DISCUSSION PAPER SERIES

A Simple Solution to the Identification Problem in Detailed Wage Decompositions

Myeong-Su Yun

July 2003

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A Simple Solution to the Identification Problem in Detailed Wage Decompositions

Myeong-Su Yun

Tulane University and IZA Bonn

ABSTRACT

A Simple Solution to the Identification Problem in Detailed Wage Decompositions

Oaxaca and Ransom (1999) show that a detailed decomposition of the coefficients effect is destined to suffer from an identification problem since the detailed coefficients effect attributed to a dummy variable is not invariant to the choice of reference groups. It turns out that the identification problem in the decomposition equation is a disguised identification problem of constant and dummy variables in a regression equation. This paper proposes a simple and natural remedy for this problem by utilizing "normalized" regressions which enable us to identify the constant and estimates of each dummy variable. The identification problem is automatically resolved once we obtain "normalized" regression equations for two comparison groups.

Myeong-Su Yun
Department of Economics
Tulane University
206 Tilton Hall
New Orleans, LA 70118
USA
Tel.: +1 504 862 8352
Fax: +1 504 865 5869
Email: msyun@tulane.edu

I. Introduction

As Oaxaca and Ransom (1999) show in this journal, the Blinder-Oaxaca decomposition

suffers from one nagging conceptual problem: the detailed Blinder-Oaxaca decomposition of wage

differentials is not invariant to the choice of reference group when a set of dummy variables is used.[1]

The problem exists even if only one binary variable (one of the two categories such as men and

women, whites and non-whites) is used in a wage equation. If we use dummy variable(s), then the

detailed coefficients effect attributed to individual variables is not invariant to the choice of left-out

group(s).[2] This invariance or identification problem is well-known to labor economists and has

plagued decomposition and discrimination analysis for a long time.[3]

Jones (1983, p. 130) strongly warns us to be cautious about the detailed decomposition since

[1] Decomposition analysis has been widely used to understand racial and gender wage differentials
since the papers of Blinder (1973) and Oaxaca (1973). Decomposition analysis explains wage
differentials in terms of differences in individual characteristics (characteristics effect) and
differences in the OLS coefficients of wage equations (coefficients effect or discrimination). Note
that the application of the decomposition analysis is not restricted to only wage differentials. For
example, Ham, Svejnar and Terrell (1998) apply the decomposition method to differences in
expected unemployment duration. Though we discuss the identification problem in the
decomposition of wage differentials, the finding in this paper can be applied to any decomposition
analysis.

[2] However, the aggregate and detailed characteristics effect and aggregate coefficients effect are
invariant to the choice of reference groups. One caveat is that even a detailed decomposition with
a continuous variable is not immune to the identification problem. For example, a "locational
transform" of the age variable, e.g., from age to age-18, will yield a different coefficients effect
attributed to the variable related to age (see Oaxaca and Ransom (1999, p.155)). However, the
identification problem related to a continuous variable cannot be resolved since there are infinitely
many transformations, unlike categorical variables. Therefore, we have to rely on customs in the
field as to how to specify continuous variables in the regression.

[3] Note that Blinder (1973) struggles futilely to prove that the choice of reference group does not
impose an identification problem for decomposition analysis. Substantial attention has been paid to
this issue in Jones (1983), Oaxaca and Ransom (1999) and Horrace and Oaxaca (2001).

"the Blinder decomposition (i.e., the detailed decomposition in this paper) is inappropriate, and the values ... are inherently arbitrary." Though economists acknowledge that detailed decomposition suffers from the identification problem, they are reluctant to completely abandon the detailed decomposition of the coefficients effect viewing that approach as too drastic. Oaxaca and Ransom (1999, p. 156) summarize this dilemma quite eloquently without prescribing a remedy or suggestion for the usage of the detailed decomposition.[4] Therefore, the decision to break down the aggregate coefficients effect into contributions of individual variables, i.e., detailed decomposition of coefficients effect, is left to researchers. Some have stopped doing detailed decompositions of the coefficients effect (e.g., Ham, Svenjnar and Terrell (1998)) while others still do detailed decompositions while keeping the identification problem in mind (e.g., Radchenko and Yun (2003)).

