CHAPTER IV
4.1. INTRODUCTION

This chapter describes the methodology followed for analysing the various aspects presented in Chapter 1. Section 4.2 describes the methodology adopted for comparison of size distributions of consumer expenditure. Section 4.3 deals with the method of decomposing inequality measurements into within-group and between group components. Section 4.4 gives an account of the model used to estimate the Engel functions. Section 4.5 describes the method of covariance analysis as applied to the model. The estimation problems encountered while working on the model and
the procedure by which these problems are overcome are presented in Section 4.6. Section 4.7 describes the grouping of commodities formed for the analysis.

4.2. COMPARISON OF SIZE DISTRIBUTIONS

The analysis involves the comparison of size distributions of consumption expenditure on the lines of Bhattacharya and Iyengar (1961)\textsuperscript{1} and Iyengar and Suryanarayana (1985). In this thesis an attempt has been made to compare the consumption distributions across the three regions of Andhra Pradesh and also of the SC/ST with that of the non-SC/ST for the rural and urban sectors separately.

Let $F_{ij}(X_t)$ be the distribution of consumer expenditure ($X_t$) in region $t$ for social group $j$ of households in sector $i$ ($i = \text{rural, urban}$; $j = \text{SC/ST, non-SC/ST}$; $t = \text{region-1, region-2, region-3}$). This distribution has a mean $\mu_t$ and median $m_t$ for every $(i,j)$. $\mu$ and $m$ have been frequently used to measure the standard of living. We have used both these measures for comparing the levels of living.

The above comparison leaves out the distributional aspect of the problem. We tried to include this aspect by estimating the percentage of SC/ST in the top 10 and 20 per cent.

\textsuperscript{1}Sailaja Raghuprasad (1986). "A Quantitative Analysis of Poverty, Inequality and Consumer Behaviour (A Case study of Scheduled Castes and Scheduled Tribes of Karnataka)", Economic Analysis Unit, Indian Statistical Institute, Bangalore.
of the general population as well as in the bottom 10 and 20 percent.

**AN ALTERNATIVE METHOD OF COMPARING SIZE DISTRIBUTIONS**

Traditional methods such as the technique of Lorenz curves have been used to analyse the changes in a given size distribution, with or without any parametric assumptions regarding the form of the distribution. Iyengar and Suryanarayana (1985) developed a measure under the assumption of log normality for analysing the changes in a given size distribution.

This method of comparing is as follows:

Let $X > 0$ be a random variable representing the value of consumption a household in group $t$. Let $F_t(X_t)$ denote the cumulative distribution function of $X_t$. Let $t = 1, 2$ represent the two distributions under study. One way of comparing the two distributions would be to consider the set $R$:

$$R = \{(x_1, x_2) : F_1(x_1) = F_2(x_2)\}$$

and compare it with the set $S$:

$$S = \{(x_1, x_2) : x_1 = x_2\}$$

Set $S$ may be regarded as the state of equality. If the relation $R$ lies above or below the equality line, then we might describe this state as one in which one group is better off than the other. However, when $R$ and the line of equality intersect each other, then it might be interpreted as a state in which some
of the households in one group are better off than some in the other group.

Studies on Indian data have shown that the distribution of $X_t$ is approximately log normal. On the assumption that $X_t$ is log normal, the relation $R$ has been obtained. $X_t$ is two parameter log normal when $\log X_t$ is distributed normally with mean $\mu_t$ and the standard deviation $\sigma_t$. The relation $R$ has been obtained as :

$$x_2 = A \cdot x_1^B \ldots (4.1)$$

where $A = \exp (\phi_2 - \phi_1 B)$ and

$$B = \lambda_2 / \lambda_1 \ldots (4.2)$$

The exponent $B$ is the ratio of the inequality parameters of the two distributions and is positive. Hence, $B > 1$ implies that the distribution of $X_2$ is more (less) unequal than the distribution of $X_1$. When $B = 1$, the degree of inequality in the two distributions is the same. Thus, the relative degree of inequality in $X_2$ compared to $X_1$ is summarised by the parameter $B$.

4.3. INEQUALITY AND DECOMPOSITION OF INEQUALITY

Any measure of inequality is concerned with the distribution of a positive random variable in relation to some suitable location parameter. From literature, we observe a number of such indices.
Several inequality indices such as Gini Ratio, Theil measures L and T, and the Variance of logarithms (Varlog) have been used in the present study.

For inter-regional and social-group comparisons the Gini Ratio figures have been used. The other inequality indices have been computed but mainly for the purposes of decomposition.

