Analysis of Expenditure in a Household

Kushal Dornahalli
230102052
Dept. of EEE
IIT Guwahati
d.kushal@iitg.ac.in

Sutheertha Enugandula
230102031
Dept. of EEE
IIT Guwahati
s.enugandula@iitg.ac.in

Neel Kedia
230103127
Dept. of Mechanical Eng.
IIT Guwahati
n.kedia@iitg.ac.in

Vansh Gandharva
230102102
Dept. of EEE
IIT Guwahati
g.vansh@iitg.ac.in

Divyaansh Ranjan
230108016
Dept. of EEE
IIT Guwahati
r.divyaansh@iitg.ac.in

Swastik Pramanik
230102094
Dept. of EEE
IIT Guwahati
p.swastik@iitg.ac.in

Chinmay Pankaj Torvi
230108014
Dept. of EEE
IIT Guwahati
c.torvi@iitg.ac.in

Sourajjal Mondal
230103098
Dept. of Mechanical Eng.
IIT Guwahati
m.sourajjal@iitg.ac.in

CONTENTS

I. INTRODUCTION

Household consumption expenditure is a fundamental indicator of economic well-being and macroeconomic dynamics. Understanding its determinants is crucial for analyzing income distribution, savings behavior, and growth patterns within an economy.

This study models total household consumption expenditure in India using an Ordinary Least Squares (OLS) regression framework. The analysis is based on data from the Indian Human Development Survey-II (IHDS-II), a nationally representative dataset collected in 2011–12. The IHDS-II captures extensive information on household income, consumption, assets, employment, education, and health, making it particularly suited for studying financial behavior and socio-economic conditions across diverse segments of the Indian population.

The regression model specifies household consumption expenditure as the dependent variable, with independent variables including State, Religion, Primary Source of Income, Current Value of Livestock, Assets Owned, and Debt Incurred over the past five years. Categorical variables such as State and Religion are converted into dummy variables to ensure numerical representation without imposing artificial order.

II. DATA SELECTION AND METHODOLOGY

The IHDS-II is a comprehensive dataset collected from over **42,000** households. This survey is conducted by National Council of Applied Economic Research (NCAER) and the University of Maryland. The data is available through the Data Sharing for Demographic Research program by ICPSR, the Interuniversity Consortium for Political and Social Research. The data from this survey yielded 14 total datasets, of which we chose to use the Household data set. This survey gives a cross-sectional data set, which means that it provides information from households at a single point of time, in this case, for the year 2011-12. Due to the large size of this dataset, it ensures a diverse analysis of household economic impacts.

A timeseries dataset is one which lets us observe the variation of trends with time. However, we opted not to use such a data set due to a lack of observations available. We required a large number of data points for a robust analysis and thus we used this cross-sectional dataset.

III. VARIABLE SELECTION

The dataset consisted of 758 variables. We had to select our outcome variable and thus appropriate independent variables and drop all the unnecessary ones. The dependent variable for this study is the total consumption expenditure of the

household (Hconsumption). This variable represents the total amount of money that a household spends on goods and services, including housing, taxes for the state in which it lies, debt incurred, education etc. Higher household expenditure indicates that the earner or earners of that house have a stable and high income so are able to spend more on various essentials for their well-being.

We have tried to incorporate a diverse set of independent variables to obtain a well-rounded analysis of our outcome variable. These variables were justified by self-intuition coupled with some economic theories. They can be broadly classified into three categories as follows:

A. Economic variables

1) **Total Income**: According to the Keynesian consumption function, expenditure is directly related to income, but the rate of increase of spending is lower than that of income. Higher income allows for increased spending, thus boosting overall expenses.
2) **Assets**: The Life-Cycle Hypothesis (Modigliani & Brumberg, 1954) suggests that households with accumulated assets will spend more as they have higher financial security.
3) **Debt**: According to the Permanent Income Hypothesis (Friedman, 1957), households with high debt prioritize repayment over expenditure for consumption, reducing their spending capacity.
4) **House Ownership**: The Wealth Effect Theory suggests that owning a house reduces the need for rent expenses and increases spending in other categories.
5) **Wedding Cost**: Based on the Precautionary Saving Hypothesis, households making large one-time expenditures, such as weddings, may adjust their future spending patterns to compensate for the large expenditure.
6) **Livestocks:** This variable is measured in Rs. This is an important economic variable for farmers or people whose income depends on animal husbandry as the amount of livestock they possess directly impacts their income.

