RESEARCH DESIGN AND METHODOLOGY

3.1 Objectives and hypotheses

In the present study, scheduled caste population of India has been selected for an enquiry into the socio-economic status of scheduled caste population in the same and particularly in Allahabad division of Uttar Pradesh. The study is based on both ‘Primary’ and ‘Secondary’ Data, the ‘Primary’ data is taken through the ‘Field Survey.’ However, ‘Secondary Data’ is also used during the analysis from various government and non-government sources like various ministries of GOI, IHDS, NSS etc. The details of the objectives, hypotheses, methodology, and methods adopted can be given as below:

Objectives:

➢ To determine the extent of discrimination in earnings and dissimilarity of occupation for scheduled castes in India.
➢ To determine the changes in the status of caste, the extent of untouchability against rural SCs in Allahabad division (Pollution purity issue).
➢ To analyse the socio-economic condition of scheduled castes in Allahabad division in particular and India in general.
➢ To analyse the role of education and occupation in overall mobility of scheduled castes in Allahabad division and India.
➢ To examine the association between migration decisions and its determining factors for male SC workers in Allahabad division.
Hypotheses:

H01: There is no pre or post labour market discrimination in earnings for scheduled castes in India.

HA1: There is pre or post labour market discrimination in earnings for scheduled castes in India.

H02: There is no occupational dissimilarity exist for scheduled castes in India.

HA2: There is occupational dissimilarity exist for scheduled castes in India.

H03: The socio-economic condition of scheduled castes is independent of their perception of caste, discrimination, and status (based on various socio-economic indicators).

HA3: The socio-economic condition of scheduled castes is dependent on their perception of caste, discrimination, and status (based on various socio-economic indicators).

H04: Earnings and economic betterment of scheduled castes are independent of socio-economic indicators.

HA4: Earnings and economic betterment of scheduled castes are dependent of socio-economic indicators.

H05: Migration decision is not associated with the value of the measurement variables (various socio-economic and demographic variables) like; age, level of education, income, possession of skill etc.

HA5: Migration decision is associated with the value of the measurement variables like; age, level of education, income, possession of skill etc.

Note: There was response lag of at least ten years (or one generation) to compare the present conditions of the scheduled castes.
3.2 Research methodology

3.2.1 Primary data collection: Selection of the area

The study analysed the socio-economic condition of SCs with reference to ‘Allahabad’ Division in ‘Uttar Pradesh.’ The SC population of Uttar Pradesh was 4,13,57,608 (Census of India, 2011), constituting 20.7% of the total population of the state. Uttar Pradesh holds 1st rank in terms of an absolute number of SC populations among all the States and UTs. The State has 66 sub-castes of scheduled castes; all of them have been enumerated at 2011 Census (Appendix IV). The overall sex ratio of the SC population in Uttar Pradesh was 908 females per 1000 males, which was very low compared to the national average sex ratio of 945 for all SCs in 2011. The overall effective literacy rate for SCs in Uttar Pradesh was 60.9% in 2011, which was also lower than the national average of 66.1% for SCs. For female SCs in Uttar Pradesh, the effective literacy rate was 48.9%, which was very low when compared to national level rates of literacy (56.5%) for SCs (Census of India, 2011).

The Scheduled Castes are predominantly rural in Uttar Pradesh as 86.3% of them live in villages. District–wise distribution of SC population shows that they have the highest concentration in percentage terms in Kaushambi (34.72%), followed by Sitapur (32.26%) and Hardoi (31.14%) districts. Baghpat has the lowest proportion of SC population (11.44%) in UP.

