

Yun, Myeong-Su

Working Paper

Selection Bias and the Decomposition of Wage Differentials

Working Paper, No. 1999-11

Provided in Cooperation with:

Department of Economics, Rutgers University

Suggested Citation: Yun, Myeong-Su (2000) : Selection Bias and the Decomposition of Wage Differentials, Working Paper, No. 1999-11, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:

<https://hdl.handle.net/10419/94256>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Selection Bias and the Decomposition of Wage Differentials

Myeong-Su Yun

Department of Economics
University of Western Ontario
London, Ontario
Canada, N6A 5C2

myun@julian.uwo.ca

July 10, 2000

Abstract

The major contribution of this paper is finding a new and general approach to decomposing log-wage differentials when selection effects are present. We divide the observed log-wage differentials between two groups into 1) *differentials in predicted log-wages* computed using observed individual characteristics and consistent coefficients while assuming both groups' stochastic component (unobserved individual characteristics) of log-wages to have the same mean zero, and 2) *differentials caused by differences in unobserved individual characteristics* (selection effects). We compute the average of the selection effects by taking sample average of the residuals of log-wages (observed log-wages minus predicted log-wages) without relying on the analytical formula(e) for computing the selection effects. Blinder-Oaxaca type decomposition can be applied for the differentials in predicted log-wages in order to find the effects of differences in observed individual characteristics and the effects of differences in coefficients. We call this approach a "generalized selection bias (GSB) approach." Our approach can be implemented with any kind estimation method as long as we can obtain consistent coefficients for the log-wage equation. We illustrate our approach by applying it to the racial wage differentials among women using data from the current population survey.

JEL codes: J71, J31, C34

Keywords: decomposition analysis of wage differentials, discrimination, selection bias, maximum likelihood estimation, Heckman's two-step method

The author wishes to thank Mark Killingsworth, Roger Klein, Hiroki Tsurumi, and Frank Vella. Special thanks to Ira Gang for his generous assistance and encouragement in writing this paper.

1 Introduction

We provide a general framework for a decomposition (discrimination) analysis of log-wage differentials when selection effects exist by answering the question, “How can we use consistent estimates of parameters of log-wages for decomposing wage differentials between two groups?”. Our framework is independent of the estimation method used to obtain the consistent estimates. As a result, it substantially broadens the scope of decomposition (discrimination) analysis.¹

The decomposition method introduced by Blinder (1973) and Oaxaca (1973) has been widely adopted in the analysis of wage differentials. Based on the regression analysis of log-wages, the average log-wage differentials are decomposed into a part explained by differences in the average individual characteristics, and a part explained by differences in coefficients, traditionally labeled “discrimination.”²

It is well recognized that sample selection causes bias in the OLS coefficients of log-wages.³ A decomposition analysis which does not take account of sample selection, therefore, could over- or underestimate “true” discrimination. Studies on wage differentials and discrimination have adopted the well-known Heckman’s two-step method to obtain consistent estimates of log-wage parameters and applied the Blinder-Oaxaca type decomposition to the log-wage differentials using

¹Though we discuss the decomposition method in terms of (hourly) wages in this paper, the Blinder-Oaxaca decomposition and our decomposition method introduced in this paper can, with minor revision, be applied to decomposing differentials of “any” continuous variables.

²See Becker (1971), Cain (1986) and chapter 2 of Joshi and Paci (1998) for a discussion of the concept of discrimination.

³We interpret sample selection in broad sense; it includes not only the self-selection issue but also any kind of endogeneity caused by censoring, truncation, etc.

second step regression estimates.⁴ However, previous studies based on Heckman's two-step method are limited to models with relatively simple selection issues, usually a single selection issue (e.g., participation vs. non-participation).

In this paper, we substantially broaden the scope of the decomposition (discrimination) analysis in order to handle numerous selection issues (e.g., the double selection model, the censored or truncated regression model, and the switching regression model). This is done by introducing a new and general framework to decomposing wage differentials which is independent of the choice of estimation method as long as we can obtain consistent estimates for the parameters of log-wage equation.

The crucial question in getting a Blinder-Oaxaca type decomposition equation when selection issues exist is "How can we compute the average effects of selection on the log-wages, i.e., the average selection bias?"⁵ Heckman's two-step method computes the average selection bias by averaging each individual's selection bias evaluated using the analytical formula(e) of selection bias. When complicated selection issues exist, it is problematic to derive the analytical formula(e), which may have limited the selection issues previous papers handled.

The key observation of our decomposition method is that the average selection bias can be easily evaluated by averaging the residuals (observed log-wages

⁴See, for example, Bloom and Killingsworth (1982), Joshi and Paci (1998), and Neuman and Oaxaca (1998). See Heckman (1979) for his two-step method.

⁵The selection bias is the difference between conditional and unconditional expectations of the log-wages (Rosen (1986), p. 654).

minus predicted log-wages) of the selected sample.⁶ The intuition behind this observation is remarkably simple. Since selection affects the distribution of the stochastic component of log-wages, the mean value of the stochastic component in the presence of selection is different from the mean value of the stochastic component where selection is absent (assumed to be zero). The average effects of the selection on log-wages (average selection bias) can be measured by computing the average value of the stochastic component of log-wages. Since empirically the residuals of log-wages represent the stochastic component of log-wages, the average of the residuals is the average selection bias. Note that our decomposition method requires only that the coefficients of log-wages be consistent for computing the predicted log-wages.⁷

We divide the average log-wage differentials between two groups into average differentials in predicted log-wages and differences in average selection bias, i.e., log-wages differentials caused by differences in unobserved individual characteristics (selection effects). We treat the differences in the average selection bias as a separate component, and apply the Blinder-Oaxaca type decomposition only to the differentials in the average predicted log-wages in order to find the effects of differences in observed individual characteristics and the effects of differences in their coefficients. We call our decomposition method a “generalized selection bias (GSB) approach” to decomposition analysis of wage differentials.

⁶In this paper, predicted log-wages mean log-wages computed using only observed individual characteristics and their consistent estimates while assuming the mean value of the stochastic component (unobserved individual characteristics) of log-wages to be zero.

