Research Design and Methodology

3.1 Objectives of the study

The purpose of this study is to analyse the impact of MGNREGA on the consumption expenditure pattern of the participants. In addition to it, the study attempts to evaluate the effect of scheme participation on food security. It also covers the issue of women participation in the scheme. The consumption expenditure pattern and female workforce participation in the scheme are analysed separately at national level as well as district level. Moreover, this study also estimates targeting errors in implementation of the scheme at macro level. The specific objectives of the study are as follows.

i. To analyse the impact of MGNREGA participation on the Monthly Per Capita Expenditure (MPCE) of beneficiaries.

ii. To study the impact of scheme participation on consumption expenditure pattern of beneficiaries.

iii. To analyse the female workforce participation in the scheme.

iv. To evaluate targeting efficiency of the scheme.

3.2. Hypotheses

After surveying the available literature and defining the gaps, we have set the following hypotheses for the study.

i. \( H_0 \): Monthly Per Capita Expenditure (MPCE) of beneficiaries and non-beneficiaries of MGNREGA is the same.
Hₐ: Monthly Per Capita Expenditure (MPCE) of beneficiaries and non-beneficiaries of MGNREGA is statistically different.

ii. H₀: Monthly per capita food expenditure of beneficiaries and non-beneficiaries of MGNREGA is the same.

Hₐ: Monthly per capita food expenditure of beneficiaries and non-beneficiaries of MGNREGA is statistically different.

iii. H₀: Monthly per capita non-food expenditure of beneficiaries and non-beneficiaries of MGNREGA is the same.

Hₐ: Monthly per capita non-food expenditure of beneficiaries and non-beneficiaries of MGNREGA is statistically different.

iv. H₀: Male and female workdays’ participation in MGNREGA is the same.

Hₐ: Male and female workdays’ participation in MGNREGA is statistically different.

v. H₀: Poor and non-poor households’ participation in MGNREGA is the same.

Hₐ: Poor and non-poor households’ participation in MGNREGA is statistically different.

vi. H₀: The per capita income from MGNREGA of poor and non-poor households is the same.

Hₐ: The per capita income from MGNREGA of poor and non-poor households is statistically different.

3.3 Data

3.3.1 Secondary data

This study uses India Human Development Survey (IHDS) unit level data. There were two rounds of IHDS; first conducted in 2004-05 and the second was conducted during
2011-12. The survey was organised by the researchers of University of Maryland and the National Council of Applied Economic Research (NCAER), New Delhi. It covered 42,152 households from 384 districts, 1,420 villages and 1,042 urban neighbourhoods throughout the country. IHDS was a multi topic nationally representative survey covering issues of employment, consumption expenditure, demography as well as social capital. The most unique feature of IHDS was its panel nature that makes it different from other Indian surveys. The first wave was conducted in 2004-05 and the second wave covering 83 percent of the wave one households was conducted in 2011-12 (Desai et al., 2015). It allows for a more efficient analysis of government and development policies exploiting the panel nature of data. This study uses both household and individual level data for cross sectional and panel analysis separately.

3.3.1.1 Weights

For all of the secondary analyses in this study, probability and frequency weights have been used accordingly that are given with IHDS data. For regression analysis, probability weights are used while for descriptive statistics and percentage analysis, frequency weights are used throughout the analysis.

3.3.2 Primary data

Primary data used in this study is based on a household survey of Allahabad district of Uttar Pradesh. This section explains, in detail, the methodology of sample selection, sample design, reference period, and household expenditure estimation procedure, among others.

3.3.2.1 Geographical area covered

Allahabad district was chosen for field survey area of our study, purposively. It is the most populous district of Uttar Pradesh, which is itself the most populous state of India. Total number of households in district was 961,120 and number of individuals was 5,954,391. Out of total population of district, 75.3 percent belong to rural areas (Census, 2011). We have considered only rural areas of the district in our survey because we aim to assess the impact of MGNREGA, a rural employment guarantee scheme.
The tehsils of district were classified as Trans-Ganga and Trans-Yamuna region and then Multi-Stage Stratified Random Sampling Method was applied to choose the sample. The strata were tehsils, blocks and villages. The administrative division of Allahabad district under tehsils, blocks and Gram Panchayats is as follows.

