Group Project 2

Group 1 QIJIA HE, Jianan Liu, Subasish Behera,Shreyansi Jain, Fanting Kai

2023-03-18

Contents

1 Introduction

The total number of people living in a household can be influenced by multiple variables. In certain cases, a specific type of variable may have a significant impact, which can be determined by different actual situations. By studying the correlation between variables, we can examine the living characteristics of local people.

In this dataset, we notice that six numeric variables are displayed, so we initially consider all of them. We plan to apply a generalized linear model to fit the data and discover the relevance of the variables.

At the same time, we should take into account the interaction between variables. The interaction between variables can sometimes provide insights that may not be evident when considering each variable independently. By including interaction terms in our generalized linear model, we can better understand the combined effects of these variables on the total number of people living in a household. This can help us develop a more accurate and comprehensive understanding of the factors that influence household size in the area under study.

This project begins with an expalnatory analysis into the data including data graphics and summaries in **??**. The formal analysis has been procured in a GLM model at which will be seen in sections **??** and **??**.

2 Exploratory Data Analysis

2.1 Data Mining

```
 Total.Household.Income Total.Food.Expenditure Household.Head.Age
 Min.   :  11988        Min.   :  6781         Min.   :17.00
 1st Qu.: 118565        1st Qu.: 51922         1st Qu.:41.00
 Median : 188580        Median : 73578         Median :52.00
 Mean   : 269540        Mean   : 80353         Mean   :52.23
 3rd Qu.: 328335        3rd Qu.: 98493         3rd Qu.:63.00
 Max.   :6042860        Max.   :327724         Max.   :99.00
 Total.Number.of.Family.members House.Floor.Area   House.Age
 Min.   : 1.000                 Min.   :  5.00   Min.   :  0.00
 1st Qu.: 3.000                 1st Qu.: 32.00   1st Qu.: 12.00
 Median : 4.000                 Median : 54.00   Median : 20.00
 Mean   : 4.669                 Mean   : 90.92   Mean   : 22.98
 3rd Qu.: 6.000                 3rd Qu.:102.00   3rd Qu.: 31.00
 Max.   :15.000                 Max.   :900.00   Max.   :100.00
 Number.of.bedrooms
 Min.   :0.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.259
 3rd Qu.:3.000
 Max.   :9.000
```

Table 1: Asian Health Data Summary Statistics for 2015.

Variable	n	Mean	SD	Min	Median	Max	IQR
Total.Household.Income	1725	269540.48	274564.17	11988	188580	6042860	139755
Total.Food.Expenditure	1725	80352.78	41194.36	6781	73578	327724	24915
Household.Head.Age	1725	52.23	14.52	17	52	99	11
Total.Number.of.Family.members	1725	4.67	2.33	1	4	15	2
House.Floor.Area	1725	90.92	99.20	5	54	900	48
House.Age	1725	22.98	15.32	0	20	100	11
Number.of.bedrooms	1725	2.26	1.44	0	2	9	1

To better understand the data at hand, various figures were created to provide visuals and illustrations. We begin with the scatterplots of each variable to understand the distribution of the data:

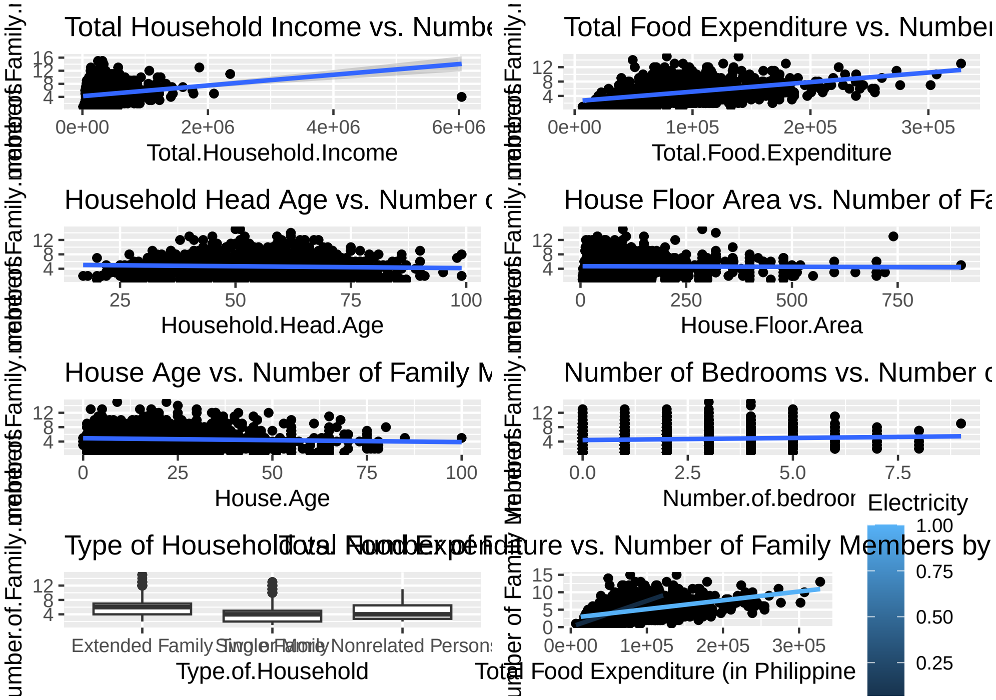


Figure 1: Scatterplots of Phillipine Family

We proceed with the correlation matrix the highlight relationships, below:

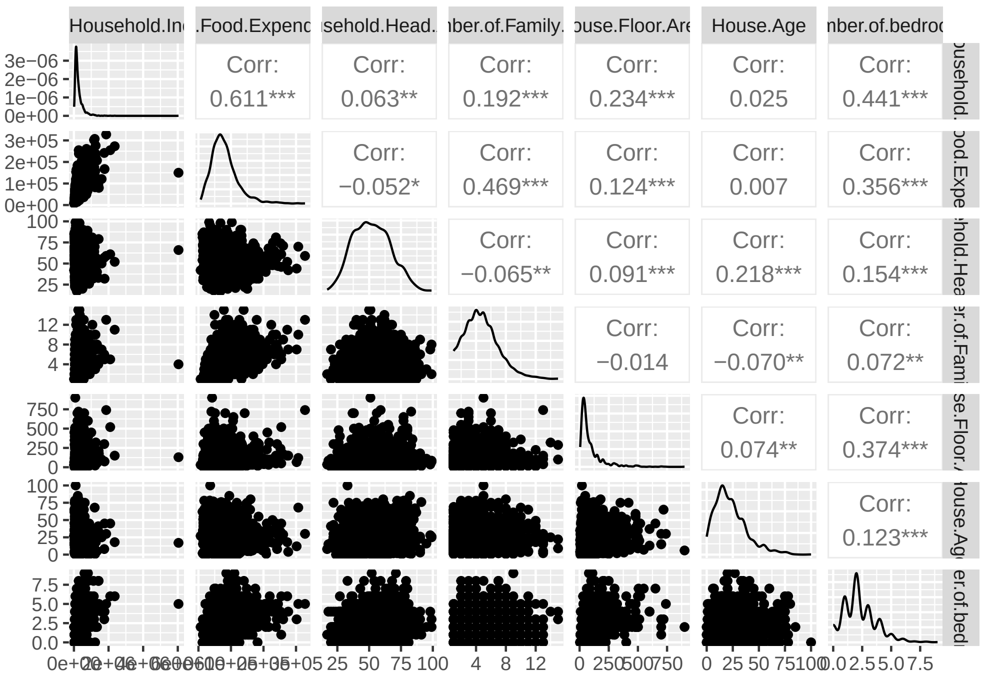


Figure 2: Correlation Matrix of Phillipine Households

It is evident from the corrmatrix plot correlation matrix that the greatest positive correlation exists between family members and food expenditure.