This paper proposes a simple and practical solution to this identification or invariance problem which has puzzled so many economists for so long. The solution is based on the simple idea that if alternative reference groups yield different estimates of the characteristics and coefficients effects of individual variables, then it is natural to obtain estimates of the two effects with every possible specification of reference groups and take the average of the estimates with various reference groups as the "true" contributions of individual variables to wage differentials.[5] This "averaging approach" resolves the identification problem in detailed Blinder-Oaxaca decompositions of wage differentials. A close inspection shows that averaging the two effects in the decomposition equation with varying reference groups is equivalent to finding average estimates of constant and

---

[4] "In any case, almost all wage regression models contain categorical variables, so it is unlikely that an application of detailed decompositions can escape the identification problem."

[5] The author thanks Ira N. Gang for suggesting this idea and regrets that it took too long before he acted on this suggestion.

dummy variables in wage equations with varying reference groups and then using the average estimates to calculate Blinder-Oaxaca decomposition equation.

Researchers may react a bit negatively to this suggestion since it could be quite cumbersome and tedious to estimate wage equations permuting the choice of reference groups. This paper shows that there is no need to estimate thousands of wage equations. One set of regression estimates is sufficient to resolve the identification problem in the detailed decompositions. That is, the average of the estimates with varying reference groups can be easily calculated by using only "one" set of regression estimates with any specific reference group(s). This paper provides a simple and practical method to resolve the identification problem by identifying the coefficients of the dummy variables and constant in the wage equation. The identification problem for the detailed decomposition disappears once the contribution of the dummy variables and the constant in the regression equations are identified.

II. Identification Problem and Invariant Decomposition: An Illustration

In order to understand what the identification problem in the detailed wage decomposition is, we will first illustrate the problem with a simple example. The generalization is a simple extension of this illustration. Suppose that there are three categories, and the regression equation has only a constant and two dummy variables on the right hand side where the reference group is the first category.[6] That is, $y_i = \alpha_i + \sum_{k=2}^{3} D_{ki} * \beta_{ki} + e_i,$ (1)

---

[6] Note that the identification problem arises even if we have only one dummy variable. However, using an equation which has only one dummy variable (i.e., a two category case) may not reveal the complete picture of the problem. See footnote 1 in Jones (1983, p. 126) for his complaint on

where y is log-wages and $i = A$ or B for two comparison groups.

Coefficients for two dummy variables taking category one as the reference group are estimated using OLS regression and are reported in the second column of Table 1 for sample A and for sample B. Note that the coefficient of the reference group (category one) is restricted to zero. Similarly, we may change our reference group specification and estimate the wage equation two more times. The estimates taking category 2 or 3 as the reference group are reported on the next two columns.

Since we have three sets of estimates with varying reference groups, we may decompose the wage differentials (0.15 = 0.76 - 0.61) into characteristics and coefficients effects three times. The characteristics effect (Δ_{X1}) and coefficients effects ($\Delta_{\beta 1}$) are specified as follows when category 1 is the reference group. Note that the residuals effect ($\overline{e}_A - \overline{e}_B$) disappears since the average residual is zero due to the OLS assumption.

$$\Delta_{X1} = \sum_{k=2}^{3} (\overline{D}_{kA} - \overline{D}_{kB}) * \beta_{kB}, \text{ and}$$

$$\Delta_{\beta 1} = \alpha_A - \alpha_B + \sum_{k=2}^{3} \overline{D}_{kA} * (\beta_{kA} - \beta_{kB}).$$

The decomposition equation can be obtained similarly when the reference group is either category 2 or 3. The results of decomposition analysis with permuting reference groups are reported in Table 2.

Blinder's use of only the two category case to prove non-existence of the identification problem in a detailed wage decomposition (footnote 13, Blinder (1973, p. 443)).

