



DIBYA BHOWMIK (AE2411)

# INTRODUCTION

In this assignment, regression models are developed for state **BIHAR,** using **STATA** statistical software from the data of **IHDS–2** (2011-2012). First, a **simple regression model** is created and then shifted to the **multiple regression model.** After that, both models are compared, and comments on the changes from simple to multiple regression are made. Then a few more variables were added to the multiple regression model and the differences were observed and explained. After that **dummy variables** are introduced with two categories named **URBAN** and **RURAL.** Following more dummy variables are introduced with more than two categories, and then discussed and interpreted the model. After this, **piecewise regression** is performed. At the end of the assignment, hypothesis testing - called the **Chow test**, has been checked on the regressed model. From this assignment, developing different linear regression models, their interpretation, and the difference in the different models are learned through the practice of this assignment.

# A Glance at Variable in Use

* **CO1X -** Total amount of money spent on rice in Rs.
* **CO2X -**Total amount of money spent on wheat/flour in Rs.
* **CO3X -** Total amount of money spent on sugar in Rs.
* **CO5X -** Total amount of money spent on sugar in Rs.
* **CO6X -** Total amount of money spent on other cereals in Rs.
* **CO7X -** Total amount of money spent on meat in Rs.
* **CO8X -** Total amount of money spent on Gue & sweets in Rs.
* **CO9X -** Total amount of money spent on Edible oil and vanaspati in Rs.
* **CO10X -** Total amount of money spent on Eggs in Rs.
* **CO11X -** Total amount of money spent on Milk in Rs.
* **CO12X -** Total amount of money spent on Milk Prod in Rs.
* **CO13X -** Total amount of money spent on Cereal products in Rs.
* **CO14X -** Total amount of money spent on Vegetables in Rs.
* **CO15 -** Total amount of money spent on Salt/spices in Rs.
* **CO16 -** Total amount of money spent on Tea & coffee in Rs.
* **CO17 -** Total amount of money spent on Processed foods in Rs.
* **CO19 -** Total amount of money spent on Fruits/nuts in Rs.
* **CO20 -** Total amount of money spent on restaurant/ outside dining services in Rs.
* **fdexp -** Sum of all the total amounts spent on consumption items. Capture the Total amount of money spent on food in Rs.

fdexp **=** CO1X + CO2X + CO3X + CO4X + CO5X + CO6X + CO7X + CO8X + CO9X + CO10X + CO11X + CO12X + CO13X + CO14X + CO15 + CO16 + CO18 + CO19

* **COTOTAL -** Total consumption expenditure incurred by a household, measured in Rs.
* **lncototal** - Logarithm of COTOTAL
* **fdshare -** Share of expenditure on food in total consumption expenditure incurred by a household.

So,

fdshare = fdexp / COTOTAL

* **NADULTM -** Number of men, age 21+, in a household
* **NELDERM -** Number of men, age 60+, in a household
* **NADULTM\_n -** Total of men, aged between 21-60, in a household NADULTM\_n = NADULTM- NELDERM
* **NADULTF -** Number of women, age 21+, in a household
* **NELDERF** - Number of women, age 60+, in a household
* **NADULTF\_n -** Total of men, aged between 21-60, in a household NADULTF\_n = NADULTF- NELDERF
* **NPERSONS (n)** - Number of people in a household
* **NCHILDF/NPERSONS (n1) -** NCHILDF is the total number of girls, aged between 0-14, and n1 captures the proportion of the number of girls, ages between 0-14, out of the total number of people in the household.
* **NCHILDM/NPERSONS (n2)** - NCHILDM is the total number of boys, aged between 0-14, and n2 captures the proportion of the number of boys, ages between 0-14, out of the total number of people in the household.
* **NTEENF/NPERSONS (n3)** - NTEENF is the total number of girls, aged between 15-20, and n3 captures the proportion of the number of girls, ages between 15-20, out of the total number of people in the household.
* **NTEENM/NPERSONS (n4)** - NTEENM is the total number of boys, aged between 15-20, and n4 capture the proportion of the number of boys, ages between 15-20, out of the total number of people in the household.
* **NADULTF\_n/NPERSONS (n5) -** n5 capture the proportion of the number of women, ages between 21-60, out of the total number of people in the household.
* **NADULTM\_n/NPERSONS (n6)** - n6 capture the proportion of the number of men, ages between 21-60, out of the total number of people in the household.
* **NELDERF/NPERSONS (n7) -** n7 capture the proportion of the number of women, aged 60+, out of the total number of people in the household.
* **NELDERM/NPERSONS (n8)** - n8 capture the proportion of the number of men, aged 60+, out of the total number of people in the household.
* **n** = NADULTM\_n + NADULTF\_n + NCHILDM + NCHILDF + NTEENM + NTEENF + NELDERM + NELDERF
* **URBAN2011**- Contain the information about region of residence.
* **GROUPS :** tells about the Caste and religion of a household. Contains the value from 1-7 representing the cast/religion as follows:

G1 - Brahmin

G2 - Forward Caste

G3 - Other Backward Caste G4 - Dalits

G5 - Adivasi G6 - Muslim

G7 - Christian, Sikh, Jain.

