Engel Curve Estimation and Inference from Customer Expenditure Survey, Maharashtra

Purushottam Saha (BS2119)
B.Stat. III year
Submitted as part of Economic Statistics Course work,
under supervision of **Prof. Nachiketa Chattopadhyay**

November 27, 2023

Abstract

The Engel curve, based on the Engel law, which states that as income rises, the share of income spent on food falls, while the share spent on other goods and services rises; shows how income affects the demand for a good or service. We estimate the Engel curve for consumers in Maharashtra, India, using the 68th household survey data of 2011 from the NSSO. We use parametric methods and economic theory to fit the Engel curve for different goods and services, such as food, clothing, education, health, etc. We analyze the same and look at the relative expenditure share vs total expenditure plot to infer the characteristics of the good/ group of goods/services.

1 Introduction

The Engel curve is named after Ernst Engel, a German statistician who first studied the income-expenditure relationship in the 19th century. He observed that as the income of a household increases, the proportion of income spent on food decreases, while the proportion spent on other goods and services increases. This observation is known as the Engel law, and it reflects the fact that food is a basic necessity that has a limited consumption capacity, while other goods and services are more discretionary and can enhance the quality of life.

The Engel curve can have different shapes depending on the nature of the good or service and the preferences of the consumer. For example, for normal goods, the Engel curve is upward sloping, meaning that the consumer buys more of the good or service as their income increases. For inferior goods, the Engel curve is downward sloping, meaning that the consumer buys less of the good or service as their income increases. For some goods, the Engel curve is linear, meaning that the consumer spends a constant proportion of their income on the good or service regardless of their income level.

In this project, we aim to estimate the Engel curve for the state of Maharashtra in India, using the 68th household survey data of 2011 from the National Sample Survey Office (NSSO). We will use parametric methods to fit the Engel curve for different categories of goods and services, such as food, clothing, education, health, etc. We will use established economic theory to model the Engel curve from the sampled NSSO data.

2 Data and Methodology

2.1 About Data

This data, Household Consumer Expenditure, is collected as a part of the 68^{th} round of the National Sample Survey, conducted in 2011. The link for the data can be found here. All the collected variables' descriptions can also be found there. For our project, we have only considered the data from the state of Maharashtra and tried to construct a sensible Consumer Price Index Basket for the aforementioned state.

For our task, only the variables named by TCV, Item Code and HHID are used, which are described below:

- **Value:** Total Consumption Value (or in short TCV), is the monetary value of an item's overall consumption during the previous few days at a household. The number of days depends on the kind of product we use; for instance, the time frame for durable goods is 365 days, or one year, while the time frame for food products is 30 days.
- **Item Code:** An item's Item Code is its special product/group of products identifier or identification number. For instance, item code 6 designates the item group Cereals.
- **HHID:** HHID is the unique household id assigned to each household in the survey.
- **Combined Multiplier:** The term Combined Multiplier refers to the population's household-based multiplier. The average number of households of that type, including that household, in the state, is revealed by sampling at the national level. The household's influence on the entire survey is represented by the multiplier, which is the same for every item in the household.

2.2 Engel Curves and Estimation

An Engel Curve expresses the relation between income for a consumer and his consumption/expenditure of a particular good. However, income is sensitive information to gather from consumers and is difficult to counter-check from other pieces of information available. Hence we try to estimate the Engel curve from consumer expenditure data: by estimating the income with expenditure.
The idea is to fit a regression model that best fits the Expenditure Share of an item or an item group (i.e. the proportion of amount spent vs total expenditure for buying items of that group) for a household vs the Total Expenditure of the household. This way the relation between the expenditure and expenditure share for an item is captured, and that relation is called the estimated Engel Curve.

Though there is a bit of discrepancy to this idea, that the sum of expenditure shares add up to 1, and this cannot show the Engel Law in full action, still it is a valid approach to look into the individual Engel curves to get a good idea about the consumer base of the concerned population and their buying habits.

A brief inspection of the literature shows certain models for the Engel curve to estimate. The Expenditure Share for item group i is denoted by w_i and the Total Expenditure is denoted by x_i. The 5 models mentioned in Yandle (1970) are

1. Linear Model: $w_i = \beta_{i1} + \beta_{i2} x_i + e_i$
2. Hyperbola Model: $w_i = \beta_{i1} + \frac{\beta_{i2}}{x_i} + e_i$
3. Sigmoid Model: $log(w_i) = \beta_{i1} + \frac{\beta_{i2}}{x_i} + e_i$
4. Log Model: $w_i = \beta_{i1} + \beta_{i2} log(x_i) + e_i$
5. Double Log model: $log(w_i) = \beta_{i1} + \beta_{i2} log(x_i) + e_i$

Here, the β_{i1} is the intercept term for the regression problem and β_{i2} is the slope term for the same. The regression problem in $\frac{1}{x_i}$ results from the simple hypothesis that the expenditure for a particular good increases linearly with an increase in total expenditure, and hence expenditure share becomes proportional to $\frac{1}{x_i}$. The log transformation seen is used to treat the generally present heteroskedasticity in the data.

