CHAPTER IV

METHODOLOGY, DATA BASE AND THE EMPIRICAL SPECIFICATION OF THE MODEL

4.1 INTRODUCTION

The wage decomposition method recently developed by Oaxaca and Ransom (1994) used to decompose the wage differentials and to identify the discrimination component by gender, sector of employment and caste in the labour market for secondary school teachers. It can be shown that the total wage differentials consist of two components: (1) due to differences in productivity characteristics and (2) a residual component which may be attributed to labour market discrimination against a particular group.

This methodology requires estimation of earnings functions separately for each group namely men, women, public sector workers, private sector workers, teachers in the forward community and teachers belonging to `other community' groups. The Mincerian semi-logarithmic earnings function is specified and estimated. Three specifications are used in the estimation of the earnings function for different groups of workers. The basic model includes education, experience and experience square variables while in the extended models, additional variables such as training, marital status and parental education are included.

The main problem in the decomposition method is to identify a correct non-discriminatory wage structure. In this work, the decomposition technique
is applied by assuming alternative wage structures as non-discriminatory wage. In the Oaxaca model (1973), the male wage and the female wage are assumed as the non-discriminatory wage structures. Since such an assumption involves the index number problem\(^1\), Cotton (1988) proposed an alternative procedure to estimate the non-discriminatory wage structure. He assumes that the wage structure that would exist in the absence of discrimination is the simple weighted average of the observed structures for the two groups. Neumark (1988) develops an alternative procedure from a particular Beckarian model. According to him, the non-discriminatory wage structure can be obtained from an earnings function estimated over the pooled sample. Both Neumark and Cotton generalise the method suggested by Oaxaca.

In the present study all the four alternative non-discriminatory wage structures are used to decompose the wage-gap and identify the discrimination component by gender, sector of employment and caste in the labour market for secondary school teachers.

This chapter is organised in the following manner. Section 4.2 discusses all the four methods, namely, the Oaxaca decomposition technique, with the female and male wage structures as the non-discriminatory wages, the Cotton method, the Neumark method and the Oaxaca-Ransom (1994) synthesised approach. The similarities and differences between these methods are also discussed in this section. The data base for the analysis is discussed in Section 4.3. The empirical specification of the model and estimation issues are given in Section 4.4.

\(^1\) An index number may be described as a specialised average designed to measure the change in a group of related variables over a period of time. (Statistical Methods, by S.P. Gupta, pp1458, 1974)
4.2 METHODOLOGY

4.2.1 Oaxaca's Method of Decomposition

The standard procedure to analyse the determinants of the two groups' earnings gap is to estimate earnings equations for sample of group one and group two separately. Here, group one may refer to male teachers, public sector or forward community and group two may refer to female teachers, private sector or to 'other community'. Specifically the procedure is to fit equation (1) to a sample of group one teachers and equation (2) to a sample of group two teachers

\[ W_m = b_m X_m + \mu_m \]  \hspace{1cm} (1)

\[ W_n = b_n X_n + \mu_n \]  \hspace{1cm} (2)

where \( W \) is hourly wage, usually measured in logarithmic terms. \( X \) is a vector of measured characteristics of workers such as education, experience and training, as well as control variables like marital status, caste and sector. The vector of regression coefficient \( b \), corresponds to the returns that the market yields to a unit change in endowments such as education and experience. The error term \( \mu \), reflects measurement error as well as the effects of unmeasured or unobserved factors by the researcher.

A property of ordinary least squares regression analysis is that the regression lines pass through the mean values of the variables so that

\[ w_m = \bar{\nu} \]  \hspace{1cm} (3)

\[ \bar{W}_n = b_n \bar{X}_n \]  \hspace{1cm} (4)
The hats denote ordinary least squares estimated values. If group \( n \)' were
given the group \( m \)' wage structure, then their average wage would be

\[
\bar{W}_n^* = \hat{b}_m \bar{X}_n
\]  

(5)

\( \bar{W}_n^* \) is the average wage for group \( n \)' that would prevail in the absence of
wage discrimination. Subtracting (5) from (3) gives the difference between
average earnings of group \( m \)' earnings and the average hypothetical earnings
of group \( n \)' that would prevail if group \( n \)' were paid according to the group \( m \)'
pay structure. This difference reflects their different endowments of wage
generating characteristics.

\[
\bar{W}_m - \bar{W}_n^* = \hat{b}_m \bar{X}_m - \hat{b}_m \bar{X}_n
\]

(6)

Subtracting (4) from (5) yields the difference between the hypothetical
"non-discriminatory" group \( n \)' wage and their actual wage. This difference
reflects the different returns to the same wage generating characteristics, i.e.

\[
\bar{W}_n^* - \bar{W}_n = \hat{b}_m \bar{X}_n - \hat{b}_n \bar{X}_n
\]

(7)

Adding (6) and (7) yields,

\[
\bar{W}_m - \bar{W}_n = \hat{b}_m (\bar{X}_m - \bar{X}_n) + (\hat{b}_m - \hat{b}_n) \bar{X}_n
\]  

(8)

That is, the overall average group \( m \)' - group \( n \)' wage gap can be
decomposed into two components. The first is the portion attributable to
differences in the endowments of wage generating characteristics \((\bar{X}_m - \bar{X}_n)\) evaluated at the group 'm' returns \((b_m)\). The second portion is attributable to differences in the return \((b_m - b_n)\) that 'm' and 'n' get for the same endowment of wage generating characteristics \((\bar{X}_n)\). This latter component is often taken as reflecting wage discrimination.