# 0: Estimation and Interpretation of simple regression model

A General Simple regression model equation is:

𝑌 = 𝛽 + 𝛽 𝑋

0 1

On regressing the simple regression model with **fdshare** as the dependent variables and

**lncototal** (Logarithm of cototal) as the independent variable, following outcomes are attained:

##### Table: 0

|  |  |
| --- | --- |
| **Estimated coefficients and standard error** | **Model 1A** |
| Intercept (Std. error) | 0.17\*\*\* (0.0045) |
| Coefficient of lncototal (Std. error) | -0.011\*\*\* (0.0004) |

The equation for the above regression is as follows:

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 17 − 0. 011𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

It follows that:

* The value of 𝛽 is equal to 0.17. This means that, when the value of lncototal equals 0,

0

the fdshare will be 0.17 that is when the total consumption expenditure of a household is very minimum that is 1 Rs (where lncototal = 0) the share of food expenditure will be

0.17 or 17%1 of the total expenditure

* For the purpose of interpreting the value of 𝛽1 following equation is derrived -

∆𝑓𝑑𝑠ℎ𝑎𝑟𝑒 =− 0. 011∆𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

The change in lncototal, when multiplied by 100, is approximately the percentage change in cototal. So, obtained results are -

∆𝑓𝑑𝑠ℎ𝑎𝑟𝑒 =− (0. 011/100)(%∆𝑐𝑜𝑡𝑜𝑡𝑎𝑙) =− 0. 00011(%∆𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

1 Here it mentions a 17% change in the variable fdshare because it is the ratio of food expenditure to total consumption expenditure, hence multiplying with 100 gives the percentage change.

So, it can be said that if total consumption expenditure is 10% higher, then, the share of food expenditure is predicted to be (0.00011\*10) = 0.0011 percentage points lower.

# 1: Multiple Regression Model

##### Estimation and comparison of the model

A Multiple regression model with two independent variables generally have the following equation -

𝑌 = 𝛽 + 𝛽 𝑋 + 𝛽 𝑋

0 1 1 2 2

On regressing the above model with an additional quadratic term **sqlncototal [=(lncototal)2].**

The following outcomes are obtained –

##### Table: 1

|  |  |  |
| --- | --- | --- |
| **Estimated coefficients and standard error** | **Model 1A** | **Model 1B** |
| Intercept (𝛽 )  0  (Std. error) | 0.17\*\*\* (0.0045) | - 0.35\*\*\* (0.044) |
| Coefficient of lncototal(𝛽1) (Std. error) | -0.011\*\*\* (0.0004) | 0.081\*\*\* (0.008) |
| Coefficient of sqlncototal (𝛽2) (Std. error) | - | -0.004\*\*\* (0.0003) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

The equation for the Model 1B regression is as follows :

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = (− 0. 35) + 0. 081𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 0. 004𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

On moving from the Model 1A to Model 1b it is found that the following results change.

* Model 1A had a positive intercept term but Model 1B has a negative intercept term which means that in case of model 1B , the share of expenditure on food is - 0.35 when

total consumption expenditure is 1 Rupee (i.e. lncototal =0). Although the value of

𝛽 is

0

statistically significant but it is not economically meaningful to have a negative share of food expenditure.

* From Table 1 above it can be said that in the case of model 1A the coefficient of

lncototal was -0.011. But in model 1B it is observed that lncototal has a diminishing effect on expenditure share of food in total consumption expenditure (i.e.) fdshare.

The total effect of total expenditure on the share of food expenditure is captured by differentiating the regressed equation with respect to Total consumption Expenditure.

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = − 0. 35 + 0. 081𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 0. 004𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

On differentiating following are the outcomes -

𝑐𝑜𝑡𝑜𝑡𝑎𝑙

𝑐𝑜𝑡𝑜𝑡𝑎𝑙

∆𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 081 ∆𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 2 × 0. 0041 × 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 ×  ∆𝑐𝑜𝑡𝑜𝑡𝑎𝑙

∆𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = ( 0.081−2×0.004×𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 )( ∆𝑐𝑜𝑡𝑜𝑡𝑎𝑙 × 100)

100

𝑐𝑜𝑡𝑜𝑡𝑎𝑙

Therefore, it can be concluded that the slope coefficient of the relationship between cototal and fdshare depends on the value of total.

.

