Journal of Economic Inequality **2:** 3–10, 2004.

Wage decompositions with selectivity-corrected wage equations: A methodological note

SHOSHANA NEUMAN[1] and RONALD L. OAXACA[2]
[1]*CEPR London, IZA Bonn, and Department of Economics, Bar-Ilan University, 52900 Ramat-Gan, Israel, E-mail: neumans@mail.biu.ac.il*
[2]*IZA Bonn and Department of Economics, 401 McClelland Hall, University of Arizona, PO Box 210108, Tucson, AZ 85721-0108, USA, E-mail: rlo@email.arizona.edu*

Received 4 December 2002; accepted 18 December 2003

Abstract. This paper examines the implications of the standard Heckman (Heckit) correction for selectivity bias in wage and earnings functions that are subsequently used in wage decompositions. Even when justified, Heckit selectivity correction introduces some fundamental ambiguities in the context of wage decompositions. The ambiguities arise from group differences in the selection term which consists of a parameter multiplied by the Inverse Mills Ratio (IMR). The parameter is identified as the product of the standard deviation of the errors in the wage equation and the correlation between the wage equation error and the selection equation error. How should group differences in these parameters be interpreted in terms of structural differences and endowment effects? The same issue arises with respect to group differences in the IMR which reflect nonlinear group differences in the determinants of selection and in the probit coefficients.

Key words: identification, selectivity bias, wage decompositions.

1. Introduction

Estimation of labor market discrimination by gender, race, and ethnicity has become routine since the popularization of the wage decomposition methodology by Blinder [2] and Oaxaca [14]. A more general approach to wage decompositions is found in Neumark [13], Oaxaca and Ransom [15, 16]. In the more general approach the nondiscriminatory wage structure is estimated from a pooled sample of the two demographic groups. This approach allows the discrimination component to be further disaggregated into overpayment (favoritism) and underpayment (pure discrimination). Another popular approach to wage decompositions is the Juhn–Murphy–Pierce decomposition [10]. The JMP decomposition attempts to identify the wage gap effects of changes in unobserved prices and unobserved skills. Applied to the gender wage gap, the JMP decomposition isolates the mean standardized wage residual for females in the male wage residual distribution [1]. Datta Gupta et al. [5] show the equivalence between the JMP decomposition applied to a pooled regression that combines the male and female samples and the Oaxaca–Ransom [16] generalized decomposition. The gender difference that would arise from gender

differences in the JMP mean standardized wage residuals in the pooled wage regression are shown to be identical to the sum of estimated male favoritism and pure discrimination against females. Suen [18] criticized the JMP decomposition by pointing to an identification problem when wage dispersion is not independent of the residual ranking. By interpreting a change in the JMP dispersion/unobserved price term as a change in the combined effects of favoritism and pure discrimination, Datta Gupta et al. [5] offer an approach that can finesse the Suen critique.

Another twist in wage decomposition methodology is occasioned by selectivity bias. In the presence of sample selection *OLS* estimation of wage equations can yield biased and inconsistent estimators [7–9]. It is widely recognized that the standard Heckit procedure is susceptible to identification problems and sensitivity of results to model specification and distributional assumption [19]. It is well known that the Heckit model can theoretically be identified by the nonlinearity of the Inverse Mills Ratio even if the selection equation and the main equation have identical regressors. However, it is also the case that relying solely on nonlinearity is generally viewed as taking the low (and risky) road to identification. Manski [11, 12] points to the inherent problems for identification in a latent variable model with exclusion restrictions such as the Heckit model. Despite these serious issues, the Heckit technique is widely used because of its simplicity. The purpose of this methodological note is to raise yet another caution that pertains to decompositions with selectivity corrected wage equations. We demonstrate below that fundamental ambiguities with decomposition terms can arise even in a correctly identified Heckit sample selection model.

2. Methodology

For purposes of illustration we will consider gender wage differentials. A simple two equation model of wage determination and employment illustrates the application. Let the employment and wage functions for individual i in gender group j be given by

$$L_{ij}^{*} = H_{ij}'\gamma_j + \varepsilon_{ij}, \tag{1}$$

$$Y_{ij} = X_{ij}', \beta_j + u_{ij}, \tag{2}$$

where L_{ij}^{*} is a latent variable associated with being employed, H_{ij}', is a vector of determinants of employment, Y_{ij} is the market wage (in logs), X_{ij}' is a vector of determinants of market wages, γ_j and β_j are the associated parameter vectors, and ε_{ij} and u_{ij} are i.i.d. error terms that follow a bivariate normal distribution $(0, 0, \sigma_{\varepsilon_j}, \sigma_{u_j}, \rho_j)$.