Equation (8) can also be expressed as follows. If group 'm' were given the group 'n' pay structures, then their average wage would be

\[
\bar{W}_m^* = b_n \bar{X}_m
\]  

(9)

\(\bar{W}_m^*\) is the average wage for group 'm' that would prevail in the absence of wage discrimination. Subtracting (9) from (4) gives the difference between group 'n' earnings and the average hypothetical group 'm' earnings that would prevail if group 'm' were paid according to group 'n' pay structure. This difference reflects their different endowments of wage generating characteristics.

\[
\bar{W}_m - \bar{W}_m^* = b_n \bar{X}_n - b_n \bar{X}_m
\]

\[
= b_n (\bar{X}_n - \bar{X}_m)
\]

(10)

Subtracting (3) from (9) yields the difference between the hypothetical non-discriminatory group 'm' wage and their actual wage. This difference reflects the different returns to the same wage generating characteristics i.e.

\[
\bar{W}_n^* - \bar{W}_n = b_n \bar{X}_m - b_m \bar{X}_m
\]

\[
= (b_n - b_m) \bar{X}_m
\]

(11)
Adding (10) and (11) yields,

\[
\bar{W}_m - \bar{W}_n = \hat{b}_n (\bar{X}_n - \bar{X}_m) + (\hat{b}_n - \hat{b}_m) \bar{X}_m
\]

(12)

The first term \((\bar{X}_n - \bar{X}_m)\) in the right-hand side of equation (12) is attributable to differences in the endowments of wage generating characteristics, evaluated at the group `n' wage structure \((\hat{b}_n)\). The other portion is attributable to the differences in returns \((\hat{b}_n - \hat{b}_m)\) that group `n' and group `m' get for the wage generating characteristics of group `m' \((\bar{X}_m)\).

4.2.1.1 Limitations of Oaxaca's Decomposition Method

The question arises as to which of the two equations, (8) or (12) is to be used in the empirical work. Oaxaca obtained estimates from both formulations, using them to establish the range within which the true values of the components presumably would fall. In Oaxaca's original article, using data on whites, equation (8) yields an estimate of 52.9 percent of the male - female log wage differential as due to discrimination, while equation (12) yields an estimate of 63.9 percent. Some subsequent researchers followed Oaxaca's example of estimating both forms, while others opted for one form or the other or some variant of both.

Yet another problem faced by the adherents of Oaxaca's decomposition method was the problem of 'omitted variables'. The residual component can become an exact measure of labour market discrimination, only if all the factors that determine the wage are taken into consideration and properly accounted for. Some of these factors may be omitted due to inadequate data. In
such cases the residual will reflect these omitted influences as well, and will therefore result in the extent of discrimination being either over or underestimated. This is a long-standing problem and one should recognize the constraint it places on the interpretation of results.

Another flaw in the construction of the Oaxaca decomposition is that suggested by Butler (1982). He argues that an attempt to measure labour market discrimination by differences in white-black regression coefficients confounds market or demand side sources of discrimination with those that originate on the non-market, or supply-side. Such coefficients, which are taken from reduced-form equations, are an amalgam of both demand and supply structural coefficients. And because of pre-market discrimination in the provision of education and other skill-acquiring opportunities, the demand for group ‘n’ labour might be more elastic than the demand for the more capital-compatible group ‘m’ labour even in the absence of discrimination. In which case, even though both the groups are identical in all other respects the group ‘m’ coefficients will be larger than the group ‘n’ coefficients and any measure of discrimination based on their differences will be overstated.

Though Butler is correct in questioning the comparisons of black (group ‘n’) and white (group ‘m’) regression coefficients, according to Cotton (1988), he is not correct to assume that these are the coefficients that would prevail in the absence of discrimination. After the elimination of discrimination in the short run, one might observe group ‘m’ and ‘n’ with different average skills, because of different opportunities in the past. However, in the long run, as the minority labourers (group ‘n’) are assured of competing on equal terms in the same
markets as the majority labourers (group 'm'), the differences in supply characteristics can be expected to diminish along with differences in the demand for minority and majority labourers. It is this convergence of the two groups that is heralded by the proponents of Oaxaca's decomposition method.

4.2.2 Cotton Method

According to Cotton(1988), in the absence of discrimination, either the male wage structure (group 'm') or the female wage structure (group 'n') would prevail. In the Oaxaca method, the group discriminated against is undervalued and the preferred group is overvalued. Thus the male and female wage structures are both functions of discrimination and either of them will prevail in the absence of discrimination.

The derivation of a more suitable decomposition formula starts with Becker's assumption that in the absence of discrimination in perfectly competitive markets, whites (Group m) and blacks (Group n) would be perfect substitutes in production. Therefore in the absence of discrimination the wage structures are assumed to be equal: \( b_m = b_n = b' \), where \( b' \) is the non-discriminatory wage structure.\(^1\)

Now consider the hypothetical term, \( \Sigma b' \left( \bar{x}_m - \bar{x}_n \right) \). This is the difference in the current average productivity characteristics of the groups 'm' and 'n'.

---

\(^1\) Cotton (1988) investigated racial wage differences. However the subsequent analysis focusses on gender, sectoral and caste differences. Here in this work, 'group 'm' is substituted for white and group 'n' for black in the reproduction of his formulas. Group 'm' may refer to male, public sector or forward caste, group 'n' may refer to female, private sector or 'other community' group.
evaluated in the absence of discrimination. It is therefore the 'true' value of the skill component of the wage differential.