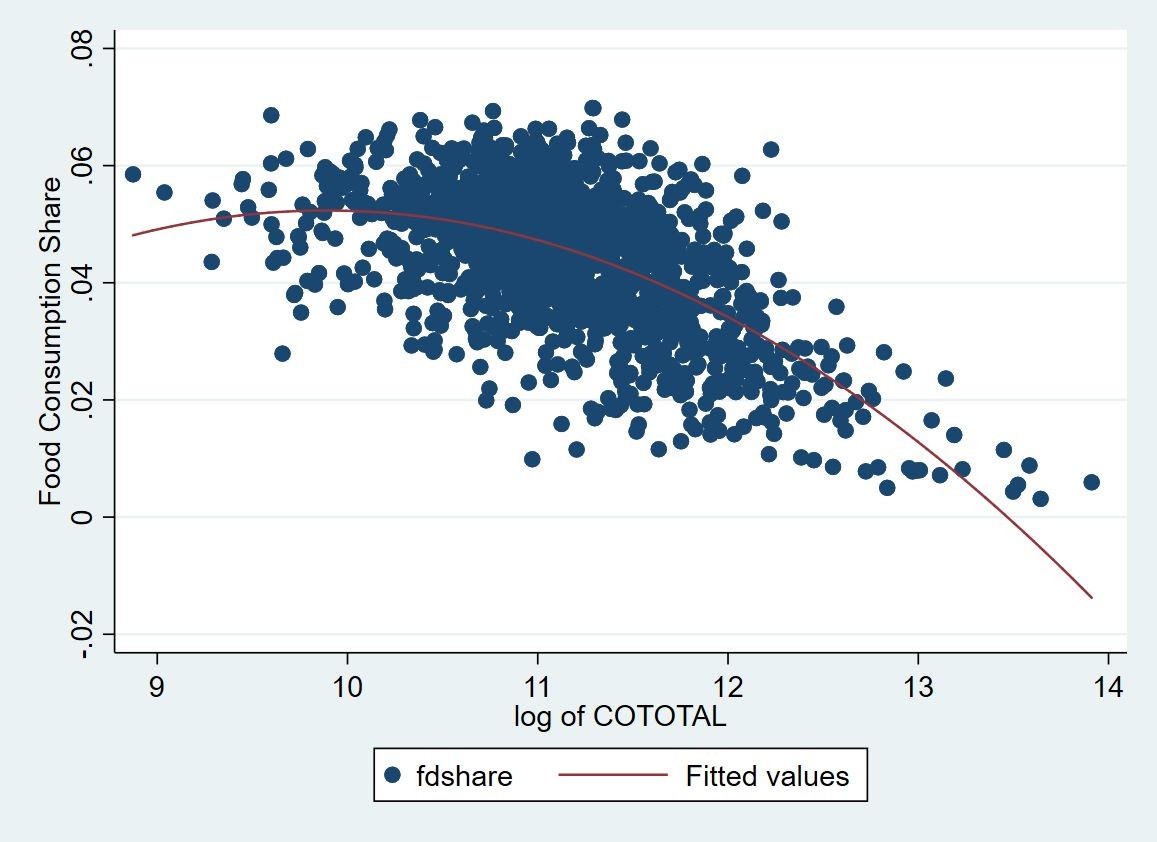
**The estimated slope is =**  0.081−2×0.004×𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

100

If, cototal =1, 0.081 = 0. 0081 can be interpreted as the approximate slope.

100

##### The Turning Point

******

**Graphic representation of Model-1B**

A quadratic function captures decreasing or increasing marginal effect. In the simplest case Y depends on a single observed factor X, but it does so in a quadratic fashion in this case.

𝑌 = β

0

 2

+ β 𝑋 + β 𝑋

1 2

It is important to remember that β does not only measure a change in Y with respect to X; it

1

makes no sense to hold X2 fixed while changing X. So the estimated equation appears as

𝑌 = β

0

 2

+ β 𝑋 + β 𝑋

1 2

As per our data set Y = fdshare, X=log of total consumption expenditure.

In general economic parlance, A person spends only a certain portion of his income on essential expenses. It is seen that share of food expenditure in total consumption expenditure increases only up to a certain point, i.e. the turning point, after which the expenditure on food does not increase proportionally with an increase in income, but rather decreases gradually as there's only a limited amount of food that one can consume irrespective of the increase in income.

As per the results derived from the empirical analysis the estimated equation as -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = − 0. 35 + 0. 081𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 0. 004𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

where, sqlncototal = (𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)2

On rewriting the equation,

2

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = − 0. 35 + 0. 081𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 0. 004 (𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

Differentiating w.r.t. lncototal gives,

### ԁ(𝑓𝑑𝑠ℎ𝑎𝑟𝑒) = 0. 081 − 0. 008 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

By first order condition,

ԁ(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

### 𝑑(𝑓𝑑𝑠ℎ𝑎𝑟𝑒) = 0. 081 − 0. 008 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 =0

⇒ 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 = 0.081

𝑑(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

0.008

⇒ 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 = 10. 125

From the second order condition,

2 𝑓𝑑𝑠ℎ𝑎𝑟𝑒)

𝑑 (

𝑑(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

2

= − 0. 008 < 0

#### So, maximum of fdshare at lncototal obtained is = 10.125. So, the turning point is 10.125.

So, It can be concluded that up to the point 10.125 fdshare increases with an increase in lncototal, and after 10.125 fdshare decreases with an increase in lncototal.

2: New Variables for Household Size and Composition.