The probability of employment is expressed as

$$\begin{aligned} \text{Prob}(L_{ij}^{*} > 0) &= \text{Prob}(\varepsilon_{ij} > -H_{ij}'\gamma_j) \\ &= \Phi(H_{ij}'\gamma_j), \end{aligned} \tag{3}$$

where $\Phi(\cdot)$ is the standard normal C.D.F. (the variance of ε_j is normalized to 1). Wages are observed for those for whom $L_{ij}^* > 0$, so that the expected wage of an employed individual is determined according to

$$\begin{aligned} E(Y_{ij} \mid L_{ij}^* > 0) &= X_{ij}'\beta_j + E(u_{ij} \mid \varepsilon_{ij} > -H_{ij}'\gamma_j) \\ &= X_{ij}'\beta_j + \theta_j \lambda_{ij}, \end{aligned} \tag{4}$$

where $\theta_j = \rho_j \sigma_{u_j}$, $\lambda_{ij} = \phi(H_{ij}'\gamma_j)/\Phi(H_{ij}'\gamma_j)$, and $\phi(\cdot)$ is the standard normal density function.[1] The estimating equation for employed individuals may be expressed as

$$Y_{ij} \mid L_{ij}^* > 0 = X_{ij}'\beta_j + \theta_j\lambda_{ij} + \text{error}. \tag{5}$$

Suppose one is interested in estimating wage discrimination between males and females in the presence of sample selectivity. For simplicity we will adopt the estimated male wage structure as the nondiscriminatory, competitive norm. The parameters of (5) would be estimated by the Heckman procedure separately for males and females.

It is clear from (5) that correction for selectivity bias requires a wage decomposition of the following sort:

$$\bar{Y}_m - \bar{Y}_f = \bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f) + (\bar{X}_m - \bar{X}_f)'\hat{\beta}_m + (\hat{\theta}_m\hat{\lambda}_m - \hat{\theta}_f\hat{\lambda}_f), \tag{6}$$

where $\bar{Y}$ is the predicted mean (log) wage, $\bar{X}'$ is the mean vector of wage determining variables (human capital variables), $\hat{\beta}$ is vector of the estimated returns to the wage determinants, $\hat{\theta}$ is an estimate of $\rho\sigma_u$, and $\hat{\lambda}$ is an estimate of the mean Inverse Mills Ratio (*IMR*). The first two terms in (6) are the familiar discrimination and human capital components. However, it is not immediately obvious how the last term in (6) should be regarded in the overall decomposition scheme. Should this term be subject to further decomposition into discrimination and human capital components, and if so, how should this be done? There is no simple answer to this question. Estimation of wage inequity in the presence of sample selectivity bias depends on assumptions as well as objectives as we show below.

One way to finesse the problem of what to do with the term $(\hat{\theta}_m\hat{\lambda}_m - \hat{\theta}_f\hat{\lambda}_f)$ is to simply net out the estimated differences in conditional means from the overall wage differential so that one is left with only the familiar decomposition terms:

$$(\bar{Y}_m - \bar{Y}_f) - (\hat{\theta}_m\hat{\lambda}_m - \hat{\theta}_f\hat{\lambda}_f) = \bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f) + (\bar{X}_m - \bar{X}_f)'\hat{\beta}_m. \tag{7}$$

Examples of this type of approach may be found in Duncan and Leigh [6], Reimers [17] and Boymond et al. [3].[2] While (7) is a decomposition of the selectivity corrected wage differential, it does not necessarily provide a decomposition of the *observed* wage differential $\bar{Y}_m - \bar{Y}_f$.