In analysing the determination and significance of a given distribution, or changes in it, one often deals with different aggregations of population. For example, the distribution by social groups, regions, races, or even age groups may be of special interest. If the total population is classified into a mutually exclusive set of such groups, then it is interesting to determine how much of the total inequality in the population is due to within-group inequality and how much is due to between-group inequality.

The term decomposition describes the total measure as a sum of two terms. The first term is the measure defined between the groups (treating each group mean as single observation) and the second term is a weighted sum of the within group measures. Additive decomposability is the property by which the total inequality of a population can be expressed as the sum of a weighted average of the inequality within subgroups of the population and the inequality existing between them. Of all the measures, only few measures are really susceptible to composition of the aggregate measure into within and between...
group inequality. Of the readily decomposable measures, the Theil measure and the log variance are seen best in terms of their decomposition properties. For both of them, one can easily determine the independent contribution of within and between group inequality to total inequality.

4.3.1. Theil's T and L measures:

Let \( Y = (Y_1, Y_2, \ldots, Y_n) \) be the income (expenditure) distribution for a population of 'n' households. Then Theil T can be written as

\[
T(Y,n) = \frac{1}{n} \sum Y_i / \mu \ln \frac{Y_i}{\mu} \quad \ldots \quad (4.3)
\]

where \( \mu \) is mean income = \( \frac{1}{n} \sum Y_i \)

The Theil L can be given as

\[
L(Y,n) = \frac{1}{n} \sum \ln \frac{\mu}{Y_i} \quad \ldots \quad (4.4)
\]

Suppose the entire population can be divided into \( k \) non-overlapping sub groups (Social/Regions) each of size \( n_j \) where \( j = 1, 2, \ldots, k \). Let \( T_j \) and \( L_j \) be the Theil measures \( T \) and \( L \) corresponding to the \( j \)th group.

Let \( Y_{ij} \) be the per capita income (expenditure) of the \( i \)th household in the \( j \)th subgroup.

The total income \( Y = \sum_{i} \sum_{j} Y_{ij} = \sum_{j} Y_{j} = n \mu \)

where \( \mu \) is the mean per capita income in the population of households.
and $Y_j = \sum_{i=1}^{n_j} Y_{ij}$ is the total income of $j$th group

$= n_j \mu_j$

where $\mu_j$ is the mean per capita income of the $j$th group.

The total number of households $n$ is given by $n = \sum n_j$

The Theil index $T$ can be given by

$$T = \sum \sum \frac{Y_{ij}}{\bar{Y}} \ln \frac{Y_{ij}/Y}{1/n}$$

$$T = \sum \sum \frac{Y_{ij}}{\bar{Y}} \cdot \frac{Y_j}{Y_{ij}} \ln \frac{Y_{ij}/Y}{1/n}$$

$$T = \sum \frac{Y_j}{\bar{Y}} \sum \frac{Y_{ij}}{Y_{ij}} \ln \frac{(Y_{ij}/Y_j) \cdot Y_j - n_j}{1/n_n}$$

$$T = \sum \frac{Y_j}{\bar{Y}} \sum \frac{Y_{ij}}{Y_{ij}} \ln \frac{Y_{ij} / Y_j}{1/n_j} + \ln \frac{Y_j / n_j}{Y/n}$$

$$T = \sum \frac{Y_j}{\bar{Y}} \sum \frac{Y_{ij}}{Y_{ij}} \ln \frac{Y_{ij} / Y_j}{1/n_j} + \sum \frac{Y_j}{\bar{Y}} \ln \frac{Y_j / Y}{n_j/n}$$
since \( \sum_{i} \frac{y_{ij}}{y_{j}} = 1 \) for all \( j \)

So that

\[
T = \sum_{j} \frac{y_{j}}{Y} T_{j} + \sum_{j} \frac{y_{j}}{Y} \ln \frac{y_{j}}{n_{j}/n}
\]

... (4.5)

where

\[
T_{j} = \sum_{j} \frac{y_{ij}}{y_{j}} \ln \frac{y_{ij}}{y_{j}}
\]

... (4.6)

The above equation says that the Theil entropy index \( T \) can be decomposed into two terms as

\[
T = T_{W} + T_{B}
\]

where \( T_{W} = \sum \frac{y_{j}}{Y} T_{j} = \sum \frac{n_{j} \mu_{j}}{n \mu} T_{j} \)

is a weighted average of within group Theil indices \( T_{j} \), the weight being equal to the income shares \( y_{j}/Y \) of the groups, and
The between group $J$ Theil index of group income and population shares $Y_j/Y$ and $n_j/n$ respectively.