The double log model has some advantages over other functional forms of the Engel curve, especially for our data, such as:

- It can capture both normal and inferior goods, depending on the sign of β_{i2}. If $\beta_{i2} > 0$, the good is normal, meaning that expenditure increases with income. If $\beta_{i2} < 0$, the good is inferior, meaning that expenditure decreases with income.
- It can capture both necessities and luxuries, depending on the magnitude of β_{i2} If $0 < \beta_{i2} < 1$, the good is a necessity, meaning that expenditure increases less than proportionally with income. If $\beta_2 > 1$, the good is a luxury, meaning that expenditure increases more than proportionally with income. (Check Appendix for more on Classification of Goods.)
- It can be easily estimated using the ordinary least squares (OLS) method on the logarithmic transformation of the data, and brings the non-linearity regression models to a generalized linear model setup.
- For the state of Maharashtra, it is expected to have unequal variances for expenditure share of particular goods between high-earning consumers and low-earning consumers, hence the log transformation helps to treat the unequal variances and hence to satisfy Linear Regression assumptions.

The double log model was used by Yandle (1970) to estimate Engel curves for meat in New Zealand. He used cross-sectional data from a household expenditure survey and found that beef and lamb were normal necessities, while pork and poultry were normal luxuries. He also compared the double log model with other functional forms and found that it performed better in terms of goodness-of-fit and economic criteria.

2.3 Method

We plot different Engel curves for four groups of goods. The groups are: Cereal products with code 6 in the survey, Oil and other processed food products with 20 as code in survey, Fuel, entertainment, toilet products and other household consumables with code 31 in the survey, and Expenditure on education, medical, clothing and durable goods with code 39 in the survey. For each household, we calculate the per capita total expenditure and the per capita expenditure for each group for a month. We then plot the scatter plot and fit a double log model to estimate the Engel curve. We also plot the relative expenditure vs total expenditure curve for analysis: to infer the nature of the goods/group of goods/services.

3 Implementation and Results

The R-code implementation can be found here.

3.1 Cereal Products

First, let's look into the both predictor and response log-transformed scatter-plot for Cereal Goods.

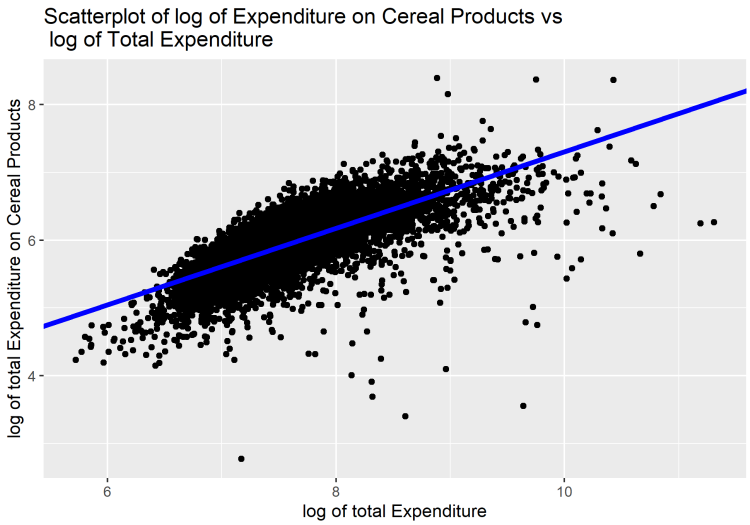

Figure 1: Log of Expenditure share vs log of total Expenditure for Cereal Goods

This inspires us to use the double-log model, as the above scatter-plot shows linearity to be captured as a simple linear regression problem. The line shown is the regression line. The R-output for the regression model is shown below.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.5495 -0.1632  0.0281  0.2064  1.7122