Consider also the hypothetical term, $\Sigma b^* \bar{x}_m$. These are the current 'm' average productivity characteristics valued in the absence of discrimination. The difference between this term and $\Sigma b_m \bar{x}_m$ [is the same as $\hat{b}_m \bar{x}_m$ in equation (3)] is solely due to differences in the way group 'm' are currently treated and the way they would be treated in the absence of discrimination.

$$\Sigma b_m \bar{x}_m - \Sigma b^* \bar{x}_m = \Sigma \bar{x}_m (b_m - b^*)$$

This is therefore that part of the treatment component of wage differential which, if positive, is due to group ms' "pure" treatment advantage. A similar situation exists with respect to group 'n', and we have

$$\Sigma b^* \bar{x}_n - \Sigma b_n \bar{x}_n = \Sigma \bar{x}_n (b^* - b_n)$$

This is the part of the treatment component which if 'positive' measures group 'n's' "pure" treatment disadvantage. The average wage differential is therefore decomposed as

$$\ln \bar{w}_m - \ln \bar{w}_n = \Sigma b^* (\bar{x}_m - \bar{x}_n) + \Sigma \bar{x}_m (b_m - b^*) + \Sigma \bar{x}_n (b^* - b_n) \quad (13)$$

In this decomposition, the treatment or discrimination component is made up of two elements. One represents the amount by which group 'm' productivity characteristics are overvalued (the benefit of being a teacher in group m) and
the other the amount by which group 'n' productivity characteristics are undervalued (the cost of being a group 'n' worker).

The major operational weakness of (13), however is the fact that the \( b' \) vector is unobserved and therefore must be estimated if the formulation is to be useful for empirical work. The non-discriminatory wage structure \( b' \) proposed by Cotton is based on a number of assumptions.

The first of these assumptions is that in the absence of discrimination, group 'm' would receive a lower average wage than they currently receive and group 'n' would receive a higher average wage. Thus,

\[
\Sigma b_m \overline{X} > \Sigma b'_n \overline{X} > \Sigma b_n \overline{X}
\]

Second, it is assumed that in the absence of discrimination the prevailing market structure will be some function of the forces that currently determine the group 'm' and group 'n' wage structures. This assumption is simplified by specifying \( b' \) as a linear function of \( b_m \) and \( b_n \), the respective group 'm' and group 'n' wage structures.

Third, it is assumed that the non-discriminatory wage structure will be closer to the current 'm' wage structure than to the current 'n' wage structure. This third assumption is operationalised by weighting the 'm' and 'n' wage structures by the respective proportions of 'm' and 'n' males in the employed male labour force. Thus the estimator of \( b^* \) used is defined as

\[
b^* = f_m b_m + f_n b_n \tag{14}
\]

where \( f_m \) and \( f_n \) are the proportions.
Finally it is assumed that neither total actual output nor the total wage bill would change in the absence of discrimination. The only effect would be a redistribution of income and jobs. The elimination of discrimination may not lead to a better utilisation of the labour force. Thurrow (1969) has observed that the total potential output would be greater in the absence than in the presence of discrimination.

4.2.3 Neumark Method

Neumark (1988) argues that the appropriate decomposition depends on the type of discrimination hypothesized. In particular, employers may practice nepotism towards men (group \(m\)) or discrimination against women (group \(n\)). Under nepotism, group \(n\) are paid the competitive wage, but men are overpaid. In such a situation, the coefficients from group n's earnings function provide an estimate of the non-discriminatory wage structure. Under discrimination, employers pay men competitive wages but underpay women. In this case the group \(m\) coefficients should be taken as the non-discriminatory wage structure. In reality employers may practice both nepotism and discrimination. Neumark proposes a general model of discrimination in which employers may have different preferences (nepotistic or discriminatory) towards different types of workers (e.g. older, uneducated, female etc.). Useful results can be obtained, given the restriction that employer preferences are homogeneous of degree zero within each type of labour; that is to say employers care only about the proportion of each type of labour employed. With such a restriction, Neumark shows that the non-discriminatory wage structure can be estimated from an earnings function estimated over the pooled sample
(that is both men and women). This non-discriminatory or pooled wage structure \( b^* \), is a weighted average of the male and female wage structures:

\[
b^* = \Omega b_m + (1-\Omega) b_n
\]

The Neumark decomposition is thus

\[
\bar{w}_m - \bar{w}_n = b^* (\bar{x}_m - \bar{x}_n) + \bar{x}_m (b_m - b^*) + \bar{x}_n (b^* - b_n)
\]

The first term is that part of the wage gap explained by differences in characteristics, given non-discriminatory returns. The second and third terms show the contribution of differences between actual and pooled returns for group 'm' and 'n' workers respectively.

4.2.4 Oaxaca - Ransom Synthesised Approach

Recently Oaxaca and Ransom (1994) suggested adopting a wage structure estimated from the pooled sample of male and female workers. This method is superior to other methods because it is derived from the economic theory of discrimination (Becker, 1957). The proposed weighting matrix is specified by

\[
\Omega = (X'X)^{-1} (X'm X_m),
\]

where \( X \) is the observation matrix for the pooled sample and \( X_m \) is the observation matrix for the white sample. This interpretation of \( \Omega \) as a weighting matrix can be understood by noting that

\[
XX = X'm X_m + X'n X_n,
\]

where \( X_n \) is the observation matrix for the group 'n' sample. Furthermore, both \( XX \) and \( X'm X_m \) are clearly positive definite matrices. However, \( \Omega \) is not
positive definite, nor even symmetric, except in special cases. It is straightforward to show that

\[ b^* = \Omega \hat{\beta}_m + (1-\Omega) \hat{\beta}_n \]  

(18)

Accordingly, this weighting scheme interprets the ordinary least square (OLS) estimates from the combined white and black groups as the estimate of the wage structure that would exist in the absence of discrimination. This estimate of the common wage structure is not in general a convex, linear combination of the separately estimated white (group 'm') and black (group 'n') wage structures. The weighting matrix \( \Omega_o \) can be compared with the weighting scheme of Cotton \( \Omega_c \). Let \( N_m \) be the number of male teachers or teachers in the public sector or teachers belonging to the forward community and \( N \) be the total number of 'm' and 'n' teachers. If

\[ (N_m/N) \ (XX) = (X_m'X_m), \]

i.e., \( I_m \ (XX) = (X_m'X_m), \)

then \( \Omega_o = \Omega_c \).

