New Estimator for Wage Discrimination between Two Groups

Abstract: The conventional estimator of wage discrimination between two groups proposed by Oaxaca(1973) and Blinder(1973) has become the basis for the empirical research of market discrimination. However, its practical value is contaminated by the "index number problem". Some subsequent studies have attempted to find out the "actual" nondiscriminatory wage structure to solve the problem. This paper proposes an alternative nondiscriminatory wage structure based on the theoretical model proposed by Neumark (1988). We claim that the proposed structure could correct the estimate bias brought about by the Neumark's method. It demonstrates that the constant coefficient in the nondiscriminatory wage structure dose not have an influence on the overall wage discrimination, which helps to increase the new estimator's practical value. Besides, the differences in estimations between the new estimator and the previous ones are highly correlated with sample features. These estimators were then compared with an application to male-female wage differentials by UHS2007 dataset in order to provide some empirical evidence.

Keywords: wage differentials, wage discrimination, index number problem, nondiscriminatory wage structure

Ⅰ INTRODUCTION

The concept of the discrimination coefficient creatively proposed by Becker(1957) provides a quantitative assessment of the discriminative components in wage differentials and establishes the foundation for the mean decomposition methods[①]. In their seminal work on wage discrimination assessment between two labor groups, developed on the basis of Becker's discrimination coefficient, Oaxaca(1973) and Blinder(1973) estimated the standard Mincerian log wage equations separately for two labor groups, and regarded any of the both vectors of regression coefficients as the index number (nondiscriminatory wage structure) to estimate the wage discrimination by their decomposition methods. The decomposition results demonstrate that the discrimination estimated by the two different index number have a large gap, and the criterion used to determine which one is more suitable under different conditions is not suggested, i.e., the decomposition results could be quite sensitive to the used wage structure, but neither is preferable to the other a priori. Ferber and Green(1982) carried out a study on the pay differentials of the professors in a U.S. university, and found a 2% wage differentials due to discrimination using the male wage structure but a 70% differentials using the female wage structure. Cotton(1988) studied the wage differentials between the white and the black, and found a 48.5% and 95.4% wage differentials due to discrimination by respectively using the white and the black wage structures. These are the so-called "index number problem". Accordingly, the market discrimination estimated by the Oaxaca-Blinder's estimator is ambiguous which suffers from the "index number problem".

The key point of solving the index number problem is to construct a nondiscriminatory wage structure which can be measured by the observed data. Reimers(1983) made an attempt and proposed a weighted average pattern. Cotton(1988) and Neumark(1988) separately proposed a nondiscriminatory wage structure from different perspectives, the estimators of wage discrimination can be then obtained by decomposition. Although large numbers of the empirical researches on wage differentials have been done by using Oaxaca-Blinder, Cotton or Neumark's method, very few researches have paid attention to the differences among these methods and the limitations among them.

The existing literature on the wage differentials decomposition methods between two groups could approximately be divided into two categories: one is concerned with the average value decomposition of the wage differences between two groups. Oaxaca(1973), Blinder(1973), Cotton(1988), Neumark(1988), Brown(1980) and Appleton(1999) belong to this category. Juhn, et al.(1991) took account of the error distribution of the wage regression equation in order to explore the unobserved component; Neuman and Oaxaca(2004) analyzed the sample selection problem in the average value decomposition, and Démurger et al.(2007) brought forward the dynamic decomposition of the average wage differentials. These studies can also be included in this category. The other is distributional decomposition which extends the wage differentials in the

[①] Becker(1957) defined the market discrimination coefficient as the simple difference between the observed wage ratio and the wage ratio that would prevail in the absence of discrimination, where the wage ratio refers to the relative wage differential between two labor groups (for example, between the females and males, the white and black, etc.). Apparently, the wage ratio in the absence of discrimination is the key point to analyze discrimination quantitatively. According to the nondiscriminatory wage structure, the wage differentials can be decomposed into the portion due to differences in individual characteristics between two labor groups and the portion attributable to discrimination.

average value to distribution (Juhn, et al., 1993; DiNardo, et al., 1996; Machado and Mata, 2005; Autor, 2005; Firpo, et al. 2007). The former mainly focuses on the estimator of wage discrimination between two groups based on the nondiscriminatory wage structure, and the latter emphasizes on the process of constructing a counterfactual wage distribution. In essence, the nondiscriminatory wage structure is also a counterfactual state. This paper aims to propose a new method to construct the nondiscriminatory wage structure and then bring forward a new estimator of wage discrimination. As a result, it just takes the decomposition methods for mean wage differentials into account.

Currently, although it is commonly understood that market discrimination in China is serious, where gender and Hukou System (people are classified into urban residents and rural residents by Hukou system) discrimination are more universal in China's labor market, opinions about the degree of the discrimination vary among different researchers. In terms of the empirical researches, a majority of the existing studies in China focus on Oaxaca-Blinder decomposition (Gustafsson and Li，2000；Meiyan Wang，2003；Xianguo Yao and Puqing Lai，2004；Quheng Deng，2007). Several researches combined Oaxaca-Blinder and Cotton decomposition, and then make a comparison between the two methods (Dandan Zhang, 2004; Sisheng Xie and Xianguo Yao, 2006). As far as we known, few studies used Neumark decomposition to estimate wage discrimination. Since the results from the various methods are always distinct and the differences in these results caused by the differences in the nondiscriminatory wage structure are sensitive to data, the ambiguity of the wage discrimination estimation could not be avoided. As is well known, the biased estimation of the labor market discrimination would be helpless for the government in making social and economic policies.

This paper attempts to bring forward a new method to construct the nondiscriminatory wage structure, based on characterizing and examining the limitations of the previous methods. The remaining is organized as follows. Section 2 begins with a brief overview of the decomposition methods for mean wage differentials with its emphasis on the nondiscriminatory wage structure. Section 3 is concerned with the theoretical model of the new nondiscriminatory wage structure. Section 4 presents the empirical study by using the urban China household survey data set in 2007, and then makes a comparison among these estimators. A concluding comment is contained in the final section.

Ⅱ Overview of Nondiscriminatory Wage Structure

Oaxaca-Blinder decomposition method introduced by Oaxaca(1973) and Blinder(1973) for the mean wage differentials, estimates wage discrimination by determining how much of the wage differentials between two groups due to the differences in the regression coefficients of respectively estimated wage equations by two sub-samples (hereafter, the two sub-samples refer to the two groups: a high-wage group and a low-wage group, respectively marked as H and L). It provides two estimators of wage discrimination, and accordingly yields two results[①]. Obviously, the two estimator of the wage discrimination depend on the selection of a nondiscriminatory wage structure ($\hat{\beta}_H$ and $\hat{\beta}_L$), which implies the aforementioned "index number problem". The choice of the index number is of great significance to estimate wage discrimination. Naturally how to solve the "index number problem" becomes the hot issue in estimating the wage discrimination.

Reimers(1983), Cotton(1988) and Neumark(1988), each of them had an attempt to solve the "index number problem". Reimers regarded the nondiscriminatory wage structure as the vector containing the weighted average of the coefficients of separately estimated wage equations by the two groups (hereafter we call the coefficients of the estimated wage equation as the wage structure), and it can be written as $\hat{\beta}_R^* = D\hat{\beta}_H + (I - D)\hat{\beta}_L$. Where $\hat{\beta}_R^*$ is the wage structure in the absence of discrimination conceived by Reimers, I is a unit matrix, D is an arbitrary diagonal matrix, $\hat{\beta}_H$ and $\hat{\beta}_L$ are the observed wage structures (also called as current wage structure) estimated by H and L sub-samples. Because of the arbitrary matrix D, Reimers's nondiscriminatory wage structure produces multiple forms in theory, so the "index number problem" is still in existence. In the empirical study, Reimers selected $D = diag(0.5, 0.5, \cdots, 0.5) = 0.5(I)$ to construct $\hat{\beta}_R^*$, but Remers said nothing with regard to the reason for the selection.