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.658512   0.045748   36.25   <2e-16 ***
log(dabs[, 1]) 0.564941   0.005957   94.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3258 on 7856 degrees of freedom
Multiple R-squared:  0.5338,    Adjusted R-squared:  0.5337
F-statistic:  8995 on 1 and 7856 DF,  p-value: < 2.2e-16
```

Figure 2: Regression output for the double-log model for Cereal Goods

The near-zero median and the same range for the first quartile to median and median to 3rd quartile (both near 0.18) of the residual shows the near-normality of the errors, satisfying model assumptions (this is further verified by the QQ-plot, though that is not shown here.) The adjusted R-squared value of 0.5337 shows a good fit for the linear model. An improvement of the R-squared value was seen in the Sigmoid model, but that didn't significantly outperform in other ways. Hence this model is kept. Also, heteroskedasticity is better treated in the approach for this model only. The F-statistic also shows a good fit of the model has been performed.
Next, we look into the Relative Expenditure share vs total expenditure plot for cereal goods.

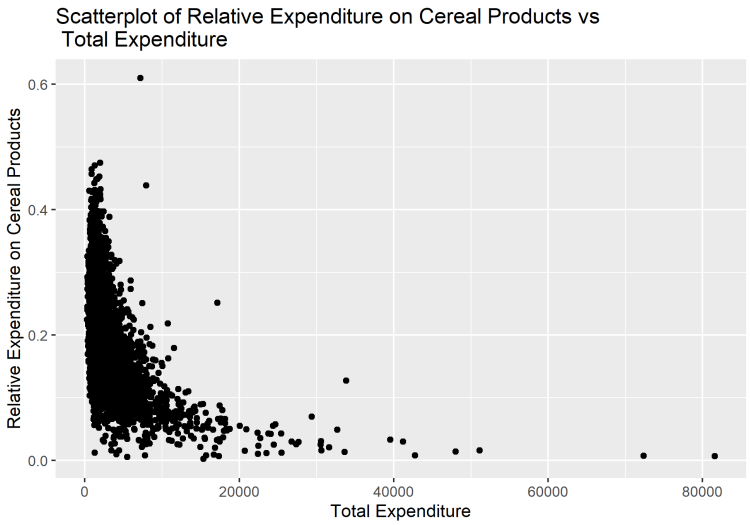

Figure 3: Relative Expenditure share vs total Expenditure for Cereal Goods

It is prominent from the plot that relative expenditure share decreases pretty quickly with an increase in total expenditure share, i.e. consumers spend less and less proportion of their total expenditure for Cereal goods as their expenditure goes up. This directly shows the good considered (Cereals) falls under the essential/necessity classification for goods.

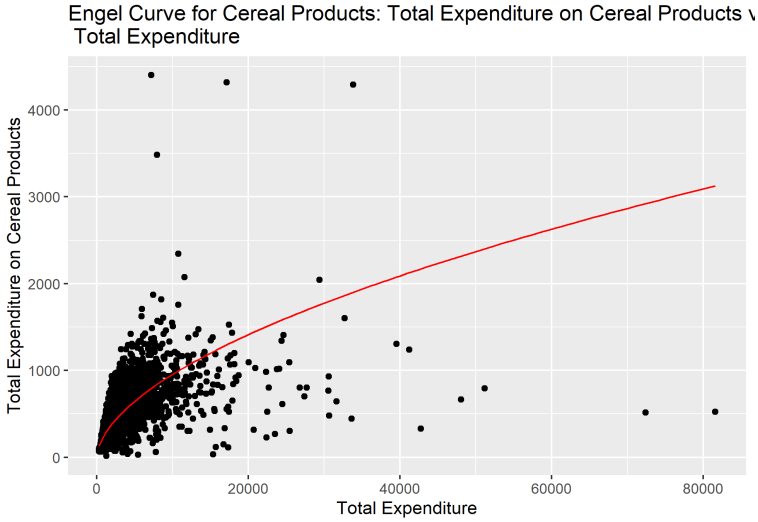

Figure 4: Engel Curve for Cereal Goods

The (estimated) Engel Curve, based on the double-log model, turns out to be $y = 5.25x^{0.56}$, shows the sub-linear relation between the expenditure for Cereal goods and the total expenditure, shown in the figure above. The sublinearity explains and confirms the previous analysis about the classification of the type of the good/ group of goods: i.e. the increase in expenditure share for Cereal items is lower than the proportionality of total expenditure for a consumer/household, making it a "Necessity" item/group, supporting **Engel Law**. The high coefficient in the curve supports the sub-linear increase in growth and explains the effect for small expenditures, that marginally more increament in expenditure share for Cereal goods will be there for low expenditure families than high expenditure ones, typical for a necessary good.

3.2 Edible Oils and Processed Foods

First, let's look into both predictor and response log-transformed scatter-plot for Edible Oils and Processed Foods data.

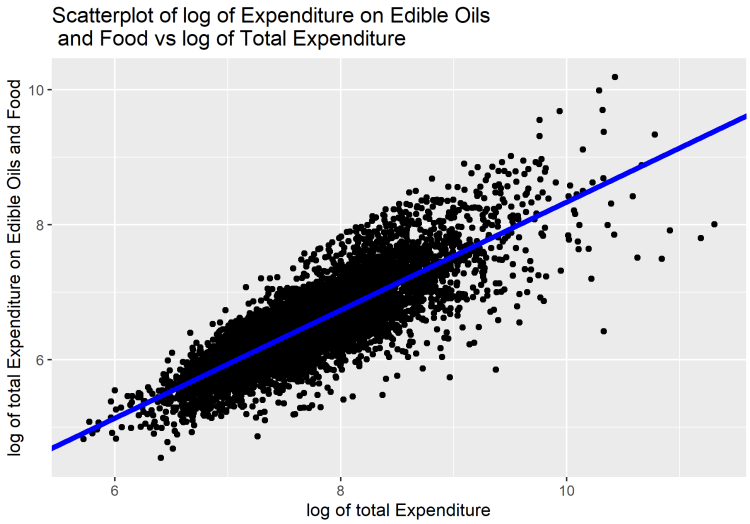

Figure 5: Log of Expenditure share vs log of total Expenditure for Edible Oils and Processed Foods

This inspires us to use the double-log model, as the above scatter-plot shows linearity to be captured as a simple linear regression problem. The line shown is the regression line. The R-output for the regression model is shown below.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.18602 -0.19875  0.00811  0.20459  1.50053

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.336553   0.046457   7.244 4.74e-13 ***
log(dabs[, 1]) 0.800662   0.006034 132.696  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3402 on 8040 degrees of freedom
Multiple R-squared:  0.6865,    Adjusted R-squared:  0.6865
F-statistic: 1.761e+04 on 1 and 8040 DF,  p-value: < 2.2e-16
```