$T_B = \sum J \frac{Y_j}{Y} \ln \frac{Y_j/Y}{n_j/n} = \sum \frac{n_j/\mu_j}{n/\mu} \ln \left(\frac{\mu_j}{\mu}\right)$

$T_W$ is called within group component and $T_B$ is called between group component. The between group contribution is then defined as the ratio of the between group component $T_B$ to the overall Theil index $T$. The within group contribution is defined as $T_W/T$.

**Theil $L$ Index:**

$L = \sum J \sum i \frac{1}{n} \ln \frac{1/n}{Y_{ij}/Y}$

which reduces to

$L = \ln \frac{Y}{n} - \sum J \sum i \frac{1}{n} \ln Y_{ij}$

which is the logarithm of the arithmetic mean income minus the logarithm of the geometric mean income.

Now consider the decomposition of the Theil measure $L$ into between and within group components.
The second measure \( L \) can be decomposed into two terms as

\[
L = L_B + L_W
\]

where \( L_W = \sum \frac{n_j}{n} L_j \) is weighted average of within group \( j \) Theil measures \( L_j \), the weights being equal to the population shares \( n_j/n \) of the groups, and

\[
L_B = \sum \frac{n_j}{n} \ln \frac{n_j/n}{Y_j/Y} = \sum \frac{n_j}{n} \ln \frac{\mu}{\mu_j}
\]

is the between group \( j \) Theil measures of group population and income shares \( n_j/n \) and \( Y_j/Y \) respectively.
$L_W$ is called the within group component and $L_B$ is called the between group component. The between group contribution is then defined as the ratio of the between group component $L_B$ to the overall Theil measure $L$. The within group contribution is defined as $L_W/L$.

The value lies between 0 and $\log n$, where $n$ is the number of observations. A value of $\log n$ is reached when all individuals earn equal income. $T$ is zero, when one individual receives all the income. Theil $L$ measure is not defined for distributions with zero incomes.

4.3.2. The Variance of Logarithms:

Log variance simply follows the statistical decomposition of variance formula. The total log variance equals the between group log variance plus a weighted average of the within group log variances.

Let $X_{1j} = \log Y_{1j}$

where $Y_{1j} =$ Per capita expenditure of $i$th individual in the $j$th group.

We have

$$V = \frac{1}{n} \sum \sum (X_{1j} - \ln \mu)^2$$

(4.9)

where $\ln \mu = \frac{1}{n} \sum \sum X_{1j} = \bar{X}_j$ is the mean of $X_{1j}$.
Let \( X_{.j} = \frac{\Sigma X_{ij}}{n_j} \) be the mean of \( X_{ij} \) over \( i \).

Then

\[
V = \frac{1}{n} \Sigma_j \Sigma_i (X_{ij} - X_{.j})^2 + (X_{.j} - X_{..})^2
\]

\[
= \frac{1}{n} \Sigma_j \Sigma_i (X_{ij} - X_{.j})^2 + (X_{.j} - X_{..})^2
\]

\[
+ (2X_{ij} - X_{.j})(X_{.j} - X_{..})
\]

\[
= \Sigma_j \frac{n_j}{n} \frac{1}{n_j} (X_{ij} - X_{.j})^2 + \Sigma_j \frac{n_j}{n} (X_{.j} - X_{..})^2
\]

\[
= \Sigma_j \frac{n_j}{n} V_j + \Sigma_j \frac{n_j}{n} (X_{.j} - X_{..})^2
\]

where

\[
V_j = \Sigma \frac{1}{n_j} (X_{ij} - X_{.j})^2
\]  

... (4.10)

The cross product term in the above expression vanishes because

\((X_{.j} - X_{..})\) is constant in the \( i \)th summation, and

\[
\Sigma_j (X_{.j} - X_{..}) = 0 \text{ for each } j \text{ by the definition of mean } X_{.j}.
\]

Thus the variance \( V \) can be decomposed into two terms \( V_W \) and \( V_B \) as
\[ V = V_w + V_B \]

where \( V_w \) is a weighted average of within group variance \( V_j \), the weights being equal to the population of shares \( n_j/n \) of the groups, and \( V_B \) is the between group variance of group means \( X_j \).

\( V_w \) is called the within group component and \( V_B \) is called the between group component The between group contribution is then defined as the ratio of the between group component \( V_B \) to the total variance \( V \). The within group contribution is defined as \( V_w/V \).

4.4. Engel Function

The consumer's demand for a specific commodity is affected by the price of that commodity, prices of other substitutes and complementary goods, the income and needs, tastes and choice of the consumer. All the factors have their respective influence on the demand for a commodity by an individual. To estimate a general demand function which takes into account all the mentioned factors is very difficult in practice.