B. Demographic Variables

1) **Family size**: According to the Engel curve, larger families require higher expenditures on necessities such as food, healthcare, and education.
2) **Marriage Status**: The Household Production Theory suggests that married households pool resources, influencing spending decisions and consumption patterns.
3) **Education Level**: The Human Capital Theory (Becker, 1964) states that education improves earning potential, thereby increasing disposable income and consumption.
4) **Health Insurance**: The Precautionary Savings Model suggests that households with health insurance face lower unexpected medical expenses, leading to different spending behaviors.
5) **Life Insurance**: According to Fischer's Consumption Smoothing Model, households with life insurance plans spend more cautiously, leading to adjustments in short-term spending.

C. Social and Environment Varaibles

1) **Migrant Work**: The Remittance Hypothesis suggests that migrant workers send money home, which alters household spending patterns.
2) **Living Conditions**: The Standard of Living Hypothesis suggests that better housing conditions reflect higher economic standing and greater consumption capacity.
3) **Income Class**: There are different types of income class that we have referred to as Poor, Middle and Comfortable.
4) **CS Knowledge (Financial Literacy)**: According to the Behavioral Economics Theory, financially literate households make informed consumption choices, reducing wasteful spending.
5) **Tragedies**: The Liquidity Constraint Theory implies that sudden financial shocks reduce disposable income, causing a decline in household spending.
6) **Crime Rates**: High-crime areas may lead to increased precautionary savings or security-related expenditures, aligning with the Risk Aversion Theory.
7) **Rooms in House**: Larger homes indicate better economic standing, aligning with the Housing Wealth Effect Hypothesis.
8) **State**: Expenditure will depend on the state in which a person lives in due to various financial state laws which will be different for each state.
9) **Religion**: Different religions have different priorities and different festivities and thus spend accordingly.

IV. Data Preprocessing

Before conducting the regression analysis on the dataset it went through various preprocessing and cleaning procedures.

A. Handling Missing values

Few variables like Income, Migrant_work, Living Conditions and Rooms contained missing values. As the number of missing values wasn't very comparable to the actual dataset, the rows with missing values were removed, reducing the dataset from 42,152 to 39,819. This ensures a consistent analysis, as regression models require numeric values for fitting and prediction.

B. Consolidation of variables

Some variables that offer similar impacts on the outcome variable were combined to strengthen their implications and thus improve the model. Government and private life insurance were combined to a single variable, life insurance. Similarly, thefts, robberies, and attacks were incorporated into crime; job loss or deaths under tragedies, etc.

C. Encoding categorical values

Since categorical variables like State, Religion, Marriage Status, and Education Level are non-numeric and do not carry any inherent order or weight, they cannot be directly used in a regression model, which requires numerical input. If we assigned arbitrary numbers (e.g., 1, 2, 3) to these categories, it would incorrectly imply that one category is "greater" or "lesser" than another, introducing false relationships.

To avoid this, we create dummy variables — binary indicators (0 or 1) — for each category. This method preserves the qualitative nature of the data without imposing any artificial ranking or weight. Dummy coding allows the regression model to interpret the presence (1) or absence (0) of a category without introducing misleading assumptions about their relative importance.

D. Outliers and Normalization

Extreme values were identified using scatter plot and standard deviation method and such rows were removed and normalization was done to reduce dominance of other variable over other for better analysis.

V. Multiple Linear Regression

A. Concept

Multiple Linear Regression (MLR) estimates the relationship between a dependent variable and multiple independent variables. The general regression equation is:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Where:

- $\hat{Y}$ = predicted HComsumption from regression
- $X_1, X_2, \cdots, X_n$ = Independent variables
- β_0 = Intercept
- $\beta_1, \beta_2, \cdots, \beta_n$ = Coefficients
- ε = Error term

The regression model aims to estimate the coefficients of the independent variables, thereby providing us with an idea of influential each variable is. This is achieved using the method of Ordinary Least Squares (OLS). It minimizes the sum of the squared differences between the observed and the predicted values from the regression model i.e. $\min(\sum(Y-\hat{Y})^2)$ where $Y =$ actual value of HConsumption at various datapoints. For minimization we differentiate wrt β:

$$\frac{\partial(\sum(Y-\hat{Y})^2)}{\partial\beta_i} = 0$$

and arrive at the coefficients.

B. Regression Outcomes

- **Coefficients**: Show the impact of each independent variable on the dependent variable, household consumption expenditure.
- **P-values**: Determine the statistical significance of each variable by measuring how extreme our sample is with respect to the null hypothesis with the help of t-values.
- **R-squared value**: Indicates the model's explanatory power i.e. how well the independent variables explain the variability of the dependent variable.
- **Adjusted R-squared**: Accounts for the number of predictors, making it more reliable than the R-squared value.