⁷Hence, we may also implement our decomposition approach with coefficients from Heckman’s two-step method.

In the next section, we discuss the econometrics of our GSB approach to decomposition (discrimination) analysis in detail. In section 3, we implement the GSB approach to racial wage differentials using the current population survey (CPS). The final section concludes the paper with comments on the possibility of using various estimation methods (semiparametric, Bayesian sampling) and their implications for decomposition analysis.

2 GSB Approach to Decomposition Analysis

2.1 Conventional Model

The foundation of conventional decomposition analysis is a regression of log-wages.⁸ We estimate the log-wage function for group g . This typically takes the following form,

$$(1) \quad Y_{gn} = X_{gn}\beta_g + e_{gn} \quad (n = 1, \dots, n_g),$$

where Y_{gn} , X_{gn} , and e_{gn} are log-wages, $1 \times K_Y$ socio-economic characteristics, and error of individual n in group g (a and b), respectively; β_g is $K_Y \times 1$ vector of parameters; $E(e_{gn}) = 0$.

Based on the OLS analysis, the conventional decomposition analysis uses a simple identity to compute the effects of differences in average individual char-

⁸The Blinder-Oaxaca type decomposition analysis without selection issues is called a conventional decomposition analysis in this paper.

acteristics and the effects of the differences in their coefficients on the log-wage differentials between group a and b . Formally, the decomposition of the difference in log-wages between group a and b can be shown to be,⁹

$$(2) \quad \bar{Y}_a - \bar{Y}_b = \begin{cases} (\bar{X}_a - \bar{X}_b)\hat{\beta}_b + \bar{X}_a(\hat{\beta}_a - \hat{\beta}_b), & \text{or} \\ (\bar{X}_a - \bar{X}_b)\hat{\beta}_a + \bar{X}_b(\hat{\beta}_a - \hat{\beta}_b), \end{cases}$$

where \bar{Y}_g , \bar{X}_g , and $\hat{\beta}_g$ are, for each group g (a and b), the sample average of log-wages ($\sum_{n=1}^{n_g} Y_{gn}/n_g$), $1 \times K_Y$ vector of the average characteristics, and $K_Y \times 1$ vector of OLS coefficients, respectively.

From equation (2), we compute a discrimination coefficient (see Oaxaca (1973) for details) as,

$$(3) \quad \hat{D}_g = \exp(\bar{X}_g(\hat{\beta}_a - \hat{\beta}_b)) - 1,$$

where $g = a$ or b , depending on which group's characteristics are used as weights.

The crucial assumption of the conventional decomposition analysis is that the sample of each group is randomly selected from the population, hence, the expectation of e_{gn} is zero. However, the expectation of e_{gn} might not be zero when the sample is not randomly selected from population. The violation of a mean zero assumption results in biased OLS estimates. This, in turn, implies that the

⁹There are two issues in the discrimination literature. One issue is that the wage differentials between two groups consists of two parts: the gain above the nondiscriminatory (competitive) wage and the loss below the nondiscriminatory wage. The nondiscriminatory wage is usually estimated using pooled data. See Oaxaca and Ransom (1988, 1994) and Neumark (1988). The other is concerning the variances for each component in equation (2). Oaxaca and Ransom (1998) introduce the delta method to compute the variances of each component.

conventional decomposition analysis which does not take account of the bias of the estimates may not correctly decompose the observed log-wage differentials.

2.2 GSB Approach: A Two Equation Model

In this section, we discuss the GSB approach and how it provides a general framework for decomposing log-wage differentials. We consider a two equation model to simplify the exposition. The extensions to more complicated selection issues are straightforward. For each group a and b , equations for individual N are,

$$(1) \quad Y_{gN}^* = X_{gN} \beta_g + e_{gN}$$

$$(4) \quad S_{gN}^* = Z_{gN} \gamma_g + v_{gN} \quad (N = 1, \dots, N_g),$$

where X_{gN} and Z_{gN} are respectively $1 \times K_Y$ and $1 \times K_S$ vectors of socio-economic characteristics of individual N in group g (a and b); coefficients β_g and γ_g are $K_Y \times 1$ and $K_S \times 1$ vectors of parameters, respectively; $E(e_{gN}) = 0$, $E(v_{gN}) = 0$, $E(e_{gN}^2) = \sigma_{e_g}^2$, $E(v_{gN}^2) = \sigma_{v_g}^2$, $E(e_{gN} v_{gN'}) = \sigma_{e_g v_g}$ if $N = N'$ and zero if $N \neq N'$.

Y_{gN}^* and S_{gN}^* are latent log-wages and selection variables, respectively. We observe a binary variable S_{gN} for every individual in group g which has a value of one if $S_{gN}^* > 0$ and zero otherwise. The sample size whose $S_{gN} = 1$ is n_g , where $N_g > n_g$. For individuals whose $S_{gN} = 1$, a continuous variable Y_{gN} is observed equal to Y_{gN}^* while, for others whose $S_{gN} = 0$, Y_{gN} is missed.¹⁰

¹⁰The conventional decomposition analysis presumes Y_{gn} is observed without any missing or censoring, that is $n_g = N_g$.

The unconditional expectation of log-wages, equation (1), can be written as

$$(5) \quad \text{E}(Y_{gN}^* | X_{gN}) = X_{gN} \beta_g,$$

since $\text{E}(e_{gN}) = 0$.

Let the conditional expectation of e_{gN} , given the value of S_{gN} , be Λ_{gN} .¹¹ Then the conditional expectation of log-wages given that the individual worker is being selected into the sample by the equation (4) may be written as

$$(6) \quad \text{E}(Y_{gN}^* | X_{gN}, S_{gN} = 1) = X_{gN} \beta_g + \Lambda_{gN},$$

where $N = 1, \dots, n_g$ because only n_g observations have data available on Y_{gN}^* .