This will be true if \( (1/N) \ (XX) = (1/N_m) \ (X_mX) \).

Table 4.1 summarises the alternative methods of constructing the weighting matrix \( \Omega \). The literature has proposed different weighting schemes, as shown in the above table, to deal with the underlying index problem. First, Oaxaca (1973) proposes either the current male wage structure, i.e. \( \Omega = I \), or the current female wage structure, i.e. \( \Omega = O \), the null matrix, as \( b^* \), suggesting that the result would bracket the "true" nondiscriminatory wage structure. Reimers (1988) implements a methodology that is equivalent to \( \Omega = 0.5I \). In other words identical weights are assigned to both men and women. Cotton (1988) argues that the non-discriminatory structure should approach the
structure that holds for a larger group. In the context of sex discrimination such weighting structure implies that an $\Omega = I_m I$, where $I_m$ is the fraction of males in the sample.

<table>
<thead>
<tr>
<th>Structure</th>
<th>$\Omega$ (weighting matrix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. White/male/group 'm'</td>
<td>I (identity matrix)</td>
</tr>
<tr>
<td>2. Black/female/group n</td>
<td>O (null matrix)</td>
</tr>
<tr>
<td>3. Cotton</td>
<td>$I_i I, i = m,w.$</td>
</tr>
<tr>
<td>4. Pooled</td>
<td>$(XX)^{-1}(X_i X_i), i = w,m$</td>
</tr>
</tbody>
</table>

Table 4.1

Construction of the Weighting Matrix

A more generalised method is provided by Neumark (1988), who shows that under certain conditions in the underlying utility function\(^2\) the correct non-discriminatory wage structure $b^*$ can be obtained by OLS estimates on the pooled sample where the model adopted is $\ln(W_i) = bX_i + \mu$, where $\ln W_i$ is the natural logarithm of the observed wage for individual $i$, $X$ is a vector of observed characteristics, $b$ is a vector of coefficients and $\mu$ is a stochastic error distributed $N(0,\sigma^2)$, i.e. without selection bias correction. As shown by Oaxaca and Ransom(1994), such result is equivalent to a weighting scheme of the form:

$$W = (XX)^{-1}(X_m X_m)$$

where $X$ is the observation matrix for the pooled sample and $X_m$ is the observation matrix for the male sample. Such a weighting scheme is not

\(^2\) Specifically the firm's utility function is homogeneous of degree zero within each type of labour (i.e. group 'm' and group 'n').
constrained to produce results that are in general a convex linear combination of the independently estimated male and female wage structures (Oaxaca and Ransom, 1994).³

4.3 THE DATA BASE

The implications of the theoretical model is tested here on a sample drawn from a primary survey conducted in the year 1999-2000. In the Indian context, there are some studies on wage differentials in the labour market. These studies show that, there is a substantial gender gap in wages and there is also evidence of discrimination against women in the labour market. However, existing studies are based on workers from all occupations. If there is occupational segregation, that is women concentrating in low paying occupations because of lack of education or labour market skills, and men entering into high paying jobs, then this would naturally lead to wage differential between men and women. It is important and interesting to study wage differentials and discrimination within an occupation. None of the previous studies for India has attempted to investigate occupation specific wage differentials. This study makes a pioneering attempt in this direction by considering wage differential and discrimination in the labour market in a particular occupation, namely teaching.

Teaching is an occupation where women are found in large number. Moreover both public and private schools prefer to employ women than men, particularly in the school level education. Further, the nature and type of work

³ Note that this is not true for Cotton scheme, which is indeed a convex linear combination of the two separate estimates.
done is the same for men and women. Hence it is decided to investigate the wage differences in the labour market for secondary school teachers.

The study requires data on the wages and information on the individual, socio-economic and demographic characteristics of the teachers. Since secondary data from the census of India or the National sample survey do not provide wage and other required information, it was decided to collect the required data through a primary survey.

Details regarding the number of schools in Chennai district as on September 2000 was obtained from the Chief educational officer, Chennai. The schools were divided on the basis of different levels, namely High school and higher secondary schools. Primary schools and middle schools (I-V and VI - VIII classes respectively) are not included in the sample because there are only few male teachers in these schools. The Government of India has recently proposed to appoint only female teachers at the primary levels. Therefore it was decided to include only high schools and higher secondary schools in the sample. These schools are identified on the basis of different types, namely, the government, corporation, aided, partly aided, unaided and matriculation schools. These schools are divided on the basis of management as public and private schools.

The schools that come under the public schools\(^4\) are the government schools, corporation schools and aided schools. The schools that come under the

\(^4\) The Adi-Dravida Welfare and social welfare schools are very small in number and hence not included in the sample.
private sector are the matriculation and the unaided. The teachers in the sample are carefully selected to represent the various caste groups namely the forward class, backward class, the most backward class and the scheduled caste. The education department do not possess a detailed, similar report or information for the private schools.

A stratified two-stage random sampling method was adopted for the collection of data. The schools were stratified on the basis of different types, viz., government, corporation, aided, unaided and matriculation schools. In the first stage 10 percent of the schools are randomly selected. Out of the total number of 486 secondary and higher secondary schools in Chennai urban area, 50 schools were selected in such a manner that all the types of schools were represented. In the second stage, 20 percent of the teachers from each sampled school were selected using simple random sampling method. The survey thus covered 800 high and higher secondary school teachers consisting of 278 (35 percent) male teachers and 522 (65 percent) female teachers. A pre-prepared questionnaire was used to collect the relevant information from the teachers through interview method. The questionnaire is given in Appendix 2.