<table>
<thead>
<tr>
<th>Place</th>
<th>Total Population</th>
<th>Number of Households</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rural</td>
<td>Urban</td>
</tr>
<tr>
<td>Allahabad</td>
<td>11,08,075</td>
<td>2,01,776</td>
</tr>
<tr>
<td>Kaushambi</td>
<td>5,32,594</td>
<td>22,803</td>
</tr>
<tr>
<td>Pratapgarh</td>
<td>6,91,199</td>
<td>18,053</td>
</tr>
<tr>
<td>Fatehpur</td>
<td>6,04,239</td>
<td>47,241</td>
</tr>
<tr>
<td><strong>U.P.</strong></td>
<td>3,56,85,227</td>
<td>56,72,381</td>
</tr>
</tbody>
</table>

Source: Compiled from Primary Census Abstract CDB-F 2011
The proportion of the SCs to the total population in the four districts of Allahabad Division was Kaushambi (34.72%), Fatehpur (24.7%), Pratapgarh (22.1%) and Allahabad (22%). Allahabad was most populated (13,09,851) among these districts of the division. The percentage of SCs in the rural areas in these districts was 95.9%, 92.7%, 97.5% and 84.6% respectively in 2011.

The study has selected 'Allahabad' Division in 'Uttar Pradesh' for the study. The study purposively has considered all the four districts of the division viz., Allahabad, Kaushambi, Pratapgarh and Fatehpur, which are dominating in terms of population, Allahabad, having the largest population of SCs in Allahabad division and Kaushambi, having the highest concentration in percentage terms (34.72%) in UP.

Allahabad Division has got distinguished historical background on the social and economic point of view, which can be found very rarely in other parts of the state. Its sanctity is manifest by references to it in Puranas, the Ramayana and the Mahabharata and a place of mixed culture. Most of the studies reviewed showed the status of the Scheduled Castes (SCs) some time ago. Hence, there is need to revisit and consider the present status of the SCs in India in general, and in Allahabad Division in particular. This study all the more necessary since no review could be found in the context of Allahabad Division.

The scope of the study is limited to the rural development blocks of both District, as classified by the Primary Census Abstract 2011. The study did not consider urban areas of the same because of low representation of SC urban population in Allahabad, Kaushambi, Pratapgarh, and Fatehpur, comprising only 15.4%, 4.1%, 2.5% and 7.3% population respectively in urban area. If would have been taken Kuashambi district would get only as many as 5 respondents from urban areas and Pratapgarh would get only 3 respondents from the same, this might showed inadequate results because of small samples.

The primary data was collected through a field survey (the respondents were the SCs only) which used structured household schedule prepared by various checks and pilot
surveys. The data pertain to the **reference period of March to August 2016**. The schedule was designed to capture the details on social, economic, demographic variables; life condition, the extent of untouchability, migration, occupation etc. The schedule has questions on different socio-economic variables and inter-generational aspect was also kept for a better understanding of change in perception, migration decisions and chronic deprivation (Appendix V).

The procedure of random sampling was followed to select one block from each district and two villages were selected from each block. The total sample size was 600 households to adhere to 5.0% margin of error. By proportionate random sampling, for 600 sample size, 213 respondents from ‘Allahabad’ district, 107 from ‘Kaushambi’ district, 147 from ‘Pratapgarh’ district and 133 from ‘Fatehpur’ district were randomly selected according to their rural population ratio of 1.47:0.8:1.1:1 respectively.

The study followed the 'Systematic Random Sampling' method to select the households. Selection of the households on the basis of either 'Occupation' or 'Income' like indicators could have been misleading analysis, hence the study preferred to select all the SC households based on random sampling in order to give realistic analysis to the study. Ultimately, the study followed Multi stage mixed sampling to choose the sample.