##### Variable omitted by the software

On Including the variables NPERSONS and each of the n1 to n8 variables in Model 1B get the following outcome-

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = γ +

1

12

∑ γ 𝑋

𝑖 𝑖

𝑖=2

Where , 𝑋

2

= 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

𝑋 = 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

3

𝑋 = 𝑁𝑃𝐸𝑅𝑆𝑂𝑁𝑆

4

And,

𝑋 𝑡𝑜 𝑋

5 12

𝑖𝑠 𝑛1 𝑡𝑜 𝑛8 respectively.

##### Table: 2.0

|  |  |  |
| --- | --- | --- |
| **Estimated coefficients and standard error** | **Model 1B** | **Model 2A** |
| Intercept (Std. error) | - 0.35\*\*\* (0.044) | -0.17\*\*\* (0.045) |

|  |  |  |
| --- | --- | --- |
| Coefficient of lncototal (Std. error) | 0.08\*\*\* (0.008) | 0.051\*\*\* (0.008) |
| Coefficient of sqlncototal (Std. error) | -0.004\*\*\* (0.0003) | -0.003\*\*\* (0.00035) |
| Coefficient of NPERSONS (Std. error) | - | 0.0015\*\*\* (0.0001) |
| Coefficient of n1 (Std. error) | - | -0.002  (0.0024) |
| Coefficient of n2 (Std. error) | - | -0.0014  (0.0023) |
| Coefficient of n3 (Std. error) | - | -0.0011  (0.003) |
| Coefficient of n4 (Std. error) | - | omitted |
| Coefficient of n5 (Std. error) | - | -0.0002  (0.003) |
| Coefficient of n6 (Std. error) | - | 0.0028  (0.0025) |
| Coefficient of n7 (Std. error) | - | -0.0041  (0.0027) |
| Coefficient of n8 (Std. error) | - | -0.002  (0.00271) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

The software has not given the result for all the variables n1 to n8, it has omitted n4. This is because regressing all the variables into the model would lead to collinearity, that is high correlation among the variables. Since NPERSONS is the sum of n1 to n8, including all

variables will lead to **perfect collinearity** since there would be a linear relationship among the variables. The omitted variable has been taken as a base of the regression model.

##### Variable omitted on Purpose

***Table: 2.1***

|  |  |  |
| --- | --- | --- |
| **Estimated coefficient and standard error** | **Model 1B** | **Model 2B** |
| Constant coefficient (Std. error) | - 0.35\*\*\* (0.044) | -0.171\*\*\* (0.045) |
| Coefficient of lncototal (Std. error) | 0.081\*\*\* (0.0079) | 0.051\*\*\* (0.008) |
| Coefficient of sqlncototal (Std. error) | -0.0041\*\*\* (0.0003) | -0.003\*\*\* (0.0003) |
| Coefficient of NPERSON (Std. error) | - | 0.0015\*\*\* (0.0001) |
| Coefficient of n1 (Std. error) | - | -0.004\*\* (0.002) |
| Coefficient of n2 (Std. error) | - | -0.004  (0.002) |
| Coefficient of n3 (Std. error) | - | -0.0012  (0.003) |
| Coefficient of n4 (Std. error) | - | -0.002  (0.003) |
| Coefficient of n5 (Std. error) | - | -0.002  (0.0025) |
| Coefficient of n6 (Std. error) | - | 0.0005  (0.0021) |
| Coefficient of n7 | - | -0.006\*\* |

|  |  |  |
| --- | --- | --- |
| **Estimated coefficient and standard error** | **Model 1B** | **Model 2B** |
| (Std. error) |  | (0.0029) |
| Coefficient of n8 (Std. error) | - | Omitted |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

The following difference in model 1B and 2B is observed -

#### Coefficient Estimates

In model 1B there are only two variables(lncototal & sqlncototal) whereas there are eleven variables In model 2B. In Model 1B the variables are confined to total expenditure incurred by households whereas In Model 2B number of people of different ages - is also taken into account.

#### Standard Errors

The Standard Errors in models 1B and 2B tell about the variability of the estimates. From Table 2.1 it is seen that the Estimators are more stable in Model 1B than in 2B.

#### Statistical significance

In Model 1B all estimators are significant at 5% and 1% levels. On the other hand in Model 2B, 4 estimators are significant at both the 1% and 5% levels of significance, 2 estimators are significant at the 5% level and the remaining are not significant neither at the 1% level nor at the 5% level.

#### Turning Point

In this case to obtain the turning point, the first order condition is set as -

### ԁ(𝑓𝑑𝑠ℎ𝑎𝑟𝑒) = 0. 051 − 0. 006 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 = 0

⇒ 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 = 8.5

ԁ(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

From the second order condition, the following is obtained -

2 

𝑑 (𝑓𝑑𝑠ℎ𝑎𝑟𝑒) = − 0. 006 < 0

2

𝑑(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)

Here, the maximum of fdshare at lncototal = 8.5. So, the turning point is at 8.5.