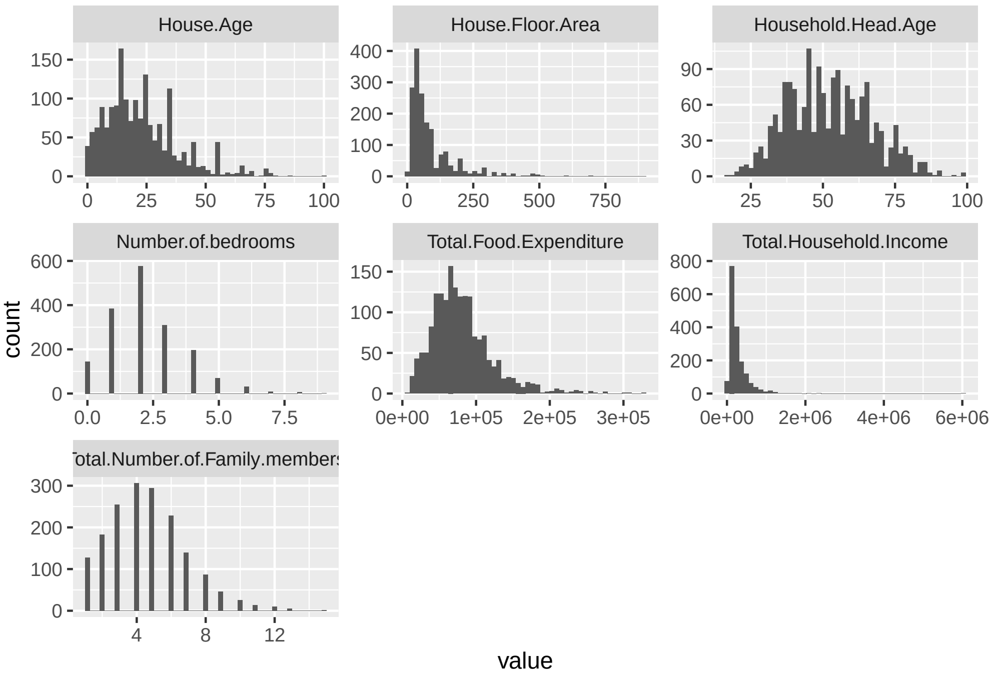


Figure 3: Plot histograms of the variables

Using the trends seen in the data exploration, our group has taken this further to examine causal relationships through generalized linear models in the Logistic regression and we bring out 2 different ways to solving this research.

3 Generalised Linear Model

3.1 Model 1: Logistic Regression Binomial Distribution

```
Type.of.Household has 3 levels
Electricity has 2 levels
Region has 1 levels



Call:
glm(formula = FamilySize_AboveMedian ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + Type.of.Household +
    House.Floor.Area + House.Age + Number.of.bedrooms, family = binomial(link = "logit"),
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0140  -0.9488  -0.4277   1.0088   2.1187

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              4.145e-01  3.104e-01
Total.Household.Income                                  -1.219e-06  3.561e-07
Total.Food.Expenditure                                   2.501e-05  2.239e-06
Household.Head.Age                                      -1.548e-02  4.171e-03
Type.of.HouseholdSingle Family                          -1.274e+00  1.271e-01
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.364e+00  7.472e-01
House.Floor.Area                                        -6.629e-04  6.156e-04
House.Age                                               -1.013e-02  3.720e-03
Number.of.bedrooms                                      -5.728e-02  4.631e-02
                                                        z value Pr(>|z|)    
(Intercept)                                               1.336 0.181708    
Total.Household.Income                                   -3.423 0.000619 ***
Total.Food.Expenditure                                   11.169  < 2e-16 ***
Household.Head.Age                                       -3.711 0.000206 ***
Type.of.HouseholdSingle Family                          -10.021  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -1.826 0.067923 .  
House.Floor.Area                                         -1.077 0.281551    
House.Age                                                -2.724 0.006444 ** 
Number.of.bedrooms                                       -1.237 0.216125    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2391.1  on 1724  degrees of freedom
Residual deviance: 1992.3  on 1716  degrees of freedom
AIC: 2010.3

Number of Fisher Scoring iterations: 4
```

In order to create a binomial logistic regression model, we needed to transform our response variable, Total.Family.Size into a binary variable. We accomplished this by dividing our data into two groups based on the median value of Total.Family.Size. We then removed any variables that only had one level from the formula and created a binary variable based on whether or not the family size was above or below the median value. Our objective was to examine how predictors relate to the possibility of having a family size above the median. After examining the summary of the model, we found that the variables House.Floor.Area and Number.of.bedrooms had high p-values, which suggests that they are not statistically significant predictors of the response variable. Consequently, we removed these variables from the model and focused only on significant predictors when re-fitting the model.