**Selection of the development blocks**

The 'Development Blocks' from the districts have been selected through the lottery method. Since the majority of the Scheduled Castes population was found in the rural area so the inter-relationship between socio-economic indicators can be stressed more effectively in rural areas, the study considered only 'rural' blocks of four districts of Allahabad division. The 'Primary Census Abstract 2011' has classified 20 Development Blocks in ‘Allahabad’ District, 8 in ‘Kaushambi’ District, 17 in ‘Pratapgarh’ District and 13 in ‘Fatehpur’ District. Out of these 20 classified rural blocks from ‘Allahabad’ districts, the study selected one block namely ‘Holagarh’ by lottery method. With the same method, one block namely ‘Kaushambi’ from ‘Kaushambi’ district, one block namely ‘Gaura’ from ‘Pratapgarh’ district and one block namely ‘Telyani’ has been selected from ‘Fatehpur’ district.
Blocks which have been selected namely, 'Holagarh' from Allahabad district, 'Kaushambi' from Kaushambi district, ‘Gaura’ from Pratapgarh district and ‘Telyani’ from Fatehpur district have a population of 47482, 71305, 41615 and 44786 respectively (Table 3.2).

Selection of the villages

**First-stage units:** A total of eight villages were surveyed in rural areas of Allahabad division as first-stage units in the scheduled study. The eight villages have been selected through the lottery method. According to this methodology, the study selected two villages from each development block in four districts.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Rural blocks</th>
<th>Total Population</th>
<th>SC Population</th>
<th>% of SC population to total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Holagarh</td>
<td>1,82,703</td>
<td>47,482</td>
<td>26.0</td>
</tr>
<tr>
<td>2</td>
<td>Kaushambi</td>
<td>1,71,933</td>
<td>71,305</td>
<td>41.5</td>
</tr>
<tr>
<td>3</td>
<td>Gaura</td>
<td>2,12,537</td>
<td>41,615</td>
<td>19.6</td>
</tr>
<tr>
<td>4</td>
<td>Telyani</td>
<td>1,44,574</td>
<td>44,786</td>
<td>31.0</td>
</tr>
</tbody>
</table>

Source: Compiled from Primary Census Abstract CDB-F 2011

**Second-stage units:** A respondent has been chosen from a household through systematic random sampling on the basis of the total SC households in the particular village.

In this exercise the study found two villages, namely, 'Girdharpur Gondwa' (1,483) and 'Sarai Madansingh urf Chanti' (883) villages in 'Holagarh' block who have population ratio of 1.7:1, dividing the sample of 213 into 134 and 79 respectively from Allahabad District (Table 3.3).

The same methodology has been applied to other districts, the study found two villages, namely, 'Kosam Inam Uparhar' (1,421) and 'Kosam Khiraj' (2,019) with a population ratio of 0.7:1, having the sample of 44 and 63 respectively in ‘Kaushambi’ block. Two villages namely, Sahpur (950) and Sigahi (700) with population ratio of 1.4:1, having the...
sample of 86 and 61 respectively in ‘Gaura’ block and two villages namely ‘Baijani’ (686) and ‘Sahli’ (503) with population ratio of 1.4:1, having the sample of 78 and 55 respectively in ‘Telyani’ block (Table 3.3). The total number of households in which the schedule was canvassed, was 600 in rural Allahabad division. Village wise number of households in which schedule was canvassed is given in table 3.3.

Table 3.3: Sample size

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Village</th>
<th>Household (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Girdharpur Gondwa</td>
<td>134 (22.33)</td>
</tr>
<tr>
<td>2</td>
<td>Sarai Madansingh urf Chanti</td>
<td>79 (13.17)</td>
</tr>
<tr>
<td>3</td>
<td>Kosam Inam Uparhar</td>
<td>44 (7.33)</td>
</tr>
<tr>
<td>4</td>
<td>Kosam Khiraj</td>
<td>63 (10.50)</td>
</tr>
<tr>
<td>5</td>
<td>Sahpur</td>
<td>86 (14.33)</td>
</tr>
<tr>
<td>6</td>
<td>Sigahi</td>
<td>61 (10.17)</td>
</tr>
<tr>
<td>7</td>
<td>Baijani</td>
<td>78 (13.00)</td>
</tr>
<tr>
<td>8</td>
<td>Sahli</td>
<td>55 (9.17)</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td><strong>600 (100)</strong></td>
</tr>
</tbody>
</table>

Source: Compiled from PCA, CDB – F, 2011. Note: Figures in parentheses are in percentage of total sample size.