Table 3.1: Administrative division of Allahabad district

<table>
<thead>
<tr>
<th>Region</th>
<th>Tehsil</th>
<th>Development Block</th>
<th>Number of Gram Panchayats</th>
<th>Number of Villages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans-Ganga</td>
<td>Soraon</td>
<td>Kaurihar</td>
<td>130</td>
<td>241</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Holagarh</td>
<td>67</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mauaima</td>
<td>56</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Soraon</td>
<td>62</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Phulpur</td>
<td>Bahria</td>
<td>106</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phulpur</td>
<td>91</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bahadurpur</td>
<td>101</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>Handia</td>
<td>Pratappur</td>
<td>83</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Saidabad</td>
<td>93</td>
<td>161</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dhanupur</td>
<td>108</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Handia</td>
<td>80</td>
<td>133</td>
</tr>
<tr>
<td>Trans-Yamuna</td>
<td>Bara</td>
<td>Jasra</td>
<td>63</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shankargarh</td>
<td>76</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>Karchhna</td>
<td>Chaka</td>
<td>59</td>
<td>133</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Karchhana</td>
<td>80</td>
<td>119</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Kaudhiyara</td>
<td>57</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Meja</td>
<td>Uruwan</td>
<td>66</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Meja</td>
<td>75</td>
<td>161</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manda</td>
<td>69</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>Koraon</td>
<td>Koraon</td>
<td>115</td>
<td>209</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>1,637</td>
<td>3,051</td>
</tr>
</tbody>
</table>

**Source:** District-wise report, Panchayati Raj Department, Government of Uttar Pradesh and Statistical bulletin of Allahabad district (2011).

One tehsil was selected randomly from each Trans-Ganga and Trans-Yamuna regions. The selected tehsils were Soraon from Trans-Ganga and Karchhna from Trans-Yamuna.
region. From each tehsil two blocks were selected. Blocks, namely, Mauaima and Kaurihar were selected from Soraon as well as Chaka and Karchhna form tehsil Karchhna. Overall, there were 1,637 Gram Panchayats (GPs) and 3,051 villages\(^5\) in Allahabad district (Panchayati Raj Department, Government of UP, 2011). Two villages were selected randomly for the survey from each block. So, in all, eight villages were survey, namely, Mohmadpur Maganpur, Gheenpur, Rampur, Sarai-Jairam, Chak Hira Nand, Chaka, Basahi and Karchhna. The sample design is explained in figure 3.1.

**Figure 3.1: Sample Design**

\(^5\) According to Section 11 F, Declaration of Panchayat area, of The U.P. Panchayat Raj Act, 1947; a Panchayat area comprises a village or a group of villages having a population of one thousand. In other words, villages, that are very small in terms of population are clubbed together to make a Panchayat. So, the number of Gram Panchayats (GP) is lesser than the number of villages.
3.3.2.2 Sample size

Eight villages were surveyed and 40 respondents; 30 MGNREGA beneficiaries and 10 non-beneficiaries, were selected from each village. So, the beneficiary sample size was 240 households and the non-beneficiary sample size was 80 households. The overall sample size was 320 households/individuals. Sample size of individual and household level data was the same because only one individual was surveyed per household regarding his/her employment and household consumption expenditure related questions.

3.3.2.3 Selection of respondents and schedule drafting

The MGNREGA beneficiaries\(^6\) were selected on the basis of ‘Systematic Random Sampling‘\(^7\) method, by the scheme participants’ list provided by the Gram Pradhan of respective villages. Due representation of female respondents was taken care of while selecting the sample. The non-beneficiaries were selected by ‘Snowball Sampling‘\(^8\) method, that is, the selection of non-beneficiaries was made by suggestion of beneficiaries. The sample of non-beneficiaries was chosen such that their socio-economic characteristics were similar to that of beneficiaries’ sample. The surveyed beneficiaries were asked about the individuals who have been looking for work under MGNREGA and could not manage to get it, or the individuals with similar occupations, in their neighbourhood. It was crucial for our sample selection because a socially and financially comparable group of non-beneficiaries was required to make a reasonable analysis for impact assessment of the scheme. The occupational distribution, income as well as caste category of both of the samples are discussed in detail in chapter 5.

For the purpose of pre-testing the schedule, a pilot survey of 30 households was conducted during September, 2015 in the Gheenpur village of Mauaima Block. On the basis of findings of the pilot survey, the schedule was revised and final schedule was prepared (Schedule is given in Appendix VI). The selected respondents of both of the

\(^6\) Here, we define a household as ‘Beneficiary household’ if at least one individual had participated in MGNREGA work in the last year from the date of survey.

\(^7\) Every \(i^{th}\) item is selected.