Ernst Engel considered the income of the consumer as a dominant factor for determining the demand for any commodity and kept all the other factors affecting the demand as constant and considered only the partial relationship between income and demand and this partial relation is known as the Engel curve or Engel function.
Empirical investigations on consumer behaviour in India start with the customary procedure of first determining suitable relationship between income (total expenditure, when used proxy for income) and particular expenditure. A vast country like India, exhibiting varying regional levels of development and prosperity, should not be treated homogeneous in respect of consumer behaviour. It has been observed that the consumer behaviour in different regions is not homogeneous.

A many researchers in the field of consumer behaviour have experimented with a number of Engel functions. Their studies indicate (1) that no unique Engel curve suffices for the complete range of consumer goods, and (11) that different mathematical forms give rise to widely differing elasticities. We thus confine our attention to the problem of finding those curves which give most adequate fit for most of the commodities.

Per capita total expenditure, which is used as proxy for income, is considered to be the most important factor affecting expenditures for most of the commodity items, and for this reason first concern is to chose an appropriate algebraic relationship between per capita total expenditure and per capita item expenditure, with other variables held constant.

The widely used functional forms to estimate the Engel elasticities can be listed as: linear, semi-log, double-log, log-inverse, hyperbolic, parabolic, semi-log-inverse,
log-log-inverse etc. The choice of an algebraic specification of the Engel function is of some importance since different equations imply different behavioural assumptions. It is well known that, the implications of these functional forms are different in terms of their economic interpretation.

In the present study we have, therefore, experimented with seven (7) alternative forms of Engel formulations for each item in both the sectors. The models with relationship and the expression for elasticity are set out in Table 4.1.

The final choice among different algebraic specifications is based on some criteria such as sign and significance of the coefficients, multiple correlation coefficient ($R^2$), adjusted $R^2$, distance function etc. The multiple correlation coefficient, $R^2$ is widely used for this purpose. The correlation coefficient ($R^2$) of linear, semi-log, exponential, hyperbolic and working-leaser, measures the degree of explanation of the variance of $C_{ij}$, where as for the remaining forms - log-inverse, double-log, and parabolic it measures the degree of explanation of the variance of log $C_{ij}$. The correlation coefficient based on grouped data is an unsatisfactory estimate$^2$ of the correlation in the given population.

$^2$ please refer Prais J. and J. Aitchison (1954) and Cramer (1964)
<table>
<thead>
<tr>
<th>S.No.</th>
<th>Algebraic Form</th>
<th>The Relationship</th>
<th>Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Linear (L)</td>
<td>$C_{ij} = \alpha_1 + \beta_1 X_j$</td>
<td>$\beta_1 \frac{(X)}{C_1}$</td>
</tr>
<tr>
<td>2.</td>
<td>Semilog (SL)</td>
<td>$C_{ij} = \alpha_1 + \beta_1 \log X_j$</td>
<td>$\beta_1 \frac{1}{C_1}$</td>
</tr>
<tr>
<td>3.</td>
<td>Log-1nverse (LI)</td>
<td>$\log C_{ij} = \alpha_1 + \beta_1 \frac{1}{X_j}$</td>
<td>$-(\beta_1 / X)$</td>
</tr>
<tr>
<td>4.</td>
<td>Doublelog (DL)</td>
<td>$\log C_{ij} = \log \alpha_1 + \beta_1 \log X_j$</td>
<td>$\beta_1$</td>
</tr>
<tr>
<td>5.</td>
<td>Exponential (EXP)</td>
<td>$\log C_{ij} = \alpha_1 + \beta_1 X_j$</td>
<td>$\beta_1 X$</td>
</tr>
<tr>
<td>6.</td>
<td>Hyperbolic (HYP)</td>
<td>$C_{ij} = \alpha_1 + \beta_1 X_j$</td>
<td>$-(\beta_1 / XC_1)$</td>
</tr>
<tr>
<td>7.</td>
<td>Working-Leser (WL)</td>
<td>$W_{ij} = \alpha_1 + \beta_1 \log X_j$</td>
<td>$1 + (\beta_1 / W_1)$</td>
</tr>
</tbody>
</table>

where $C_{ij}$ = per capita expenditure on specific (ith) commodity by the jth household

$X_j$ = per capita monthly total expenditure by the jth household

$C_1$ = Average monthly consumption expenditure on ith commodity
\[ W_{ij} = \text{ith item's share of expenditure in the total outlay (}X_{ij}\text{) of the }j\text{th household} \]

\[ X = \text{Average monthly total expenditure} \]

\[ W_i = \text{Average ith item's share of expenditure in the total outlay} \]

\( \alpha_i \) and \( \beta_i \) are parameters of relation pertaining to the \( i \)th commodity in question.