---

① According to the aforementioned, the high-wage group and low-wage group are respectively noted as group H and group L. Firstly, the wage equations of the two groups can be separately written as: $\ln w_H = X_H\beta_H + u_H$, $\ln w_L = X_L\beta_L + u_L$, where β_H and β_L are the vectors of coefficients; X_H and X_L are the matrices of individual characteristics pertain to group H and group L. The mean wage differentials in total between H and L group is given as: $\ln \overline{w}_H - \ln \overline{w}_L = \overline{X}'_H\hat{\beta}_H - \overline{X}'_L\hat{\beta}_L$, where $\hat{\beta}_H$ and $\hat{\beta}_L$ are the estimated vectors of coefficients; $\overline{X}_H$ and $\overline{X}_L$ are vectors containing the mean of the individual characteristics' variables for H and L group. Secondly, assuming that $\hat{\beta}_H$ is the wage structure which prevails in the absence of discrimination, the decomposition equation can be written as: $\ln \overline{w}_H - \ln \overline{w}_L = (\overline{X}'_H - \overline{X}'_L)\hat{\beta}_H + \overline{X}'_L(\hat{\beta}_H - \hat{\beta}_L)$, where the first term $\overline{X}'_L(\hat{\beta}_H - \hat{\beta}_L)$ is the part of the log wage differentials attributable to discrimination. Given that $\hat{\beta}_L$ is the nondiscriminatory wage structure, the decomposition equation is described as: $\ln \overline{w}_H - \ln \overline{w}_L = (\overline{X}'_H - \overline{X}'_L)\hat{\beta}_L + \overline{X}'_H(\hat{\beta}_H - \hat{\beta}_L)$, where the first term $\overline{X}'_H(\hat{\beta}_H - \hat{\beta}_L)$ can be interpreted as the part due to discrimination. Obviously, the estimator of the wage discrimination depends on the selection of a nondiscriminatory wage structure ($\hat{\beta}_H$ and $\hat{\beta}_L$).

Cotton(1988) found that Oaxaca-Blinder nondiscriminatory wage structure was derived by the assumption of employers' discriminatory tastes that employers impose neither pure discrimination against group L nor pure nepotism towards group H. Instead, Cotton supposed that not only was the group L discriminated against undervalued, but the group H preferred was overvalued. Given that, Cotton concluded that the observed wage structure of the group H and L were both functions of discrimination, i.e., each wage structure was contaminated by discrimination, which implied that Oaxaca-Blinder nondiscriminatory wage structure was problematic. An alternative weighted average of the group H and L wage structures with a specific weighting matrix conditional on that in the absence of discrimination the prevailing wage structure will be some function of the forces currently determined the group H and L wage structures, as well as given that the nondiscriminatory wage structure will be closer to the majority rather than the minority, was then introduced by Cotton. With the simplified assumption that in the absence of discrimination, the prevailing market wage structure would be a linear function of $\hat{\beta}_H$ and $\hat{\beta}_L$, Cotton's nondiscriminatory wage structure can be written as $\hat{\beta}_C^* = f_H\hat{\beta}_H + f_L\hat{\beta}_L$, where f_H and f_L are respectively the proportion of group H and group L in labor force. Apparently, $\hat{\beta}_C^*$ is just a special case of $\hat{\beta}_R^*$ by letting $D = diag(f_H, f_H, \cdots, f_H) = f_H I$. If the index number problem is completely interpreted as the inconclusive results aroused by the multiple index numbers, Cotton's method could be a solution for the "index number problem". It is also shown that, Cotton's index number is likely to be accepted more widely than Reimer's and Oaxaca-Blinder's since the reality of market discrimination is taken into account by the former.

Given a nondiscriminatory wage structure $\hat{\beta}^*$, the average wage differentials can be decomposed as follows:

$$\ln \bar{w}_H - \ln \bar{w}_L = (\bar{X}_H' - \bar{X}_L')\hat{\beta}^* + \bar{X}_H'(\hat{\beta}_H - \hat{\beta}^*) + \bar{X}_L'(\hat{\beta}^* - \hat{\beta}_L) \qquad [1]$$

The first term on the right-hand side is the portion of wage differentials due to differences in endowments (individual characteristics) between group H and L, which is interpreted as nondiscriminatory part. The remainder is attributable to differences in returns to those endowments regarded as the overall discrimination, where the first term of the overall discrimination is overvalued part due to employers' nepotism towards group H, the second one is undervalued due to employers' discrimination against group L.

Using the two discrimination estimators of Oaxaca-Blinder decomposition respectively by regarding $\hat{\beta}_H$ and $\hat{\beta}_L$ as the index number to establish the range, it is shown that Cotton's discrimination estimator will fall within the range.

$$f_L\bar{X}_H'(\hat{\beta}_H - \hat{\beta}_L) + f_H\bar{X}_L'(\hat{\beta}_H - \hat{\beta}_L) \in \left(\bar{X}_L'(\hat{\beta}_H - \hat{\beta}_L), \bar{X}_H'(\hat{\beta}_H - \hat{\beta}_L)\right) \qquad [2]$$

Neumark(1988) proposed a theoretical model for constructing nondiscriminatory wage structure based on employers' discriminatory tastes, which is an extension of the employers' discrimination model of Arrow(1972) and Becker(1957). Neumark's theoretical framework of the nondiscriminatory wage structure is capable of handling aforementioned Oaxaca-Blinder's and Cotton's. Given the assumption that within each type of labor, the utility function capturing employers' discriminatory tastes is homogeneous of degree zero with respect to labor inputs from each of the two groups, it turns out that wage structure in the absence of discrimination is the coefficient vector of the wage regression equation over the pooled sample (that is a sample combined by both H and L group sub-samples). It is then noted as $\hat{\beta}_N^* = (X'X)^{-1}X'Y$. Oaxaca and Ransom(1994) demonstrated that Neumark's "pooled" wage structure can also be transformed into a weighted average written as $\hat{\beta}_N^* = \Omega\hat{\beta}_H + [I - \Omega]\hat{\beta}_L$, and the weighting matrix $\Omega = (X'X)^{-1}X_H'X_H$, Where X is the matrix of individual characteristics for the pooled sample and X_H is the corresponding matrix for group H only. It is clearly seen that Neumark's nondiscriminatory wage structure is similar to Cotton's in terms of weighting form, whereas Neumark's seem to be slightly better than Cotton's shown as follows: (1) in terms of weighting matrix, Neumark's $\Omega = (X'X)^{-1}X_H'X_H$ covers more information than Cotton's $D = f_H I$, because the former not only includes the structure feature of group H and L, but implies the distributional information of individual characteristics. The more sample information is used in empirical study, the better results would be produced. (2) in terms of the estimators, Cotton's estimator is within the range from $\bar{X}_L'(\hat{\beta}_H - \hat{\beta}_L)$ to $\bar{X}_H'(\hat{\beta}_H - \hat{\beta}_L)$, while Neumark's estimator is not absolutely within the range. According to equation [1], Letting $\hat{\beta}^* = \hat{\beta}_N^*$, Neumark's estimator of wage discrimination can be then written as $[\bar{X}_H'(X'X)^{-1}X_L'X_L + \bar{X}_L'(X'X)^{-1}X_H'X_H](\hat{\beta}_H - \hat{\beta}_L)$. Whether it is within the range or not depends on data itself[①]. The range established by the two estimators of Oaxaca-Blinder decomposition appears to be impossible to include the true discrimination estimation for all time.

Although Neumark's estimator is more attractive than these alternatives, its restrictions should be taken into account cautiously. It is known that Neumark's theoretical model is derived conditional on some strict assumptions and few evidences could demonstrate that the zero-homogeneity restriction on employers is a valid hypothesis. Besides, Neumark's theoretical model still doesn't take the effect of discrimination on labor force supply into consideration. Thirdly, even though the theoretical framework is accepted by the realistic labor market, Neumark's "pooled" wage structure would be confronted with endogenous problem without controlling group dummy variable in the wage regression equation over the pooled sample because it is always impossible to

① In fact, an empirical study carried out by Neumark(1988) presented that discrimination explained 57% of wage differentials, the corresponding figures were 69% and 70% by both estimators of Oaxaca-Blinder. It is clearly shown that Neumark's estimate is lower than either estimate of Oaxaca-Blinder.

control all group-specific characteristics. Thus the purpose of this paper is to present a new estimator of the nondiscriminatory wage structure based on Neumark's estimator.