Our task is to estimate wage inequity from an appropriately defined wage decomposition that includes the effects of selectivity such as (6). One can pursue an

exact decomposition of the gender difference in the conditional mean error terms. Estimates of the contributions of human capital (endowments) and discrimination to the wage differential can be obtained from (6) in a number of alternative ways that derive from a decomposition of the gender difference in selectivity effects. In keeping with our adoption of the male (dominant group) wage structure as the standard, we introduce the following decomposition of the gender difference in the conditional mean error terms for the wage equations:

$$\begin{aligned}&\bar{E}(u_m \mid \varepsilon_m > -H_m'\hat{\gamma}_m) - \bar{E}(u_f \mid \varepsilon_f > -H_f'\hat{\gamma}_f)\\ &\quad = \hat{\theta}_m\hat{\lambda}_m - \hat{\theta}_f\hat{\lambda}_f = \hat{\theta}_m(\hat{\lambda}_f^0 - \hat{\lambda}_f) + \hat{\theta}_m(\hat{\lambda}_m - \hat{\lambda}_f^0) + (\hat{\theta}_m - \hat{\theta}_f)\hat{\lambda}_f, \qquad (8)\end{aligned}$$

where $\hat{\lambda}_f^0 = \sum_{i=1}^{N_{1f}} \hat{\lambda}_{if}^0/N_f$, and $\hat{\lambda}_{if}^0 = \phi(H_{if}'\hat{\gamma}_m)/\Phi(H_{if}'\hat{\gamma}_m)$. The term $\hat{\lambda}_f^0$ is the mean value of the *IMR* if females faced the same selection equation that the men face. The term $\hat{\theta}_m(\hat{\lambda}_f^0 - \hat{\lambda}_f)$ measures the effects of gender differences in the parameters of the probit selectivity equation on the male/female wage differential. The effects of gender differences in the variables that determine employment are measured by the term $\hat{\theta}_m(\hat{\lambda}_m - \hat{\lambda}_f^0)$. Finally, the effects of gender differences in the observed wage response to selection are captured by the term $(\hat{\theta}_m - \hat{\theta}_f)\hat{\lambda}_f$. Given that $\hat{\theta}_j = \hat{\rho}_j\hat{\sigma}_{u_j}$ and that the parameters $\hat{\rho}_j$ and $\hat{\sigma}_{u_j}$ are identified, further decomposition of $\hat{\theta}_m - \hat{\theta}_f$ is possible:

$$\hat{\theta}_m - \hat{\theta}_f = \hat{\rho}_m(\hat{\sigma}_{u_m} - \hat{\sigma}_{u_f}) + (\hat{\rho}_m - \hat{\rho}_f)\hat{\sigma}_{u_f} \qquad (9)$$

$$= (\hat{\rho}_m - \hat{\rho}_f)\hat{\sigma}_{u_m} + \hat{\rho}_f(\hat{\sigma}_{u_m} - \hat{\sigma}_{u_f}). \qquad (10)$$

The decompositions derived from (9) and (10) measure the effects of gender differences in wage error term variances and correlations between unobserved errors in the selection and wage equations. Decompositions (9) and (10) correspond to standardizing on the male correlation coefficient (female wage error variance) or on the female correlation coefficient (male wage error variance).

How should the components of (8) and (9) or (10) be allocated to discrimination and endowments? The most straight forward approach would be to use (6) and simply identify the overall selection component as a category apart from discrimination and endowment effects:

$$\bar{Y}_m - \bar{Y}_f = \underbrace{\bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f)}_{\text{discrimination}} + \underbrace{(\bar{X}_m - \bar{X}_f)'\hat{\beta}_m}_{\text{endowments}} + \underbrace{(\hat{\theta}_m\hat{\lambda}_m - \hat{\theta}_f\hat{\lambda}_f)}_{\text{selectivity}}. \qquad (11)$$

Reimers [17] uses a similar decomposition. Apart from the selectivity correction, the Reimers decomposition is a special case of the methodology presented in Oaxaca and Ransom [16] in which the nondiscriminatory wage structure is a weighted average of the separately estimated wage structures.

We will briefly explore alternative decompositions that could be considered but that require stronger assumptions and perhaps value judgements about what constitutes inequity. If one believed that gender differences in the probit selection

parameters for employment represented discrimination and that gender differences in personal attributes that determine the probability of employment are simply endowment differences, the resulting decomposition would be:

$$\begin{aligned}\bar{Y}_m - \bar{Y}_f \;=\; & \underbrace{\bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f) + \hat{\theta}_m(\hat{\lambda}_f^0 - \hat{\lambda}_f)}_{\text{discrimination}} + \\ & + \underbrace{(\bar{X}_m - \bar{X}_f)'\hat{\beta}_m + \hat{\theta}_m(\hat{\lambda}_m - \hat{\lambda}_f^0)}_{\text{endowments}} + \\ & + \underbrace{(\hat{\theta}_m - \hat{\theta}_f)\hat{\lambda}_f}_{\text{selectivity}}. \end{aligned} \tag{12}$$

