average wage difference into two parts, one due to differences in the specification parameters and the other due to differences in endowments. The former is the discrimination measure. Often they are reported as a percentage of their sum.

When producing equation (1) above we could have subtracted and added $\overline{\mathrm{X}}_{\mathrm{f}}\hat{\beta}_m$ instead of $\overline{\mathrm{X}}_{\mathrm{m}}\hat{\beta}_f$, obtaining an alternative measure

$$\overline{\mathrm{w}}_{\mathrm{m}} - \overline{\mathrm{w}}_{\mathrm{f}} = \overline{\mathrm{X}}_{\mathrm{f}}\left(\hat{\beta}_m - \hat{\beta}_f\right) + \left(\overline{\mathrm{X}}_{\mathrm{m}} - \overline{\mathrm{X}}_{\mathrm{f}}\right)\hat{\beta}_m \qquad (2)$$

Unfortunately these two discrimination calculations, $\overline{\mathrm{X}}_{\mathrm{m}}\left(\hat{\beta}_m - \hat{\beta}_f\right)$ and $\overline{\mathrm{X}}_{\mathrm{f}}\left(\hat{\beta}_m - \hat{\beta}_f\right)$, do not produce the same numbers, requiring researchers to report both measures.[1]

The discrimination measure, $\overline{\mathrm{X}}_{\mathrm{m}}\left(\hat{\beta}_m - \hat{\beta}_f\right)$, can be calculated as follows. Run the male and female regressions to get $\hat{\beta}_m$ and $\hat{\beta}_f$, respectively, and subtract them to obtain the K×1 vector $\left(\hat{\beta}_m - \hat{\beta}_f\right)$. Calculate the average of all the male explanatory variable observations and place them in a 1×K row vector, namely $\overline{\mathrm{X}}_{\mathrm{m}}$. Finally, multiply this row vector by the vector $\left(\hat{\beta}_m - \hat{\beta}_f\right)$. The variance of the discrimination measure can be calculated as follows. The K×K variance-covariance matrix V of $\left(\hat{\beta}_m - \hat{\beta}_f\right)$ is the sum of the variance-covariance matrices of $\hat{\beta}_m$ and of $\hat{\beta}_f$ because they are estimated independently using different data. So $\hat{\mathrm{V}}$ is found by adding the variance-covariance matrix estimates from the two regressions. The variance of $\overline{\mathrm{X}}_{\mathrm{m}}\left(\hat{\beta}_m - \hat{\beta}_f\right)$ is estimated as $\overline{\mathrm{X}}_{\mathrm{m}}\hat{\mathrm{V}}\overline{\mathrm{X}}_{\mathrm{m}}'$. Its standard error is the square root of this. For those adept at using software to perform matrix calculations these computations are straightforward, but for those without this skill these calculations are burdensome. The next section describes a computational trick that simplifies estimation of both the discrimination measure and its standard error.

3. A Computational Trick

The computational trick of this paper is to estimate the male and female regressions simultaneously, while introducing an artificial observation along with an observation-specific dummy to produce automatically the discrimination measure. This is shown in the following matrix specification:

$$\begin{bmatrix} \mathrm{w}_{\mathrm{m}} \\ \mathrm{w}_{\mathrm{f}} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathrm{X}_{\mathrm{m}} & 0 & 0 \\ 0 & \mathrm{X}_{\mathrm{f}} & 0 \\ -\overline{\mathrm{X}}_{\mathrm{m}} & \overline{\mathrm{X}}_{\mathrm{m}} & 1 \end{bmatrix} \begin{bmatrix} \beta_m \\ \beta_f \\ \theta \end{bmatrix} + \begin{bmatrix} \varepsilon_m \\ \varepsilon_f \\ \varepsilon_\theta \end{bmatrix} \tag{3}$$

The first two rows represent the male and female regressions using the N_{m} male and N_{f} female observations, respectively; it mimics the setup for SURE (seemingly unrelated regression estimation). The bottom row is an extra, artificial observation, structured to capture the discrimination measure, and the third column in the regressor matrix is an observation-specific dummy with coefficient θ. When running regression (3), the estimate of θ will be whatever is necessary to create a perfect fit on this last, artificial observation, a well-known result first noted by Salkever (1976). In this case we have $\hat{\theta} = \overline{X}_m\left(\hat{\beta}_m - \hat{\beta}_f\right)$, the desired discrimination measure. Expressing this as a fraction

[1] Subsequent literature has addressed this and other issues. For example, Oaxaca and Ransom (1994) discuss means of breaking the discrimination portion of the difference between blacks and whites into an advantage to whites and a disadvantage to blacks.

of $\overline{\mathrm{w}}_{\mathrm{m}} - \overline{\mathrm{w}}_{\mathrm{f}}$ gives the percentage of the sample male/female average wage difference due to discrimination.