Ⅲ A New Nondiscriminatory Wage Structure

According to Mincerian equation, the regression model of natural log wage can be expressed as:

$$\ln w = \beta_0 + X_1\beta_1 + \varepsilon = X\beta + \varepsilon \qquad [3]$$

Where β_0 is a constant coefficient, β_1 is a vector containing regression coefficients except intercept and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$; X_1 is a matrix containing individual characteristics and $X = (I, X_1)$, in which I is a unit vector denoting constant; $\ln w$ denotes the natural log wage.

Neumark's nondiscriminatory wage structure is the "pooled" wage structure, namely the estimated vector of regression coefficients from [3] with *X* just containing individual perceived endowments. From the viewpoint of model specification in practice, it is well known that conventional earning functions are likely to be miss-specified, omitting a number of unobserved but important variables that influence productivity. Cotton(1988) claimed that these omitted variables, e.g., school quality or family background, may be themselves the results of past discrimination. Neumark's nondiscriminatory wage structure is subjected to criticism because Neumark's regression equation used to estimate the nondiscriminatory wage structure suggests serious endogenous problem induced by ignoring group-specific characteristics that are unobserved but significantly influence productivity, i.e., the pooled regression coefficients are biased by discrimination.

From the viewpoint of the theoretical framework, Neumark's estimator is derived by some rather strong assumptions, especially the core hypothesis that within each type of labor the utility function capturing employers' discrimination tastes is homogenous of degree zero with respect to labor inputs from each of group H and L (in Neumark's paper, group H and L refer to the male and female group respectively). Given that within each type of labor the portion due to employers' nepotism toward group H is equal to that due to employers' discrimination against group L, viz., $M_j \cdot (w_{Mj} - f_j) = F_j \cdot (f_j - w_{Fj})$, where j indicates the type of labor force, M_j and F_j separately denote the inputs of male and female workers of each type of labor (respectively the amount of male and female workers of each type of labor); $(w_{Mj} - f_j)$ and $(f_j - w_{Fj})$ separately represent portion attributable to employers' nepotism toward males (overvalued part) and that due to employers' discrimination against females (undervalued part) of each type of labor.

Conditional on the above assumption, it is concluded that discrimination leads to the redistribution of the total amount of wage between group H and L within each type of labor, but has no effect on the actual output of the labor market, namely, discrimination makes no impact on the aggregate wage of the two groups within each type of labor. Appleton et al.(1999) declared that there is no evidence that the zero-homogeneity restriction on employer preferences is valid. The realistic labor market also seems to be suspicious of this assumption. Employers seem to discriminate against the group L rather than to be nepotism towards group H. Besides, discrimination would

affect individual characteristics attainment because of individual self-selection. For example, if females discover that discrimination is more serious in high-skilled labor market than that in low-skilled one, which discourages the female to improve skill level, and indirectly influences the productivity of the whole labor market. Butler(1982) and Reimers(1983) made some insightful comments on it, and concluded that discrimination would result in labor self-selection.

Generally, discrimination should not only affect wage redistribution between groups within each type of labor, but also influence the total amount of wage. It is convinced again that discrimination should induce biased estimate of Neumark's nondiscriminatory wage structure. All these mentioned above imply that Neumark's nondiscriminatory wage structure marked as $\hat{\beta}_N^* = \begin{pmatrix} \hat{\beta}_{0N} \\ \hat{\beta}_{1N} \end{pmatrix}$ is a biased estimator, which then leads to a biased estimate of wage discrimination.

Although Neumark's estimator appears to be biased unless the strict assumption is satisfied and the regression model is completely specified, the theoretical model proposed by Neumark(1988) builds a bridge between the nondiscriminatory wage structure and the "pooled" wage structure. This paper tries to correct the bias of Neumark's estimator based on the bridge.

In order to correct the bias, a group indicator G, used to control the effect of unobserved discrimination on the current labor market[①], enters the regression equation of log wage expressed as:

$$\ln w = \beta_0 + \delta G + X_1\beta_1 + u = \delta G + X\beta + u \qquad [4]$$

The new nondiscriminatory wage structure is expressed as $\hat{\beta}_G^* = \begin{pmatrix} \hat{\beta}_{0G} \\ \hat{\beta}_{1G} \end{pmatrix}$. $\hat{\beta}_{1N}$ and $\hat{\beta}_{1G}$ separately denote Neumark's and this paper's nondiscriminatory wage structure excluding intercept, respectively estimated by equation [3] and [4].

The two errors of equation [3] and [4] are correlated with each other: $\varepsilon = \delta G + u$, and it is well known that the relation between $\hat{\beta}_{1N}$ and $\hat{\beta}_{1G}$ can be given as: $\hat{\beta}_{1N} = \hat{\beta}_{1G} + \delta \dfrac{Cov(X_1, G)}{Var(X_1)}$.

① Oaxaca and Ransom(1994) measured the part of wage differences due to discrimination by adding a group indicator variable (a dummy variable), and directly regarded the coefficient of the indicator variable as discrimination. It is clearly seen that the estimator proposed by Oaxaca and Ransom(1994) ignores the effect of discrimination on returns to individual characteristics. Fortin(2008) argued that a better alternative to Neumark(1988) and Oaxaca and Ransom(1994) solution to the choice of nondiscriminatory wage structure is to include group-indicator variable. However, Fortin(2008) does not revise the constant coefficient and so that the "index number problem" still exists.

Let $\hat{\beta}_H$ and $\hat{\beta}_L$ represent vectors of coefficients estimated for group H and L respectively with $\hat{\beta}_H = \begin{pmatrix} \hat{\beta}_{0H} \\ \hat{\beta}_{1H} \end{pmatrix}, \hat{\beta}_L = \begin{pmatrix} \hat{\beta}_{0L} \\ \hat{\beta}_{1L} \end{pmatrix}$, where $\hat{\beta}_{0H}$ and $\hat{\beta}_{0L}$ are the constant coefficients; $\hat{\beta}_{1H}$ and $\hat{\beta}_{1L}$ are vectors containing coefficients of independent variables. X_{1j} is the vector of the observed characteristics by group j. The current wage equations for group H and L are given as:

$$\ln w_j = \beta_{0j} + X_{1j}\beta_{1j} + u_j, j = H, L \quad [5]$$

Thus Neumark's estimator of the explained and unexplained part (discrimination part) of wage differentials could be expressed by the new estimator.

$$\begin{aligned} D_E &= (\overline{X}_{1H} - \overline{X}_{1L})\hat{\beta}_{1N} = (\overline{X}_{1H} - \overline{X}_{1L})(\hat{\beta}_{1G} + \delta\frac{Cov(X_1, G)}{Var(X_1)}) \\ D_{UE} &= \overline{X}_{1H}(\hat{\beta}_{1H} - \hat{\beta}_{1N}) + \overline{X}_{1L}(\hat{\beta}_{1N} - \hat{\beta}_{1L}) \\ &= \overline{X}_{1H}(\hat{\beta}_{1H} - \hat{\beta}_{1G} - \delta\frac{Cov(X_1, G)}{Var(X_1)}) + \overline{X}_{1L}(\hat{\beta}_{1G} + \delta\frac{Cov(X_1, G)}{Var(X_1)} - \hat{\beta}_{1L}) \end{aligned} \quad [6]$$

Equation [6] clearly illustrates the differences in decomposition result between Neumark's and ours.