A second alternative is to add the effects of gender differences in ρ to the estimated endowment (human capital) effects on the grounds that the gender difference in the error correlation coefficient is a justifiable structural source of gender wage gaps. It is difficult to know where to assign the wage gap effects of gender differences in the wage error variances. Differences in wage dispersion might or might not reflect direct labor market discrimination. Therefore, we include wage dispersion effects in the neutral category of selection effects. To illustrate, we standardize on the male wage error variance so that the overall wage decomposition becomes

$$\begin{aligned}\bar{Y}_m - \bar{Y}_f \;=\; & \underbrace{\bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f) + \hat{\theta}_m(\hat{\lambda}_f^0 - \hat{\lambda}_f)}_{\text{discrimination}} + \\ & + \underbrace{(\bar{X}_m - \bar{X}_f)'\hat{\beta}_m + \hat{\theta}_m(\hat{\lambda}_m - \hat{\lambda}_f^0) + (\hat{\rho}_m - \hat{\rho}_f)\hat{\sigma}_{u_f}}_{\text{endowments}} + \\ & + \underbrace{\hat{\rho}_m(\hat{\sigma}_{u_m} - \hat{\sigma}_{u_f})}_{\text{selectivity}}. \end{aligned} \tag{13}$$

The most encompassing (and least defensible) view of discrimination is to regard both gender differences in the estimated γ parameters from the probit selection equation for employment and gender differences in the wage effects of selectivity (θ) as manifestations of discrimination. Gender differences in the values of the employment determining variables (H') would continue to be treated as nondiscriminatory endowment effects. These assumptions lead to the following decomposition:

$$\begin{aligned}\bar{Y}_m - \bar{Y}_f \;=\; & \underbrace{\bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f) + \hat{\theta}_m(\hat{\lambda}_f^0 - \hat{\lambda}_f) + (\hat{\theta}_m - \hat{\theta}_f)\hat{\lambda}_f}_{\text{discrimination}} + \\ & + \underbrace{(\bar{X}_m - \bar{X}_f)'\hat{\beta}_m + \hat{\theta}_m(\hat{\lambda}_m - \hat{\lambda}_f^0)}_{\text{endowments}} \end{aligned}$$

$$= \underbrace{\bar{X}_f'(\hat{\beta}_m - \hat{\beta}_f) + \hat{\theta}_m\hat{\lambda}_f^0 - \hat{\theta}_f\hat{\lambda}_f}_{\text{discrimination}} +$$

$$+ \underbrace{(\bar{X}_m - \bar{X}_f)'\hat{\beta}_m + \hat{\theta}_m(\hat{\lambda}_m - \hat{\lambda}_f^0)}_{\text{endowments}}. \tag{14}$$

Decomposition (11) is noncommittal about the assignment of gender differences in the components of the selection effects to discrimination and endowment. On the other hand the decompositions expressed in (12), (13), and (14) involve varying degrees of assignment of selection effect decompositions to discrimination and endowment components. What these last three decompositions have in common is that they treat gender differences in the effects (γ) of employment determination variables as products of labor market discrimination. Unfortunately, an identification problem may now arise because gender differences in γ actually reflect gender differences in σ_ε. This problem exists because of the normalization of γ as the ratio of the original probit parameters to σ_ε which is equivalent to normalizing the nonidentified parameter σ_ε to 1. Should gender differences in σ_ε be counted as discriminatory?

3. Concluding remarks

Even from the limited number of selection decompositions explored in this paper, one can appreciate that different assignments to discrimination and endowments can potentially generate quite different estimates of labor market inequity. Some of these decompositions may induce identification problems even beyond those normally associated with the Heckit correction for sample selection. Although (14) is the most inclusive of the decompositions as far as measuring discrimination is concerned, it would not necessarily yield the largest estimate of discrimination. None of what has been presented here authoritatively identifies the "correct" decomposition. The determination of what really constitutes inequity rests upon opinions about which parameter differences constitute discrimination. While this issue could in principle apply to any of the β parameters, it is particularly relevant to the selection parameters ρ, σ_u, and γ. The choice of which selectivity corrected decomposition to use is largely judgmental because it inevitably reflects value judgments about what constitutes labor market inequity. Under what theoretical framework would group differences in the correlation parameters, the wage dispersion parameters, or the probit selection weights constitute labor market discrimination?