Another reason for restricting the survey to urban locality is the limitation imposed by time and personnel. Even though a comparison between urban and rural areas may be desirable, one cannot obviously expect a small study such as this to do full justice in all directions, especially considering the

---

5 The Anglo Indian Schools, that come under the private sector form only 2 percent of the total number of schools in Chennai and therefore not included in the sample.

6 The sample does not have any teacher belonging to 'Scheduled Tribe' group.
nature of the study demanding good amount of information. Moreover in India there is no micro economic data either at national or regional level yielding particulars of the number of hours worked, remuneration in different types of schools, different levels of training, their demographic characteristics, job characteristics and a host of other factors which are essential for the study.

4.3.1 Sampling Distribution

Table 4.2 and 4.3 show the distribution of male and female teachers in the public school on the basis of type of school and sex. While Table 4.2 shows the teachers in high school and Table 4.3 shows the teachers working in higher secondary schools.

The total number of teachers from high school is 123 of which 41 teachers are men (33 percent) and 82 teachers are women (67 percent). The number of female teachers are double when compared with male teachers.

Table 4.2

Distribution of High School Teachers on the basis of Sex and Type for the Year 2000

<table>
<thead>
<tr>
<th>Type</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Percentage</td>
<td>Total</td>
</tr>
<tr>
<td>Government</td>
<td>5</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Corporation</td>
<td>9</td>
<td>32</td>
<td>19</td>
</tr>
<tr>
<td>Aided</td>
<td>27</td>
<td>38</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>33</td>
<td>82</td>
</tr>
</tbody>
</table>
Table 4.3

Distribution of Higher Secondary School Teachers on the basis of sex and type of School

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Percentage</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>39</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>99</td>
<td>42</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>138</td>
<td>40</td>
<td>207</td>
</tr>
</tbody>
</table>

The total number of Higher secondary school teachers account for 345 out of which 138 (40 percent) are males and 207 (60 percent) are females. This distribution is more or less similar to the population figures which is 41 percent and 59 percent for males and females respectively. The following table shows the distribution of teachers in the Higher secondary school on the basis of sex and type.

Though the Directorate of School Education in Chennai has information regarding the total number of unaided and matriculation schools in the city, they do not have information regarding the number of male and female teachers in the unaided and matriculation schools. It was also observed that the number of male teachers in the private schools is less compared to female teachers. One main reason for the above situation is that the teachers in the private schools do not get an attractive pay; their pay scale is fixed by the private management. Most of the male teachers do not stay
permanently in such jobs where there is no security of job and the pay is also less. Women teachers on the other hand do not change their jobs frequently and generally in any developing country women are less mobile. Moreover, for most of the women, their earnings constitute only a secondary source of income to their family. Therefore, they prefer to work in such private schools which are relatively flexible enabling them to divide their time between the household work and teaching. The following table 4.6 shows the distribution of male and female teachers in the matriculation and unaided schools.

From Table 4.4 it is clear that the number of male teachers in the matriculation and unaided schools both in the high school level and higher secondary level are very low. 80 percent of the teachers in the unaided high schools are women. Their percentage in the matriculation high school in 76. The corresponding figures for the male teachers are 20 and 24 percent respectively.

Among higher secondary schools the percentage of women teachers in unaided and matriculation schools are 72 and 62 respectively while the percentage of men teachers are 28 and 38 respectively. Teaching profession is therefore apparently considered as "women's job".
Table 4.4

Distribution of Teachers in the Private Sector on the
Basis of Gender, School Level and School Type

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th></th>
<th>Female</th>
<th></th>
<th>Total</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Percentage</td>
<td>Total</td>
<td>Percentage</td>
<td>Total</td>
<td>Percentage</td>
</tr>
<tr>
<td>High School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaided</td>
<td>7</td>
<td>20</td>
<td>28</td>
<td>80</td>
<td>35</td>
<td>100</td>
</tr>
<tr>
<td>Matriculation</td>
<td>17</td>
<td>24</td>
<td>54</td>
<td>76</td>
<td>71</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>23</td>
<td>82</td>
<td>77</td>
<td>106</td>
<td>100</td>
</tr>
<tr>
<td>Higher Secondary School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaided</td>
<td>31</td>
<td>28</td>
<td>78</td>
<td>72</td>
<td>109</td>
<td>100</td>
</tr>
<tr>
<td>Matriculation</td>
<td>44</td>
<td>38</td>
<td>73</td>
<td>62</td>
<td>117</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>33</td>
<td>151</td>
<td>67</td>
<td>226</td>
<td>100</td>
</tr>
</tbody>
</table>

4.4 EMPIRICAL SPECIFICATION OF THE EARNINGS FUNCTION AND ESTIMATION ISSUES

The efficiency of the productivity of an individual is reflected in terms of the earnings he receive in a particular job. According to Mincer (1974) the two main factors that increase the stock of human capital are education and on the job training. In addition to these variables the present study includes the dummy variables for training, namely, M.Ed., and B.Ed., the family background variables, dummy variables for sex, sector and caste and the marital status.
The estimating equations of the wage functions for group m and group n are specified as,

\[ \ln W_{ij} = \beta_0 + \beta_1 P_{Gij} + \beta_2 E_{ij} + \beta_3 I_{Eij} + \beta_4 M_{Edij} + \beta_5 B_{Edij} + \beta_6 F_{EDUij} + \beta_7 I_{MEDUij} + \beta_8 I_{MSij} + \beta_9 I_{Miij} + \beta_{10} I_{PUBij} + \beta_{11} I_{FORij} \] \quad (19)

Where,

\[ \ln W = \text{logarithm of hourly wage of the ith group (}=m,n) \]

P.G. refers to dummy variable for post graduate level of education measure as years of completed education (reference group is U.G. degree).