Table 1. Regression Estimates with Varying Reference Group: Illustration

	Means	Coefficients with Different Reference Groups (Sample A)			
	Means	Category 1	Category 2	Category 3	Average $(\beta_{kA} - \overline{\beta}_A)$
D_{1A}	0.2	0	-0.2	-0.4	-0.2 = 0 - (0+0.2+0.4)/3
D_{2A}	0.3	0.2	0	-0.2	0 = 0.2 - (0+0.2+0.4)/3
D_{3A}	0.5	0.4	0.2	0	0.2 = 0.4 - (0+0.2+0.4)/3
Constant	1	0.5	0.7	0.9	0.7 = 0.5 + (0+0.2+0.4)/3
Log Wage	0.76				
		Coefficients with Different Reference Groups (Sample B)			
	Means	Category 1	Category 2	Category 3	Average $(\beta_{kB} - \overline{\beta}_B)$
D_{1B}	0.1	0	-0.3	0.6	0.1 = 0 - (0+0.3-0.6)/3
D_{2B}	0.5	0.3	0	0.9	0.4 = 0.3 - (0+0.3-0.6)/3
D_{3B}	0.4	-0.6	-0.9	0	-0.5 = -0.6 - (0+0.3-0.6)/3
Constant	1	0.7	1	0.1	0.6 = 0.7 + (0+0.3-0.6)/3
Log Wage	0.61				

Note: $k = 1$, 2 or 3; $D_{1i} = 1 - D_{2i} - D_{3i}$; $\overline{\beta}_i = \sum_{k=1}^{3} \beta_{ki}/3$. The average is calculated using estimates when category 1 is the reference group. Note that the value of the average is independent of the choice of reference group.

Table 2. Decomposition with Varying Reference Group: Illustration

| | Reference Group | | | | | | Average | |
| | Category 1 | | Category 2 | | Category 3 | | | |
	Char.	Coeff.	Char.	Coeff.	Char.	Coeff.	Char.	Coeff.
D_1	0	0	-0.03	0.02	0.06	-0.2	0.01	-0.06
D_2	-0.06	-0.03	0	0	-0.18	-0.33	-0.08	-0.12
D_3	-0.06	0.5	-0.09	0.55	0	0	-0.05	0.35
Constant	0	-0.2	0	-0.3	0	0.8	0	0.1
SUM	-0.12	0.27	-0.12	0.27	-0.12	0.27	-0.12	0.27

Note: Char. and Coeff. are characteristics and coefficients effect.

From the Table 2, we can confirm what Oaxaca and Ransom (1999, p. 156) find regarding identification issues with decomposition analysis. First, aggregate characteristics and coefficients effects are invariant to the choice of the left-out group as the last row (SUM) shows. Second, the sum of the coefficients effect of the two categories (e.g., D_2 and D_3 when category one is the reference group) is not invariant to the choice of the reference group.[7] This can be easily verified by the fact that the coefficients effect of the constant changes in this illustration. Third, the sum of the characteristics effect for the two categories (e.g., D_2 and D_3 when category one is the reference

---

[7] They also state that the sum of the coefficients effect of the constant and categorical variables (e.g., D_2 and D_3 when category one is the reference group) is the contribution of categorical variables. This definition may be misleading since if there are multiple sets of categorical variables, e.g., dummy variables of industry and occupation are included jointly as independent variables, then the sum of the contribution of each set of categorical variables may exceed the aggregate coefficients effect since the constant will be counted several times. The solution in this paper provides a natural way to obtain the contribution of each set of categorical variables.

group) is invariant to the choice of reference group.

Note that the invariance or identification problem arises because there is no agreement on which category should be the reference group. Therefore, we have a number of choices for the reference group. Our "averaging approach" suggests using the average of the characteristics and coefficients effects with varying reference groups as the contribution of individual variables to the wage differentials. The benefit of doing this is that we can identify characteristics and coefficients effects for every category in addition to the constant term.

Indeed, the root source of our identification problem is the well-known identification problem of constant and dummy variables in regression analysis (Suits (1984)). It is intuitive to assign the contribution of each categorical variable in regression analysis by averaging the measured contributions (coefficients) with the different specifications of the reference group. Once we identify the contribution of each variable including constant and all categorical variables (including the reference group) and compute the decomposition equation using the identified estimates, the contribution of each variable (including constant and categorical variables) can be easily identified.

One "apparent" drawback of this intuitive averaging approach is that it requires several estimation runs which might be tedious and cumbersome. Fortunately, we do not have to estimate wage equations several times with varying reference groups. The reason is because all information necessary to calculate the average characteristics and coefficients effects can be obtained by just one regression run.