**So, it can be concluded that, by introducing variables like NPERSONS, n1, n2, etc. Different consumption patterns of different age groups are taken into account, who may also have a different standard of living, which affects the share of food**

**expenditure in total consumption expenditure consequently. So, the rate of change of the turning point is further affected by more conditions/factors, which is preventing the turning point from increasing more and pushing it towards the left. That is why the turning point value is lesser than the previous one.**

# 3: Qualitative Variable with two categories

##### Model 3A -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = α + α 𝐷

0 1 𝑖

Where , 𝐷 = 1 , i **ϵ Rural**

𝑖

= 0 , i **ϵ Urban**

In model 3A, a dummy variable model is regressed on the share of food expenditure. The dummy variable showcases the characteristic of whether the household belongs to an urban or rural region. Here urban is taken as the base category -

##### Table: 3.0

|  |  |
| --- | --- |
| **Estimated coefficient and standard error** | **Model 3A** |
| Intercept (Std. error) | 0.041\*\*\* (0.0005) |
| Coefficient of urban (Std. error) | 0.0046\*\*\* (0.0006) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

Yes, The variables are significant at 5% and 1% level. The estimated equation looks as follows -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 041 + 0. 0046𝐷

𝑖

From the above equation it can be said that -

* Average share of food expenditure in total consumption expenditure for the people in urban areas is 0.041 (i.e.) 4.1% of total consumption expenditure.
* Difference between the average food expenditure share in total consumption expenditure in people of urban and rural areas is 0.0046.
* So, it may be concluded that the Average food expenditure share on total consumption expenditure for the people in the rural area is (0.041+ 0.0046) = 0.046 (i.e.) 4.6%

##### Model 3B -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = β + β 𝐷

0 1 𝑖

Where, 𝐷 = 1 , i **ϵ Urban**

𝑖

= 0, i **ϵ Rural**

In Model 3B the same variables are regressed as in Model 3A but by taking rural as a base category of the model. The values are recorded in the table below -

##### Table: 3.1

|  |  |  |
| --- | --- | --- |
| **Estimated coefficient and standard error** | **Model 3A** | **Model 3B** |
| Intercept (Std. error) | 0.041\*\*\* (0.0005) | 0.046\*\*\* (0.00035) |
| Coefficient of dummy variable (Std. error) | 0.0046\*\*\* (0.0006) | -0.0046\*\*\* (0.0006) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

**Comparison**

#### Qualitative Variable

The difference in qualitative variables in both the models is discussed below -

In model 3A the urban average food expenditure to total consumption expenditure (𝝰0) is obtained as the intercept term and in model 3B the intercept term is the average food expenditure to total consumption expenditure for rural areas (𝞫0). It is observed that

α (= 0. 041)+ α (= 0. 0046) ≈ β (=0.046). Also, it is observed that the absolute value of the

0 1 0

difference between the intercept and slope coefficient is the same for the two models the sign

changes only. So it can also be said that β (= 0. 046)+ β (=− 0. 0046) ≈ α (=0.041).

0 1 0

**It can be concluded from the given data that the rural share of food expenditure in total consumption expenditure is higher than the urban share of food expenditure in total**

**consumption expenditure, it can be reasoned out that in urban areas people spend more on non-essential (other than food share expenses), to maintain their standard of living, while in rural areas, as people do not spend as much in non-essential items, the standard of living being relatively lower, comparatively, the proportionate expenditure on food seems higher.**

1. **Standard Error**
   * **Model 3A -**

It can be seen that the standard error of the intercept term is 0.0005 and the slope coefficient is 0.006

#### Model 3B -

It can be seen that the standard error of the intercept term is 0.00035 and the slope coefficient is 0.0006

It can be concluded from the values of standard errors that they are almost equal to zero. So, the values of the coefficient in the sample are not very different from the population. But for both the models the standard error has changed for the intercept and slope coefficient.

#### T- Statistic

* + **Model 3A**

The t-statistic for intercept term (α ) = 74.74

0

The t-statistic for slope coefficient (α ) = 6.96

1

#### Model 3B

The t-statistic for intercept term (β ) = 127.42

0

The t-statistic for slope coefficient (β ) = -6.96

1

So, it can be concluded that, **t-statistic of**

#### α = -( t-statistic of β )

1 1

#### Adjusted R2

The adjusted R2 for both models is = 0.0298. (i.e.) it can be concluded that the explanatory power of the independent categories does not change in the two models, so the adjusted R2 also does not change.

##### Model 3C and Model 3D -

The variables lncototal and sqlncototal are added in Model 3A and Model 3B respectively to get the results below -

##### Table: 3.1

|  |  |  |
| --- | --- | --- |
| **Estimated coefficient and standard error** | **Model 3C** | **Model 3D** |
| Intercept (Std. error) | -0.35\*\*\* (0.045) | -0.35\*\*\* (0.044) |
| Coefficient of lncototal (Std. error) | 0.08\*\*\* (0.008) | 0.08\*\*\* (0.008) |
| Coefficient of sqlncototal (Std. error) | -0.004\*\*\* (0.0003) | -0.004\*\*\* (0.0003) |
| Coefficient of dummy variable (Std. error) | 0.0009  (0.0005) | -0.0009  (0.0005) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

There is no change in intercept term of lncototal in this model as it can be considered to be the model for rural and urban areas, as, even if it is considered as a dummy variable, it covers the whole demography, so there is no change in values.