The squares of the correlation coefficients estimated from different Engel curves are compared by using the adjusted coefficients of multiple determination \( (\bar{R}^2) \) given by

\[ \bar{R}^2 = 1 - (1-R^2)(n-1)/(n-k-1) \quad \text{... (4.11)} \]

where \( n = \text{Number of observations} \)

\( k = \text{Number of explanatory variables} \)

Keeping in view some of the limitations of the multiple correlation coefficient, to choose the "best" fitting Engel function, the following Distance function \( "D^2" \) is computed.

\[ D_i^2 = \Sigma K_i (C_i - \hat{C}_i) \quad \text{... (4.12)} \]

where \( C_i \) = observed level of average per capita consumer expenditure of given item in the \( i \)th class

\( \hat{C}_i \) = expenditure on that item in the \( i \)th class as estimated from the fitted Engel curve
\( K_i = \text{proportion of population in the } i\text{th class and} \)

\( n = \text{number of items.} \)

The Engel function with minimum value of Distance function \( (D^2) \) giving due consideration to sign and magnitude is chosen as the best one. The selected seven forms of Engel curves are fitted using weighted least square method with population proportions as weights and then the value of Distance function \( D^2 \) is computed for each function\(^3\). In this study, it is observed that \( D^2 \) is minimum with working-leser form for all commodity groups both in rural and urban sectors. Hence the Working - Leser form is chosen to estimate the elasticities for various commodity groups.

**WORKING-LESER MODEL DESCRIPTION**

Working-Leser form satisfies the criteria of additivity without any cross equation restrictions.

\(^3\) In some cases where discrimination is not possible on the basis of distance function and adjusted multiple correlation coefficient of determination, the final choice about appropriateness of an Engel function can be drawn on considering the sign and significance of the coefficient and the \( F \)-value associated with \( R^2 \) and simplicity of the form.
The model can be given as

$$W_{ij} = \alpha_1 + \beta_1 \ln X_j \quad \ldots (4.13)$$

where $W_{ij}$ = share of $i$th item in the total outlay $X_j$ of $j$th household

i.e. $W_{ij}$ = expenditure on $i$th item / total outlay of $j$th household

$X_j$ = per capita monthly total expenditure by the $j$th household

$\alpha_1$ and $\beta_1$ are the parameters to be estimated.

Equation (4.13) does not have any variable pertaining to household size which is another important explanatory variable. The model can be improved by including $\ln N$ and $(\ln X)^2$ as additional variables. The improved model can be written as

$$W_i = \alpha_1 + \beta_1 \ln X + \gamma_1 \ln N + \delta_1 (\ln X)^2 + u_i \quad \ldots (4.14)$$

This model (4.14) allows separate description of the effect of total output ($X$) and household size ($N$) on share of expenditure on particular item ($W_i$).

The model satisfies the property of additivity and hence the estimates of the parameters are obtained by applying ordinary least squares (OLS) equation.
The expenditure elasticity for the specific ith item \( (\eta_i) \) based on model (4.14) is given by

\[
\eta_i = 1 + (\beta_1 + 2\delta_1 \ln X_i) / W_1 ... (4.15)
\]

\( \eta_i \)'s have been calculated with overall regional means of the explanatory variables and for social group specific regional means.

The effect of an increase in total expenditure \( X \) from \( X_1 \) and \( X_2 \) keeping the \( N \) constant is determined in the following way.

Let

\[
W_1 = \alpha + \beta \ln X_1 + \gamma \ln N + \delta (\ln X_1)^2 \ldots (4.16)
\]

Let

\[
W_2 = \alpha + \beta \ln X_2 + \gamma \ln N + \delta (\ln X_2)^2 \ldots (4.17)
\]

Then

\[
W_2 - W_1 = (\ln X_2 - \ln X_1) [\beta + \delta (\ln X_1 + \ln X_2)] \ldots (4.18)
\]

\( W_2 - W_1 > 0 \) only if \( [\beta + \delta (\ln X_1 + \ln X_2)] > 0 \ldots (4.19) \)

\( W_2 - W_1 < 0 \) if \( [\beta + \delta (\ln X_1 + \ln X_2)] < 0 \ldots (4.20) \)

i.e. the share of expenditure increases if \( [\beta + \delta (\ln X_1 + \ln X_2)] > 0 \) and decreases if \( [\beta + \delta (\ln X_1 + \ln X_2)] < 0 \).
4.5. THE METHOD OF COVARIANCE ANALYSIS

To analyse the changes in the consumption pattern across regions and between social groups the method of covariance analysis is used.