\(^8\) Present respondents recruit the future subjects.
sub-samples, namely, beneficiaries and non-beneficiaries were enquired about their household consumption expenditure pattern, employment and functioning of MGNREGA related questions. On the basis of type of questions asked and consequently constructed variables, we have prepared two datasets, household and individual level datasets. Both of the datasets are used for separate analysis in chapters 5 and 7. Apart from the survey schedule, Focused Group Discussions (FGDs) were also performed to understand the problems faced by the scheme participants as well as by non-participants. The Gram Pradhan and Gram Rojgar Sevak\(^9\) were also interviewed to get their perspective on MGNREGA.

3.3.2.4 Estimation of Consumption Expenditure

NSSO Consumer Expenditure Survey (CES) reports MPCE estimates obtained by three methods; Uniform Reference Period (URP), Mixed Reference Period (MRP) and Modified Mixed Reference Period (MMRP). In the first method the reference period is ‘last 30 days’ for all commodity/service items, ‘last 30 days’ for some items and ‘last 365 days’ for others in the second method, and in the third method a reference period of ‘last 7 days’, ‘last 30 days’ and ‘last 365 days’ are separately used as recall periods for different item groups.

The questions in our survey schedule were a shortened form of NSSO CES Schedule type 1\(^\text{10}\) with required revisions as emerged during the pilot survey. For most of the commodities, respondents were asked about quantity consumed during reference period and the consumption expenditure was estimated by getting that quantity multiplied by market price of that commodity. In the case of other commodities and services, total value of expenditure was asked to the respondents.

\(^9\) Under the guidelines provided under the Act, at each Gram Panchayat, there should be appointment of a full time Gram Rojgar sewak, whose responsibility will be the implementation of the scheme (MGNREGA Guidebook for Gram Panchayats, 2010).

\(^\text{10}\) NSSO 68\(^\text{th}\) round CES was conducted by two types of schedules with same questions but different reference/recall periods. Schedule type 1 was based on ‘Mixed reference period’ while schedule type 2 was based on ‘Modified mixed reference period’.
3.3.2.5 Reference period

Different reference/recall period for separate items is considered for household consumption expenditure estimation to minimize the recall error. While 68th round of NSSO CES used all of the three MPCE estimation methods, namely, URP, MRP and MMRP, we have used MRP method for consumption expenditure estimation in our survey. The survey was conducted during September and October, 2016.

Table 3.2: Items and reference periods considered in survey

<table>
<thead>
<tr>
<th>Items</th>
<th>Reference period</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Food:</strong></td>
<td></td>
</tr>
<tr>
<td>Cereal, pulses, oil, spices, sweeteners, vegetables, meat/fish</td>
<td>Last 30 days</td>
</tr>
<tr>
<td><strong>Non-food:</strong></td>
<td></td>
</tr>
<tr>
<td>Fuel, transportation, paan/tobacco/intoxicants, other monthly non-food expenses</td>
<td></td>
</tr>
<tr>
<td><strong>Non-food:</strong></td>
<td></td>
</tr>
<tr>
<td>Medical, education, durable items, clothing, bedding, footwear</td>
<td>Last 365 days</td>
</tr>
</tbody>
</table>

Source: Reference period considered by the author in primary data collection

Throughout our analysis, we have used variables in per capita monthly expenditure format. So, all the variables having annual reference period (365 days), have been divided by 12 to transform them into monthly expenditure form. Further, the variables were divided by household size to get per capita expenditures, except for education, where it is per child expenditure. We get per child education expenditure by dividing monthly education expenses by number of children in the household.

3.3.2.6 Zero expenditure responses

For some commodities and services, a fraction of households had reported zero expenditure in the survey. Under non-food expenditure, these categories were household medical, education, durable expenditure and expenses incurred on paan, tobacco and intoxicants. However for food expenditure components, there were no non-zero responses for cereal, pulses and other food expenditure, with some households stating no expenses on non-vegetarian food items.
Table 3.3: Percentage of household reporting zero expenditure

<table>
<thead>
<tr>
<th>Items</th>
<th>Households with zero expenditure (Percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meat/Fish</td>
<td>26.25</td>
</tr>
<tr>
<td>Paan/Tobacco/Intoxicants</td>
<td>0.63</td>
</tr>
<tr>
<td>Medical</td>
<td>6.25</td>
</tr>
<tr>
<td>Education (Overall)</td>
<td>55.31</td>
</tr>
<tr>
<td>Education (with at least one child in household)</td>
<td>28.75</td>
</tr>
<tr>
<td>Durables</td>
<td>10.63</td>
</tr>
</tbody>
</table>

Source: Author’s calculations from primary survey

Zero expenditure is not absence of consumption for all items though, in the present case. Some of the households, who reported their medical expenditure to be zero, reported that they consider Government Prathmik Chikitsa Kendra for medical services and medicines while a fraction of households with no expenditure of education said that their children go to Government Primary Schools. So, their expenditure on medical and education was noted as zero in spite of that their consumption in these items was not nil. Here, we make the presumption that if a household gets better off in financial terms, they will shift their services from government sources to private clinics and private schools and incur non-zero expenditure on these items.