# 4: Qualitative Variable with more than two categories

##### Model 4A -

|  |  |
| --- | --- |
| **Estimated coefficient and standard error** | **Model 4A** |
| Intercept (Std. error) | 0.041\*\*\* (0.0008) |
| Coefficient of Groups (Std. error) | 0.0008\*\*\* (0.0002) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

From the above table the regression is the equation as follows:

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 041 + 0. 0008(𝐺𝑅𝑂𝑈𝑃𝑆)

* From the above regression equation, the intercept term is obtained as 0.041 and the slope coefficient of the group is obtained as 0.0008.
* 0.0008 captures the factor change in the share of food expenditure in total consumption expenditure of the groups.

##### Model 4B -

***Table 4.1***

|  |  |  |
| --- | --- | --- |
| **Estimated coefficient and standard error** | **Model 4A** | **Model 4B** |
| Intercept(𝞪0) (Std. error) | 0.041\*\*\* (0.0008) | 0.032\*\*\* (0.012) |
| Coefficient of Groups(𝞪) (Std. error) | 0.0008\*\*\* (0.0002) | - |
| Coefficient of Bhramin(𝞪1) (Std. error) | - | 0.005  (0.011) |
| Coefficient of Forward Caste(𝞪2) (Std. error) | - | 0.011  (0.011) |
| Coefficient of OBC(𝞪3) (Std. error) | - | 0.0116  (0.012) |
| Coefficient of Dalit(𝞪4) (Std. error) | - | 0.015  (0.012) |
| Coefficient of Adivasi(𝞪5) (Std. error) | - | 0.013  (0.012) |
| Coefficient of Muslim(𝞪6) (Std. error) | - | 0.012  (0.012) |
| Coefficient of Sikh and Christian (Std. error) | - | omitted |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

#### Comparison of model

* + From the above table it can be observed that in Model 4A a single variable-group, captures the information of all the groups and gives an aggregate change for all groups

which makes it difficult to analyze the food expenditure pattern for an individual group, whereas model 4B have different variables for different groups and hence it easy to analyse a particular group’s food expenditure pattern.

* + Model 4A doesn't capture the individual group's consumption pattern on the other hand it

does not provide a lucid interpretation of data from the values, as in the table because it is a single variable group. In Model 4B, such data can be obtained from a simple overview of the values corresponding to the variables in the table

Hence, model 4B can be considered a better estimator of the share of food expenditure out of total consumption expenditure.

#### Hypothesis testing

From the model developed the following equation is obtained.:

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = α + α 𝐺 + α 𝐺 + α 𝐺 + α 𝐺 + α 𝐺 + α 𝐺 + 𝑢

0 1 1 2 2 3 3 4 4 5 5 6 6 𝑖

Taking expectation of the equation given all the 𝐺 = 0 i.e i belongs to the G7

𝑖

𝐸[𝑓𝑑𝑠ℎ𝑎𝑟𝑒|𝑏𝑎𝑠𝑒𝑔𝑟𝑜𝑢𝑝 = 𝐺 ] = α

7 0

Taking expectation of the equation given all the 𝐺 = 1 i.e i belongs to the G1

1

𝐸[𝑓𝑑𝑠ℎ𝑎𝑟𝑒|𝑐𝑜𝑚𝑝𝑎𝑟𝑖𝑠𝑜𝑛 𝑔𝑟𝑜𝑢𝑝 = 𝐺 ] = α + α

1 0 1

Stating the Hypothesis for the test to be followed:

H : α + α = α , ⇒ 𝐻 : α = 0

0

0 1 0 0 1

𝐻 : α ≠ 0

𝑎 1

Calculating the t value -

α − α α

𝑡 = 1 1 = 1

𝑐𝑎𝑙

𝑆.𝐸.(α )

1

𝑆.𝐸.(α )

1

𝑡

𝑐𝑎𝑙

= 0.005 = 0. 42

Since the P-value obtained from our data is 0.68 which is greater than 0.05. Hence the null hypothesis may be accepted at a 95% level of significance so the average share of food expenditure in total consumption expenditure for a group of 1 is not different from group 7.

0.012

#### One-sided Hypothesis testing

From the regression model (table 4.1) it is observed that the coefficients of the G5 and G6 are the closest to each other, which are Aadivasi and Muslim respectively.

On the basis of the given query, the hypothesis will be set as

Ho: 𝞪5 = 𝞪6 or Ho: 𝞪5 - 𝞪6 = 0 Ha: 𝞪5 > 𝞪6 or Ha: 𝞪5 - 𝞪6 > 0

( α − α )

Calculating the t value, 𝑡

= 5 6

𝑐𝑎𝑙

𝑆𝐸(α

5

− α )

6

Since the P-value obtained from the use of software STATA equals to 0.92 which is greater than 0.05, the null hypothesis may be accepted at 95% level of significance. This implies that there will be no difference in the share of food expenditure in the total consumption expenditure.