C. STATA code for regression

In STATA, we can obtain the regression of our data using the following code. STATA utilises the OLS method in its estimation by default.

```
reg HComsumption Income Migrant_Work Livestocks
Living_Conditions CS_Knowledge House_Ownership
Income_Class Assets Debt Marriage_Status Rooms
Family_Size Education Life_Insurance
Health_Insurance Crime Wedding_Costs Tragedies
```

D. Expalnation of code

This piece of code detects the household consumption using MLR. A set of ecconomic, demographic and scocial environmental variables were taken. Additionally, the technique of partialling out was applied to isolate the impact of one variable (State) while controlling for all other factors.

The fitted values (HConsumptionHat) were obtained to represent the predicted household consumption based on the estimated model from MLR. The residuals (uHat) captured the deviation of actual household consumption from the predicted values. To isolate the effect of State on Hconsumption, State was regressed on all other independent variables to obtain residuals (ehat), which represent the variation in State that is not explained by other factors. Then, HConsumption was regressed on ehat to estimate the direct relationship between State and HConsumption, independent of other influences.

E. Regression diagnostics

1) Multicollinearity Check
   - Variance Inflation Factor (VIF) was used to detect multicollinearity.
   - Variables with VIF > 10 were examined for redundancy.
2) Heteroskedasticity Test
   - The Breusch-Pagan test was conducted to check for heteroskedasticity (unequal variance in residuals).
   - If detected, robust standard errors were applied to correct it.
3) Residual Normality Check
   - Residuals were plotted to confirm a normal distribution.

VI. Verification of Regression Results

A. Goodness of Fit

The R-squared value is a measure of the goodness of fit or how well the regression fits the data. Initially, we obtained the value of adjusted R-squared as 0.3293. To improve this value we analysed the graphs and arrived at the best case, when log(HConsumption) and log(WeddingCosts) is used in regression, without changing other variables. Adjusted R-squared value reached upto 0.599.

Source	SS	df	MS	Number of obs = 40414
				F(57, 40356) = 1185.99
Model	13418.5673	57	235.413462	Prob > F = 0.0000
Residual	8010.46534	40356	.198495028	R-squared = 0.6262
				Adj R-squared = 0.6257
Total	21429.0326	40413	.530250975	Root MSE = .44553

log_HConsumption	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
State_1	.1794058	.0606764	2.96	0.003	.0604787	.2983329
State_2	-.3896305	.0592542	-6.58	0.000	-.50577	-.273491
State_3	-.1717304	.0598344	-2.87	0.004	-.2890072	-.0544535
State_4	-.0987787	.0762162	-1.30	0.195	-.2481642	.0506069
State_5	-.4228724	.0616682	-6.86	0.000	-.5437436	-.3020013
State_6	-.169155	.0590591	-2.86	0.004	-.2849122	-.0533978
State_7	-.056813	.0600825	-0.95	0.344	-.1745762	.0609501
State_8	-.2158139	.0587413	-3.67	0.000	-.3309482	-.1006795
State_9	-.2143917	.0586057	-3.66	0.000	-.3292603	-.0995232
State_10	-.2845347	.0592984	-4.80	0.000	-.4007609	-.1683086
State_11	-.2221103	.0735835	-3.02	0.003	-.3663357	-.077885
State_12	.0076886	.0695106	0.11	0.912	-.1285538	.143931
State_13	.1771375	.0751398	2.36	0.018	.0298617	.3244132
State_14	.1440767	.0756371	1.90	0.057	-.0041737	.2923271
State_15	.1363681	.0795701	1.71	0.087	-.019591	.2923272
State_16	-.1441176	.0655615	-2.20	0.028	-.2726197	-.0156155
State_17	-.1453383	.0716769	-2.03	0.043	-.2858266	-.00485
State_18	-.1094768	.0600608	-1.82	0.068	-.2271974	.0082437
State_19	-.3413659	.0588372	-5.80	0.000	-.4566882	-.2260437
State_20	-.3827386	.060179	-6.36	0.000	-.5006908	-.2647864
State_21	-.4364434	.0589479	-7.40	0.000	-.5519826	-.3209042
State_22	-.471718	.0594611	-7.93	0.000	-.5882632	-.3551728
State_23	-.2390689	.0586803	-4.07	0.000	-.3540835	-.1240542
State_24	-.1231325	.0589847	-2.09	0.037	-.2387439	-.007521
State_25	-.0586383	.0820895	-0.71	0.475	-.2195356	.1022591
State_26	0	(omitted)				
State_27	-.2927151	.0586366	-4.99	0.000	-.4076442	-.177786
State_28	-.1520439	.0589156	-2.58	0.010	-.2675198	-.0365679
State_29	-.1919873	.0585814	-3.28	0.001	-.3068081	-.0771665
State_30	-.3206721	.0676276	-4.74	0.000	-.4532238	-.1881204
State_31	-.4036545	.0593955	-6.80	0.000	-.520071	-.287238
State_32	-.4664148	.0589891	-7.91	0.000	-.5820347	-.3507949