3.2.2 Secondary data source

For the purpose of the study, it utilises individual data from the nationally representative unit level secondary data of India Human Development Survey (IHDS). In the first wave of fieldwork (2004-05), a sample of 41,554 households was surveyed enumerating 2,15,754 individuals and in the second wave (2011-12), a sample of 42,152 households was surveyed enumerating 2,04,569 individuals, of which most of the household reinterviewed. The IHDS is a nationally representative survey of the population across all states and Union Territories of the country except the Andaman and Nicobar Islands and the Lakshadweep. The survey sampling and more information about the survey may be found in Desai et al. (2010). The survey has detailed demographic information (e.g., age, gender, marital status, household size, religion, social group, sector, and place of residence) with socio-economic position (e.g., land ownership, educational attainment,
occupation and industry, type of job, and wages and earnings) among several other characteristics.

Appropriate sampling and frequency weights have been used during the analysis of secondary data because it was necessary for obtaining accurate and generalised point estimates. The IHDS had already constructed sampling weights based on multiple stages of sampling, non-response, post-stratification etc. and use of weights resolved the problem of over and under representation of some or the other groups. The descriptive statistics and definition of the variables used in the whole study are given in Appendix II & III.

The analysis of occupational dissimilarity is based upon the secondary source of data. The study used data from four successive rounds of NSS Employment and Unemployment Situation among social groups (EUS) in India from 1993-94 to 2009-10 to analyse patterns of occupation choices and level of occupational segregation for both SC/ST and non-SC/ST households. The study uses *Duncan’s Dis-Similarity Index* to know the level of occupational segregation in both rural and urban areas. It uses the latest EUS of 2011-12, to identify the present situation of social groups in terms of poverty ratios, educational attainment, literacy rates and Worker Population Ratio (WPR). The main findings of the study are presented in chapter 4 and tables used in the analysis are given in Appendix VI.

Here the study has taken only four levels of education, viz. not literate, higher secondary, diploma/certificate, and graduates.

The work force, according to the usual status (ps+ss), includes persons who (i) either worked for a relatively longer part of the 365 days preceding the date of survey and (ii) also those persons from among the remaining population who had worked at least for 30 days during the reference period of 365 days preceding the date of survey. The number of persons employed per 1000 persons is referred to as work-force participation rate (WFPR) or worker population ratio (WPR).
3.2.3 Econometric tools and techniques applied

For the analysis of data various statistical tools and methods have been used, e.g., simple percentage method, Karl Pearson’s Correlation Coefficient (r), chi square test of association of attributes, simple linear regression analysis, logistic regression analysis, various decomposition techniques like Oaxaca Blinder, Oaxaca Ransom, Cotton/Neumark decomposition and Duncan Dissimilarity Index methods which are discussed below in detail. Other than that probability distribution plots and cumulative distribution plots have been used wherever necessary. Appropriate modeling (equations) based on objectives of the study has been given in respective chapters of the data analysis.

Chi square test

The statistical tool of chi-square test has been used for the analysis. One can use the Pearson’s chi-square test if one wishes to see the association between two categorical attributes (Fisher, 1922; Pearson, 1900). This statistic is based on the simple idea of comparing the frequencies we observe in certain categories to the frequencies you might expect to get in those categories by chance the relation of statistical summary of data will be defined to visualise the data. The two basic assumptions of the test are:

- The assumption about the independence of data. For the chi square test to relevant, it is necessary that each person, item or entity contributes to only one cell of the contingency table. Therefore, you cannot use a chi-square test in a repeated-measures design.

- The expected frequencies should be greater than 5. Although it is acceptable in larger contingency tables to have up to 20% of expected frequencies below 5, the result is a loss of statistical power (so, the test may fail to detect a genuine effect). Even in larger contingency tables, no expected frequencies should be below 1 (Field, 2009).