The alternative measure of discrimination, namely $\overline{X}_f\left(\hat{\beta}_m - \hat{\beta}_f\right)$, could be computed automatically at the same time by adding an additional artificial observation and running the following regression:

$$\begin{bmatrix} \mathrm{w}_{\mathrm{m}} \\ \mathrm{w}_{\mathrm{f}} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \mathrm{X}_{\mathrm{m}} & 0 & 0 & 0 \\ 0 & \mathrm{X}_{\mathrm{f}} & 0 & 0 \\ -\overline{\mathrm{X}}_{\mathrm{m}} & \overline{\mathrm{X}}_{\mathrm{m}} & 1 & 0 \\ -\overline{\mathrm{X}}_{\mathrm{f}} & \overline{\mathrm{X}}_{\mathrm{f}} & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_m \\ \beta_f \\ \theta \\ \phi \end{bmatrix} + \begin{bmatrix} \varepsilon_m \\ \varepsilon_f \\ \varepsilon_\theta \\ \varepsilon_\phi \end{bmatrix} \tag{4}$$

From running this regression the estimate of ϕ, the slope of the second observation-specific dummy, is the desired alternative discrimination measure.[2]

4. Calculating Confidence Intervals

It is tempting to conclude that calculation of confidence intervals for the two discrimination measures can be undertaken by using the estimated standard errors associated with the estimates of θ and ϕ. Unfortunately this is legitimate only when the error variances of the male and female error terms are equal. When they are not equal the standard errors of all the estimates are biased because estimating these two regressions together induces the computer to estimate standard errors using an overall error variance rather than two individual variances. The coefficient estimates are not affected because of the special structure of the specification; the two artificial observations drop out of the minimization process because they are fit perfectly, and the male/female regressor observations enter without being associated with one another (because of the zero submatrices in the regressor matrix above).

Whenever we are not comfortable assuming that the male/female error variances are equal, an adjustment needs to be made to regressions (3) and (4) if we wish to produce confidence intervals. The simplest way to do this is to perform a traditional transformation for heteroskedasticity. Run the male and female regressions separately and obtain $\hat{\sigma}_m$ and $\hat{\sigma}_f$, the estimated standard errors of their respective error terms (called the standard error of the regression in most software). Then divide all the male observations on w and on X by $\hat{\sigma}_m$ and all the female observations on w and on X by $\hat{\sigma}_f$. Replace the original observations in the first two rows of equations (3) and (4) by these transformed observations and run ordinary least squares. Do not adjust the artificial observations. The resulting estimates of θ and ϕ will have attached to them appropriate standard errors that can be used to create confidence intervals or undertake t tests.

[2] There is a simple way of checking this calculation for errors, by exploiting the result of equation (1). In the final row of equation (4) replace $-\overline{\mathrm{X}}_{\mathrm{f}}$ with 0 and replace $\overline{\mathrm{X}}_{\mathrm{f}}$ with $\left(\overline{\mathrm{X}}_{\mathrm{f}} - \overline{\mathrm{X}}_{\mathrm{m}}\right)$. Run the regression and check that the estimates of θ and ϕ sum to $\overline{\mathrm{w}}_{\mathrm{m}} - \overline{\mathrm{w}}_{\mathrm{f}}$.

5. An Example

To illustrate this computational procedure, a random sample of 550 individuals was drawn from the 1978 Current Population Survey, with observations on log of wage in dollars per hour (WAGE), years of education (ED), age (AGE), a dummy equal to one for females (FEMALE), otherwise zero, a dummy equal to one for nonwhites (NONWHITE), otherwise zero, and a dummy equal to one for union membership (UNION), otherwise zero. Two regressions were run, one on the male observations and one on the female observations, producing results shown in Table I.

Table I: Regression Results, Dependent Variable WAGE

	Male Regression N=343		Female Regression N=207	
Variable	Coefficient	Std. Error	Coefficient	Std. Error
ED	0.063622	0.007243	0.060458	0.012562
AGE	0.014086	0.001628	0.008621	0.002357
NONWHITE	-0.175836	0.059823	0.019184	0.070277
UNION	0.183493	0.043854	0.275639	0.067062
INTERCEPT	0.461148	0.120865	0.325996	0.189029
Mean of Dep. Var.	1.812921		1.462412	
S.E. of regression	0.379628		0.397303	

A “sort” command was used to order the observations such that the 343 male observations appeared before the 207 female observations. A male dummy (MALE) equal to one, otherwise zero, was created by subtracting FEMALE from one. New variables MED, MAGE, MNONWHITE, and MUNION were created by multiplying ED, AGE, NONWHITE, and UNION by MALE, respectively. New variables FED, FAGE, FNONWHITE, and FUNION were created by multiplying ED, AGE, NONWHITE, and UNION by FEMALE, respectively. This created the four upper left-hand elements of equation (4).