The log-wages for the selected sample may be written using consistent coefficients of log-wages, denoted by tilde ($\tilde{}$), as

$$(7) \quad Y_{gN} = X_{gN} \tilde{\beta}_g + \tilde{\Lambda}_{gN} + \tilde{\varepsilon}_{gN},$$

where $\tilde{\varepsilon}_{gN} = \tilde{e}_{gN} - \tilde{\Lambda}_{gN}$ and $\text{E}(\tilde{\varepsilon}_{gN} | X_{gN}, \tilde{\Lambda}_{gN}, S_{gN} = 1) = 0$.

Note that we have defined predicted log-wages as log-wages computed using only observed individual characteristics and their consistent estimates while assuming the mean value of the stochastic component (unobserved individual characteristics) of log-wages to be zero. Hence the predicted log-wages are equal to the

¹¹It is called the “generalized residuals” (Gourieroux, Monfort, Renault, and Trognon (1987)).

unconditional expectation of log-wages using consistent coefficients ($X_{gN}\tilde{\beta}_g$). The selection bias is reduced to the conditional expectation of e_{gN} , i.e., $\tilde{\Lambda}_{gN}$, which is equal to the difference between conditional and unconditional expectation of log-wages.

We divide the average log-wage differentials into differences in average predicted log-wages and differences in average selection bias. We can apply the Blinder-Oaxaca type decomposition to the differences in predicted log-wages to find the effects of differences in observed individual characteristics and the effects of differences in coefficients. The decomposition equation (2) is modified as follows using only the selected sample ($S_{gN} = 1$),

$$(8) \quad \bar{Y}_a - \bar{Y}_b = \begin{cases} (\bar{X}_a - \bar{X}_b)\tilde{\beta}_b + \bar{X}_a(\tilde{\beta}_a - \tilde{\beta}_b) + (\tilde{\Lambda}_a - \tilde{\Lambda}_b), & \text{or} \\ (\bar{X}_a - \bar{X}_b)\tilde{\beta}_a + \bar{X}_b(\tilde{\beta}_a - \tilde{\beta}_b) + (\tilde{\Lambda}_a - \tilde{\Lambda}_b), \end{cases}$$

where \bar{Y}_g , \bar{X}_g , and $\tilde{\Lambda}_g$ are, for each group g (a and b), the sample average of log-wages ($\sum_{N=1}^{n_g} Y_{gN}/n_g$), $1 \times K_Y$ vector of the average characteristics of individuals, and the average selection bias ($\sum_{N=1}^{n_g} \tilde{\Lambda}_{gN}/n_g$), respectively. $\tilde{\beta}_g$ is $K_Y \times 1$ vector of the consistent estimates of log-wage parameters for group g (a and b).

The corresponding discrimination coefficient using consistent estimators is

$$(9) \quad \tilde{D}_g = \exp(\bar{X}_g(\tilde{\beta}_a - \tilde{\beta}_b)) - 1$$

where $g = a$ or b , and $\tilde{\beta}$ is $K_Y \times 1$ vector of consistent estimates of log-wage

parameters.

2.3 GSB Approach: Evaluation

The GSB approach is developed to provide a general framework for the decomposition analysis, because previous papers based on Heckman’s two-step method are limited to very simple selection issues.¹² Compared to previous studies based on Heckman’s two-step method, there are two factors which make our GSB approach a general framework for decomposing wage differentials. First, our GSB approach does not compute the selection bias ($\tilde{\Lambda}_{gN}$) itself for each individual when we calculate the average selection bias ($\overline{\tilde{\Lambda}}_g$). This is the main difference compare to previous papers based on Heckman’s two-step method. The fact that we don’t evaluate the selection bias for each individual eliminates the burden of deriving the analytical formula(e) of the selection bias. Previous papers have had to analytically derive the formula(e) of the selection bias or selection bias correction term (λ_{gN}) for each individual because they use Heckman’s two-step method.¹³ Previous papers based on Heckman’s two-step method compute the average selection bias as the product of the average of λ_{gN} and its OLS coefficient (θ_g). In the GSB approach, the average of the selection bias used in the decomposition equation (8) is measured by the sample average of the residuals

¹²It is useful to have a general framework which can be applied to any kind selection issue. One example is Yun (2000) which studies the gender wage gap when two wages (full-time and part-time wages) are offered using the piecewise-linear budget constraint model for estimation.

¹³The analytical formula for the selection bias in the previous section is $\rho_{e_g v_g} \sigma_{e_g} \lambda_g$. In Heckman’s two-step method, the product $\rho_{e_g v_g} \sigma_{e_g}$ is estimated as the second-step OLS coefficient (θ_g) for the $\lambda_{gN} = \phi(-Z_{gN} \gamma_g / \sigma_{v_g}) / \Phi(Z_{gN} \gamma_g / \sigma_{v_g})$.

of log-wages ($\tilde{e}_{gN} = Y_{gN} - X_{gN}\tilde{\beta}_g$).¹⁴ We can easily show the equality, that is, $\sum_{N=1}^{n_g} \tilde{e}_{gN}/n_g = \sum_{N=1}^{n_g} (\tilde{\Lambda}_{gN} + \tilde{\varepsilon}_{gN})/n_g = \bar{\Lambda}_g$, since $\sum_{N=1}^{n_g} \tilde{\varepsilon}_{gN}/n_g = 0$.

Second, we divide the observed log-wage differentials into differences in average predicted log-wages (differences in unconditional expectations of log-wages) and differences in average selection bias. We apply the Blinder-Oaxaca type decomposition to the differences in average predicted log-wages to find the effects of differences in observed individual characteristics and the effects of differences in coefficients. Some previous papers based on Heckman’s two-step method consider the selection bias correction term (λ_{gN}) as another individual characteristic, and decompose the differences in selection bias into the effects of differences in the selection bias correction term and the effects of differences in its OLS coefficients (see Neuman and Oaxaca (1998) for details).

Theoretically, we might justify our focusing on the differences in unconditional expectations of log-wages by interpreting Y_{gN}^* and Y_{gN} as “offered” log-wages and “observed” log-wages (Reimers (1983)). Though this interpretation is very attractive, its implicit assumption that the firms and the economists observe the same individual characteristics is problematic. It is possible that the firms observe more worker’s characteristics than economists do from the data (Heckman (1998)).