The model used to estimate the Engel function is given as:

\[ W_{ij} = \alpha_i + \beta_i \ln (X_j) + \gamma_j \ln (N_j) + \delta_j (\ln X_j)^2 + u_{ij} \]

where \( W_{ij} \) = share of the \( i \)th item in the monthly budget of \( j \)th household

\( X_j \) = monthly total expenditure of the \( j \)th household

\( N_j \) = household size of the \( j \)th household

\( u_{ij} \) = random disturbance term (this term follows the general assumptions of a linear regression model)

In the present study, we have considered eleven (11) broad group of commodities. These commodity groups are described in the section 4.7. Therefore the subscript \( i \) represents \( i = 1, 2, \ldots 11 \). The subscript \( j \) represents the households in the sample for each analysis.

Three basic aspects will be examined under covariance analysis.
(1) To investigate the possibilities of complete homogeneity of Engel curves between social groups and across regions.

(2) To investigate the possibilities of group/regional effects in favour of differential intercepts with the assumption of same slopes.

(3) To investigate the possibilities of group/regional effects taking the form of differential slopes and intercepts.

We examine the stability of a relation across regions and between social groups using the following hierarchy of three models. Let there be $k$ variables and $p$ classes (groups).

MODEL - 1 : Common Intercept, Common Slope Vector in all $p$ classes.

\[
\begin{bmatrix}
  y_1 \\
y_2 \\
  \vdots \\
y_p
\end{bmatrix}
= \begin{bmatrix}
  1 & x_1^* \\
  1 & x_2^* \\
  \vdots & \vdots \\
  1 & x_p^*
\end{bmatrix}
\begin{bmatrix}
  \alpha \\
  \beta^*
\end{bmatrix}
+ u
\]
MODEL - II : Differential Intercepts, Common Slope Vector

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_p
\end{bmatrix} =
\begin{bmatrix}
i_1 & 0 & \ldots & 0 & \mathbf{x}_1^* \\
0 & i_2 & \ldots & 0 & \mathbf{x}_2^* \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & i_p & \mathbf{x}_p^*
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_p
\end{bmatrix} + u
\]

MODEL - III : Differential Intercepts, differential slope

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_p
\end{bmatrix} =
\begin{bmatrix}
i_1 & 0 & \ldots & 0 & \mathbf{x}_1^* \\
0 & i_2 & \ldots & 0 & \mathbf{x}_2^* \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \ldots & i_p & \mathbf{x}_p^*
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_p
\end{bmatrix} + u
\]

Here \( \mathbf{i}_i \) = column vector of \( n_i \) units (\( i = 1, 2, \ldots, p \))

and \( \mathbf{x}_i^* = n_i \times (k-1) \) matrix observations on the explanatory variables in the class \( i \) (\( i = 1, 2, \ldots, p \))
Fitting of each model by ordinary least squares method produces a residual sum of squares. The residual sum of squares associated with each model have the degrees of freedom as indicated below:

\[
\begin{align*}
\text{MODEL - I} & \quad n-k \quad \text{degrees of freedom} \\
\text{MODEL - II} & \quad n-p-k+1 \quad \text{degrees of freedom} \\
\text{MODEL - III} & \quad n-pk \quad \text{degrees of freedom}
\end{align*}
\]

where \( n = n_1 + n_2 + n_3 + \ldots + n_p \) denotes the total number of sample observations.

Let \( \text{RSS}_1, \text{RSS}_2 \) and \( \text{RSS}_3 \) be the residual sum of squares for the models I, II and III respectively. Then there are basic tests on the differences between the various residual sum of squares may be conducted. They are

(1) Test of differential regressions - Model I contrasted with Model III
\( (H_0(1) : \text{Homogeneity of slopes and intercepts}) \)

(2) Test of differential slope coefficients - Model II contrasted with Model III
\( (H_0(2) : \text{Slope coefficients in various groups/regional regressions are same}) \)

(3) Test of differential intercept - Model I contrasted with Model II
\( (H_0(3) : \text{Intercepts are common}) \)
The hypotheses $H_0(1)$, $H_0(2)$ and $H_0(3)$ are then tested by contrasting residual sum of squares. The test statistics for various hypothesis are then as follows:

**Hypothesis $H_0(1)$**: overall homogeneity of regressions

\[
F_1 = \frac{(RSS_1 - RSS_2)/k(p-1)}{RSS_3/(n-p-k)} \sim F [k(p-1), n-pk] \quad \ldots (4.21)
\]

**Hypothesis $H_0(2)$**: Common slope coefficients

\[
F_2 = \frac{(RSS_2 - RSS_3)/(k-1)(p-1)}{RSS_3/(n-pk)} \sim F [k-1(p-1), n-pk] \\
\quad \ldots (4.22)
\]

**Hypothesis $H_0(3)$**: Common Intercepts

\[
F_3 = \frac{RSS_1 - RSS_2/(p-1)}{RSS_2/(n-p-k+1)} \sim F [(p-1), n-pk] \quad \ldots (4.21)
\]

If on testing the hypothesis $H_0(1)$, the groups/regions turns out to be statistically homogeneous. We stop further testing and group/regional factor can be neglected and efficient estimates of the parameters can be obtained by combining the data on the groups/regions into a single homogeneous sample. On the other hand, if the hypothesis is rejected i.e., the groups/regions are found to be statistically significant, the
test for $H_0(2)$ is undertaken. If $H_0(2)$ cannot be rejected we go as to test $H_0(3)$.