E- years of labour market experience; \( E^2 \) experience square.

MEd., - Master of education, Dummy variable for training with no training as the omitted category.

B.Ed., - Bachelor of Education, Dummy variable for training with no training as the omitted category.

F.Edu- Father's education (in years)

M.Edu -- Mother's education (in years)

M.S – Marital Status measured as dummy variable with Not married as omitted category. This aspect is further discussed in section 4.6.

M-Dummy variable for male with female as the omitted category.

Pub- Dummy variable for public school with private school as the omitted category.
For- Dummy variable for Forward community with “Other castes” as the Omitted category.

For the present study, the earnings function specified is based on Mincer's post-school investment model.

4.5 DESCRIPTION OF VARIABLES

The logarithm of hourly wages is used as the dependent variable. The hourly wages of a teacher is calculated from the annual earnings (Gross) and the total number of hours worked in a year by the teacher.

\[
\text{hourly wage of a teacher} = \frac{\text{Gross annual earnings}}{\text{Total number of hrs/annum}}
\]

Annual earnings are calculated by multiplying monthly earnings by twelve. No adjustments are made for taxes. Annual hours worked are given by the product of weekly hours of work and the number of weeks worked in a year. The dependent variable is expressed in natural logarithms, as the semi-logarithmic specification of the earnings function, is stated to have a better statistical fit than the one based on absolute earnings (Dougherty and Jimenez, 1991) (Duraisamy & Duraisamy, 1997).

The independent or explanatory variables that are used in the alternative specification of the earnings function are: the level of education, actual experience, experience square, training, family background variables like the education of the parents', marital status and the sex dummy, sector
dummy and caste dummy variables. A brief note on these variables are given below. The explanatory variables and their mean values are given in Table 4.5.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Mean</th>
<th>S.D</th>
</tr>
</thead>
<tbody>
<tr>
<td>In W - logarithm of hourly wage</td>
<td>3.8315</td>
<td>0.7166</td>
</tr>
</tbody>
</table>

**Table 4.5**

**Definition and Measurement of Variables**

**Independent Variables**

<table>
<thead>
<tr>
<th>Education Level of education dummy variables</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P.G. Master or other equivalent degree holders.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If the individual holds post graduate degree = 1; 0 otherwise</td>
<td>0.6063</td>
<td>0.4839</td>
</tr>
<tr>
<td>U.G. Bachelors degree (the reference category)</td>
<td>0.3937</td>
<td>0.4889</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Work Experience</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp - Total experience (in years)</td>
<td>12.87</td>
<td>9.68</td>
</tr>
<tr>
<td>Exp.Sq. - Total experience square</td>
<td>259.0788</td>
<td>329.9028</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Training Dummy Variables</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M.Ed. Master of Education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If the individual holds M.Ed. = 1; 0 otherwise</td>
<td>0.2425</td>
<td>0.4289</td>
</tr>
<tr>
<td>B.Ed. Bachelor of education</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N.T No Training (Omitted category)</td>
<td>0.1100</td>
<td>0.3131</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family Background Variables</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Father Father's Education (in years)</td>
<td>14.81</td>
<td>1.79</td>
</tr>
<tr>
<td>Mother Mother's Education (in years)</td>
<td>11.58</td>
<td>1.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital Status</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Marry Married</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If the individual is married = 1; 0 otherwise</td>
<td>0.8913</td>
<td>0.3115</td>
</tr>
<tr>
<td>N_marry Not married (the omitted category)</td>
<td>0.1088</td>
<td>0.3115</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discrimination variables</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex dummy Male</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If the individual is male = 1; 0 otherwise</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>Female Female (the omitted category)</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>Sector dummy variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Public - Public school</td>
<td>0.58</td>
<td>0.49</td>
</tr>
<tr>
<td>If the individual is working in public school = 1; 0 otherwise.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private - Private school (the basic category)</td>
<td>0.41</td>
<td>0.49</td>
</tr>
<tr>
<td>Caste Dummy Forward - Forward community</td>
<td>0.5087</td>
<td>0.5002</td>
</tr>
<tr>
<td>If the individual is forward = 1; 0 otherwise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others - Other community (the basic category)</td>
<td>0.4913</td>
<td>0.5002</td>
</tr>
</tbody>
</table>

Education increases the earnings capacity since it is a potential factor that increases the human capital in an individual. While the signaling hypothesis states that education does not directly enhance the productivity (Spence, 1974), however the positive association between education and earnings for both sexes can not be denied. In this study education is measured as a dummy variable. Since the minimum educational qualifications required for a high school and higher secondary school teacher are a graduate degree and post-graduate degree respectively, education variable has only two levels in our study. Higher the level of education, higher is the expected earnings. This variable in general is expected to have a positive sign, since more education will fetch more earnings for a teacher.

Experience refers to the actual years of experience gained in the current job, after the completion of schooling. Post-schooling experience has been observed to influence wages positively. To capture any concavity in the experience - wage profile, an experience squared term is introduced. The experience square term is expected to have a negative sign.
Training in our study refers to professional or teacher training. Two levels of training are considered in this study. One is called the Bachelor of education (B.Ed.) and the other, Master of education (M.Ed.). B.Ed is generally a two-year training course (recently shortened to one year), undertaken after a Bachelor's degree. The basic qualification for a secondary school teacher in a public school is a Bachelor's degree with B.Ed. Though this is not essential in private schools, a teacher with a B.Ed. or M.Ed. is preferred to the one without any training. M.Ed. is one-year training course for those who have already completed B.Ed. course. However, in all public and many private schools, for entry at the higher secondary level a teacher must possess a minimum of B.Ed. degree.