# 5: Piecewise linear regression model

Piecewise regression model helps us to address the discontinuity in any model where at the point of discontinuity the value of the dependent variable is the same.

For estimating a piecewise linear regression model regression for the value less than 10.125 and greater than 10.125 has been followed. This is because of the turning point at 10.125 of Model 1B. (Which is the threshold level taken here).

The outcome of regressing the model for less than 10.125 the following output is obtained -

##### Table 5.1

|  |  |  |
| --- | --- | --- |
| **Estimated coefficient and standard error** | **for value** ≤**10.125** | **for value >10.125** |
| Intercept(𝞪0) | 0.034 | 0.19 |
| (Std. error) | (0.036) | (0.005) |
| Coefficient of Groups(𝞪1) | 0.0018 | -0.013 |
| (Std. error) | (0.0037) | (0.0004) |

The estimated equations for the above models are -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 034 + 0. 0018(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙), For lncototal ≤ 10.125

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 19 − 0. 013(𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙), For lncototal > 10.125

On estimating the piecewise regression, the following outputs are attained:

##### Table 5.2

|  |  |
| --- | --- |
| **Estimated coefficient and standard error** | **Values** |
| lncototal\_3 (Std. error) | 0.0018  (0.004) |
| lncototal\_4 (Std. error) | -0.013  (0.00046) |
| dum\_3 (Std. error) | 0.05  (0.0018) |
| dum\_4 (Std. error) | 0.058  (0.00056) |

From the above outcome the equation of Econometric model can be stated as :

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 0. 052 + 0. 0018𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + 𝐷 (0. 0056 − 0. 015𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙)2

𝑖

for, Di = 0 for

𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 ≤ 10. 125

Di = 1 for 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 > 10. 125

From the equation above followings are the drawn interpretation:

* Di is the dummy variable that helps to differentiate the trend of the model. when lncototal is less than 10.125, the Di = 0, and therefore the intercept term will be 0.052 and the slope coefficient will be 0.0018.
* 0.052 will be the initial intercept term of the model, i.e when lncototal will be zero,

implying that the cototal is near or equal to one, the share of food expenditure will be 5.2%. The Di will be 0 because the lncototal is less than 10.125.

* When the lncototal will be greater than 10.125 the intercept term will be 0.0576 and the

slope coefficient will be -0.0132.

2 Since here the turning point is an approximated value, the difference in intercept term can be observed between the individual model and piece wise model. If the actual turning point value is obtained, then no difference will be observed.

* The model will show an upward trend in the share of food expenditure till the point where the lncototal will be equal to 10.125, after reaching the maximum, the model will show a downward trend in the share of food expenditure out of total consumption expenditure.

# 6: Chow-test and two separate models for a region of residence

There are two regions of residence for the data containing information about Bihar state - rural and urban. Regressing through both of them gives the pooled model.

##### Pooled model -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = 𝛽 + 𝛽 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 𝛽 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

0 1 2

##### Table 6.1

|  |  |
| --- | --- |
| **Estimated coefficient and standard error** | **Model 1B** |
| Intercept (𝛽 )  0  (Std. error) | - 0.35\*\*\* (0.044) |
| Coefficient of lncototal(𝛽1) (Std. error) | 0.081\*\*\* (0.008) |
| Coefficient of sqlncototal (𝛽2) (Std. error) | -0.004\*\*\* (0.0003) |
| RSSpooled | 0.138 |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

## Chow test -

To test if there is any difference between the people of two regions of residence the Chow test is pursued.

To do the Chow test the homoscedasticity assumption of the error term must hold and the parameters of the two models of two groups should be less than the number of observations.

#### The model with Group 1 (people residing in Rural areas of Bihar) -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = β + β 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 𝛽 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

10 11 12

The Residual Sum of squares of this model is defined as RSS1

#### Model with Group 2 (people residing in Urban areas of Bihar) -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = β + β 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 𝛽 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

20 21 22

The Residual Sum of squares of this model is defined as RSS2**.** Now, to do the Chow test, the null hypothesis is defined as -

H : β = β ; β = β ; β = β

0

10 20 11 21 12 22

H1: At least one of them is not equal

Now,

[𝑅𝑆𝑆 −(𝑅𝑆𝑆 +𝑅𝑆𝑆 )]/𝐾

𝐹

=

𝑐𝑎𝑙

𝑝𝑜𝑜𝑙𝑒𝑑 1 2

(𝑅𝑆𝑆 +𝑅𝑆𝑆 )/ 𝑛 +𝑛 −2𝐾

1 2 1 2

Where, equality of K coefficients = K - restrictions with n1 + n2 = n total no. of observations and K+K = 2K total no. of parameters put together.

Now, from the dataset, the following is obtained-

𝐹

𝑐𝑎𝑙

= [0.1383−(0.1057+0.0317)]/3 = 2. 873

(0.1057+0.0317)/ 1085+461−6

Here the degrees of freedom are k (no. of restrictions) = 3 and (n-2K) = 1540, where n is the number of observations, here n1+n2.=n

Now that the tabulated value of F is:

𝐹 = 0. 035

𝑡𝑎𝑏, 3, 1540

Therefore, .