In order to get the average estimate of the two effects, first, we simply transform the wage equation (1) as follows:

$$y_i = \alpha_i + \overline{\beta}_i + \sum_{k=1}^{3} D_{ki} * (\beta_{ki} - \overline{\beta}_i) + e_i, \tag{1'}$$

where $\overline{\beta}_i = (\beta_{1i} + \beta_{2i} + \beta_{3i})/3$, and $\beta_{1i} = 0$. Equation (1') may be called the "normalized"

regression equation where the estimate is simply the average of three sets of estimates with varying

reference groups as shown in the Table 1.[8] The estimate in the "normalized" regression equation,

called the average in Table 1, is simply the deviation of the OLS estimates of the dummy variable

from the mean coefficients ($\beta_k - \overline{\beta}$).[9] Using (1') and the mean characteristics of three categories

and the constant, it is simple manipulation to identify the characteristics and coefficients effects for

both aggregate and detailed decompositions. The characteristics and coefficients effects are specified

as follows:

$$\Delta'_X = \sum_{k=1}^{3} (\overline{D}_{kA} - \overline{D}_{kB}) * (\beta_{kB} - \overline{\beta}_B), \text{ and}$$

$$\Delta'_\beta = \alpha_A - \alpha_B + \overline{\beta}_A - \overline{\beta}_B + \sum_{k=1}^{3} \overline{D}_{kA} * (\beta_{kA} - \beta_{kB} - \overline{\beta}_A + \overline{\beta}_B).$$

---

[8] Interestingly, Suits (1984) also finds the same "normalized" equation as (1') by solving a restriction on the regression:
$$\sum_{k=1}^{K} (\beta_k + c) = 0, \text{ where the solution of } c \text{ is } -\overline{\beta}.$$
Of course, the average estimate can be obtained by using the share of reference group k as a weight. This is identical to having the constraint weighted by the share of category k. That is,
$$\sum_{k=1}^{K} w_k * (\beta_k + c) = 0, \text{ where } \sum_{k=1}^{K} w_k = 1, \text{ and the solution of } c \text{ is the weighted mean}$$
of the coefficients, $-\sum_{k=1}^{K} w_k * \beta_k$.

This weighted mean of the coefficients is used in Kennedy (1986), Krueger and Summers (1988), Greene and Seaks (1991), Edin and Zetterberg (1992), and Haisken-DeNew and Schmidt (1997). As the constraint indicates, using a weighted average has an undesirable consequence for decomposition analysis: the contribution of the set of dummy variables to mean wages should be zero, hence the characteristics and coefficients of dummy variables should be the same in magnitude but have the opposite sign. Therefore, using a simple average is preferable to using a weighted average.

[9] Only one set of regression estimates is required to get the average estimates. It does not matter which reference group is chosen to get the OLS estimates.

Note that the decomposition using the estimates of the "normalized" regression equation as shown above is identical to the detailed decomposition computed by averaging the three estimates of characteristics and coefficients effects with varying reference groups.[10] Therefore, the "apparent" drawback of the averaging approach of running large number of regression equations in order to exhaust all possible specifications of reference groups disappears. The next section will generalize what we have found in this illustration for applications of detailed decompositions when the regression equation includes several sets of categorical variables (e.g., industry, occupation, region, etc.) as independent variables.

III. General Solution and Conclusion

The previous section illustrates how to identify characteristics and coefficients effects in both aggregate and detailed decompositions using normalized regression. The normalized regression can be obtained from our averaging approach, or equivalently from setting a constraint as Suits (1984). Can the finding from the simple illustration be generalized to more complicated specifications of the regression equation? The answer is"Yes".