Pearson’s chi-square test examines whether there is an association between two categorical attributes. The Pearson chi-square tests whether the two attributes/variables
are independent. If the calculated value of chi statistics is more than the table value then we reject the hypothesis that the variables are independent and gain confidence in the hypothesis that they are in some way related. Chi-square statistic can be calculated as follows;

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]

Where \( \chi^2 \) is chi square statistics, \( O_i \) is observed values and \( E_i \) is expected values.

The degree of freedom for the chi square statistics is \((r-1) \times (c-1)\), where \( r \) is the number of rows and \( c \) is the number of columns in the contingency table of attributes.

**Simple percentage method**

This method is mainly used for obtaining different rates from the absolute figure of respondents for different variables. On the basis of which trends and patterns of employment are obtained. This calculation has also provided the basis for the application of higher statistical techniques like Karl Pearson’s Correlation Coefficient \((r)\), linear regression and chi square tests.

\[ S.P. = \left( \frac{N^{th}}{n} \right) 100 \]

Where,

\( S.P. \) = simple percentage.

\( n \) = the numerical value of the particular group

\( N^{th} \) = the part of that group \( n \).

**Correlation analysis**

Correlation is a statistical technique used for finding a relationship between two variables. The range of correlation coefficient is between -1 to +1, where negative and positive values show negative and positive correlation respectively between any two variables. In the condition when all actual values are on the regression line, the prediction is exact and the relationship between the two variables is perfect. When it is not so then
the relationship between the variables is not perfect, and the correlation coefficient may lie between \(-1\) and \(+1\).

**Multiple correlation coefficients**, for linear association among all the variables both dependent and independent, it is always greater than any simple correlation expressing the degree of linear association between the dependent variable \((Y)\) and any of the independent variable (the X’s). The multiple correlation coefficient \((r)\) based on Karl Pearson’s method is given below:

\[
r = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sqrt{\left(\sum X_i^2 - n \bar{X}^2\right) \left(\sum Y_i^2 - n \bar{Y}^2\right)}}
\]

Where, \(r\) = co-efficient of correlation \(X, Y\) = the two given variables, \(n\) = Number of observations.

**Linear regression analysis**

Simple or single equation linear regression analysis is concerned with the study of the dependence of one variable, the *dependent variable*, on one or more other variables, the *explanatory variables*, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter *(Gujarati et al., 2013)*. The regression analysis is different from the correlation analysis in a manner that in regression analysis, there is an asymmetry in the way the dependent and explanatory variables are treated. The dependent variable is assumed to be statistical, random, or stochastic, that is, to have a probability distribution. The explanatory variables, on the other hand, are assumed to have fixed values (in repeated sampling). On the other hand, in correlation analysis, we treat any (two) variable symmetrically and there is no distinction between the dependent and explanatory variables.

If we are studying the dependence of a variable on only a single explanatory variable, the study will be known as *simple, or two-variable, regression analysis*. However, if we are
studying the dependence of one variable on more than one explanatory variable, it is known as **multiple regression analysis**. The linear regression analysis (linear in parameters) can be described by regression equations as follows;

The linear regression model describes how the dependent variable is related to the independent variable(s) and the error term:

\[ Y_i = X_i' \beta + \varepsilon_i \]

Where \( E(\varepsilon_i) = 0 \) and \( i \in \{(\text{any individual or group})\} \)

Where \( X' \) is a vector containing the predictors and a constant, \( \beta \) contains the slope parameters and the intercept, and \( \varepsilon_i \) is the error term.

To calculate the predicted values of the dependent variable using the values of the independent variable(s), the estimated regression equation is being found as follows;

\[ \hat{Y} = X_i' \hat{\beta} \]

Where hat (\( \wedge \)) means estimated values of the same. Interpretation of the coefficients is like; one unit increase in \( X \) will increase the dependent variable \( Y \) by \( \beta \) units. It is important to note that there is no error term when we predict the value of the depended variable.