In fact, the reason why we do not decompose the selection bias stems from practical considerations. For simple selection issues, applying Heckman’s two-

¹⁴In practice, we don’t have to compute the residuals either to compute the average of the selection bias. It is because we can compute the average selection bias as the average observed log-wages minus the average predicted log-wages.

step method by deriving the formula(e) of the selection bias is not a problem. However, it becomes very difficult to derive the selection bias analytically if the selection issues are complicated.¹⁵ Furthermore, it is possible that Heckman's two-step method cannot be applied due to conceptual difficulties (e.g., truncation issue).¹⁶

Another reason is related to the nature of the selection bias. The selection bias represents the effects of unobserved characteristics of an individual on the log-wages. The unobserved characteristics will be the combination of many factors, not just a single characteristic of the person. If the λ_{gN} is considered as another exogenous variable, then it is not clear whether λ_{gN} and its coefficient can be treated like any other observed exogenous variables and their coefficients because each observed exogenous variable is presumed to represent only one aspect of the individual's characteristics.

To summarize, the GSB approach can be implemented as follows; first, compute the means of exogenous variables (\bar{X}_g) and the observed log-wages (\bar{Y}_g); second, compute the average predicted log-wages as product of the mean of the exogenous variables and their consistent coefficients ($\bar{X}_g\tilde{\beta}_g$); third, compute the average selection bias as $\bar{Y}_g - \bar{X}_g\tilde{\beta}_g$; and finally compute the each component of the decomposition equation (8). The GSB approach computes the average se-

¹⁵Of course, it is true that obtaining consistent coefficients will be difficult using other estimation methods when the selection issues are complicated. However, there are many papers which deal with complicated selection issues using various estimation methods. In those cases, our GSB approach will be very useful.

¹⁶See Bloom and Killingsworth (1985), Hausman and Wise (1977), and chapter 6 of Maddala (1983) for truncation issue.

lection bias by averaging the residuals for selected sample, and decomposes the differences in average predicted log-wages into the effects of differences in individual characteristics and the effects of differences in coefficients. Since the GSB approach is not restricted to a specific estimation method as long as we can obtain consistent coefficients, it frees us from the limited choices of estimation method previous papers employed. By extending the choices of estimation method, we may be able to substantially broaden the scope of the decomposition analysis when there are selection issues.

3 Racial Wage Discrimination Among Women

We apply the GSB approach to racial wage differentials between white and other races using the female sample from March 1995 CPS. This illustration is a direct application of the two equation model we introduced in the previous section.

3.1 Data

The female sample used in our empirical study is drawn from the March 1995 CPS. The data comes from the outgoing rotation group only, and the responses to questions about the survey month are used rather than those for last year.¹⁷ The sample includes females aged between 25 and 60 who were not in school, retired, disabled or self-employed. For married women, we exclude those whose

¹⁷The information on last year's earnings are used to compute non-labor income. For details, see Table 1.

husbands are under 25 years old. We also exclude women whose hourly wage rate is greater than \$40, or whose working hours are top-coded (99 hours per week). Table 1 describes the variables used for our study.

Table 2 shows the means and standard deviations of variables used in the decomposition analysis. The characteristics of working women are different from those of non-working women. Working women are older and have more years of education than non-working women. Non-working women have a higher marriage rate among white women but there is little difference among other race women. Non-working women have more children (for both age under 6 and between age 6 and 18) and larger family size. Non-working women also have a higher non-labor income among white women, but there is not much difference among other race women, which might be related to the marriage rates.

White women are more educated than other race women in both the non-working and working samples. Though white women have a higher rate of marriage, family size of white women is smaller than that of other race women. The number of children is not significantly different between white and other race women. White women, especially among non-working women, have higher non-labor income than other race women do.

Though both white and other race women are working similar hours, their wages (measured both in level and log) are significantly different from each other according to the t-test at 5% (level wage) and 1% level (log-wage), respectively. We explain racial wage differentials in terms of differences in individual charac-

teristics and differences in the coefficients using the GSB approach proposed in section 2.

3.2 Selection Bias Due to Participation Choice

The selection bias due to the participation decision is a well-studied subject in the area of both labor supply and wage determination.¹⁸ Most papers on wage differentials, especially those on the gender gap, address sample selection bias arising from the participation decision using Heckman's two-step method.¹⁹

Women are partitioned into two groups according to their race, whites ($g = w$) and other races ($g = o$). Hence group a and b in section 2 are whites (w) and other races (o), respectively. For the selection equation, we have a participation equation. Equations (1) and (4) are respectively latent log-wages and participation equations. Women will participate in the labor market if S_{gN}^* in equation (4) has a positive value.

For the GSB approach, we estimate log-wages and participation equations (equations (1) and (4), respectively) jointly using a full information maximum likelihood estimation (MLE) method to obtain consistent coefficients. For comparison purposes, we also use Heckman's two-step method for the decomposition

¹⁸Participation is usually defined to include employment or unemployment. However most studies of labor supply do not count unemployment in the definition of participation. We treat unemployment as non-participation to keep the analysis simple. Blundell, Ham and Meghir (1987) is a rare exception. They include unemployment in the definition of participation.

¹⁹There are numerous studies which use Heckman's two-step method to correct selection bias caused by the participation decision; for example, Reimers (1983), Hoffman and Link (1984), Dolton and Makepeace (1986, 1987), Blau and Beller (1988), Dolton, Makepeace, and van Der Klaauw (1989), Wright and Ermisch (1991), Choudhury (1993), Wellington (1993), Baker, Benjamin, Cegep, and Grant (1995), and Joshi and Paci (1998).

analysis following previous papers.²⁰ The MLE method is implemented using both Gauss CML (constrained maximum likelihood) program and the SAS NLP (non-linear programming) procedure (SAS Institute, 1997). The likelihood function for joint estimation of the log-wages and participation equations is omitted since this model is well known.²¹

Tables 3 and 4 show the estimates of log-wages and participation equations, respectively. Most of the estimates of the log-wage parameters for white women are significant, but estimates for other race women are not very significant. Nonetheless, the estimates for both groups of women have the expected signs. The determinants of participation, shown in Table 4, also show the expected signs; education increases participation, and the presence of children decreases participation. Only the marriage variable has an unexpected positive sign. The estimate of the marriage coefficient is not significant in white women but is significant at 5% in other race women.