Possible extensions of our work include any decomposition that seeks to measure explained and unexplained influences. While the points raised here are presented in the context of the ever popular two-stage Heckit method, they in principle apply to other sample selection methods that are less restrictive in their assumptions.

Acknowledgements

We gratefully acknowledge the helpful comments of David Macpherson, participants at the Ph.D. workshop on the econometrics of labor market discrimination and at the Centre for Labor Market and Social Research seminar, University of Aarhus, and participants at the Conference on Labor Economics, Gstaad, Switzerland. Any remaining errors are the sole responsibility of the authors.

Notes

[1] While researchers euphemistically refer to the labor force participation selection outcome, what is actually used is either the employment/nonemployment or the employment/unemployment dichotomy. A more complicated model that could be considered is the double hurdle type model in which the labor force participation and subsequent employment/unemployment outcomes are taken into account (see [4]).

[2] Duncan and Leigh estimated separate selectivity-corrected wage equations for union and nonunion workers and presented estimates of the union/nonunion wage differential with and without the weighted difference in the mean values of the λ's for union workers and nonunion workers. Conceptually, their context differs from ours in that a single selection equation is estimated for endogenous union status while our application is conditional on gender status and involves estimation of separate selection equations for males and females.

References

1. Blau, F.D. and Kahn, L.M.: Swimming upstream: Trends in the gender wage differential in the 1980's, *J. Labor Economics* **15** (1997), 1–42.
2. Blinder, A.S.: Wage discrimination: Reduced form and structural estimates, *J. Human Resources* **8** (1973), 436–455.
3. Boymond, M., Flückiger, Y. and Silber, J.: Wage discrimination and occupational segregation by gender: Some evidence from Swiss data, mimeo, Geneva, 1994.
4. Cragg, J.: Some statistical models for limited dependent variables with applications to the demand for durable goods, *Econometrica* **39** (1971), 829–844.
5. Datta Gupta, N., Oaxaca, R.L. and Smith, N.: The Danish gender wage gap and wage determination in the private and public sectors, In: S. Gusstafson and D. Meulders (eds.), *Gender and the Labour Market: Econometric Evidence on Obstacles in Achieving Gender Equality*, AEA Macmillan Series, 2000.
6. Duncan, G.M. and Leigh, D.E.: Wage determination in the union and nonunion sectors: A sample selectivity approach, *Industrial and Labor Relations Review* **34** (1980), 24–34.
7. Gronau, R.: Wage comparisons: A selectivity bias, *J. Political Economy* **82** (1974), 1119–1143.
8. Heckman, J.: The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement* **5** (1976), 475–492.
9. Heckman, J.: Sample selection bias as a specification error, *Econometrica* **47** (1979), 153–163.
10. Juhn, C., Murphy, K.M. and Pierce, B.: Accounting for the slowdown in black-white wage convergence, In: M.H. Kosters (ed.), *Workers and Their Wages: Changing Patterns in the United States*, American Enterprise Institute Press, Washington, 1991.
11. Manski, C.: Anatomy of the selection problem, *J. Human Resources* **24** (1989), 343–360.
12. Manski, C.: *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge, MA, 1995.

13. Neumark, D.: Employers' discriminatory behavior and the estimation of wage discrimination, *J. Human Resources* **23** (1988), 279–295.
14. Oaxaca, R.L.: Male–female wage differentials in urban labor markets, *International Economic Review* **14** (1973), 693–709.
15. Oaxaca, R.L. and Ransom, M.: Searching for the effect of unionism on the wages of union and nonunion workers, *J. Labor Research* **9** (1988), 139–148.
16. Oaxaca, R.L. and Ransom, M.: On discrimination and the decomposition of wage differentials, *J. Econometrics* **61** (1994), 5–21.
17. Reimers, C.: Labor market discrimination against hispanic and black men, *The Review of Economics and Statistics* **65** (1983), 570–579.
18. Suen, W.: Decomposing wage residuals: Unmeasures skill or statistical artifact, *J. Labor Economics* **15** (1997), 555–566.
19. Vella, F.: Estimating models with sample selection bias: A survey, *J. Human Resources* **33** (1998), 127–169.