It has been hypothesised that corresponding slopes coefficients for regressors are not significantly different than zeros and the alternative hypothesis is that the coefficients are significantly different from zero as:

\[ H_0: b_j = 0 \quad & \quad H_A: b_j \neq 0 \]

**The least squares method (OLS: ordinary least squares)**

The least squares method is used to calculate the coefficients so that the errors are as small as possible.

We minimise the sum of squared residuals:
Goodness of fit

The coefficient of determination (R-squared or R\(^2\)) provides a measure of the goodness of fit for the estimated regression equation.

\[ R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \]

Where SSR = sum of squares due to regression, SST = sum of squares total and SSE = sum of squares due to error.

The values of R\(^2\) close to 1 indicate perfect fit and values close to zero indicate a poor fit. R\(^2\), that is greater than 0.25 is considered good in the field of economics.

R\(^2\) always increases when a new independent variable is added. This is because the SST is still the same but the SSE declines and SSR increases. To correct this problem, the concept of Adjusted R-squared is used that corrects for the number of independent variables and is preferred to R-squared.

\[ \text{Adjusted } R^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \]

Where p is the number of independent variables, and n is the number of observations.

F-test for overall significance of all coefficients

Testing whether the relationship between y and all x variables is significant or not the F value is being used as overall significance of all coefficients. Here the hypothesis is that:

H0: The coefficients are not jointly significantly different from zero and

H1: The coefficients are jointly significantly different from zero.

The test statistic is;
F = MSM/MSE
Where MSM = Mean of squares for model and MSE= Mean of squares for error

The critical values are from the F distribution and it is an upper one-tail test.

Classical linear regression model (CLRM) is used with certain assumptions like; model should be linear in parameters, explanatory variables must be exogenous, normally distributed error term with zero mean and constant variance, no multicollinearity, no heteroskedasticity and no serial correlation should be there to get best linear unbiased estimates (BLUE).

**Logistic model (binary outcome models)**

The study also uses logistic regression analysis and as we know in this model we should have a dichotomous dependent variable. Since the regressand is dichotomous, we cannot predict a numerical value for it using logistic regression, so the usual ordinary least square (OLS) regression deviations criterion for the best-fit approach of minimising error around the line of best fit is inappropriate. For that purpose, logistic regression employs binomial probability theory in which there are only two values to predict: that probability (p) is 1 or 0, i.e. the event/person belongs to one group or the other.

\[ P(L_i = 1) = \Phi(X_i\beta) \]

Where \( L_i \) is a set of dummies measuring individual \( i \) occupational status/economic betterment/migration etc., the vector \( X_i \) denotes individual and household characteristics, \( \beta \) is a vector of parameters to be estimated, and \( \Phi \) is the standard normal/logistic cumulative distribution function.

Logit and probit models are estimated using the maximum likelihood method. The coefficients are interpreted as; an increase in X increases/decreases the likelihood that outcome \( Y=1 \) (makes that outcome more/less likely). In other words, an increase in X makes the outcome of 1 more or less likely. Here we interpret the sign of the coefficient but not the magnitude. The magnitude cannot be interpreted using the coefficient because different models have different scales of coefficients.
Other than that, the independent variables in logistic regression need not be interval, nor
normally distributed, nor linearly related, nor of equal variance within each group, categories (groups) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one of the groups. It needs larger samples than the linear regression because maximum likelihood coefficients are large sample estimates.

The outcome of the regression is not a prediction of a dependent variable’s (Y) value, as in linear regression, but a probability of belonging to one of two conditions of Y, which can take on any value between 0 and 1 rather than just 0 and 1. Log transformation is used to normalise the distribution, which is of the p values to a log distribution enables us to create a link with the normal regression equation. The log distribution (or logistic transformation of p) is also called the logit of p or logit (p). Logit (p) is the log (to base e) of the odds ratio or likelihood ratio that the dependent variable is 1. In symbols, it is defined as:

\[
\text{logit} (p) = \log \left( \frac{p}{1 - p} \right) = \ln \left( \frac{p}{1 - p} \right)
\]