---

[10] This identity is not surprising considering that the sets of coefficients in Table 1 can be expressed as $\beta_{ki} - \beta_{1i}$, $\beta_{ki} - \beta_{2i}$, and $\beta_{ki} - \beta_{3i}$ when the reference group is the category 1, 2 and 3 respectively, where $k = 1, 2,$ and 3, and $i = A$ and B. Similarly the constant can be expressed as $\alpha_i + \beta_{1i}$, $\alpha_i + \beta_{2i}$, and $\alpha_i + \beta_{3i}$ when the reference group is the category 1, 2 and 3 respectively. The average characteristics and coefficients effect are:

$$\overline{\Delta}_X = \sum_{k=1}^{3} \sum_{j=1}^{3} (\overline{D}_{kA} - \overline{D}_{kB}) * (\beta_{kB} - \beta_{jB})/3 = \sum_{k=1}^{3} (\overline{D}_{kA} - \overline{D}_{kB}) * (\beta_{kB} - \overline{\beta}_B) = \Delta'_X,$$

and

$$\overline{\Delta}_\beta = \sum_{j=1}^{3} (\alpha_A - \alpha_B + \beta_{jA} - \beta_{jB})/3 + \sum_{k=1}^{3} \sum_{j=1}^{3} \overline{D}_{kA} * (\beta_{kA} - \beta_{kB} - \beta_{jA} + \beta_{jB})/3$$

$$= \alpha_A - \alpha_B + \overline{\beta}_A - \overline{\beta}_B + \sum_{k=1}^{3} \overline{D}_{kA} * (\beta_{kA} - \beta_{kB} - \overline{\beta}_A + \overline{\beta}_B) = \Delta'_\beta.$$

9

Note that we identify the two decomposition effects by identifying coefficients of the dummy variables in regression equations. Suppose we have following regression equation:

$$y = \alpha + \sum_{l=1}^{L} X_l * \delta_l + \sum_{m=1}^{M} \sum_{k_m=2}^{K_m} D_{m\,k_m} * \beta_{m\,k_m} + e, \qquad (2)$$

where there are L continuous variables (X) and M sets of categorical variables (D); the m th set has K_m categories and $K_m - 1$ dummy variables in the equation; without loss of generality, the reference group is the first category of each set of dummy variables; note that the group subscript i is suppressed.

After simple manipulation, the equation can be transformed into a "normalized" regression equation which enables us to identify the coefficients of the dummy variables and the constant. The "normalized" equation is:

$$y = \alpha + \sum_{m=1}^{M} \overline{\beta}_m + \sum_{l=1}^{L} X_l * \delta_l + \sum_{m=1}^{M} \sum_{k_m=1}^{K_m} D_{m\,k_m} * (\beta_{m\,k_m} - \overline{\beta}_m) + e. \qquad (2')$$

Using the "normalized" equation and mean characteristics of every variable including the reference groups, computing the decomposition equation which can identify characteristics and coefficients effects of "each" category including the reference group in equation (2) is a simple manipulation.[11]

Interaction terms may cause some complications. There are two types of interaction terms, either between dummy variables or between dummy and continuous variables. Interaction terms of sets of dummy variables may be treated as another set of categorical variables. For example, there

---

[11] Also note the decomposition can provide another answer to the endeavor of researchers for identifying discrimination in each category including the reference group (e.g., Horrace and Oaxaca (2001)).

is only one interaction variable of race and gender included where the race and gender interaction dummy variable has a value of one when race is white and gender is male. The reference group is then non-whites or females or both. This implies that there are two categories according to this interaction variable. The interaction variables between dummy variables can be treated the same as the "usual" set of dummy variables.

The other possible interaction is between categorical variables and continuous variables. For example, we may have a interaction variable between industries and age. The "normalized" regression equation will transform $AGE * \delta_{age} + \sum_{ind=2}^{I} D_{ind} * AGE * \beta_{ind*age}$ to

$$AGE * (\delta_{age} + \overline{\beta}_{ind*age}) + \sum_{ind=1}^{I} D_{ind} * AGE * (\beta_{ind*age} - \overline{\beta}_{ind*age}).$$

Economists have innocently ignored the identification problem when applying decomposition analysis empirically or have simply given up the detailed decomposition of the coefficients effect. The solution of this identification problem has long eluded economists. This has been frustrating since the identification problem looks deceivingly simple. Careful examination reveals that the identification problem in the decomposition equation is a disguised identification problem of constant and dummy variables in the regression equation. This paper proposes a simple remedy for this problem by utilizing "normalized" regression through an "averaging approach." The identification problem is automatically resolved once we identify the constant and estimates of the dummy variables through "normalized" regression equations, since the source of the identification is located in the regression equation.