Letting the dummy variable G indicate group H, and then the regressive equation [4] is written as $\ln w = \hat{\beta}_0 + \hat{\delta}_H G + X_1\hat{\beta}_1$, where $\hat{\delta}_H$ represents the effect of discrimination on wage, and therefore the wage equation in the absence of discrimination should be expressed as $\ln w = \hat{\beta}_0 + X_1\hat{\beta}_1$. From the other point of view, now letting the dummy variable G indicate group L, and then the estimate of equation [4] is $\ln w = \hat{\alpha}_0 + \hat{\delta}_L G + X_1\hat{\beta}_1$, where $\hat{\delta}_L$ denotes discrimination impact, and consequently the nondiscriminatory wage equation should be described as $\ln w = \hat{\alpha}_0 + X_1\hat{\beta}_1$ with $\hat{\alpha}_0 + \hat{\delta}_L = \hat{\beta}_0, \hat{\alpha}_0 = \hat{\delta}_H + \hat{\beta}_0$ and $\hat{\delta}_L = -\hat{\delta}_H$.

It is clearly shown that there are two constant coefficients of nondiscriminatory wage structure with $\hat{\alpha}_0 \neq \hat{\beta}_0$. Combining the two aforementioned conditions, in line with Cotton's assumption that the wage structure in the absence of discrimination is closer to the majority than to the minority, constant coefficient is thus expressed as:

$$\hat{\beta}^*_{0G} = f_L\hat{\beta}_0 + f_H\hat{\alpha}_0 \quad [7]$$

Where f_H and f_L denote the percent of group H and group L in labor market respectively.

If plugging $\hat{\alpha}_0 = \hat{\delta}_H + \hat{\beta}_0$ into [7] and letting $\hat{\delta}_H = \hat{\delta}$, then $\hat{\beta}_{0G}$ becomes $\hat{\beta}_{0G} = \hat{\beta}_0 + f_H \hat{\delta}$.

If plugging $\hat{\beta}_0 = \hat{\delta}_L + \hat{\alpha}_0$ into [7] and letting $\hat{\delta}_L = -\hat{\delta}$, then $\hat{\beta}_{0G}$ becomes $\hat{\beta}_{0G} = \hat{\alpha}_0 - f_L \hat{\delta}$.

Accordingly, the nondiscriminatory wage structure is given as with G indicating H group:

$$\hat{\beta}_G^* = \begin{pmatrix} \hat{\beta}_0 + f_H \hat{\delta} \\ \hat{\beta}_1 \end{pmatrix}; \qquad [8]$$

When G indicates L group, the nondiscriminatory wage structure becomes $\hat{\beta}_G^* = \begin{pmatrix} \hat{\alpha}_0 - f_L \hat{\delta} \\ \hat{\beta}_1 \end{pmatrix}$.

The wage equations prevailing in the absence of discrimination with the aforementioned two patterns of nondiscriminatory wage structure can then be written as $\ln w = (\hat{\beta}_0 + f_H \hat{\delta}) + X_1 \hat{\beta}_1$ and $\ln w = (\hat{\alpha}_0 - f_L \hat{\delta}) + X_1 \hat{\beta}_1$. Figure 1 provides the curve of log wage equation with our nondiscriminatory wage structure.

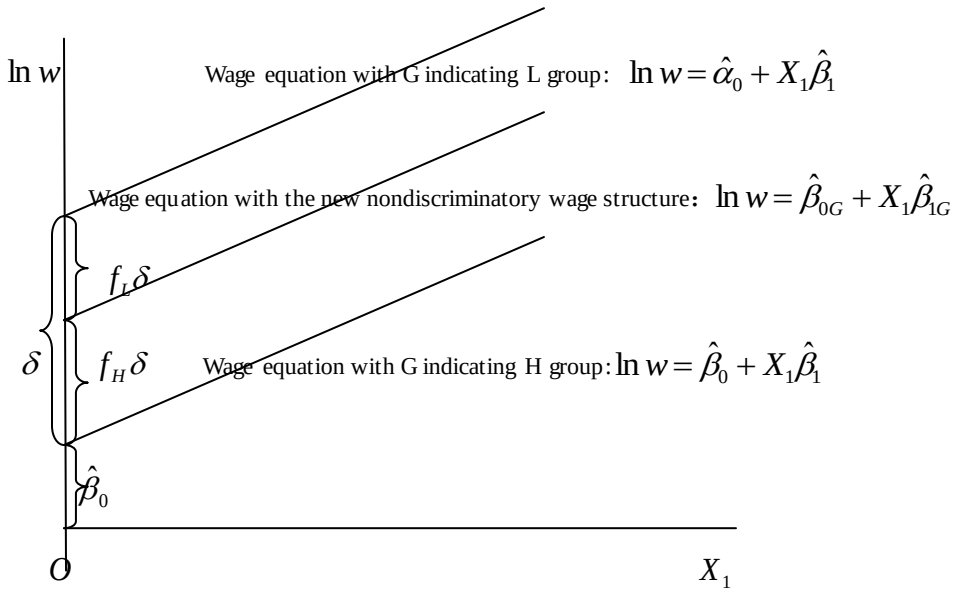

Figure 1: wage equation with $\hat{\beta}_G^* = \begin{pmatrix} \hat{\beta}_{0G} \\ \hat{\beta}_{1G} \end{pmatrix}$

Two propositions are then derived from the nondiscriminatory wage structure and the comparison

between the new one and the others (hereafter taking the fist pattern for instance).

Proposition 1: Constant coefficient in the nondiscriminatory wage structure has no impact on the overall component of wage differentials due to discrimination. However, it influences the composition of overvalued and undervalued portions.

Proof: take $\hat{\beta}^* = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ as the nondiscriminatory wage structure, according to [5], the current mean log wage of group H and L can be expressed as:

$$\ln \overline{w}_H = \hat{\beta}_{0H} + \overline{X}_{1H}^{'}\hat{\beta}_{1H} = \begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0H} \\ \hat{\beta}_{1H} \end{pmatrix}, \ln \overline{w}_L = \hat{\beta}_{0L} + \overline{X}_{1L}^{'}\hat{\beta}_{1L} = \begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0L} \\ \hat{\beta}_{1L} \end{pmatrix}.$$

Thus, group H-L wage differentials can be decomposed by equation [1].

$$\ln \overline{w}_H - \ln \overline{w}_L = \begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0H} \\ \hat{\beta}_{1H} \end{pmatrix} - \begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0L} \\ \hat{\beta}_{1L} \end{pmatrix} = \left(\begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'} - \begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'} \right)\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

$$+ \begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'}\left(\begin{pmatrix} \hat{\beta}_{0H} \\ \hat{\beta}_{1H} \end{pmatrix} - \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \right) + \begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'}\left(\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} - \begin{pmatrix} \hat{\beta}_{0L} \\ \hat{\beta}_{1L} \end{pmatrix} \right)$$

$$= \begin{pmatrix} 0 \\ \overline{X}_{1H} - \overline{X}_{1L} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} + \begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0H} - \hat{\beta}_0 \\ \hat{\beta}_{1H} - \hat{\beta}_1 \end{pmatrix} + \begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_0 - \hat{\beta}_{0L} \\ \hat{\beta}_1 - \hat{\beta}_{1L} \end{pmatrix}$$

$$= (\overline{X}_{1H}^{'} - \overline{X}_{1L}^{'})\hat{\beta}_1 + (\hat{\beta}_{0H} - \hat{\beta}_{0L}) + \overline{X}_{1H}^{'}(\hat{\beta}_{1H} - \hat{\beta}_1) + \overline{X}_{1L}^{'}(\hat{\beta}_1 - \hat{\beta}_{1L}) \qquad [9]$$

Where, the portion due to differences in observed characteristics is $(\overline{X}_{1H}^{'} - \overline{X}_{1L}^{'})\hat{\beta}_1$, and $(\hat{\beta}_{0H} - \hat{\beta}_{0L}) + \overline{X}_{1H}^{'}(\hat{\beta}_{1H} - \hat{\beta}_1) + \overline{X}_{1L}^{'}(\hat{\beta}_1 - \hat{\beta}_{1L})$ is the portion attributable to discrimination in total.