Table 5 shows the decomposition results. Some earlier papers based on Heckman's two-step method treat λ_{gN} and its OLS coefficient (θ_g) as simply another individual characteristic and its coefficient (albeit one pertaining to unobserved characteristics).²² We treat the difference in the average selection bias as a sepa-

²⁰Dolton and Makepeace (1986, 1987) estimate the log-wages and participation equations using both Heckman's two-step and the MLE methods. To the best of our knowledge, Dolton and Makepeace (1986, 1987) are the only papers which use the MLE method in the context of decomposition analysis. However, they use the MLE estimates to compute λ_{gN} and its coefficient.

²¹See Heckman (1974), Mroz (1987), and Zabel (1993) for details.

²²Dolton and Makepeace (1986), and Joshi and Paci (1998), among others.

rate component in decomposition equation (8).²³ We have already discussed the reasons for not pursuing further decomposition in the previous section. Our approach guarantees a consistent decomposition equation regardless of the selection issues. We can find the analytical form of the selection bias easily in this illustration ($\Lambda_{gN} = \rho_{e_g v_g} \sigma_{e_g} \lambda_{gN}$), but we may have difficulty in deriving selection bias analytically in many cases. This means that we cannot decompose the selection term into coefficient and unobserved characteristics.

We compute the selection bias following analytical formula using Heckman’s two-step method. The sample average of the selection bias ($\widehat{\theta}_g^T \overline{\widehat{\lambda}_g^T} = \widehat{\theta}_g^T \sum_{N=1}^{n_g} \widehat{\lambda}_{gN}^T / n_g$) is 0.03832 and 0.07575 for white and other race women, respectively. From the GSB approach using MLE method, the sample average of the selection bias ($\overline{\widetilde{\Lambda}_g} = \sum_{N=1}^{n_g} \widetilde{e}_{gN} / n_g$) is 0.02360 and 0.02374 for white and other race women, respectively.²⁴

As shown in Table 5, decomposition results from the conventional analysis using simple OLS estimates and the GSB approach show very similar patterns. Differences in characteristics explain half of log-wage differentials (about 57% (46%) when the coefficients of log-wage parameters of other race (white) women

²³Reimers (1983), Wright and Ermisch (1991), and Ermisch and Wright (1992) follow this direction.

²⁴For illustration purposes, we compute the selection bias following the analytical formula ($\rho_{e_g v_g} \sigma_{e_g} \lambda_{gN}$) using the MLE estimates. The sample average of the selection bias following the analytical formula using the MLE estimates ($\overline{\widetilde{\rho}_{e_g v_g} \widetilde{\sigma}_{e_g} \widetilde{\lambda}_g} = \widetilde{\rho}_{e_g v_g} \widetilde{\sigma}_{e_g} \sum_{N=1}^{n_g} \widetilde{\lambda}_{gN} / n_g$) is 0.02363 and 0.02375 for white and other race women, respectively. The discrepancy between the average selection bias computed by averaging residuals and the selection bias measured using the analytical formula is 0.00003 and 0.00001 for white and other race women, respectively. The discrepancy might be caused either because the MLE method fails to obtain the “true” optimization, or because there is a precision problem in computing the λ_{gN} , since computing λ_{gN} in extreme area where probability is very close to zero or 1 might be problematic. See McCullough and Vinod (1999).

are used), and differences in coefficients explain other half of the log-wage differentials (about 43% (54%) when the average characteristics of white (other race) women are used as weights). This is because the difference in the average selection bias for the GSB approach is virtually non-existent (whites: 0.02360; other races: 0.02374). The discrimination coefficients (in percentage) from the conventional analysis (\widehat{D}_g) and the GSB approach (\widetilde{D}_g) are about 3.4% and 4.3% when characteristics of white and other race women are, respectively, used as weights. In contrast, decomposition results based on Heckman’s two-step method shows larger discrimination than do those from both the conventional analysis and the GSB approach. The discrimination coefficients (in percentage) based using Heckman’s two-step method (\widehat{D}_g^T) are 7.0% and 8.1% when characteristics of white and other race women are, respectively, used as weights.²⁵ This is because differences in the average selection bias between white and other race women are large and negative (whites: 0.03832; other races: 0.07575) which means log-wages of other race women are more likely increased due to their unobserved characteristics.

4 Conclusion

In this paper, we answer the question, “How can we use consistent estimates of parameters of log-wages for decomposing wage differentials between two groups?”.

By answering this question, we provide a new and general framework for decom-

²⁵If we include the difference in the coefficients of the λ_{gN} as part of discrimination, the discrimination coefficients (in percentage) become 5.1% and 5.6%, still higher than those from both the conventional analysis and the GSB approach.

posing log-wage differentials when selection issues exist. To answer this question, we find a new and flexible way to measure the average effects of selection on the log-wages (selection bias). The average selection bias is measured by averaging the residuals of log-wages (observed log-wages minus predicted log-wages) for the selected sample. We divide the average log-wage differentials into differences in average predicted log-wages and differences in average selection bias. We can apply the Blinder-Oaxaca type decomposition to the differences in predicted log-wages to find the effects of differences in observed individual characteristics and the effects of differences in coefficients. We called this approach the “generalized selection bias approach.” Since the GSB approach is not restricted to a specific estimation method as long as we can obtain consistent coefficients, it gives us many possible choices of estimation method. By extending the choices of estimation method, we may be able to decompose log-wage differentials with virtually any kind of selection issue. We have illustrated GSB approach by applying it to the racial wage differentials among women using data from the CPS.