Whereas p lies between 0 and 1, logit (p) scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero). The form of the logistic regression equation is:

\[
\text{logit}(p(x)) = \log \left( \frac{p(x)}{1 - p(x)} \right) = a + X_i\beta
\]

And p can be calculated with the following formula:

\[
p = \frac{e^{a + X_i\beta}}{1 + e^{a + X_i\beta}}
\]

Where p = the probability that a case is in a particular category,

\(e\) = the base of natural logarithms, \(a\) = the constant of the equation and \(X_i\beta\) = the coefficients of the predictor variables.
The null hypothesis is that the probability of a particular value of the dependent variable is not associated with the value of the measurement variable (regressors); in other words, the line describing the relationship between the regressors and the probability of dependent variable has a slope of zero.

**The Oaxaca Blinder decomposition**

This approach parts the observed wage gap into two “endowment” and a “coefficient” component and called “decomposition techniques”. The “endowment” part is such that, it shows the component of wage differentials explained by individual ‘characteristics’ like education, age, and others. And the later part is derived as an unexplained residual and shows the component of wage differentials explained by ‘discrimination’. This method was first developed by Blinder (1973) and Oaxaca (1973), which is called The Blinder Oaxaca decomposition method. The details on how this decomposition is done may be seen there in their research papers mentioned earlier and Jann (2008). The interpretations are done accordingly to capture the separate effect of individual/household characteristics as ‘endowment’ and differences in the effectiveness of these characteristics as ‘discrimination’ in deciding wage gaps between groups.

The Blinder-Oaxaca decomposition technique can be explained as follows:

Here for this study, we have two groups general (GEN) and scheduled caste (SC) (any two groups may be taken here like; male and female to capture the gender wage differentials), an outcome variable, Y (log wages) and a set of variables like; education, age etc. Now the difference of mean outcome is to be computed:

$$D = E(Y_{GEN}) - E(Y_{SC})$$  \hspace{2cm} (3.1)

Where E (Y) denotes the expected value of outcome variable, is accounted for by the group differences in the regressors. The linear model is as follows:

$$Y_i = X_i' \beta_i + \varepsilon_i$$, where E ($\varepsilon_i$) = 0 and $i \in$ (GEN, SC)
Where \( X \) is a vector containing the predictors and a constant, \( \beta \) contains the slope parameters and the intercept, and \( \varepsilon_i \) is the error term, the mean outcome difference can be expressed as the difference in the linear prediction at the group-specific means of the regressors. That is,

\[
D = E(Y_{GEN}) - E(Y_{SC}) = E(X_{GEN})'\beta_{GEN} - E(X_{SC})'\beta_{SC}
\]  

(3.2)

Because

\[
E(Y_i) = E(X_i'\beta_i + \varepsilon_i) = E(X_i'\beta_i) + E(\varepsilon_i) = E(X_i'\beta_i)
\]

Where \( E(\beta_i) = \beta_i \) and \( E(\varepsilon_i) = 0 \) by assumption

To identify the contribution of group differences in predictors to the overall outcome difference, (3.2) can be rearranged, for example, as follows:

\[
D = \{E(X_{GEN}) - E(X_{SC})\}'\beta_{SC} + E(X_{SC})'(\beta_{GEN} - \beta_{SC}) + \{E(X_{GEN}) - E(X_{SC})\}'(\beta_{GEN} - \beta_{SC})
\]  

(3.3)

Here what we get is called the ‘threefold’ decomposition, that is, the mean wage difference (D) is divided into three components:

\[
R = E + C + I
\]

The first component, \( E = \{E(X_{GEN}) - E(X_{SC})\}'\beta_{SC} \) amounts to the part of the differential that is due to group differences in the regressors (the “endowments effect”). The second component, \( C = E(X_{SC})'(\beta_{GEN} - \beta_{SC}) \) measures the contribution of differences in the coefficients (including differences in the intercept, “discrimination” component). And the third component \( I = \{E(X_{GEN}) - E(X_{SC})\}'(\beta_{GEN} - \beta_{SC}) \) is an interaction term accounting for the fact that differences in endowments and coefficients exist simultaneously between the two groups.