Two new observations were added to the existing 550 observations. Observations 551 and 552 for WAGE were both zero. Observations 551 and 552 for MALE were both minus one and for FEMALE were both plus one. Observations 551 for FED, FAGE, FNONWHITE, and FUNION were the averages of the male observations for ED, AGE, NONWHITE and UNION, respectively. For MMED, MAGE, MNONWHITE, and MUNION they were the negative of these, respectively. Observations 552 for FED, FAGE, FNONWHITE, and FUNION were the averages of the female observations for ED, AGE, NONWHITE and UNION, respectively. For MMED, MAGE, MNONWHITE, and MUNION they were the negative of these, respectively. This created the bottom four left-hand elements of equation (4).

Two observation-specific dummies were created, DISCRIM1 with a one in row 551 and zeros elsewhere, and DISCRIM2 with a one in row 552 and zeros elsewhere. This created the two right-hand columns of equation (4).

Finally, the first 343 observations (the male observations) were divided through by the estimated standard error of the male error term, reported in Table I as 0.379628 and the next 207 observations were divided through by 0.397303, the estimated standard error of the female error term. Note that in addition to the explanatory variables the associated WAGE observations were also transformed in this way, as well as the MALE and

FEMALE dummies. Observations 551 and 552 were not transformed. A T is added to the beginning of each variable name to emphasize that it has been transformed.

Equation (4) was estimated by regressing these augmented (with two extra observations) and transformed (by dividing by appropriate standard errors) data, 552 observations. TWAGE was regressed on TMALE, TMED, TMAGE, TMNONWHITE, TMUNION, TFEMALE, TFED, TFAGE, TFNONWHITE, TFUNION, DISCRIM1, and DISCRIM2. No intercept is included because the TMALE and TFEMALE variables play this role. The results are reported in Table II.

Table II: Computational Trick Results, Dependent Variable TWAGE, N=552

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TMED	0.063622	0.007243	8.783875	0.0000
TMAGE	0.014086	0.001628	8.651919	0.0000
TMNONWHITE	-0.175836	0.059823	-2.939266	0.0034
TMUNION	0.183493	0.043854	4.184219	0.0000
TMALE	0.461148	0.120865	3.815405	0.0002
TFED	0.060458	0.012562	4.812749	0.0000
TFAGE	0.008621	0.002357	3.658297	0.0003
TFNONWHITE	0.019184	0.070277	0.272982	0.7850
TFUNION	0.275639	0.067062	4.110220	0.0000
TFEMALE	0.325996	0.189029	1.724585	0.0852
DISCRIM1	0.316044	1.000647	0.315840	0.7522
DISCRIM2	0.304865	1.000616	0.304677	0.7607

The first section of Table II provides the results of running the male regression, and the second section provides the results from running the female regression. These results are identical to those shown in Table I. The DISCRIM1 coefficient 0.316044 estimates that part of the sample average (log) wage difference due to male/female parameter differences, as measured using the male endowments. The DISCRIM2 coefficient 0.304865 estimates that part of the sample average (log) wage difference due to male/female parameter differences, as measured using the female endowments. From Table I the sample average (log) wage difference is 1.812921 - 1.462412 = 0.350509. Using the former measure 90.2% of this difference is due to discrimination; using the latter measure 87.0% of this difference is due to discrimination. A full reporting requires that an indication of the accuracy of these estimates be provided. As seen in Table II the standard errors of the DISCRIM1 and DISCRIM2 measures are both such that at traditional significance levels the nulls that each is zero cannot be rejected.

References

Blinder, A. (1973) “Wage discrimination, reduced form, and structural estimates” *Journal of Human Resources* **8**, 436-55.

Murray, M. (2006) *Econometrics: A Modern Introduction*, Pearson: Boston.

Oaxaca, R. L. (1973) “Male-female wage differences in urban labor markets” *International Economic Review* **14**, 693-709.

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

Salkever, D. (1976) “The use of dummy variables to compute predictions, prediction errors, and confidence intervals” *Journal of Econometrics* **4**, 393-7.