The Identification Problem in Detailed Wage Decompositions: Revisited

ChangHwan Kim

University of Kansas

August 2011

OLS: Mean Wage Gap between Two Groups

$$y = a + \sum_{j=1}^{J} \sum_{k=1}^{K} b_{jk} x_{jk} + e \tag{1}$$

$$\begin{aligned} \bar{y}^W &= a^W + \sum_{j=1}^{J} \sum_{k=1}^{K} b_{jk}^W \bar{x}_{jk}^W \\ \bar{y}^B &= a^B + \sum_{j=1}^{J} \sum_{k=1}^{K} b_{jk}^B \bar{x}_{jk}^B \end{aligned} \tag{2}$$

$$\bar{y}^W - \bar{y}^B = (a^W - a^B) + \sum_{j=1}^{J} \sum_{k=1}^{K} (b_{jk}^W \bar{x}_{jk}^W - b_{jk}^B \bar{x}_{jk}^B) \tag{3}$$

Blinder-Oaxaca Decomposition

Blinder-Oaxaca Decomposition

$$\bar{y}^W - \bar{y}^B = \underbrace{\underbrace{\left(a^W - a^B\right)}_{D1A} + \underbrace{\sum_{j=1}^{J}\sum_{k=1}^{K}(b_{jk}^W - b_{jk}^B)\bar{x}_{jk}^B}_{D1B}}_{D1} + \underbrace{\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{x}_{jk}^W - \bar{x}_{jk}^B)b_{jk}^W}_{D2} \tag{4}$$

D1A: intercept effect
D1B: coefficients effect
D1 (D1A + D1B): total coefficients effect
D2: endowment effect

Identification Problem

- $\bar{y}^W - \bar{y}^B$ is a constant, therefore D1 + D2 is a constant. It is evident that D1 and D2 are also constants.
- As the choices of reference groups change, the estimate of intercept changes, so do other coefficients estimated. As a result, D1A and D1B are not constant, but variant by the choices of reference groups.

An Example: BO Decompositions

	White	Black	Decomposition ($\Delta = .265$)	
	b^W	b^B	D1	D2
I-A. Original BO Decomposition (Ref=LTHS)				
LTHS (=ref)	–	–	–	–
HSG	.251	.223	.010	-.018
SC	.353	.361	-.003	-.008
BA	.706	.673	.005	.054
Grad	.934	1.001	-.005	.049
[Σ Edu Effect]			[.008]	[.077]
Intercept	2.555	2.376	[.179]	
I-B. Original BO Decomposition (Ref=BA)				
LTHS	-.706	-.673	-.003	.025
HSG	-.454	-.450	-.001	.032
SC	-.353	-.312	-.013	.008
BA (=ref)	–	–	–	–
Grad	.229	.328	-.007	.012
[Σ Edu Effect]			[-.025]	[.077]
Intercept	3.261	3.049	[.212]	

A Solution: Averaging Method?

- Gardeazabal and Ugidos (2004) suggest a normalization of the coefficients of dummy variables by imposing a restriction of $\sum \beta_{jk} = 0$ for each factor j.
- This restriction requires to compute the average of the coefficients obtained from all possible reference-group permutations.
- To circumvent this cumbersome procedure, Yun(2005) proposes an averaging method as follows:

Averaging Method

$$
\begin{aligned}
y &= \left(a + \sum_{j=1}^{J} \bar{b}_j\right) + \sum_{j=1}^{J}\sum_{k=1}^{K}(b_{jk} - \bar{b}_j)x_{jk} + e \\
&= a' + \sum_{j=1}^{J}\sum_{k=1}^{K} b'_{jk}x_{jk} + e
\end{aligned}
\tag{5}
$$

$\bar{b}_j = \frac{\sum_{k=1}^{K} b_{jk}}{K}$.

Both the new coefficients for independent variables, $(b_{jk} - \bar{b}_j)$ and the new intercept, $a + \sum_{j=1}^{J} \bar{b}_j$, are invariant to the choice of reference groups. Since the coefficient of a reference group, b_{j0}, becomes $-\bar{b}_j$, there is no omitted group.

Averaging Method Decomposition

	White	Black	Decomposition ($\Delta = .265$)	
	b^W	b^B	D1	D2
I-C. Averaging Method Decomposition				
LTHS	-.449	-.452	.000	.016
HSG	-.198	-.229	.011	.014
SC	-.096	-.091	-.002	.002
BA+	.257	.221	.006	.020
Grad	.485	.549	-.005	.025
[Σ Edu Effect]			[.011]	[.077]
Intercept	3.004	2.828	[.176]	

The Hidden Identification Problems in the Averaging Method

- The intercept is the expected wage when all xs is $1/K$. That is, $E[y|(x_{jk} = 1/K)] = a'$. The difference of the intercepts between two groups, $a'^W - a'^B$, presents the expected wage difference between group W and group B when all xs are distributed evenly by $1/K$ across k for both groups.
- As K changes, so does the intercept.
- Furthermore, the averaging method is not only sensitive to the number of groups, but also sensitive to the ways of grouping.