The decomposition shown in (3.3) is formulated from the viewpoint of SC. That is, the group differences in the regressors are weighted by the coefficients of SC (\( \beta_{SC} \)) to determine the endowments effect (\( E \)) and similarly for the C component, the differences

---

*Research Design and Methodology*
in coefficients are weighted by SCs predictor levels. Naturally, the differential can also be expressed from the viewpoint of GEN, yielding the reverse ‘threefold’ decomposition (eq. 3.4),

\[
D = \{E(X_{\text{GEN}}) - E(X_{\text{SC}})\}'\beta_{\text{GEN}} + E(X_{\text{GEN}})'(\beta_{\text{GEN}} - \beta_{\text{SC}}) + \{E(X_{\text{GEN}}) - E(X_{\text{SC}})\}'(\beta_{\text{GEN}} - \beta_{\text{SC}}) \tag{3.4}
\]

Scholars used either of these two equations (equation 3.3 or 3.4) based on their assumptions about the existing market wage structure. It can be argued that, under discrimination, GENs are paid competitive wages but pay GENs more than the SCs. Coefficient should be used as the non-discriminatory wage structure. Therefore, the issue in literature is how to determine the wage structure, would prevail in the absence of discrimination. This choice poses the well-known index number problem given that we would use either the GEN or the SC wage structure as the non-discriminatory benchmark.

To over this problem and to extend the wage discrimination component further Cotton (1988), Neumark (1988) and Oaxaca and Ransom (1994) have proposed an alternative decomposition, prominent in the discrimination literature results from the concept that there is a nondiscriminatory coefficient vector that should be used to determine the contribution of the differences in the predictors. Let \( \beta^* \) be such a nondiscriminatory coefficient vector. The outcome difference can then be written as;

\[
D = \{E(X_{\text{GEN}}) - E(X_{\text{SC}})\}'\beta^* + E(X_{\text{GEN}})'(\beta_{\text{GEN}} - \beta^*) + E(X_{\text{SC}})'(\beta^* - \beta_{\text{SC}}) \tag{3.5}
\]

Where the first term on the right-hand side (RHS) of Equation (3.5) is the part of the wage differential that is explained by group differences in the regressors (skill difference), the second term is overpayment to GENs due to favouritism and the third term is underpayment to SCs due to discrimination. The equation (3.5) is operationalised under the assumption of non-discriminatory wage structure by assigning proportions of GEN \( P_{\text{GEN}} \) and SC \( P_{\text{SC}} \) weights to the wage structure and \( \beta^* \) is defined as;

\[
\beta^* = P_{\text{GEN}}\beta_{\text{SC}} + P_{\text{SC}}\beta_{\text{GEN}} \tag{3.6}
\]
Duncan’s Dis-Similarity Index

Here a good measure is used to identify the occupational segregation among social groups is the Duncan dis-similarity index, which is defined as:

\[
D = \left(\frac{1}{2}\right) \sum |X_i - Y_i|
\]

Where \(X_i\) is the proportion of households in occupational category \(i\) among social group \(X\), and \(Y_i\) is the proportion of households in occupational category \(i\) among social group \(Y\). The Duncan Dis-similarity Index measures the segregation of any one social group from any other mutually exclusive social group. It captures in a simple way the degree of dis-similarity in occupational structure between two social groups. The Index (\(D\)) ranges from zero to 100. If the value of \(D\) is zero, it shows complete integration, which indicates that the distribution of one social group across occupations is identical to that of the comparator social group, and if the value of \(D\) is 100, it indicates complete occupational segregation and shows that one social group is in occupations that are not populated at all by the comparator social group. The value of this index is statistically independent of the relative size of the groups used in its computation.

We have calculated this index for each pairing using the four rounds of EUS of our social groups ST, SC, and Non-SC/ST (see Appendix VI).