REFERENCES

Blinder, Alan S, "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources* 8:4 (1973), 436-455.

Edin, Per-Anders and Jonny Zetterberg, "Interindustry Wage Differentials: Evidence from Sweden and a Comparison with the United States," *American Economic Review* 82:5 (1992), 1341-1349.

Greene, William H. and Terry G. Seaks, "The Restricted Least Squares Estimator: A Pedagogical Note," *Review of Economics and Statistics* 73:3 (1991), 563-567.

Haisken-DeNew, John P. and Christoph M. Schmidt, "Interindustry and Interregion Differentials: Mechanics and Interpretation," *Review of Economics and Statistics* 79:3 (1997), 516-521.

Ham, John C., Jan Svejnar and Katherine Terrell, "Unemployment and the Social Safety Net During Transitions to a Market Economy: Evidence from the Czech and Slovak Republics," *American Economic Review* 88:5 (1998), 1117-1142.

Horrace, William C. and Ronald L. Oaxaca, "Inter-Industry Wage Differentials and the Gender Wage Gap: an Identification Problem," *Industrial and Labor Relations Review* 54:3 (2001), 611-618.

Jones, F. L., "On Decomposing the Wage Gap: A Critical Comment on Blinder's Method," *Journal of Human Resources* 18:1 (1983), 126-130.

Kennedy, Peter, "Interpreting Dummy Variable,"*Review of Economics and Statistics* 68:1 (1986), 174-175.

Krueger, Alan B. and Lawrence H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica* 56:2 (1988), 259-293.

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

Oaxaca, Ronald. L. and Michael R. Ransom, "Identification in Detailed Wage Decompositions," *Review of Economics and Statistics* 81:1 (1999), 154-157.

Radchenko, Stanislav I. and Myeong-Su Yun, "A Bayesian Approach to Decomposing Wage Differentials," *Economics Letters* 78:3 (2003), 431-436.

Suits, Daniel B.,. "Dummy Variables: Mechanics v. Interpretation," *Review of Economics and Statistics* 66:1 (1994), 177-180.

IZA Discussion Papers

No.	Author(s)	Title	Area	Date
822	D. Weichselbaumer R. Winter-Ebmer	The Effects of Competition and Equal Treatment Laws on the Gender Wage Differential	6	07/03
823	A. Filippin	Discrimination and Workers' Expectations	5	07/03
824	A. Filippin	Discrimination and Workers' Expectations: Experimental Evidence	5	07/03
825	A. Filippin A. Ichino	Gender Wage Gap in Expectations and Realizations	5	07/03
826	K. T. Hansen J. J. Heckman K. J. Mullen	The Effect of Schooling and Ability on Achievement Test Scores	6	07/03
827	H. Buddelmeyer E. Skoufias	An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA	6	07/03
828	D. Checchi T. Jappelli	School Choice and Quality	3	07/03
829	J. J. Heckman X. Li	Selection Bias, Comparative Advantage and Heterogeneous Returns to Education: Evidence from China in 2000	6	07/03
830	T. J. Hatton	Emigration from the UK, 1870-1913 and 1950-1998	1	07/03
831	J. H. Abbring G. J. van den Berg	Analyzing the Effect of Dynamically Assigned Treatments Using Duration Models, Binary Treatment Models, and Panel Data Models	6	07/03
832	P.-C. Michaud	Joint Labour Supply Dynamics of Older Couples	3	07/03
833	H. Gersbach	Structural Reforms and the Macroeconomy: The Role of General Equilibrium Effects	2	07/03
834	T. Boeri J. I. Conde-Ruiz V. Galasso	Protecting Against Labour Market Risk: Employment Protection or Unemployment Benefits?	3	07/03
835	G. Joseph O. Pierrard H. R. Sneessens	Job Turnover, Unemployment and Labor Market Institutions	3	07/03
836	M.-S. Yun	A Simple Solution to the Identification Problem in Detailed Wage Decompositions	3	07/03

An updated list of IZA Discussion Papers is available on the center's homepage www.iza.org.