Averaging Method and Number of K

	White	Black	Decomposition ($\Delta = .265$)	
	b^W	b^B	D1	D2
I-C. Averaging Method Decomposition				
LTHS	-.449	-.452	.000	.016
HSG	-.198	-.229	.011	.014
SC	-.096	-.091	-.002	.002
BA+	.257	.221	.006	.020
Grad	.485	.549	-.005	.025
[Σ Edu Effect]			[.011]	[.077]
Intercept	3.004	2.828	[.176]	
II-A. Averaging Method Using Four Educational Groups: LTHS, HSG, SC and BA+				
LTHS	-.347	-.339	-.001	.012
HSG	-.096	-.117	.008	.007
SC	.006	.021	-.005	.000
BA+	.437	.435	.000	.056
[Σ Edu Effect]			[.002]	[.075]
Intercept	2.902	2.715	[.189]	

Averaging Method and Grouping

	White	Black	Decomposition ($\Delta = .265$)	
	b'^W	b'^B	D1	D2
II-A. Averaging Method Using Four Educational Groups: LTHS, HSG, SC and BA+				
LTHS	-.347	-.339	-.001	.012
HSG	-.096	-.117	.008	.007
SC	.006	.021	-.005	.000
BA+	.437	.435	.000	.056
[Σ Edu Effect]			[.002]	[.075]
Intercept	2.902	2.715	[.189]	
II-B. Averaging Method Using Four Educational Groups: <HSG, SC, BA, and Grad				
<HSG	-.337	-.373	.016	.036
SC	-.199	-.193	-.002	.004
BA	.154	.119	.006	.012
Grad	.382	.447	-.005	.020
[Σ Edu Effect]			[.015]	[.072]
Intercept	3.107	2.930	[.178]	

Issues with Continuous Variables

- As the starting point changes, so does the intercept.
  E.g., age; age-18; age-25
- Oaxaca and Ransom (1999:156) discuss the problem with continuous variables, but they consider this "not necessarily an identification problem."
- Yun (2005:766) simply recommends "to rely on customs" because "the identification problem related to a continuous variable cannot be resolved bacause there are infinitely many transformations."
- Kim (2010) recommends to use a discrete grouping with multiple dummy variables instead of using age as a continuous variable.

Identification Problems with Continuous Variables

	White	Black	Decomposition($\Delta = .265$)	
	b^W	b^B	D1	D2
III-A. Decomposition with Age				
Age	.101	.070	1.233	.051
Age-squared	-.001	-.001	-.586	-.052
[Σ Age Effect]			[.647]	[-.001]
Intercept	.802	1.184	[-.382]	
III-B. Decomposition with Age: Centered to Age 18				
Age	.063	.045	.418	.032
Age-squared	-.001	-.001	-.213	-.033
[Σ Age Effect]			[.205]	[-.001]
Intercept	2.277	2.216	[.060]	
III-C. Averaging Method Decomposition Using Age Groups				
18-24	-.549	-.386	-.019	.002
25-34	-.043	-.058	.004	.099
35-44	.184	.126	.015	-.003
45-54	.226	.174	.013	.000
55-64	.182	.144	.005	.004
[Σ Age Effect]			[.018]	[.004]
Intercept	2.957	2.715	[.242]	

A Suggestion: The Grand-Mean Centering (GMC) Method

- Should we have generally agreeable choices of reference groups, detailed decompositions will become feasible.
- Transform the independent variables x to $(x - \bar{\bar{x}})$ where $\bar{\bar{x}}$ refers to the grand-mean for both group W and group B. The $\bar{\bar{x}}$ is not a simple arithmetic mean between $\bar{x}^W$ and $\bar{x}^B$, but a mean computed using all observations.

GMC Methods

$$
\begin{aligned}
y &= a^* + \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}(x_{jk} - \bar{\bar{x}}_{jk}) + \sum_{l=1}^{L} d_l(c_l - \bar{\bar{c}}_l) + e \\
&= a^* + \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}x^*_{jk} + \sum_{l=1}^{L} d_l c^*_l + e
\end{aligned}
\tag{6}
$$

After estimating equation 6, conduct the usual BO decompositions.

Why Grand Mean Centering?

- The reason why it should be the grand-mean, not the group-specific mean (or other weighting factors), is because the determination of wage will be affected by the demand and the supply of whole labor forces in a society, not only by the demand and supply of a specific group.
- If the currently observed labor market situation is a reflection of an equilibrium condition of employment which affects the wage rates, the most reasonable and practical assumption on the current status of labor market would be $\bar{\bar{x}}$.