𝐹 > 𝐹

𝑐𝑎𝑙 𝑡𝑎𝑏, 3, 1540

#### So, the null hypothesis is rejected.

It is found that there is a difference between the share of food expenditure out of total consumption expenditure of urban and rural regions, correctly so, due to various factors including the standard of living, so, the null hypothesis, that both, being same, does not hold.

## Chow test with dummy variables -

On regressing the specified econometric model the following outcomes are obtained-

##### Table 6.2

|  |  |
| --- | --- |
| **Estimated coefficient and standard error** | **Model 1B** |
| Intercept (𝛽 )  0  (Std. error) | - 0.43\*\*\* (0.057) |
| Coefficient of lncototal(𝛽1) (Std. error) | 0.096\*\*\* (0.0103) |
| Coefficient of sqlncototal (𝛽2) (Std. error) | -0.005\*\*\* (0.0005) |
| lnX (Std. error) | -0.038\*\* (0.017) |
| (lnX)2 (Std. error) | 0.0017\*\* (0.0008) |
| URBAN2011  (Std. error) | 0.21\*\* (0.097) |

\*\* - denotes the statistical significant at 5% level

\*\*\* - denotes the statistical significant at 1% level

lnX and (lnX)2 in the above table are the interaction terms between dummy, lncototal and sqlncototal respectively, that captures the difference in intercept and slope coefficient the regression of share of food expenditure between rural and urban

Here, the values of coefficients as well as the dummy variable are statistically significant, it can be reasoned out as the difference between urban and rural region is reasonable. If it was not the

case, the Joint hypothesis would have been considered, instead of relying on individual t statistics.

The model is as follows on the introduction dummy variable-

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = β

/

+ β 𝐷

+ β 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + β

/

(𝐷 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙) + β 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + β

/

(𝐷 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙) + 𝑢

1 1 𝑖 2

2 𝑖 3

3 𝑖 𝑖

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 =

− 0. 43 + 0. 21𝐷 + (0. 096 − 0. 038𝐷 )𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + (− 0. 005 + 0. 0017𝐷 )𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

𝑖

𝑖

𝑖

#### For Group 1 (People residing in rural areas): when, 𝐷

𝑖

becomes:

= 0, the regression equation

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = β + β 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙+ β 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + 𝑢

1 2 3 𝑖

After regressing the following is obtained -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = − 0. 43 + 0. 096𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 0. 005𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

#### For Group 2 (People residing in urban areas): when, 𝐷

𝑖

becomes:

= 1, the regression equation

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = (β

1

/

+ β ) + (β

1 2

/

+ β )𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + (β

2 3

/

+ β )𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 + 𝑢

3 𝑖

**(1)**

#### This can be considered as our unrestricted model.

The Regression gives following outcome -

𝑓𝑑𝑠ℎ𝑎𝑟𝑒 = − 0. 22 + 0. 058𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙 − 0. 0033𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙

To test that there if there is any difference between the share of food expenditure out of total consumption expenditure of people of two regions of residence the null and alternative hypothesis is defined as -

𝐻 : β = β = β = 0

/

/

/

0 1 2 3

Ha: At least one of them is non-zero

So, the restricted model from (1) becomes -

fdshare = β + β 𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙+ β 𝑠𝑞𝑙𝑛𝑐𝑜𝑡𝑜𝑡𝑎𝑙+𝑢

1 2 3 𝑖

Therefore the F-statistic is -

(𝑅𝑆𝑆 − 𝑅𝑆𝑆 )/(𝑘 −𝑘 )

𝐹 = 𝑅 𝑈𝑅 𝑈𝑅 𝑅

𝑐𝑎𝑙

𝑅𝑆𝑆 /(𝑛−𝑘 )

𝑈𝑅 𝑈𝑅

𝑅𝑆𝑆 ⇒

𝑅

Residual sum of squares of restricted model

𝑅𝑆𝑆 ⇒

𝑈𝑅

Residual sum of squares of unrestricted model

𝑘 ⇒ No. of parameters in the unrestricted model

𝑈𝑅

𝑘 ⇒ No. of parameters in the restricted model

𝑅

n ⇒ No. of total observations

Here, the degrees of freedom are 3 and 1540.

The p-value obtained from the data is 0.0351 at 3 and 1540 degrees of freedom, which is less than 0.05 hence null hypothesis may be rejected at a 95% level of significance. But under a 99% level of significance, the null hypothesis may be accepted. This implies that rural and urban areas have different consumption patterns.

Further t-test for testing the statistical significance of

/

β 𝑜𝑟 β

1

/

𝑜𝑟 β

2

/ any two or all three will

3

inform which of the independent variable(s) has a differential effect on the dependent variable fdshare keeping everything else constant.

#### The Advantage of econometric analysis between B) and C) -

When the Chow test is used, the data from both the sets are pooled, i.e. urban and rural, which will allow us to visualize the region-wise structural break, and identify stability, but not individually. The Dummy variable analysis acts as an extension of this existing model, allowing us to identify the stability of a coefficient or component.