3.4 Econometric tools and methods

In this study two types of analysis are conducted; cross sectional as well as panel, therefore the methods are different for both of analysis. Although the empirical strategy and modelling is elaborated in the respective chapters, a brief definition of the tools and methods used throughout the study is given below.

3.4.1 Cross-sectional data analysis

3.4.1.1 Linear regression

A regression equation represents a behavioural model that estimates the causal effects of an independent variable or a set of independent variables on the dependent variable of
interest (Winship & Radbill, 1994). The Ordinary Least Squares (OLS) method is the most popular and widely used estimation method for regression analysis. If there are only two variables, one dependent and another independent, in the regression model; it is the two-variable regression model. However, a model with more than two variables or in other words, more than one independent variable is a multiple regression model. A multiple regression model can be written as follows.

\[ Y_i = \beta_1 + \beta'X'_i + u_i \]

Here, \( Y_i \) is the dependent variable, \( X' \) represents the set of independent variables and \( u_i \) is the error term. \( \beta_1 \) is the constant term and \( \beta' \) is the vector of slope coefficients.

In the linear regression models, linearity in terms of parameters is a mandate; however, they may or may not be linear in variables. So, the functional forms of the regression equations are transformable to make them suitable to be estimated with OLS or to analyse elasticity.

i. Log-Log model

ii. Log-Lin model

iii. Lin-Log model

In the Log-Log model, both the dependent and independent variables are log transformed. In the Log-Lin and Lin-Log models, the dependent variable and independent variables are log transformed, respectively (Gujarati et al., 2012). These models are used in analysis and are explained in respective chapters in this study.

### 3.4.1.2 Linear regression with interactions

Interaction effects with dummy variables and categorical variables are of great interest to obtain differential slope coefficients. The model can be written as follows.

\[ Y_i = \beta_1 + \beta_2X_{1i} + \beta_3X_{2i} + \beta_4(X_{1i} \ast X_{2i}) + u_i \]

Here, the independent variables \( X_{1i} \) and \( X_{2i} \) are either dummy or some categorical variable. The coefficient of term \( (X_{1i} \ast X_{2i}) \) gives differential slope coefficient (Gujarati et al., 2012).
et al., 2012). Suppose, X_{1i} is a categorical variable ‘occupation’ and X_{2i} is a dummy variable ‘sex’; and the objective is to estimate if ‘occupation’ has the same effect on Y for different values of ‘sex’. In this case, X_{1i} is ‘focal independent variable’ and X_{2i} is the ‘effect moderator’. The differential slope coefficient(s) will allow estimating this effect.

3.4.1.3 Logistic regression

In the case, when dependent variable is an indicator variable or a dummy variable, the OLS estimates pose several problems because many of the assumptions are not fulfilled (Gujarati et al., 2012). Because of it, the discrete response models are estimated with Maximum Likelihood Estimation. The binary response models are of the form,

\[ P(Y = 1|x) = G(x\beta) \]

A logit model is of the form,

\[ G(z) = \Lambda(z) = \frac{e^z}{1 + e^z} \]

Here, G has a standard logistic distribution and \( z = \beta_1 + \beta_2X_i \) (Wooldridge, 2008).

If \( P_i \) is the probability of success and \( P_i \) is of the form,

\[ P_i = \frac{e^z}{1 + e^z} \]

Then, the probability of failure will be,

\[ 1 - P_i = \frac{1}{1 + e^z} \]

Thus, the odds ratio in favour of a success will be \( P_i / (1 - P_i) \) and taking natural log of it gives the logit function of the form,

\[ L_i = \ln \left( \frac{P_i}{1 - P_i} \right) = \beta_1 + \beta_2X_i \]

Where, \( L_i \) is the log of odds ratio (Gujarati et al., 2012).
3.4.1.4 Coefficient of Determination (R-squared)

The coefficient of determination or the R-squared is a measure to see to the extent regression model is a good replication of observed outcomes i.e. to what extent the model fits the data. The value of $R^2$ lies between 0 to 1; and 1 indicates that the independent variables in regression model explain 100 percent of variation in the dependent variable.