Decomposition with the GMC Method: Ref Group

	White	Black	Decomposition ($\Delta = .265$)	
	b^W	b^B	D1	D2
I-D. GMC Method Decomposition (Ref=LTHS)				
LTHS (=ref)	–	–	–	–
HSG	.251	.223	.002	-.018
SC	.353	.361	.000	-.008
BA	.706	.673	-.002	.054
Grad	.934	1.001	.003	.049
[Σ Edu Effect]			[.003]	[.077]
Intercept	3.013	2.828	[.185]	
I-E. GMC Method Decomposition (Ref=BA)				
LTHS	-.706	-.673	-.001	.025
HSG	-.454	-.450	.000	.032
SC	-.353	-.312	-.001	.008
BA (=ref)	–	–	–	–
Grad	.229	.328	.005	.012
[Σ Edu Effect]			[.003]	[.077]
Intercept	3.013	2.828	[.185]	

Decomposition with the GMC Method: Grouping

	White	Black	Decomposition ($\Delta = .265$)	
	b^W	b^B	D1	D2
II-C. GMC Method Using Four Educational Groups: LTHS, HSG, SC and BA+				
LTHS	-.784	-.774	.000	.028
HSG	-.532	-.551	.001	.037
SC	-.431	-.413	.000	.009
BA+	–	–	–	–
[Σ Edu Effect]			[.001]	[.075]
Intercept	3.013	2.824	[.189]	
II-D. GMC Method Using Four Educational Groups: <HSG, SC, BA, and Grad				
<HSG	–	–	–	–
SC	.138	.180	-.001	-.003
BA	.490	.493	.000	.038
Grad	.719	.821	.005	.037
[Σ Edu Effect]			[.004]	[.072]
Intercept	3.013	2.825	[.189]	

GMC Method: Continuous Variable

	White	Black	Decomposition ($\Delta = .265$)	
	b^W	b^B	D1	D2
III-D. GMC Method Decomposition with Age				
Age	.101	.070	-.014	.051
Age-squared	-.001	-.001	.015	-.052
[Σ Age Effect]			[.001]	[-.001]
Intercept	3.019	2.755	[.265]	
III-E. GMC Method Decomposition with Age-18:				
Age-18	.063	.045	-.009	.032
Age-18-squared	-.001	-.001	.010	-.033
[Σ Age Effect]			[.001]	[-.001]
Intercept	3.019	2.755	[.265]	
III-F. GMC Method Decomposition Using Age Groups				
18-24	—	—	—	—
25-34	.506	.328	.001	-.002
35-44	.733	.512	.003	-.011
45-54	.774	.559	.000	.000
55-64	.730	.529	-.004	.017
[Σ Age Effect]			[-.001]	[.004]
Intercept	3.019	2.758	[.261]	

Modified GMC Method

Because there are omitted values (i.e., the coefficients for reference groups are set to zero by definition), a detailed decomposition by factor levels (e.g., LTHS, HSG, SC, Married, Not-married) appears still not feasible with the GMC method. However, an application of the averaging method to the GMC method helps to make the detailed decomposition by each variable viable.

Modified GMC Method

$$y = a^{\dagger} + \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}x_{jk} + \sum_{l=1}^{L} d_l(c_l - \bar{\bar{c}}_l) + e \tag{7}$$

$$\begin{aligned} y &= \left(a^{\dagger} + \sum_{j=1}^{J} \bar{b}_j^*\right) + \sum_{j=1}^{J}\sum_{k=1}^{K}(b_{jk} - \bar{b}_j^*)x_{jk} + \sum_{l=1}^{L} d_l(c_l - \bar{\bar{c}}_l) + e \\ &= a^* + \sum_{j=1}^{J}\sum_{k=1}^{K} b_{jk}^* x_{jk} + \sum_{l=1}^{L} d_l c_l^* + e \\ &\text{where} \quad \bar{b}_j^* = \sum_{k=1}^{K} b_{jk}\bar{\bar{x}}_{jk} \quad \text{for each factor } j. \end{aligned} \tag{8}$$

Summary

- Detailed decompositions of BO techniques are problematic b/c of identification problems.
- To solve this problem, Yun(2005) proposes the averaging methods.
- However, the averaging methods is not free from identification problems. The decomposition results of averaging methods are sensitive to the number of factor levels and ways of grouping.
- To resolve these problems, I suggest the grand-mean centering (GMC) methods and the modified GMC method.

Conclusion

The modified GMC method resolves all identification issues, provides a clear meaning of the intercept term, and makes the detail decomposition feasible with a reasonable assumption.

However,

- The modified GMC method is not the ultimate solution of the identification problems. There are no such methods that can ultimately solve the identification problems.
- Whatever methods–the BO decomposition, the averaging methods, the GMC methods, or any other methods with the constraints of $\sum_{k=1}^{K} b_k' = 0$–are utilized, the detail decompositions are mathematically correct.
- The different choices of model specifications for detail decompositions can be accepted depending on theoretical or practical reasonings.

Thank you!

chkim@ku.edu