Equation [9] shows that the constant coefficient of the nondiscriminatory wage structure is not required in calculating the overall discrimination. However, it affects the further decomposition of the overall discrimination, which is composed of overvalued and undervalued part. In light of the decomposition equation [1], the overvalued part to group H is $\begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0H} - \hat{\beta}_0 \\ \hat{\beta}_{1H} - \hat{\beta}_1 \end{pmatrix} = (\hat{\beta}_{0H} - \hat{\beta}_0) + \overline{X}_{1H}^{'}(\hat{\beta}_{1H} - \hat{\beta}_1)$, and the undervalued part to group L is $\begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_0 - \hat{\beta}_{0L} \\ \hat{\beta}_1 - \hat{\beta}_{1L} \end{pmatrix} = (\hat{\beta}_0 - \hat{\beta}_{0L}) + \overline{X}_{1L}^{'}(\hat{\beta}_{1L} - \hat{\beta}_1)$. The former should be regarded as employers' nepotism toward to group H or reverse discrimination toward group L; and the latter

should be called as direct discrimination against group L.

Accordingly, the constant coefficient in our nondiscriminatory wage structure with $\hat{\beta}_{0G}^{*} = \hat{\beta}_0 + f_H\hat{\delta}$ is lack of theoretical base, because assumption of Cotton(1988) indicates that nondiscriminatory wage structure is closer to the majority than to the minority. Whereas, Proposition 1 implies that the choice of weights in estimating constant coefficient has no influence on the overall discrimination, which alleviates the contamination of imperfect estimate of constant coefficient to the whole nondiscriminatory wage structure.

Proposition 2: The nondiscriminatory wage structures proposed by Reimer(1983), Cotton(1988) and Neumark(1988) can be written as a weighted average of $\hat{\beta}_H$ and $\hat{\beta}_L$. In essence, the new nondiscriminatory wage structure should also be expressed as a function of $\hat{\beta}_H$ and $\hat{\beta}_L$. However, it should not be given as a simple weighted average of $\hat{\beta}_H$ and $\hat{\beta}_L$. Among these nondiscriminatory wage structures, only Reimer's and Cotton's should lie between the wage structure of group H and L. Oaxaca-Blinder decomposition provides two estimators for the discrimination component respectively taking $\hat{\beta}_H$ and $\hat{\beta}_L$ as the nondiscriminatory wage structures, which establish a range from $X_L^{'}(\hat{\beta}_H - \hat{\beta}_L)$ and $X_H^{'}(\hat{\beta}_H - \hat{\beta}_L)$. Similarly, Reimer's and Cotton's estimator for wage discrimination should be within the range of $[X_L^{'}(\hat{\beta}_H - \hat{\beta}_L), X_H^{'}(\hat{\beta}_H - \hat{\beta}_L)]$, while whether the estimators of Neumark and this paper are within the range or not is highly correlated with sample characteristics.

Proof: Take $I_h = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{h\times 1}$, $I_l = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{l\times 1}$, where h and l separately represent sub-sample observations of group H and L, then let dependent variable $Y = \begin{pmatrix} \ln w_H \\ \ln w_L \end{pmatrix}$. Since estimated wage of group H and L can be measured by equation [5], thus $Y = \begin{pmatrix} \ln w_H \\ \ln w_L \end{pmatrix} = \begin{pmatrix} X_H\hat{\beta}_H \\ X_L\hat{\beta}_L \end{pmatrix}$. For the sake of simplicity, equation [4] is converted into $\ln w = \delta G + \beta_0 + X_1\beta_1 + u$ where G indicates group H, and thus

$$\begin{pmatrix}\hat{\delta}\\ \hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}=(X'X)^{-1}X'Y=\left(X^{'}X\right)^{-1}\begin{pmatrix}I_h^{'}X_H\hat{\beta}_H\\ I_h^{'}X_H\hat{\beta}_H+I_l^{'}X_L\hat{\beta}_L\\ X_{1H}^{'}X_H\hat{\beta}_H+X_{1L}^{'}X_L\hat{\beta}_L\end{pmatrix} \quad [10]$$

With $\mathrm{X}=\begin{pmatrix}I_h & I_h & X_{1H}\\ 0 & I_l & X_{1L}\end{pmatrix}$. Equation [10] shows that $\hat{\beta}_G^*=\begin{pmatrix}\hat{\beta}_0+f_H\hat{\delta}\\ \hat{\beta}_1\end{pmatrix}$ should also be regarded as a function of $\hat{\beta}_H$ and $\hat{\beta}_L$. Apparently, constant coefficient has no influence on explained portion.

So, let $\hat{\beta}_2=\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$ and $X=\begin{pmatrix}I_h & X_H\\ 0 & X_L\end{pmatrix}$, thus

$$\begin{pmatrix}\hat{\delta}\\ \hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}=\begin{pmatrix}\hat{\delta}\\ \hat{\beta}_2\end{pmatrix}=(X^{'}X)^{-1}X^{'}Y=\begin{pmatrix}h & h\overline{X}_H^{'}\\ h\overline{X}_H & X^{'}X\end{pmatrix}^{-1}\begin{pmatrix}h\overline{X}_H^{'}\hat{\beta}_H\\ X_H^{'}X_H\hat{\beta}_H+X_L^{'}X_L\hat{\beta}_L\end{pmatrix}$$

By basic matrix operation, $\hat{\beta}_2$ is obtained as the following equation.

$$\hat{\beta}_2=(X^{'}X-h\overline{X}_H\overline{X}_H^{'})^{-1}(X_H^{'}X_H-h\overline{X}_H\overline{X}_H^{'})\hat{\beta}_H+(X^{'}X-h\overline{X}_H\overline{X}_H^{'})^{-1}X_L^{'}X_L\hat{\beta}_L,$$

Thus the portion can be explained is given as:

$$\begin{aligned}(\overline{X}_H^{'}-\overline{X}_L^{'})\hat{\beta}_2=(\overline{X}_H^{'}-\overline{X}_L^{'})[&(X^{'}X-h\overline{X}_H\overline{X}_H^{'})^{-1}(X_H^{'}X_H-h\overline{X}_H\overline{X}_H^{'})\hat{\beta}_H\\ &+(X^{'}X-h\overline{X}_H\overline{X}_H^{'})^{-1}X_L^{'}X_L\hat{\beta}_L]\end{aligned} \quad [11]$$

Obviously, whether $(\overline{X}_H^{'}-\overline{X}_L^{'})\hat{\beta}_G^*$ is within the range of $[(\overline{X}_H^{'}-\overline{X}_L^{'})\hat{\beta}_L,(\overline{X}_H^{'}-\overline{X}_L^{'})\hat{\beta}_H]$ or not lies on weighting matrices of $(X^{'}X-h\overline{X}_H\overline{X}_H^{'})^{-1}(X_H^{'}X_H-h\overline{X}_H\overline{X}_H^{'})$ and $(X^{'}X-h\overline{X}_H\overline{X}_H^{'})^{-1}X_L^{'}X_L$, which indirectly proves that whether the portion due to discrimination is within the range constituted by the two estimator of discrimination by Oaxaca-Blinder(1973) or not is also highly associated with the sample characteristics. Subsequently, let's make comparison between Neumark's decomposition and ours as an example. In terms of Neumark's decomposition, differentials in wage due to differences in individual characteristics is $(\overline{X}_H^{'}-\overline{X}_L^{'})(X^{'}X)^{-1}(X_H^{'}X_H\hat{\beta}_H+X_L^{'}X_L\hat{\beta}_L)$, the overvalued and undervalued part respectively are $\overline{X}_H^{'}(X^{'}X)^{-1}X_L^{'}X_L(\hat{\beta}_H-\hat{\beta}_L)$ and $\overline{X}_L^{'}(X^{'}X)^{-1}X_H^{'}X_H(\hat{\beta}_H-\hat{\beta}_L)$,

and the overall part due to discrimination is $[\overline{X}_L^{'}(X^{'}X)^{-1}X_H^{'}X_H + \overline{X}_H^{'}(X^{'}X)^{-1}X_L^{'}X_L](\hat{\beta}_H - \hat{\beta}_L)$. The corresponding estimators of this paper separately are

$(\overline{X}_H^{'} - \overline{X}_L^{'})\hat{\beta}_1$, $\begin{pmatrix} 1 \\ \overline{X}_{1H} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_{0H} - \hat{\beta}_0 - f_H\hat{\delta} \\ \hat{\beta}_{1H} - \hat{\beta}_1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ \overline{X}_{1L} \end{pmatrix}^{'}\begin{pmatrix} \hat{\beta}_0 + f_H\hat{\delta} - \hat{\beta}_{0L} \\ \hat{\beta}_1 - \hat{\beta}_{1L} \end{pmatrix}$

and $(\hat{\beta}_{0H} - \hat{\beta}_{0L}) + \overline{X}_{1H}^{'}(\hat{\beta}_{1H} - \hat{\beta}_1) + \overline{X}_{1L}^{'}(\hat{\beta}_1 - \hat{\beta}_{1L})$. These expressions show that the differences between two decomposition results is uncertain, mainly depends on the sample characteristics.