Figure 6: Regression output for the double-log model for Edible Oils and Processed Foods

The near-zero median and the same range for the first quartile to median and median to 3rd quartile (both near 0.20) of the residual shows the near-normality of the errors, satisfying model assumptions. The adjusted R-squared value of 0.6865 shows a good fit for the linear model. Heteroskedasticity is treated well in this model, which is seen from Fitted-residual plots (not attached). The F-statistic also shows a good fit of the data has been performed.
Next, we look into the Relative Expenditure share vs total expenditure plot for Edible Oils and Processed Foods.

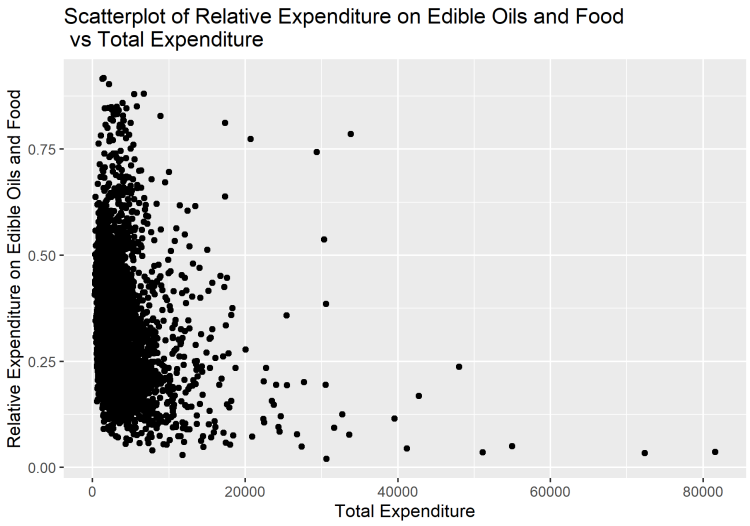

Figure 7: Relative Expenditure share vs total Expenditure for Edible Oils and Processed Foods

It is prominent from the plot that relative expenditure share decreases with an increase in total expenditure share, i.e. consumers spend less and less proportion of their total expenditure for Edible Oils and Processed Foods as their expenditure goes up. This directly shows the good considered (Edible Oils and Processed Foods) falls under the essential/necessity classification for goods. Still, there are some values that are still larger than a lot of expenditure for processed food, and we shall not forget about the multipliers that come with the survey, so those points have weights, still the overall pattern repeats and hence it doesn't have significant effects.