$$R^2 = 1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares}}$$

Although theoretically 0.5 or higher value of $R^2$ is considered to be the threshold; there is no consensus on the acceptable value of $R^2$ in empirical studies. Choi et al. (2016) analysing 315 peer-reviewed economics articles based on OLS estimation found the distribution of unadjusted $R^2$ from the articles concerned to be skewed to left. It means that most of the articles reported $R^2$ to be below 0.4. In addition to it, the study found that only in 6 percent of the papers there was any comment on the $R^2$ statistic within the text and 46 percent of the empirical studies did not even report the $R^2$ statistic. Thus, in empirical research there is no minimum acceptable or unacceptable value of $R^2$.

3.4.2 Panel data analysis

3.4.2.1 Random Effects Logistic Regression

Random effects Logistic Regression is the estimation model for binary outcomes with panel data. The model is as follows.

$$P(y_{it} = 1| x_i) = \Lambda(x_{it}\beta), \quad t = 1, 2, \ldots, T$$

Here $\Lambda$ is the logistic function and $x_{it}$ can contain a variety of factors, including time dummies, time-constant and time-varying variables, and lagged dependent variables. The main assumption of this model is indicated below.

$$P(y_{it} = 1| x_{i}, c_i) = P(y_{it} = 1| x_{it}, c_i) = \Lambda(x_{it}\beta + c_i), \quad t = 1, 2, \ldots, T$$

Here $c_i$ is the unobserved effect and $x_i$ contains $x_{it}$ for all $t$. The first equality says that $x_{it}$
is strictly exogenous conditional on $c_i$: once $c_i$ is conditioned on, only $x_{it}$ appears in the response probability at time $t$. It rules out lagged dependent variables in $x_{it}$, as well as certain kinds of explanatory variables whose future movements depend on current and past outcomes on $y$ (Wooldridge, 2008).

### 3.4.2.2 Hausman Specification Test

To decide the more appropriate model between ‘Fixed effects’ and ‘Random effects’ models according to the objective at hand, **Hausman (1978)** suggested a test. The test was based on the difference between fixed and random effect estimators or the model coefficients (Wooldrige, 2008). The null hypothesis underlining Hausman test is that the random effect estimates are consistent and efficient. On the basis of the test estimates the null hypothesis is rejected or accepted and the random effect or fixed effect model is considered for estimation.

### 3.4.2.3 Difference-in-Differences with covariates

Contrary to experimental studies, estimation of causality is very difficult in the case of observational studies. It requires a precise estimation of a counterfactual to get the treatment effects. The Difference in Differences (Diff-in-Diff) is a method widely used to estimate the causal effects in observational studies, particularly policy impact evaluation. One of the most popular studies using this method is the study by **Card and Krueger (1994)** analysing minimum wages. The simplest setting of Diff-in-Diff model in two time points’ case can be written as follows.

\[
Y_{it} = \alpha + \gamma T_{it} + \lambda d_t + \delta (T_{it} * d_t) + \beta' X'_{it} + \varepsilon_{it} \quad \text{for } t = 1, 2
\]

In this model, $Y$ is the outcome variable. $T$ represents the treatment dummy and $d_t$ is the year dummy for post treatment time point. $X'$ indicates the set of covariates. The subscripts $i$ and $t$ represent individuals and time points respectively. Coefficient of interaction of these two dummy variables is ‘$T_{it} * d_t$’ is the Diff-in-Diff estimator, $\delta$. 
\[ \delta = (\bar{Y}_{T,2} - \bar{Y}_{T,1}) - (\bar{Y}_{NT,2} - \bar{Y}_{NT,1}) \]

Here, ‘T’ and ‘NT’ stand for treated and non-treated individuals respectively. 1 represents pre and 2 stands for post treatment time periods. It allows for estimating the effect of treatment, net of time effect. The Diff-in-Diff estimation can be explained in a three step method; first to calculate the difference in outcome in the before and after situation for treated group, second to calculate the outcome in before and after situation for the control group, and third to estimate the difference in the differences of before and after situation of treated as well as control groups (Gertler et al., 2011).

All the assumptions of the OLS apply to the Diff-in-Diff model as well (Doanld & Lang, 2007). In addition, ‘Parallel Trends Assumption’ is also crucial for Diff-in-Diff estimation. It implies that the composition of individuals in the treated and non-treated groups remain the same throughout different time points considered in analysis.