In our illustrations, the fully parametric MLE method has been used to estimate log-wages and selection equations jointly. However, the GSB approach is not restricted to the fully parametric MLE method. It is not crucial which estimation technique is used for the estimation of log-wages and selection equations as long as the estimation provides the consistent estimates of log-wages. We will now briefly discuss two other methods which have their own merits.

A distribution free estimation method, a semiparametric method, has been

studied to avoid the misspecification of the error distribution.²⁶ Since semiparametric methods usually estimate parameters up to a scale factor, it is difficult to estimate the constant term. Hence, the difference in the constants, which is thought to be a part of discrimination, cannot be calculated. One solution is to include the difference in the constant coefficients into the selection bias. The decomposition formula will be the same, only the concept of the selection bias is extended to include the difference in the constant terms.

Bayesian methods may also be used for the estimation. Since Bayesian estimation gives us the (posterior) distribution of coefficients, mean values of coefficients could be used for computing the selection bias and the decomposition analysis (8), presuming that mean values are consistent estimates of population parameters. An interesting development in recent Bayesian estimation is the Bayesian sampling method. It estimates the (posterior) distribution of coefficients of highly complicated models by utilizing Markov Chain Monte Carlo (MCMC) simulation methods. We can estimate the distribution of each component of the decomposition analysis (8), by evaluating them from the sampled estimates in each sampling round.²⁷

The GSB approach can be implemented in conjunction with any kind estimation method; MLE, semiparametric, and Bayesian estimation, etc. With the progress of computing technology, the GSB approach is able to handle virtually

²⁶See Vella (1998) and Pagan and Ullah (1999), chapter 8 for a survey of semiparametric methods. The majority of the semiparametric methods have adopted parametric Heckman's two-step method, but MLE type semiparametric methods also have been developed (see Galland and Nychka (1987) and Ai (1997)).

²⁷See Tanner (1996) and Chib and Greenberg (1996).

any selection issue. The GSB approach is a basic tool for wage differentials and discrimination analysis. It is conceptually simple, and practically versatile.

REFERENCES

- Ai, Chunrong (1997): "A Semiparametric Maximum Likelihood Estimator," *Econometrica*, 65, 933-963.
- Baker, Michael, Dwayne Benjamin, Andrée Desaulniers Cegep, and Mary Grant (1995): "The Distribution of the Male/Female log-wages Differential, 1970-1990," *Canadian Journal of Economics*, 28, 479-501.
- Becker, Gary S. (1971): *The Economics of Discrimination*, second edition. Chicago: University of Chicago Press.
- Blau, Francine D. and Andrea H. Beller (1988): "Trends in log-wages Differentials by Gender, 1971-1981," *Industrial and Labor Relations Review*, 41, 513-529.
- Blinder, Alan S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436-455.
- Bloom, David E. and Mark R. Killingsworth (1982): "Pay Discrimination Research and Litigation: The Use of Regression," *Industrial Relations*, 21, 318-339.
- (1985): "Correction for Truncation Bias Caused by A Latent Truncation Variable," *Journal of Econometrics*, 27, 131-135.
- Blundell, Richard, John Ham, and Costas Meghir (1987): "Unemployment and Female Labour Supply," *Economic Journal*, 97, 44-64.
- Cain, Glen G. (1986): "The Economic Analysis of Labor Market Discrimination: A Survey," in *Handbook of Labor Economics, Vol. I*, ed. by Orley Ashenfelter and Richard Layard. Amsterdam: Elsevier Science B.V., 693-785.
- Chib, Siddhartha and Edward Greenberg (1996): "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, 12, 409-431.
- Choudhury, Sharmila (1993): "Reassessing the Male-Female Wage Differential: A Fixed Effects Approach," *Southern Economic Journal*, 60, 327-340.
- Dolton, Peter J. and Gerald H. Makepeace (1986): "Sample Selection and Male-Female Log-Wages Differentials in the Graduate Labour Market," *Oxford Economic Papers*, 38, 317-341.

- (1987): “Marital Status, Child Rearing and Log-Wages Differentials in the Graduate Labour Market,” *Economic Journal*, 97, 897-922.
- Dolton, Peter J., Gerald H. Makepeace, and W. van Der Klaauw (1989): “Occupational Choice and Log-Wages Determination: The Role of Sample Selection and Non-Pecuniary Factors,” *Oxford Economic Papers*, 41, 573-594.
- Ermisch, John F. and Robert E. Wright (1992): “Differential Returns to Human Capital in Full-time and Part-time Employment,” in *Issues in Contemporary Economics Vol. 4 Women’s Work in the World Economy* ed. by Nancy Folbre, Barbara Bergmann, Bina Agarwal and Maria Floro. New York: New York University Press, 195-212.
- Gallant, A. Ronald, and Douglas W. Nychka (1987): “Semi-nonparametric Maximum Likelihood Estimation,” *Econometrica*, 55, 363-390.
- Gourieroux, Christian, Alain Monfort, Eric Renault, and Alain Trognon (1987): “Generalized Residuals,” *Journal of Econometrics*, 34, 5-32.
- Hausman, Jerry A. and David A. Wise (1977): “Social Experimentation, Truncated Distributions, and Efficient Estimation,” *Econometrica*, 45, 919-938.
- Heckman, James (1974): “Shadow Prices, Market Wages, and Labor Supply,” *Econometrica*, 42, 679-694.
- (1979): “Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153-161.
- (1998): “Detecting Discrimination,” *Journal of Economic Perspectives*, 12, 101-116.
- Hoffman, Saul D. and Charles R. Link (1984): “Selectivity Bias in Male Wage Equations: Black-White Comparisons,” *Review of Economics and Statistics*, 66, 320-324.
- Joshi, Heather and Pierella Paci (1998): *Unequal Pay for Women and Men: Evidence from the British Birth Cohort Studies*, Cambridge: MIT press.
- Maddala, G. S. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.
- McCullough, B. D. and H. D. Vinod (1999): “The Numerical Reliability of Econometric Software,” 37, 633-665.
- Mroz, Thosmas A. (1987): “The Sensitivity of an Empirical Model of Marred Women’s Hours of Work to Economic and Statistical Assumptions,” *Econometrica*, 55, 765-799.