4.6. ESTIMATION PROBLEMS

For the analysis, we have used the following model

$$W_i = \alpha_1 + \beta_1 \ln (X) + \gamma_1 \ln (N) + \delta_1 (\ln X)^2 + u_i$$

This model was estimated using the ordinary least squares method. The General Linear Model specifies the relationship between a dependent variable $Y$ and $k-1$ explanatory variables ($X_2$, $X_3$, ..., $X_k$) and disturbance term "u". The model can be written in matrix notation as

$$Y = X\beta + u$$

... (6.24)

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

where $X_{ki}$ denotes the $i$th observation on the variable $K$. This means that the subscript in the "X" matrix follow the reverse of normal pattern, where the first subscript usually indicates the
row, and the second column, of the matrix. The following assumptions ensures that the estimates of the parameters are BLUE.

1. \( E(u) = 0 \). The first assumption implies that variables excluded from the model give a positive or negative value to the disturbance term, \( u \), and in repeated sampling these effects tend to cancel out. That is the \( u_i \) s are variables with zero expectation.

2. \( E(uu') = \sigma^2 I_n \). This shows that the variance of \( u \) is constant and independent of the set of explanatory variables \( X \) (Homoscedastic) and the values of \( u \) are drawn independently of one another.

3. \( X \) is a set of fixed numbers. This means that in repeated sampling the sole source of variation in the \( Y \) vector is variation in the \( u \) vector and the properties of the estimators and tests are conditional upon \( X \).

4. \( X \) has a rank \( k < n \). This implies that the number of observations exceeds the number of parameters to be estimated and that no exact linear relations exist between any of the explanatory variables \( X \) (absence of multicollinearity).

The model described above is an extremely powerful and widely used statistical tool. As in all statistical applications, however, the power of the method depends on the underlying assumptions being fulfilled for the particular application in
question. The usual estimation problems of Autocorrelation, Multicollinearity and Heteroscedasticity in the case of our fitted model and data are discussed below.

Autocorrelation:

The important assumption about the general linear model \( Y = \alpha \beta + u \) is that of the zero covariance for the disturbance term i.e., \( E(ww') = \sigma^2 I_n \) in which the off diagonal terms give \( E(u_t u_{t+s}) = 0 \) for all \( t \) and \( s \). They are pair wise independent. In the present analysis we have used the cross-section data. For this type of data generally the disturbance value that is "drawn" for any one unit is uninfluenced by the values drawn for other units. The problem of autocorrelation is not a serious problem in the cross-section studies where as the problem of correlated errors posses a serious problem in the time series analysis.

Multicollinearity:

The general linear model assumes the data matrix \( X \), which is of order \( n \times k \), has rank \( k \). This implies that no linear dependence exists between the set of explanatory variables. The reason for this is that the least square estimates requires the inversion of \( XX' \), which is impossible if the rank of \( X \), and hence the rank of \( XX' \) is less than \( k \).

The independent (explanatory) variables in the used model are \( \ln X \) and \( \ln N \) We have calculated the simple correlation coefficients for the rural and urban sectors and noticed that the simple correlations are all low. This suggests
that in the case of our data set multicollinearity is not a serious problem.

Heteroscedasticity:

The assumption about the random variable "u" in the model is that its probability distribution remains the same over all observations of X, and in particular that the variance of each \( u_1 \) is the same for all values of the explanatory variable. 

\[
\text{Var}(u) = E\{(u_1 - E(u))^2\} = E(u_1)^2 = \sigma_u^2 \text{ constant.}
\]

This assumption is known as the assumption of homoscedasticity. If it is not satisfied in any particular case (\( \sigma_u^2 \) not constant), we say that u's are heteroscedastic.

In the presence of heteroscedasticity the OLS estimates do not have the minimum variance property in the class of unbiased estimators, i.e., they are inefficient. The prediction based on this estimates from the original data, would have high variance, that is, the prediction would be inefficient. Due the over estimation or underestimation of the sampling variances of the estimated regression coefficients depending on the type of heteroscedasticity, one may have unduly short or large confidence intervals. Larger confidence intervals would mean greater difficulty in rejecting the null hypothesis and shorter confidence interval implies a Type - I error (probability of rejecting a true hypothesis) higher than the assumed value.
In cross-section data, one may expect the problem of heteroscedasticity to occur. Hence we have tested the assumption of homoscedasticity using the standard method, Goldfeld and Quandt test.