(a)

State_33	-.5662755	.0724833	-7.81	0.000	-.7083445	-.4242065
Religion_1	.035533	.0729584	0.49	0.626	-.1074671	.1785331
Religion_2	.0620207	.0732117	0.85	0.397	-.0814759	.2055172
Religion_3	.0873766	.0741643	1.18	0.239	-.0579872	.2327404
Religion_4	.0835589	.0752463	1.11	0.267	-.0639255	.2310433
Religion_5	-.0320711	.078508	-0.41	0.683	-.1859485	.1218064
Religion_6	-.0821779	.0859481	-0.96	0.339	-.2506382	.0862823
Religion_7	.0230257	.0789505	0.29	0.771	-.131719	.1777704
Religion_8	0	(omitted)				
Religion_9	.1456824	.1597557	0.91	0.362	-.1674425	.4588073
INCOMEPC	3.12e-10	4.08e-08	0.01	0.994	-7.96e-08	8.02e-08
Migrant_Work	.0584365	.0092756	6.30	0.000	.0402561	.0766169
Livestocks	5.30e-07	4.63e-08	11.46	0.000	4.40e-07	6.21e-07
CS_Knowledge	.136704	.0063634	21.48	0.000	.1242316	.1491763
House_Ownership	.0464425	.0054504	8.52	0.000	.0357596	.0571254
Income_Class	.1079259	.0047652	22.65	0.000	.0985859	.1172658
Assets	.0442855	.0005937	74.59	0.000	.0431218	.0454491
Debt	.1335476	.0048167	27.73	0.000	.1241067	.1429885
Marriage_Status	.0618232	.004871	12.69	0.000	.0522759	.0713705
Rooms	.0085487	.0015879	5.38	0.000	.0054364	.0116611
Family_Size	.0820932	.0010519	78.05	0.000	.0800316	.0841549
Education	.0072671	.0006005	12.10	0.000	.0060901	.0084442
Life_Insurance	.1588215	.0055292	28.72	0.000	.1479842	.1696589
Health_Insurance	.042883	.0074545	5.75	0.000	.0282721	.0574939
Crime	.0842839	.0098819	8.53	0.000	.0649151	.1036527
log_Wedding_Costs	.0902805	.0033651	26.83	0.000	.0836848	.0968763
Tragedies	.0814265	.0046138	17.65	0.000	.0723833	.0904696
_cons	8.82808	.1015714	86.92	0.000	8.628998	9.027162

(b)

Fig. 1: Regression Table

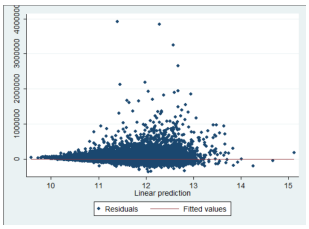

Fig. 2: uHat vs yHat: uHat = difference between predicted and actual value, yHat = predicted value

B. Multicollinearity Check

Multicollinearity is a situation in which independent variables are highly correlated with each other, which makes it difficult to obtain their individual impacts. It causes inaccurate estimates and inflated standard errors.

- Variance Inflation Factor (VIF) was used to detect multicollinearity.
- Our VIF values were in the range of 1 to 3, with a mean of 1.29.
- Variables with VIF > 10 are theoretically examined for redundancy, but obviously not needed in our case.

C. Heteroscedasticity and Homoscedasticity

Homoscedasticity is when the variance of the residuals in a regression model remains constant across all levels of the independent variables. Heteroscedasticity is when the variance of the residuals varies. A key assumption of our regression method, OLS method, is homoscedasticity. Thus it is important to ensure that the assumption holds true.

- Breusch-Pagan Test - We obtained p-values greater than 0.05, indicating a lack of heteroscedasticity which is preferred.
- White Test - White test is an extension of the Brueusch-Pagan test where the squared residuals are further regressed on the independent variables, squared independent variables, and cross-product interactions. The p-values are re-examined.
- Graphical Method - We ensured that the regression scatter plot indicated no heteroscedasticity.
  When we graphed log_HConsumption and Wedding cost the heteroscedasticity was on the higher side and so we used log of wedding to reduce it.