- Neuman, Shoshana and Ronald L. Oaxaca (1998): "Estimating Labor Market Discrimination with Selectivity Corrected Wage Equations: Methodological Considerations and An Illustration from Israel," paper presented at CEPR European Summer Symposium in Labour Economics and Migration, July 12, 1998.
- Neumark, David (1988): "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination," *Journal of Human Resource*, 23, 279-295.
- Oaxaca, Roland (1973): "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 14, 693-709.
- Oaxaca, Ronald L. and Michael R. Ransom (1988): "Searching for the Effect of Unionism on the Wages of Union and Nonunion Workers," *Journal of Labor Research*, 9, 139-148.
- (1994): "On discrimination and the Decomposition of Wage Differentials," *Journal of Econometrics*, 61, 5-21.
- (1998): "Calculation of Approximate Variances for Wage Decomposition Differentials," *Journal of Economic and Social Measurement*, 24, 55-61.
- Pagan, Andrian and Aman Ullah (1999): *Nonparametric Econometrics*, Cambridge, U.K.: Cambridge University Press.
- Reimers, Cordelia W. (1983): "Labor Market Discrimination Against Hispanic and Black," *Review of Economics and Statistics*, 65, 570-579.
- Rosen, Sherwin (1986): "The Theory of Equalizing Differences," in *Handbook of Labor Economics, Vol. I*, ed. by Orley Ashenfelter and Richard Layard. Amsterdam: Elsevier Science B.V., 641-692.
- SAS Institute (1997): *SAS/OR Technical Report: The NLP Procedure*, Cary, NC: SAS Institute Inc.
- Tanner, Martin A. (1996): *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag.
- Vella, Francis (1998): "Estimating Models with Sample Selection Bias: A Survey," *Journal of Human Resources*, 33, 127-169.
- Wellington, Alison J. (1993): "Changes in the Male/Female Wage Gap, 1976-85," *Journal of Human Resources*, 28, 383-411.
- Wright, Robert E. and John F. Ermisch (1991): "Gender Discrimination in the British Labour Market: A Reassessment," *Economic Journal* 101, 508-521.
- Yun, Myeong-Su (2000): "Gender Wage Gap and Part-Time Work," Working Paper, Department of Economics, Rutgers University.

Zabel, Jeffrey E. (1993): "The Relationship between Hours of Work and Labor Force Participation in Four Models of Labor Supply Behavior," *Journal of Labor Economics*, 11, 387-416.

Table 1: VARIABLES USED IN THE ANALYSIS

Variables	Definition and Note
Age	Aged 25 – 60 years.
Age ² /100	Age squared in hundreds.
Education	Number of years of schooling.
Marriage	Married = 1, Single = 0. Married but spouse absent is treated as single.
Children < 6	Number of children under age 6.
Children 6–18	Number of children age 6 – 18.
Family Size	Number of family members.
Non-Labor Inc.	Sum of last year’s survivor’s income, interest income, dividends income, rent income, child support payment, alimony. If married, husband’s annual wage of last year is added. Unit is \$1000.
MSA	Metropolitan statistical areas = 1, Else = 0.
West	West region = 1, Else = 0.
South	South region = 1, Else = 0.
Midwest	Midwest region = 1, Else = 0.
Northeast	Reference region.
Wages (\$)	Hourly wage rate (level) = usual weekly earnings / usual weekly hours of work.
Hours	Usual weekly hours of work.

Table 2: MEAN CHARACTERISTICS OF THE SAMPLE

	Whites		Others	
	Mean	(s.d.)	Mean	(s.d.)
Whole Sample				
Age	39.694	(9.371)	39.088	(9.368)
Education	13.290	(2.642)**	12.832	(2.648)
Marriage	0.615	(0.487)**	0.403	(0.491)
Children < 6	0.294	(0.609)	0.318	(0.656)
Children 6–18	0.618	(0.925)*	0.706	(0.967)
Family Size	2.919	(1.439)**	3.125	(1.600)
Non-Labor Inc.	25.154	(27.218)**	13.267	(20.746)
Sample size	3829		894	
Non-Working Sample				
Age	38.576	(9.434)	37.701	(9.134)
Education	12.326	(2.933)*	11.814	(2.996)
Marriage	0.755	(0.430)**	0.398	(0.491)
Children < 6	0.622	(0.842)	0.534	(0.882)
Children 6–18	0.799	(1.017)	0.928	(1.122)
Family Size	3.555	(1.483)	3.561	(1.743)
Non-Labor Inc.	33.479	(31.286)**	14.142	(23.850)
Sample size	695		221	
Working Sample				
Age	39.942	(9.340)	39.544	(9.406)
Education	13.503	(2.524)**	13.166	(2.435)
Marriage	0.584	(0.493)**	0.404	(0.491)
Children < 6	0.222	(0.517)	0.247	(0.544)
Children 6–18	0.578	(0.899)	0.633	(0.900)
Family Size	2.778	(1.390)**	2.982	(1.524)
Non-Labor Inc.	23.308	(25.876)**	12.979	(19.631)
MSA	0.737	(0.440)**	0.816	(0.388)
West	0.185	(0.388)	0.198	(0.399)
South	0.280	(0.449)**	0.441	(0.497)
Midwest	0.268	(0.443)**	0.160	(0.367)
Wages	11.723	(6.120)*	11.070	(6.229)
Log-wages	2.331	(0.526)**	2.254	(0.576)
Hours	37.594	(9.024)	37.960	(7.482)
Sample size	3134		673	

^a ** and * imply that the null hypothesis, mean of white women is equal to that of other race women, is rejected at 1% and 5% level of significance, respectively.