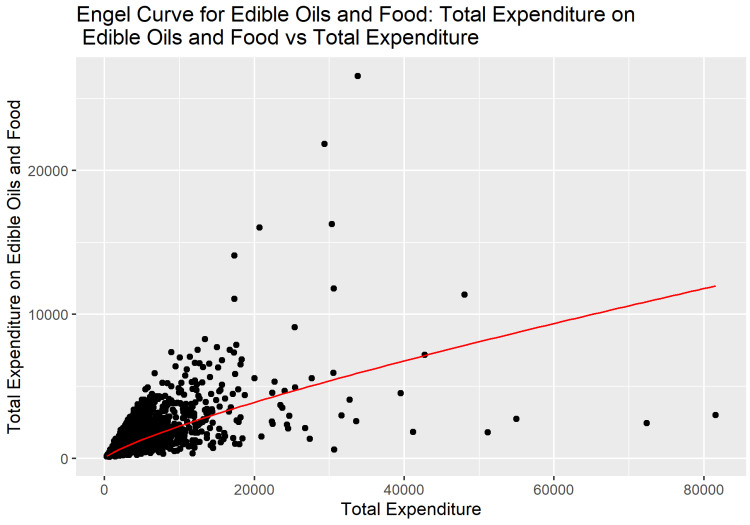

Figure 8: Engel Curve for Edible Oils and Processed Foods

The (estimated) Engel Curve, based on the double-log model, turns out to be $y = 1.4x^{0.8}$, shows the just sub-linear relation between the expenditure for Cereal goods and the total expenditure, shown in the figure above. The sub-linearity signifies the expenditure share for Edible Oils and Processed Foods increases less than proportionally to the increase in expenditure, supporting the previous analysis about the classification of the type of good/group of goods or services, i.e. it belongs under a normal good, close to necessary goods.

3.3 Fuel, Entertainment etc

First, let's look into the both predictor and response log-transformed scatter-plot for Fuel, Entertainment etc data.

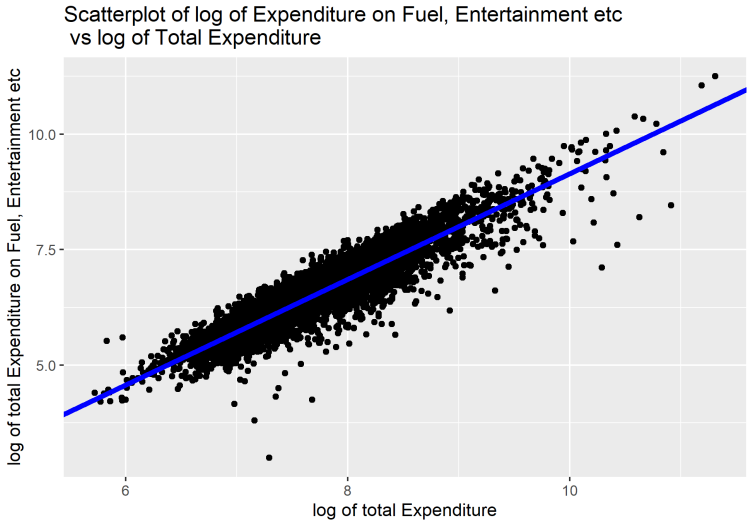

Figure 9: Log of Expenditure share vs log of total Expenditure for Fuel, Entertainment etc

This inspires us to use the double-log model, as the above scatter-plot shows linearity to be captured as a simple linear regression problem. The line shown is the regression line. The R-output for the regression model is shown below.

```
Residuals:
     Min       1Q   Median       3Q      Max
-3.06369 -0.16897  0.01794  0.19655  1.13710

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.27842    0.04135   -55.1   <2e-16 ***
log(dabs[, 1])  1.14305    0.00537   212.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3031 on 8042 degrees of freedom
Multiple R-squared:  0.8492,    Adjusted R-squared:  0.8492
F-statistic: 4.53e+04 on 1 and 8042 DF,  p-value: < 2.2e-16
```

Figure 10: Regression output for the double-log model for Fuel, Entertainment etc

The near-zero median and the same range for the first quartile to median and median to 3rd quartile (both near 0.18) of the residual shows the near-normality of the errors, satisfying model assumptions (this is further verified by the QQ-plot, though that is not shown here.) The adjusted R-squared value of 0.8492 shows an impressively good fit for the linear model, which is also seen from the plot before. Heteroskedasticity is also treated well in this model, which is seen from Fitted-residual plots (not attached). The F-statistic also shows that a good fit of the data has been performed.
Next, we look into the Relative Expenditure share vs total expenditure plot for Fuel, Entertainment etc.