Ⅳ Applications

In this section, the estimators are compared empirically in an application to estimate gender discrimination in the labor market based on the Chinese data to highlight the discrepancies between these estimators of wage discrimination, especially between the new estimator and the others. The data used are taken from an urban household survey (UHS) in 2007 in China conducted by a city social and economic survey organization (i.e., the Urban Survey Organization) of the National Bureau of Statistics. The survey includes household-based and individual-based parts. The selections of cities and towns, and of households, both are based on the principle of random and representative sampling. Samples from six provinces are included, which include Beijing, Liaoning, Zhejiang, Guangdong, Sichuan and Shaanxi. They cover 16079 households, and 58430 individuals in total.

The applications to estimate gender discrimination, i.e., to decompose male-female wage differentials using a decomposition method, require us to select samples in line with the following criteria: (1) Only people who were in the labor market were included. These samples with missing data and samples whose log wage less than zero were excluded and the samples who are likely to be part-timers, and those who are not working for a wage, including students, household workers, re-employed retired workers and self-employed people were also excluded. (2) The samples were restricted to include only the urban residents; (3) The 16- and 60-year-old female and male workers were respectively selected[①]. This results in the final sub-sample sizes of 10301 males and 7138 females, and the pooled (full) sample size was 17439. The Mincerian wage equation was applied with an extension of controlling more variables, including employment, demographic characteristics, some household characteristics, industry, occupation and region. Tables A-1, A-2 and A-3 separately present regression coefficients of natural log yearly wage by female, male and pooled samples, description and definition of basic variables.

Decomposition results of Oaxaca-Blinder, Cotton, Neumark and the new one are presented in Tables 4-1, 4-2, 4-3 and 4-4. They indicate that the estimators are consistent with each other in the following results: (1) Education and the constant are the most important factors that influence male-female wage differentials, where education has a significantly negative effect on gender wage gap while the constant makes a rather great positive impact on it. All decomposition results reach an agreement that educational return substantially reduces gender wage discrimination, whereas difference in constant coefficient is the leading reason for the gender wage discrimination aggravation; Besides, as it is shown that female wage premium of ownership is more significant that male's, viz., compared with the basement, female's coefficients of ownership dummy variables are relatively bigger than male's (see, Table A-1). Therefore differences in ownership coefficients alleviate the gender discrimination. (2) Four decomposition results show that the discrimination accounts for a larger part of gender wage differentials than male-female endowment differences. More than 77% (the lowest estimate by Neumark) of gender wage differentials is accounted for by the discrimination in China's labor market in 2007.

Table 4-5 gives a summary of four decomposition results. Among these results, the lowest degree of discrimination is estimated by Neumark's estimator, and the highest one is estimated by one of

① It refers to the definition of labor fore age given by the National Bureau of Statistics.

Oaxaca-Blinder estimator (regarding the female current wage structure $\hat{\beta}_f$ as the nondiscriminatory wage structure). They show that the discrimination accounts for 77.01% and 92.02% of wage differentials respectively. The new estimator provides the second highest estimate of 89.73%. The estimates of the new one and Cotton are both between the ranges established by the two estimates by Oaxaca-Blinder's method. The comparisons between the various estimates yield the following findings:

1. Generally, the differences in the estimates of gender discrimination among these various estimators of wage discrimination are insignificant. As stated in Section Ⅲ, differences in estimates among these estimators mainly depend on the sample features. In the light of proposition 1 stated in Section Ⅲ, if the difference in the constant coefficient $(\hat{\beta}_{0H} - \hat{\beta}_{0L})$ plays a dominant role in wage discrimination, the differences among these estimators will be small.

2. Other than Oaxaca-Blinder, these various decomposition methods can subdivide the entire wage discrimination into overvalued portion (employers are always nepotistic toward males) and undervalued portion (employers are always discriminatory against female workers). As is shown in rows 5 and 6 of Table 4-5, Cotton's subdivided results are similar to the new one: the overvalued portion of new one is smaller than that of Cotton's estimator by 0.43%, and the undervalued portion is opposite: the former is 1.33% larger than the latter. As is previously stated, Cotton's estimator is conditional on that the nondiscriminatory wage structure would be closer to the current wage structure of the majority than to that of the minority. Though the new estimator is not based on this assumption, the weighting pattern of the "pooled" wage structure implies that the nondiscriminatory wage structure is also partial to the majority .There is some differences in the definition of the majority. Cotton defined the group which dominates the other group in terms of the amount of labor force as the majority, and the new one defined the group which dominates the other group in terms of the synthesis of the amount and the endowments. So there is no doubt that undervalued portion exceeds overvalued one by both of the two estimators. Besides, since both estimators are partial to the majority, the result of the new one seems to be well approximate to Cotton's by this dataset. In terms of the results of the Oaxaca-Blinder decomposition, the estimates of Cotton and new decomposition are both within the range of 88.06% and 92.02%.

3. Compared with the results of new one, Neumark appears to underestimate gender wage discrimination in China, including overvalued portion and undervalued portion. The relative large gap between the estimates of the new one and Neumark's suggests that the effect of discrimination on labor market is of great significance. The new estimator focuses on alleviating endogenous problem of Neumark's estimator by adding a group-indicator variable to control those unobserved group-specific features induced by discrimination or to control the effect of discrimination on labor market. The fact of labor supply exceeding labor demand in urban China of 2007 reinforces employers' discrimination against the female, and considering that in China a large number of employers are profit-oriented, which drives them to discriminate against females substantially. These facts imply that undervalued portion

should be greater than the overvalued one. Decomposition results of the new one are likely to be more believable than Neumark's.

Table 4-1 Oaxaca-Blinder Decomposition of Male-female Wage Differentials

Explanatory variables	Wage differentials	Male		Female	
		Portion due to differences in characteristics	Portion due to discrimination	Portion due to differences in characteristics	Portion due to discrimination
Education	-0.27994	-0.02522	-0.25471	-0.03250	-0.24744
Experience	0.11667	0.03009	0.08658	0.04310	0.07356
Occupation	0.00812	0.00942	-0.00131	0.00287	0.00525
Industry	0.06575	0.01233	0.05342	0.01237	0.05338
Ownership	-0.00900	0.00736	-0.01636	0.00499	-0.01399
Province	0.00310	-0.00368	0.00678	-0.00856	0.01166
Constant	0.33716	0.00000	0.33716	0.00000	0.33716
In total	**0.24490**	0.03031	0.21155	0.02227	0.21959

Table 4-2 Cotton Decomposition of Male-female Wage Differentials

Explanatory variables	Wage differentials	Portion due to differences in characteristics	Portion due to discrimination	
			Undervalued part	Overvalued part
Education	-0.27994	-0.02819	-0.15071	-0.10103
Experience	0.11667	0.03540	0.05123	0.03004
Occupation	0.00812	0.00675	-0.00077	0.00214
Industry	0.06575	0.01235	0.03161	0.02180
Ownership	-0.00900	0.00639	-0.00968	-0.00571
Province	0.00310	-0.00567	0.00401	0.00476
Constant	0.33716	0.00000	0.19950	0.13766
In total	0.24186	0.02703	0.12518	0.08966