D. Hypothesis Testing

We do the hypothesis testing of whether to reject the null hypothesis, which whether to remove any column, or not by 2 methods.

- T-test : Comparing the t statistic and t critical value and we found no column was removed.
- P-value : Here we took the threshold as 5% and then checked the $p > |t|$ from regression table and found that living condition column must be removed as it was rejecting null hypothesis.
- F-test : Similar to the above we found Probability ¿ f to be zero which is less than 5 and NH was accepted.

CONCLUSION

1) Summary of Key Findings
   - Household final consumption expenditure is a significant component of GDP and varies considerably across income groups.
   - Higher income households allocate a larger share of their consumption to non-essential goods and services, while lower income households focus on necessities.

Variable	VIF	1/VIF
Religion_1	607.81	0.001645
Religion_2	418.38	0.002390
Religion_3	110.39	0.009059
Religion_4	95.52	0.010469
State_9	56.55	0.017684
State_29	55.90	0.017888
State_23	49.63	0.020149
State_27	48.24	0.020730
State_8	41.80	0.023926
State_19	38.34	0.026084
State_28	34.45	0.029023
State_21	33.15	0.030165
State_32	30.65	0.032627
State_24	29.53	0.033862
State_3	28.55	0.035031
State_6	28.11	0.035578
Religion_5	26.57	0.037630
State_31	25.31	0.039506
State_2	24.86	0.040229
State_10	24.70	0.040480
State_22	22.68	0.044092
Religion_7	21.10	0.047399
State_7	15.45	0.064739
State_20	14.66	0.068211
State_18	14.46	0.069140
State_1	12.39	0.080727
Religion_6	10.77	0.092822
State_5	8.69	0.115012
Religion_8	4.58	0.218549
State_16	4.56	0.219441
State_30	3.73	0.268192
State_12	3.61	0.276716
Assets	3.16	0.316369
State_17	3.04	0.328923
State_33	2.82	0.354595
State_13	2.60	0.384125
State_11	2.58	0.387175
State_4	2.42	0.413291
State_14	2.35	0.426346
State_15	2.05	0.488125

(a)

State_26	2.00	0.499983
Education	1.90	0.525736
Income_Class	1.64	0.608117
log_Weddin~s	1.63	0.613730
Rooms	1.48	0.677772
CS_Knowledge	1.47	0.681131
Life_Insur~e	1.34	0.749006
Family_Size	1.21	0.827262
Debt	1.17	0.853170
Health_Ins~e	1.10	0.911909
House_Owne~p	1.08	0.924617
Livestocks	1.07	0.931298
Tragedies	1.07	0.938350
Migrant_Work	1.07	0.938908
Living_Con~s	1.05	0.951075
Marriage_S~s	1.05	0.955107
Crime	1.02	0.977449
INCOMEPC	1.00	0.997808
Mean VIF	34.30	

(b)

Fig. 3: VIF value Table for verification

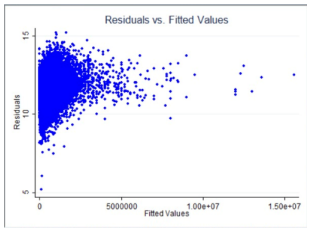

Fig. 4: This is a scatter plot showing the relation between log_HConsumption vs Wedding_Costs.

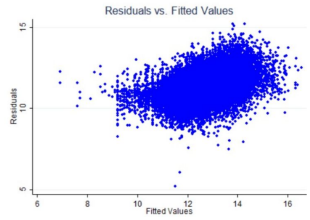

Fig. 5: After applying the log transformation to wedding costs, the level of heteroscedasticity has decreased.

- Economic shocks and inflation significantly influence household spending behavior.

2) Policy Implications
   - **Enhancing financial literacy** can help households manage expenditure efficiently.
   - **Improving access to credit services** can reduce financial constraints on consumption.
   - **Social security schemes** can help mitigate the impact of unexpected financial shocks.
3) Scope for Future Research
   - Future studies could explore regional variations in household consumption and the effect of macroeconomic variables (e.g., inflation, GDP growth) on spending patterns.

REFERENCES

- Household Consumption Expenditure Patterns - SAGE Journals
- Impact of COVID-19 on Household Spending - PubMed Central (PMC)
- Economic Perspectives on Household Consumption - MDPI