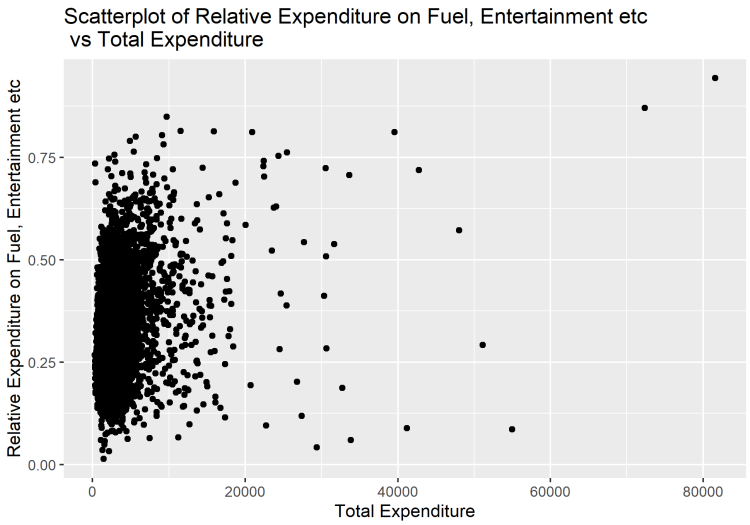

Figure 11: Relative Expenditure share vs total Expenditure for Fuel, Entertainment etc

The plot above shows that the Relative Expenditure share has increased in value with an increase in the total expenditure. This may be the consequence of the fact that Fuel and Entertainment are really not necessary goods, but luxurious items that poor consumers generally can not, and need not afford; and consumers with high income and hence high expenditure on average, spend more in Fuel, Entertainment etc (owns more cars: requires more fuel to run, high-end lifestyle spends more on different lavish entertainments). Hence these group of goods/services comes under the classification of Luxurious goods.

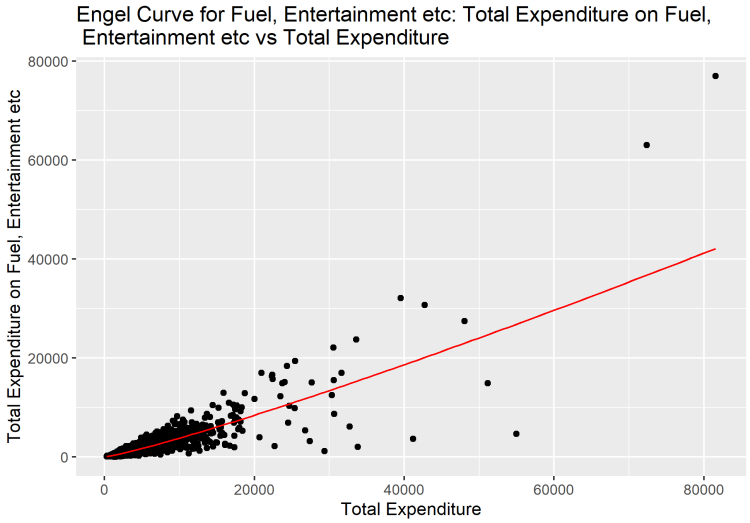

Figure 12: Engel Curve for Fuel, Entertainment etc

The (estimated) Engel Curve, based on the double-log model, turns out to be $y = 0.1x^{1.14}$, shows the super-linear relation between the expenditure for Fuel, Entertainment etc and the total expenditure, shown in the figure above. The super-linearity signifies the expenditure share for Fuel, entertainment etc increases more or less (in this case, just a bit more) proportionally to the increase in expenditure, supporting the previous analysis about the classification of the type of good/group of goods/services, that it's close to being a Luxury good, supporting **Engel Law**. The low coefficient is typical for Luxury goods, explaining the marginal increase in expenditure share for fuel, entertainment etc is smaller for low-expenditure consumers/households than high-expenditure ones.

3.4 Medical, Education, Clothing etc Durable goods

First, let's look into the both predictor and response log-transformed scatter-plot for Medical, Education etc data.

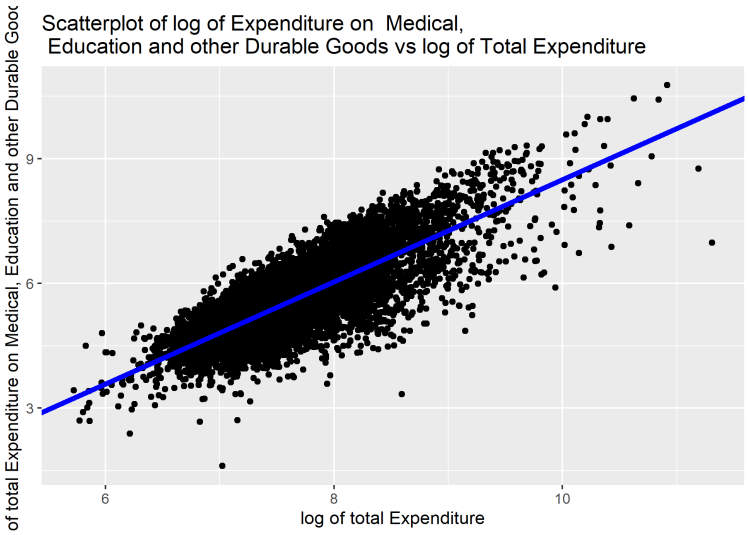

Figure 13: Log of Expenditure share vs log of total Expenditure for Medical, Education etc