The null hypothesis for the test is

\[ H_0 : \sigma_1 = \sigma_2 = \ldots = \sigma_n \]

where \( n \) is the number of observations against the alternative

\[ \sigma_j^2 \sim (\ln(X_j / N_j))^b \]

where \( b \) is the strength of heteroscedasticity assumed to be present (\( b \) is unknown).

The Goldfeld-Quandt Test: It is a simple and finite sample test, which is applicable if it is thought that one of the \( X \) variables is the basic explanation of heteroscedasticity. This method involves the calculation of two least square regressions, one using the subset of the data thought to be associated with low variance errors and the other using the subset of the data thought to be associated with high variance errors. If the residual variances associated with each regression are approximately equal, then the homoscedasticity assumption cannot be rejected. The test is as follows:

Step 1: Data is ordered by the magnitude of the variable which is thought to be related to the error variance (\( \ln(X_j / N_j) = Z_j \), per capita expenditure of the each household).
Step 2: Omit 'c' central observations. The power of the test depends on c. For large values of c the power will be small. If c is reduced the residual variances will move closer which will tend to offset the increase in power because of added observations.

Step 3: Two separate regressions were fitted by ordinary least squares method, the first for the portion of the data associated with low values of z and the other for the second associated with high values of z. Each regression has (n-c)/2 observations.

(In our test we have not omitted any observations c = 0).

Step 4: Compute the residual sum of squares for the two regressions

Let

\[ \text{RSS}_1 = \text{residual sum of squares for the first regression (low z's)} \]

\[ \text{RSS}_2 = \text{residual sum of squares for the second regression (high z's)} \]

Step 5: Compute the statistic \( R = \frac{\text{RSS}_2}{\text{RSS}_1} \)

Considering the error process to be normally distributed, \( R \) value is compared with F-statistic with \((n-4)/2\) degrees of freedom.
If the hypothesis of homoscedasticity is rejected, then we transform the variables by dividing the model through $Z_{1}^{b/2}$ and then reestimated the coefficients. Then this transformed model will be used for the covariance analysis.

4.7. GROUPING OF COMMODITIES

The National Sample Survey data provide information on consumption expenditure for most of the commodities entering the consumption basket. For the analysis we classify items of expenditure into the following 11 broad categories. The broad groups are:

1. Cereals
2. Pulses
3. Milk and Milk Products
4. Edible oil
5. Meat, Egg and Fish
6. Vegetables, Fruits and Dry fruit
7. Sugar
8. Other Food items
9. Fuel and Light
10. Clothing
11. Other Nonfood items

The details of these consumption items that come under a specific category are presented in the Table 4.2.
<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cereals</td>
<td>Cereals, cereal substitutes, wheat, ragi, maize, bajra, jowar, barley, gram and their products.</td>
</tr>
<tr>
<td>2. Pulses</td>
<td>All varieties of pulses and its products.</td>
</tr>
<tr>
<td>3. Milk and Milk Products</td>
<td>Milk and milk products like ghee, butter, curd, ice cream etc.</td>
</tr>
<tr>
<td>4. Edible oil</td>
<td>Mustard oil, vanaspati oil, coconut oil, palm oil, groundnut oil etc.</td>
</tr>
<tr>
<td>5. Meat, Egg and Fish</td>
<td>Poultry, other birds, eggs, egg products, fish-fresh and dry</td>
</tr>
<tr>
<td>6. Vegetables, Fruits and Dry Fruits</td>
<td>All varieties of vegetables including root vegetables and chillies, fruits fresh and dry</td>
</tr>
<tr>
<td>7. Sugar</td>
<td>Mill and Khandasri sugar, gur and shakkar</td>
</tr>
<tr>
<td>8. Other Food items</td>
<td>Salt, species, beverages, refreshment and all processed food items</td>
</tr>
<tr>
<td>9. Fuel and Light</td>
<td>Coal, firewood, electricity, gas, dung cake, kerosene oil, candles, match-sticks</td>
</tr>
<tr>
<td>10. Clothing</td>
<td>All types of clothing (silken, cotton, woollen, terrycot), readymade garments, knitted garments etc.</td>
</tr>
<tr>
<td>11. Other Nonfood items</td>
<td>Expenditures on entertainment, education, medicine, conveyances, consumer rents, all types of durable goods</td>
</tr>
</tbody>
</table>