The influence of family background variables are also important because earnings of individuals with a given level of education vary depending on the parental background of individuals. The parental background or the socio-economic status of a family is usually measured by father's education, occupation and/or earnings as well as mother's education, occupation and/or earnings. In this study, father's education and mother's education(in years) represent the parental or the family background of the teacher.

Married women with children have family commitments that make it difficult for them to pursue career aspirations whereas married men with the same attribute are more ambitions and are able to focus better on their career because there is a spouse to take care of the house. As a result of the family role specialisation, marriage and career became complementary for men and competitive for the women (Greenhalgh, 1980). The employers then perceive
marital status as a positive indicator of productivity for men and the opposite for the women (Polachek, 1975). Thus the expected signs on the coefficient of the marital status dummy in the earnings function is positive for men and negative for women. If one were to ignore the opposite effects of marital status on earnings and include it in the earnings specification, it would largely overstate the amount of discrimination (Malkiel and Malkiel, 1973).

In most foreign countries, part-time jobs are readily available, hence when a woman gets married and is tied down with young children she would switch over to a part-time job that would lower her earnings. Here in India, labour market conditions and family formations are different; part-time jobs are not so common. If a working woman leaves a job upon marriage she would have to most probably withdraw from the labour force. However, this is not necessarily the alternative for the Indian women who in all probability would have an elderly relative, the parents or in-laws to take care of her family, which is unlikely in the foreign households. In this context, it is of research interest to examine the effect of marital status for the two groups (group ‘m’ and group ‘n’) in the Indian context. An earnings specification including marital status is also considered in the present study and it is measured as a dummy variable with the unmarried as the reference category.

In addition to the above mentioned human capital variables we have included as potential determinants of earnings, a number of variables viz., sex, sector of employment, and caste. Sex is measured as dummy variable with females being the reference group. Smith (1978), Gunderson (1979), and others have revealed that the sector of employment exerts a considerable effect on
earnings differential. Public sector employees are assumed to enjoy non-monetary benefits like security of employment, pension schemes etc. and hence accept a lower salary in monetary terms, than their private sector counterparts. It is also possible to argue that public sector does not operate on the basis of profit considerations and it is likely that public school teachers are paid more than private school teachers. *A priori*, it is difficult to say whether public school teachers get a relatively higher or lower pay and this has to be empirically tested. Therefore sector of employment is entered as a dummy variable with the public sector taking the value one and the private sector zero.

Indian society is segmented by caste. It significantly influences the education and employment activities of the individuals and families. Since independence, the centre and state governments have been taking several affirmative actions to improve the economic conditions of the persons from the lower caste such as the scheduled caste, scheduled tribes and backward classes. The governments adopt the caste reservation policy in the recruitment of workers for the government and public enterprises. 67 percent of the jobs are reserved for the backward class and SC/ST groups. Studies on racial discrimination in the United States indicate the prevalence of wage gap between white and the black workers. Unfortunately the wage differential among communal group is not given priority in most of the studies for India. Hence it would be interesting and useful to study the wage gap between the different community groups in the teacher's labour market. There are two main group of castes; 'Forward caste' and the 'other caste'. The 'other caste' includes teachers belonging to the backward community, 'most backward
community and the SC/ST groups. We measure this variable dichotomously with 'other caste' being the left out category.

Following the statistical methodology, outlined here, separate wage regressions are estimated for male female teachers and for the combined sample of male and female teachers. Three specifications are tried in order to estimate the wage differentials in the teacher's labour market. The first specification (Model I) includes the human capital productivity variables like the level of education, experience and its square term. The second specification (Model II) includes more control variables, viz., training (M.Ed. and B.Ed.) with 'no training' as the omitted category. These two levels of training are expected to capture the premium earned by the teacher due to training over the omitted category. This model also includes the effect of the 'family background' variables measured by the father's and mother's education in terms of years. These above two variables, viz., training and the parental education are expected to exert a positive influence on the teacher's earnings in general. The third specification (Model III) includes dummy variables for sex, sector and community. The wage functions are estimated separately for male and female sub samples and also for the pooled sample.

ESTIMATION ISSUES

In estimating the earnings function one has to encounter the following issues. The first issue is with regard to selectivity bias. The sample data consisting of men and women teachers do not represent all population. This is because only those who reported wages are included in the sample. This may
ead to bias in sample selection. Several methods are suggested in the literature to overcome these shortcomings. One approach is to estimate the reduced form equations with instruments for wage. Some studies have restricted the estimation of the wage function to working women and use a predicted instead of observed wage for them while others estimate the wage equation for all women using predicted wage based on wage equation estimates from working women subsample. These procedures in turn lead to selectivity bias in wage equation. If this bias is ignored, then the estimates of wage equation introduce biased estimates of the parameters of wage equation.

Two approaches are followed to correct for sample selectivity bias. One is the multistage estimation method proposed by Heckman (1980) in which a participation decision equation is first estimated by probit, bivariate probit or logit depending on the underlying assumptions about the error term in the equation. If it is assumed that the disturbance term is normally distributed, then the maximum likelihood probit method is used to obtain the estimates of the inverse of Mills ratio which is introduced in wage equation as an explanatory variable and is estimated by OLS method. Inclusions of this variable then ensures that the coefficient produced by the wages model are unbiased and consistent which increases the reliability of the findings. The other method suggested in literature is the full information maximum likelihood (FIML) method.