Table 3: LOG-WAGES

	<u>Whites</u>					
	<u>OLS</u>		<u>Two-Step</u>		<u>MLE</u>	
	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Constant	-0.214	(0.161)	-0.357*	(0.178)	-0.303	(0.165)
Age	0.059**	(0.008)	0.060**	(0.008)	0.060**	(0.008)
Age ² /100	-0.067**	(0.009)	-0.069**	(0.010)	-0.068**	(0.010)
Education	0.098**	(0.003)	0.103**	(0.004)	0.101**	(0.003)
MSA	0.089**	(0.019)	0.090**	(0.017)	0.089**	(0.019)
West	-0.023	(0.025)	-0.024	(0.025)	-0.024	(0.025)
South	-0.104**	(0.022)	-0.104**	(0.023)	-0.104**	(0.022)
MidWest	-0.105**	(0.022)	-0.106**	(0.022)	-0.105**	(0.022)
λ			0.136*	(0.057)		
σ_e					0.457**	(0.006)
ρ_{ev}					0.184*	(0.072)
Adjusted R ²	0.254		0.255			
	<u>Others</u>					
	<u>OLS</u>		<u>Two-Step</u>		<u>MLE</u>	
	Est.	(s.e.)	Est.	(s.e.)	Est.	(s.e.)
Constant	-0.102	(0.373)	-0.391	(0.509)	-0.192	(0.390)
Age	0.039*	(0.018)	0.042*	(0.021)	0.040*	(0.018)
Age ² /100	-0.036	(0.022)	-0.037	(0.025)	-0.036	(0.022)
Education	0.103**	(0.008)	0.113**	(0.013)	0.106**	(0.009)
MSA	0.105*	(0.051)	0.106*	(0.045)	0.105*	(0.051)
West	0.053	(0.063)	0.047	(0.071)	0.051	(0.063)
South	-0.101	(0.053)	-0.104	(0.056)	-0.102	(0.053)
MidWest	-0.032	(0.066)	-0.040	(0.069)	-0.035	(0.065)
λ			0.199	(0.217)		
σ_e					0.504**	(0.014)
ρ_{ev}					0.124	(0.162)
Adjusted R ²	0.229		0.230			

^a ** and * mean statistically significant at 1% and 5%, respectively.

Table 4: PARTICIPATION

<u>Whites</u>				
	<u>Two-Step</u>		<u>MLE</u>	
	Est.	(s.e.)	Est.	(s.e.)
Constant	-1.626**	(0.525)	-1.646**	(0.525)
Age	0.086**	(0.026)	0.085**	(0.026)
Age ² /100	-0.108**	(0.032)	-0.107**	(0.032)
Education	0.119**	(0.010)	0.120**	(0.010)
Marriage	0.054	(0.077)	0.071	(0.076)
Children < 6	-0.475**	(0.050)	-0.478**	(0.049)
Children 6–18	-0.072	(0.040)	-0.067	(0.040)
Family Size	-0.053	(0.030)	-0.049	(0.030)
Non-Labor Inc.	-0.009**	(0.001)	-0.010**	(0.001)
<u>Others</u>				
	<u>Two-Step</u>		<u>MLE</u>	
	Est.	(s.e.)	Est.	(s.e.)
Constant	-1.418	(0.945)	-1.411	(0.945)
Age	0.025	(0.046)	0.022	(0.047)
Age ² /100	-0.015	(0.057)	-0.010	(0.057)
Education	0.129**	(0.020)	0.132**	(0.021)
Marriage	0.293*	(0.147)	0.296*	(0.147)
Children < 6	-0.239**	(0.082)	-0.244**	(0.082)
Children 6–18	-0.076	(0.063)	-0.065	(0.065)
Family Size	-0.024	(0.042)	-0.021	(0.042)
Non-Labor Inc.	-0.011**	(0.003)	-0.012**	(0.004)

^a ** and * mean statistically significant at 1% and 5%, respectively.

Table 5: DECOMPOSITION ANALYSIS

Observed Log-Wage Differentials	
$\bar{Y}_w - \bar{Y}_o$	0.077 (100.00% ^a)
<u>OLS</u>	
Component	Logarithm (%)
(a) Difference in Characteristics	
(a.1) $(\bar{X}_w - \bar{X}_o)\hat{\beta}_o$	0.043 (56.24%)
(a.2) $(\bar{X}_w - \bar{X}_o)\hat{\beta}_w$	0.035 (45.22%)
(b) Difference in Coefficients	
(b.1) $\bar{X}_w(\hat{\beta}_w - \hat{\beta}_o)$	0.034 (43.75%)
(b.2) $\bar{X}_o(\hat{\beta}_w - \hat{\beta}_o)$	0.042 (54.78%)
<u>Two-Step</u>	
Component	Logarithm (%)
(a) Difference in Characteristics	
(a.1) $(\bar{X}_w - \bar{X}_o)\hat{\beta}_o^T$	0.047 (60.94%)
(a.2) $(\bar{X}_w - \bar{X}_o)\hat{\beta}_w^T$	0.036 (47.31%)
(b) Difference in Coefficients	
(b.1) $\bar{X}_w(\hat{\beta}_w^T - \hat{\beta}_o^T)$	0.067 (87.69%)
(b.2) $\bar{X}_o(\hat{\beta}_w^T - \hat{\beta}_o^T)$	0.078 (101.32%)
(c) Difference in Selection Bias	
$\hat{\theta}_w^T \bar{\lambda}_w^T - \hat{\theta}_o^T \bar{\lambda}_o^T$	-0.037 (-48.63%)
<u>MLE</u>	
Component	Logarithm (%)
(a) Difference in Characteristics	
(a.1) $(\bar{X}_w - \bar{X}_o)\tilde{\beta}_o$	0.044 (57.75%)
(a.2) $(\bar{X}_w - \bar{X}_o)\tilde{\beta}_w$	0.036 (46.54%)
(b) Difference in Coefficients	
(b.1) $\bar{X}_w(\tilde{\beta}_w - \tilde{\beta}_o)$	0.033 (42.43%)
(b.2) $\bar{X}_o(\tilde{\beta}_w - \tilde{\beta}_o)$	0.041 (53.64%)
(c) Difference in Selection Bias	
$(\bar{\Lambda}_w - \bar{\Lambda}_o)$	-0.0001 (-0.18%)

^a Percentage of observed differentials contributed by component is in parentheses.