Table 4-3 Neumark Decomposition of Male-female Wage Differentials

Explanatory variables	Wage differentials	Portion due to differences in characteristics	Portion due to discrimination	
			Undervalued part	Overvalued part
Education	-0.27994	-0.02863	-0.13559	-0.11572
Experience	0.11667	0.04673	0.04442	0.02552
Occupation	0.00812	0.01413	-0.02917	0.02316
Industry	0.06575	0.02164	0.00715	0.03696
Ownership	-0.00900	0.00728	-0.01236	-0.00391
Province	0.00310	-0.00555	0.01141	-0.00276
Constant	0.33716	0.00000	0.22436	0.11280
In total	0.24186	0.05560	0.11021	0.07605

Table 4-4 New Decomposition of Male-female Wage Differentials

Explanatory variables	**Wage differentials**	**Portion due to differences in characteristics**	**Portion due to discrimination**	
			Undervalued part	Overvalued part
Education	-0.27994	-0.02791	-0.16074	-0.09129
Experience	0.11667	0.03333	0.04605	0.03729
Occupation	0.00812	0.00721	0.00130	-0.00039
Industry	0.06575	0.01127	0.03786	0.01661
Ownership	-0.00900	0.00649	-0.01088	-0.00461
Province	0.00310	-0.00557	0.00604	0.00263
Constant	0.33716	0.00000	0.20878	0.12839
In total	0.24186	0.02484	0.12840	0.08862

Table 4-5 Decomposition for Male-female Wage Differentials by Four Methods

Decomposition method		Portion due to differences in characteristics	Portion due to discrimination		
Oaxaca-Blinder Decomposition	Male	12.53%	88.06%		
	Female	9.21%	92.02%		
			Overvalued portion	Undervalued portion	Overall discrimination
Cotton Decomposition		11.17%	37.07%	51.76%	88.83%
Neumark Decomposition		22.99%	31.45%	45.57%	77.01%
New Decomposition		10.27%	36.64%	53.09%	89.73%

Ⅴ Conclusions

The data used to estimate wage discrimination is from the current labor market in the presence of discrimination, so, it is impossible to obtain the nondiscriminatory wage structure directly, which is of the essence for wage discrimination estimation. Oaxaca(1973) and Blinder(1973) proposed the conventional estimator of wage discrimination. However, the estimator is confronted with an "index number problem". The essence of the answer to the "index number problem" is to construct a counterfactual wage structure prevailing in the absence of discrimination, i.e., to obtain the nondiscriminatory wage structure by current wage structure. Reimer(1983) and Cotton(1988) both regarded the weighted average of the current wage structures of group H and L as the nondiscriminatory wage structure. The weighted averages by synthesizing the two indices of Oaxaca-Blinder seem to get rid of the effect of the arbitrary selection on estimate of wage discrimination, however, it still could not avoid the "index number problem" caused by the arbitrary choice of the weights essentially.

Neumark(1988) built a theoretical framework on wage level in absence of discrimination, and then suggested that the "pooled" wage structure is equal to the wage structure that would prevail in the absence of discrimination. Compared with the previous ones, the theoretical model renders Neumark's nondiscriminatory wage structure more brilliantly. Whereas, it is important to note that the strict assumption in the theoretical model may lead to a biased estimate of the nondiscriminatory wage structure in practice and finally result in a biased estimate of wage discrimination. We believe that the "pooled" wage structure estimated from the regression equation after controlling the group-specific indicator should to some extent correct Neumark's estimate bias. It is acknowledged that there is some arbitrariness in the constant coefficient in our nondiscriminatory wage structure with $\hat{\beta}_{0G}^{*} = (\hat{\beta}_0 + f_H\hat{\delta})$. While fortunately, it is implied by the Proposition 1 that constant coefficient has no effect on the overall wage discrimination, and it only affects the discrimination composition. Consequently, we believe that our effort to correct Neumark's estimate bias would further improve the estimate of wage discrimination and the limitation of our estimator should be taken up in future research.

Proposition 2 demonstrates that the specific differences in the estimators of wage discrimination by the various nondiscrimatory wage structure are highly correlated with the sample characteristics. The empirical study on gender wage differences in urban China using UHS dataset in 2007 provides us abundant evidences.

Furthermore, the idea of the counterfactual construction used in nondiscriminatory wage structure should be extended to Brown decomposition. Brown(1988) put forward another decomposition method by introducing occupational segregation, while Brown decomposition still could not escape from the "index number problem", and moreover, it involves dual "index number problem". Appleton, et al. (1999) has made an attempt to solve the problem. We believe that Brown decomposition would make a further improvement by the method of the paper.

Appendix

Table A-1: Regression Coefficient of important variables

Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

Independent variable	Male sample	Female Sample	Pooled Sample	Pooled Sample
edu	0.0671***	0.0865***	0.0762***	0.0743***
exp	0.0414***	0.0294***	0.0334***	0.0364***
expsqr	-0.0739***	-0.0423***	-0.0491***	-0.0618***
_Ioccu_2	-0.0932***	-0.0769**	-0.104***	-0.0851***
_Ioccu_3	-0.179***	-0.206***	-0.233***	-0.193***
_Ioccu_4	-0.391***	-0.316***	-0.405***	-0.347***
_Ioccu_6	-0.276***	-0.298***	-0.283***	-0.276***
_Ioccu_8	-0.407***	-0.336***	-0.409***	-0.376***
_Iindus_2	0.0559	0.142	0.0916	0.09
_Iindus_3	-0.0265	-0.117	-0.0761*	-0.0563
_Iindus_4	0.106*	0.0452	0.0907*	0.0903*
_Iindus_5	0.00075	-0.00595	0.00169	-0.0055
_Iindus_6	-0.04	0.0221	-0.0264	-0.0237
_Iindus_7	0.047	0.03	0.0512	0.0435
_Iindus_8	-0.0974	-0.0863	-0.108**	-0.0787
_Iindus_9	0.143**	0.150*	0.128***	0.160***
_Iindus_10	0.104*	0.043	0.0753	0.0881*
_Iindus_11	-0.139**	-0.248***	-0.245***	-0.195***
_Iindus_12	-0.0166	0.0428	-0.0321	0.0365
_Iindus_13	0.0243	-0.00836	-0.0388	0.0247
_Iindus_14	0.119**	0.126	0.103**	0.121**
_Iindus_15	0.0507	-0.0625	-0.0197	0.00551
_Iindus_16	0.0672	0.0321	0.043	0.057
_Iownership_2	-0.262***	-0.178***	-0.255***	-0.231***
_Iownership_3	-0.106***	-0.0946***	-0.107***	-0.106***
_Iownership_5	-0.271***	-0.202***	-0.239***	-0.245***
_Iprovin_21	-0.201***	-0.328***	-0.240***	-0.247***
_Iprovin_33	0.307***	0.339***	0.333***	0.322***
_Iprovin_44	0.334***	0.294***	0.326***	0.317***
_Iprovin_51	-0.440***	-0.331***	-0.388***	-0.391***
_Iprovin_61	-0.494***	-0.490***	-0.495***	-0.494***
gender	––	––	––	0.217***
Constant	7.843***	7.505***	7.730***	7.586***
Observations	10253	7075	17328	17328
R-squared	0.391	0.402	0.388	0.406