Since the objective of this study is to estimate the wage differentials in the teacher’s labour market in the urban city of Chennai, the collection of
primary data was confined to one occupation, viz., teaching. The entire teaching population is in the paid employment (both gender) and report a wage rate. Due to the non-availability of information of the wages of non-teachers it is not possible to test for selectivity bias in our sample.

The next issue is concerned with the functional form of the earnings function. The dependent variable in the earnings function can be expressed either in natural logarithm or in absolute values. Most of the studies use a semi-logarithmic specification of the human capital earnings function proposed by Mincer (1974), that is the dependent variable is expressed in logarithmic form and the independent variables in non-log form. James Heckman and Solomon Polachek (1974) and more recently Christopher Dougherty and Emmanuel Jimenez (1991), using the Box-Cox transformation test being semi-logarithmic form against the linear form of the earnings function. The interesting feature of this transformation is that the parameter \( \lambda \) which is estimated along with \( \beta \)'s and \( c2 \). If the estimated \( \lambda \) is equal to one then the earning function is linear and if the \( \lambda \) is equal to zero then the appropriate form is semi-logarithmic. Since this functional form as been widely used in the empirical studies an attempt is not made to test its validity in the present study.

The next issue is with regard to the functional form of the relationship between the regressand and the regressor. The units and scale in which the regressand and the regressor are expressed is very important because the interpretation of the regression coefficient critically depends on them. Some of
the important functional forms include the log-linear or constant elasticity model, semi-log regression model and the reciprocal models. In choosing the various functional forms great attention should be paid to the stochastic disturbance term. In the linear regression models it is explicitly assumed that the disturbance term has zero mean and constant (homoscedastic) variance and that it is uncorrelated with the regressor. Under these assumptions the OLS estimates are the best liner unbiased estimates. Further the OLS estimators are normally distributed in such models. Since it is difficult to collect the absolute values of the investment variables if log earnings are used, as done in this study, the investment variables can be expressed in unit of time - years of schooling and training. The time measures are more readily available than absolute values of investment in human capital. The Semi-logarithmic function is less skewed and less heteroscedastic when applied and the formulation fits the earnings data better than linear function or other power functions of earnings (Heckman & Pollachek, 1974 and Thurow, 1969). More over when the dependent variable is expressed in logarithmic form its variance gives a direct measure of the inequality.

The validity of the semi logarithmic form against the linear form of the earnings function is tested by many researchers using the Bo-Cox transformation. The Box-Cox transformation of the dependent variable takes the following form (J Jonston 1989):
\[ W_{ij}(\lambda) = \begin{cases} \frac{(W_{ij} - 1)/\lambda}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(W_{ij}) & \text{for } \lambda = 0 \end{cases} \]

The functional form is dictated by the parameter \( \lambda \) which is estimated along with \( \beta \)s and \( \sigma^2 \) (variance of the error term). If the estimated \( \lambda \) is equal to 1 then the earnings function is Linear and if the \( \lambda \) is equal to zero then the appropriate function is semilogarithmic. Since the semi logarithmic function is the most popular method, it has been adopted in this study.

The next issue is with regard to the problem of simultaneity. In contrast to single-equation models, in simultaneous-equation models more than one dependent or endogenous variable is involved, necessitating as many equations as the number of endogenous variables. A unique feature of simultaneous-equation models is that the endogenous variable (i.e., regressand) in one equation may appear as an explanatory variable (i.e., regressor) in another equation of the system. As a consequence, such an endogenous variable becomes stochastic and is usually correlated with the disturbance term of the equation in which it appears as an explanatory variable. This leads to the simultaneous equation bias. In this situation the classical OLS method may not be applied because the estimators thus obtained are not consistent: that is they do not converge to their true population values no matter how large the sample size is.

If in a system of simultaneous equations containing two or more equations if it is not possible to obtain numerical values of each parameters in
each equation because the equations are observationally indistinguishable then we have the identification problem. It is therefore essential to resolve the identification problem before we proceed to estimation because if we do not know what we are estimating, estimation per se is meaningless.

Assuming that an equation in a Simultaneous equation model is identified there are several methods to estimate it. These methods fall into two broad categories: single equation methods and systems methods. Three commonly used single equation methods are Ordinary least squares (OLS), Indirect Least Squares (ILS) and Two Stage Least Squares (2SLS). Although OLS is in general inappropriate in the context of simultaneous equation model it can be applied to the so called recursive models where there is a definite but uni-directional cause-and-effect relationship among the endogenous variables. The method of ILS is suited for just or exactly identified equation. In this method OLS is applied to the reduced for equation and it is from the reduced form coefficients that one estimates the original structural coefficients. The method of 2SLS is specially defined for over identified equations although it can also be applied to exactly identified equation. But then the results of 2SLS and ILS are identical. The basic idea behind 2SLS is to replace the (Stochastic) endogenous explanatory variable by linear combination of the predetermined variables in the model and use this combination as the explanatory variable in lieu of the original endogenous variable. The 2SLS method is thus resembles the instrumental variable method of estimation in that the linear combination of the predetermined variable serves as an instrument or proxy for the endogenous regressor. A noteworthy feature of both ILS and 2SLS is that the
estimates obtained are consistent that is as the sample size increases indefinitely the estimates converge to their true population values. The estimates may not satisfy small sample properties such as unbiasedness and minimum variance. Therefore the results obtained by applying these methods to small samples and the inferences drawn from them should be interpreted with due caution.

Marital status is an endogenous variable which is used as an explanatory variable in the regression equation. Since there was no instrumental variables suggested in the literature for marital status, I could not find a suitable variable for it and therefore marital status was used as an explanatory variable in the earnings function for male and female teachers.