This inspires us to use the double-log model, as the above scatter-plot shows linearity to be captured as a simple linear regression problem. The line shown is the regression line. The R-output for the regression model is shown below.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4430 -0.3661  0.0272  0.3966  1.6615

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.79024    0.08107  -46.75   <2e-16 ***
log(dabs[, 1]) 1.22927    0.01053  116.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5942 on 8041 degrees of freedom
Multiple R-squared:  0.6289,    Adjusted R-squared:  0.6289
F-statistic: 1.363e+04 on 1 and 8041 DF,  p-value: < 2.2e-16
```

Figure 14: Regression output for double-log model for Medical, Education etc

The near-zero median and the same range (but a bit larger, characterised by the fact that luxurious items may be involved) for the first quartile to median and median to 3rd quartile (both near 0.38) of the residual shows the near-normality of the errors, satisfying model assumptions (this is further verified by the QQ-plot, though that is not shown here.) The adjusted R-squared value of 0.6289 shows an good fit for the linear model. Heteroskedasticity is also well treated in this model, which is seen

from Fitted-residual plots (not attached). The F-statistic also shows that a good fit of the data has been performed.
Next, we look into the Relative Expenditure share vs total expenditure plot for Medical, Education etc.

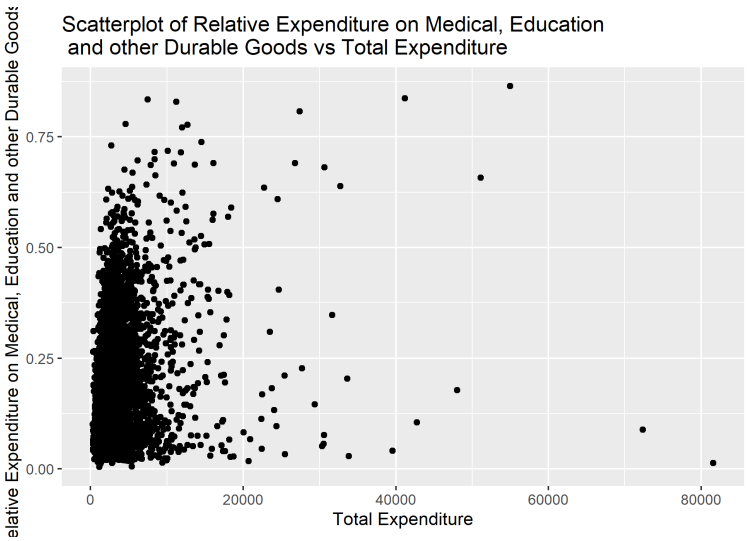

Figure 15: Relative Expenditure share vs total Expenditure for Medical, Education etc

The figure above shows that consumers with more expenditures tend to spend on average more on Medical, education etc expenses, though some extremely high expenditure consumers spent quite low (as outliers) in proportion of their total expenditure for the same. A possible explanation for the two very low Relative expenditure shares can be that for large size households (big families), durable goods are required in less quantity per capita (or result in bulk purchases), which may result in an outlier like this. The average increase in Relative Expenditure share with an increase total Expenditure for Medical, Education etc characterizes this as a Luxury good.

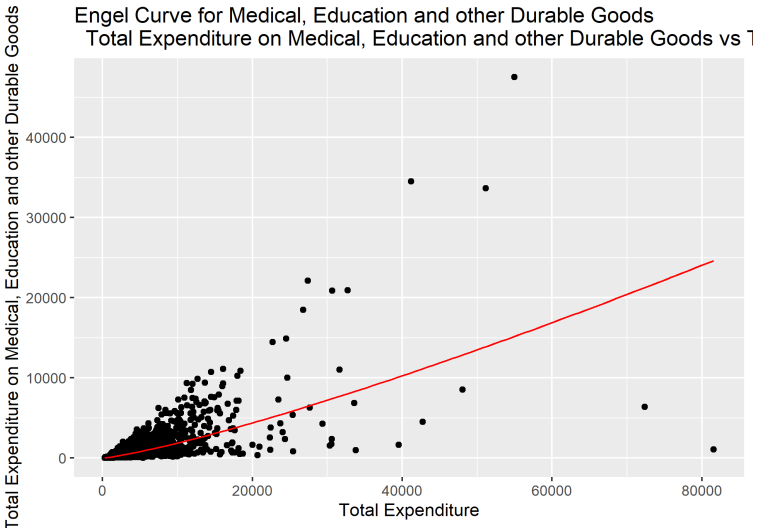

Figure 16: Engel Curve for Medical, Education etc

The (estimated) Engel Curve, based on the double-log model, turns out to be $y = 0.02x^{1.22}$, shows the super-linear relation between the expenditure for Medical, Education, and the total expenditure, shown in the figure above. The super-linearity signifies the expenditure share for Medical, Education and other durable goods increases more than proportionally to the increase in expenditure, supporting

the previous analysis about the classification of the type of good/group of goods or services being Luxury goods. This Engel curve is steeper than Fuel, Entertainment etc, explaining Medical, Education, Clothing and other durable goods expenses is more luxurious, supporting intuition. The low coefficient is typical for Luxury goods, explaining the marginal increase in expenditure share for fuel, entertainment etc is smaller for low-expenditure consumers/households than high-expenditure ones.

4 Conclusion

In this project, we have estimated the Engel curve for the state of Maharashtra in India, using the 68th household survey data of 2011 from the NSSO. We have used parametric methods and economic theory to fit the Engel curve for different categories of goods and services, such as food, clothing, education, health, etc. We have found that the Engel curve varies significantly across different goods and services.

The main findings of our project are:

- Cereal items is a basic necessity that has a low-income elasticity of demand, while other goods and services are more discretionary and have higher income elasticities of demand. This is consistent with the Engel law and the theory of consumer choice.
- The Engel curve for education, health, and clothing is more steep than the Engel curve for entertainment, fuel, and toilet products. This suggests that education, health, and clothing are more income-elastic, or more responsive to income changes, than entertainment, fuel, and toilet products. This also indicates that education, health, and clothing are more luxury goods, while entertainment, fuel, and toilet products are more necessity goods.