Table A-2: Description of Basic Variables

Variable	Male sample	Female Sample	Difference
ywage	9283.59	7416.76	1866.83
lnywage	8.90	8.65	0.25
edu	12.78	13.16	-0.35
exp	22.82	18.29	4.49
expsqr	6.32	4.19	2.11
occu basement	8.11%	4.33%	3.64%
_Ioccu_2	22.86%	23.87%	-1.18%
_Ioccu_3	33.59%	41.90%	-8.55%
_Ioccu_4	5.90%	13.28%	-7.40%
_Ioccu_6	27.16%	13.91%	0.78%
_Ioccu_8	2.38%	2.70%	12.84%
indus basement	1.27%	0.77%	0.49%
_Iindus_2	1.46%	0.77%	0.68%
_Iindus_3	25.62%	19.84%	5.56%
_Iindus_4	4.64%	2.49%	2.09%
_Iindus_5	4.83%	2.24%	2.55%
_Iindus_6	1.70%	1.09%	0.61%
_Iindus_7	12.86%	5.42%	7.38%
_Iindus_8	2.30%	4.06%	-1.78%
_Iindus_9	3.29%	4.78%	-1.51%
_Iindus_10	2.11%	1.91%	0.18%
_Iindus_11	4.36%	10.52%	-6.16%
_Iindus_12	3.63%	8.14%	-4.51%
_Iindus_13	6.78%	12.03%	-5.26%
_Iindus_14	2.83%	2.02%	0.84%
_Iindus_15	17.80%	19.19%	-0.96%
_Iindus_16	4.52%	4.72%	-0.21%
ownership basement	66.16%	63.15%	3.27%
_Iownership_2	4.82%	7.89%	-3.10%
_Iownership_3	17.50%	17.81%	-0.44%
_Iownership_5	11.52%	11.15%	0.28%
provin basement	19.99%	19.25%	0.83%
_Iprovin_21	22.84%	19.80%	3.17%
_Iprovin_33	23.03%	23.47%	-0.57%
_Iprovin_44	11.03%	12.48%	-1.48%
_Iprovin_51	12.86%	14.16%	-1.33%
_Iprovin_61	10.25%	10.84%	-0.63%
Observations	10301	7138	–––

Table A-3: Definition of Basic Variables

Variable	Definition
ywage	Yearly wage
lnywage	Natural log of yearly wage
edu	Education attainment
exp	Experience
expsqr	Experience square/100
occu basement	Executive, administrative and managerial
_Ioccu_2	Professional specialty
_Ioccu_3	Clerical workers
_Ioccu_4	Commercial workers and service workers
_Ioccu_6	Production and transportation and related workers
_Ioccu_7	Soldier
_Ioccu_8	Others
indus basement	Farming, forestry, animal husbandry and fishing
_Iindus_2	Mining
_Iindus_3	Manufacturing
_Iindus_4	Production and supply of electricity, gas and water
_Iindus_5	Building industry
_Iindus_6	Water conservation, management of environment and public utility
_Iindus_7	Transportation, storage and postal service
_Iindus_8	Lodging and catering services
_Iindus_9	Finance
_Iindus_10	Real estate
_Iindus_11	Resident services and other services
_Iindus_12	Health care, social security and social welfare
_Iindus_13	Education
_Iindus_14	Scientific research, polytechnic services and geological exploration
_Iindus_15	Public administration and social organization
_Iindus_16	Others(including leasing, business services, information transfer and international organizations)
ownership basement	Employed in state-owned enterprise
_Iownership_2	Employed in urban collective enterprise
_Iownership_3	Employed in other economic enterprises
_Iownership_5	Employed in microeconomic or private enterprise (basement)
provin basement	Beijing
_Iprovin_21	Liaoning
_Iprovin_33	Zhejiang
_Iprovin_44	Guangdong
_Iprovin_51	Sichuan
_Iprovin_61	Shaanxi (basement)

Reference

Appleton, S., J. Hoddinott and P. Krishnan, 1999, "The Gender Wage Gap in Three African Countries", *Economic Development and Cultural Change*, 47(2), 289-312.

Autor, D. H., Lawrence, F. K., and Melissa, S. K., 2005, "Rising Wage Inequality: The Role of Composition and Prices". *Harvard Institute of Economic Research Working Papers No. 2096*, .

Blau, F. and L. Kahn, 1996, "Wage Structure and Gender Earning Differentials: An International Comparison", *Econometrica*, 63(250), 3-8.

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

Butler, R., 1982, "Estimating Wage Discrimination in the Labor Market", *Journal of Human Resources*, 17(4), 606-21.

Brown, R. S., M. Moon and B. S. Zoloth, 1980, "Incorporating Occupational Attainment in Studies of Male-Female Earnings Differentials", *The Journal of Human Resources*, 15(1), 3-28.

Ben Jann, 2008, "A Stata Implementation of Blinder-Oaxaca Decomposition", ETH Zurish Sociology Working Paper No.5.

Cotton, J., 1988, "On the Decomposition of the Wage Differentials", The Review of Economics and Statistics, 70(2), 236-243.

DiNardo, J., Fortin, N. M., and Lemieux, T., "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach", *Econometrica*, 1996, 64(5), 1002-1044.

Démurger, S., M. Fournier and Yi Chen, 2007, "The Evolution of Gender Earnings Gaps and Discrimination in Urban China, 1988-95", *The Developing Economies*, XLV-1, 97-121.

Ferber, M. A., and C. A. Green, 1982, "Traditional or Reverse Sex Discrimination? A Case Study of a Large Public University", *Industrial and Labor Relations Review*, 35(4), 550-564.

Firpo, S., Fortin, N., and Lemieux, T., "Decomposing Wage Distributions Using Recentered Influence Function Regressions", Mimeo, Department of Economics, University of PUC-RIO. 2007b.

Fortin, N.2006, "Greed, Altruism, and the Gender Wage Gap", University of British Columbia. *Available from http://www.econ.ubc.ca/nfortin/Fortinat8.pdf.*

Gustafsson, B. and Li, Shi, 2000, "Economic Transformation and the Earnings Gap in Urban China", *Journal of Population Economics*, 13(2), 305-329.

Juhn, C., Murphy, K., and Pierce, B., "Accounting for the Slowdown in Black-White Wage Convergence", In *Workers and Their Wages: Changing Patterns in the United States*, edited by Marvin H. Kosters, Washington: American Enterprise Institute Press, 1991.

Juhn, C., K. M. Murphy and B. Pierce. 1993, "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy*, 101(3), 410 – 442.

Machado, J., and J. Mata. 2005, "Counterfactual Decompositions of Changes in Wage Distributions Using Quantile Regression." *Journal of Applied Econometrics*, 20(4), 445-465.

Neuman S., and R. Oaxaca, 2004, "Wage Decomposition with Selectivity-corrected Wage Equation: A Methodological Note", *Journal of Economic Inequality*, 2, 3-10.

Neumark, D., 1988, "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination", *The Journal of Human Resources*, 23(3), 279-295.

Oaxaca, R., 1973, "Male-Female Wage Differentials in Urban Labor Markets", *International Economic Review*, 14(3), 693-709.

Oaxaca. R., and M. Ransom, 1994, "On Discrimination and the Decomposition of Wage Differentials", *Journal of Econometrics*, 61, 5-21.

Reimers, C. W., 1983, "Labor Market Discrimination Against Hispanic and Black Men", *The Review of Economics and Statistics*, 65(4), 570-579.

Deng Quheng, 2007, "Earnings Differential between Urban Residents and Rural Migrants: Evidence from Oaxaca-Blinder and Quantile Regression Decompositions", *Chinese Journal of Population Science*, 2, 8-16.

Wang Meiyan, 2005, "Employment Opportunities and Wage Gaps in the Urban Labor Market: A Study of the Employment and Wages of Migrant Laborers", *Social Sciences in China*, 5, 35-46.

Xie Sisheng and Yao Xianguo, 2006, "Empirical Study on Wage Discrimination of Migrant Workers", *Chinese Rural Economy*, 4, 49-55.

Yao Xianguo and Lai Puqing, 2004, "Urban-rural Hukou Differentials in Chinese Labor Relations", *Economic Research Journal*, 7, 82-90.

Yao Xianguo and Li Xiaohua, 2007, "Rising Wage Inequality: Composition Effect and Price Effect", *Chinese Journal of Population Science*, 1, 36-43.

Zhang Dandan, 2004, "Marketization and Gender Wage Differentials", *Chinese Journal of Population Science*, 1, 32-41.