5 Limitations and Possible Improvements

The limitations of our project are:

1. We have used cross-sectional data from the 68th household survey of 2011, which may not capture the dynamic and temporal changes in the income and expenditure of the consumers. The income and expenditure of the consumers may vary over time due to various factors, such as inflation, economic growth, business cycles, shocks, and crises. The use of more recent and longitudinal data can improve the accuracy and reliability of the Engel curve estimation.
2. We have used parametric methods to fit the Engel curve, which may impose some restrictive assumptions on the functional form and the error term of the model. The parametric methods may not be flexible enough to capture the non-linear and complex relationships between income and expenditure on different goods and services. The use of non-parametric or semi-parametric methods can relax some of the assumptions and allow for a more general and robust estimation of the Engel curve.
3. We have used established economic theory to model the Engel curve, which may not account for some of the behavioural and psychological factors that influence the consumption choices of the consumers. The economic theory assumes that consumers are rational and utility-maximizing agents who make consistent and optimal decisions based on their budget constraints. However, in reality, consumers may be influenced by various factors, such as habits, preferences, social norms, emotions, biases, and heuristics, that may deviate from rational and utility-maximizing behaviour. The use of behavioural economics or psychology can enrich and complement economic theory and provide a more comprehensive and realistic explanation of the consumption behaviour of consumers.
4. The survey has an important structure named as the multiplier, imposing weights on the data points we get. It may occur as an improvement to over-sampling the data based on that multiplier value or similar actions to consider the same variable to better treat the sampling scheme.

5. It appears that the trend in saving = income - expenditure is neither constant, i.e. it's often observed that poor consumers are not able to save money, and rich consumers don't use their total income to fulfil needs. A better estimate of income other than just raw expenditure (may involve other variables such as social status, and number of children in households) may reflect significantly.

The directions for future research are:

1. To use more recent and longitudinal data to estimate the Engel curve and capture the dynamic and temporal changes in the income and expenditure of the consumers.
2. To use non-parametric or semi-parametric methods to estimate the Engel curve and allow for more flexible and robust modelling of the income-expenditure relationship. See works of Blundell, R. for example.
3. To use behavioural economics or psychology to model the Engel curve and incorporate some of the behavioural and psychological factors that affect the consumption choices of the consumers.

6 References

1. Deaton A. The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. Baltimore MD: Published for the World Bank by Johns Hopkins University Press; 1997.
2. Hausman JA, Newey WK, Powell JL. Nonlinear errors in variables Estimation of some Engel curves. Journal of Econometrics. 1995;65(1):205-233.
3. Yandle CA. The Theory and Estimation of Engel Curves: some estimates for meat in New Zealand. Published online 1970.

7 Appendix

7.1 Classification of Goods

We can classify goods in this way, which have been used in the model discussion and inference:

- **Inferior Goods**: The consumption of which declines both relatively and absolutely to income, as total expenditure rises.
- **Necessities**: The consumption of which declines only relatively as total expenditure rises.
- **Luxurious Goods**: The consumption of which increases both relatively and absolutely to income, as total expenditure rises.