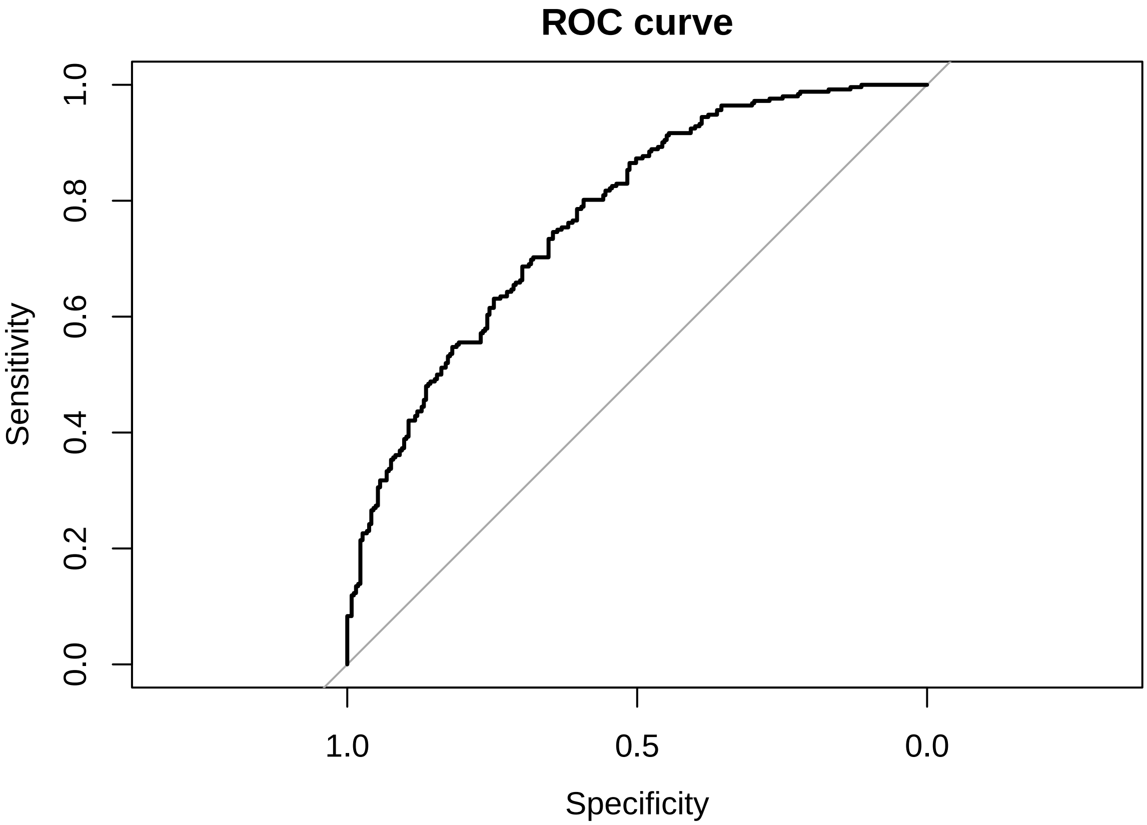
```
Call:
glm(formula = FamilySize_AboveMedian ~ Total.Household.Income + 
    Total.Food.Expenditure + Household.Head.Age + Type.of.Household + 
    House.Age, family = binomial(link = "logit"), data = dataset)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.9901  -0.9370  -0.4291   1.0019   2.1486  

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              3.433e-01  3.079e-01
Total.Household.Income                                  -1.459e-06  3.363e-07
Total.Food.Expenditure                                   2.498e-05  2.234e-06
Household.Head.Age                                      -1.627e-02  4.140e-03
```

```
Type.of.HouseholdSingle Family                          -1.269e+00  1.269e-01
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.338e+00  7.461e-01
House.Age                                               -1.075e-02  3.708e-03
                                                        z value Pr(>|z|)    
(Intercept)                                               1.115  0.26482    
Total.Household.Income                                   -4.338 1.44e-05 ***
Total.Food.Expenditure                                   11.183  < 2e-16 ***
Household.Head.Age                                       -3.931 8.48e-05 ***
Type.of.HouseholdSingle Family                          -10.000  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -1.793  0.07293 .  
House.Age                                                -2.899  0.00375 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2391.1  on 1724  degrees of freedom
Residual deviance: 1996.0  on 1718  degrees of freedom
AIC: 2010

Number of Fisher Scoring iterations: 4
```

This AIC is similar (2010) to the AIC of the original model (2010.3), indicating that the model's performance has not been negatively affected by removing these variables. And since the larger the absolute value of a z-value, the more significant the variable is in the model.So in this model Total.Food.Expenditure: 11.183 has the most significant influence on the response variable (FamilySize_AboveMedian), followed by Type.of.HouseholdSingle Family, and then Total.Household.Income.

We then to evaluate the performance of this logistic regression model; the AUC-ROC is commonly utilized to evaluate the effectiveness of a binary classification model, with a value of 1 denoting perfect classification and 0.5 suggesting a random guessing. By plotting the true positive rate against the false positive rate for differing predicted probability thresholds, the ROC curve provides a visual representation illustrating the sensitivity-versus-specificity trade-offs of the model for various points. Overall, evaluating the AUC-ROC value allows us to determine how adequate the model performs for its intended purpose. A higher value suggests that the model is proficient in distinguishing between the two classes.

```
Area under the curve: 0.7727
```



The ROC value is 0.7727 higher than 0.5 close to 1 indicates a good performance of the model. The plot below shows the predicted probabilities for different combinations of predictor variable values. It illustrate relationship between the predictor variables and the predicted probabilities of the response variable FamilySize_AboveMedian being equal to 1 (the probability of family size above median). The y-axis of the plot represents the predicted probability of the response variable being equal to 1, and the x-axis of the plot represents the values of the predictor variables.

	2.5 %	97.5 %
(Intercept)	-0.2601364	0.9475078
Total.Household.Income	-0.0000021	-0.0000008
Total.Food.Expenditure	0.0000207	0.0000294
Household.Head.Age	-0.0244323	-0.0081959
Type.of.HouseholdSingle Family	-1.5192969	-1.0217675
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-2.8522921	0.1692814
House.Age	-0.0180642	-0.0035168

```
$Total.Household.Income
```

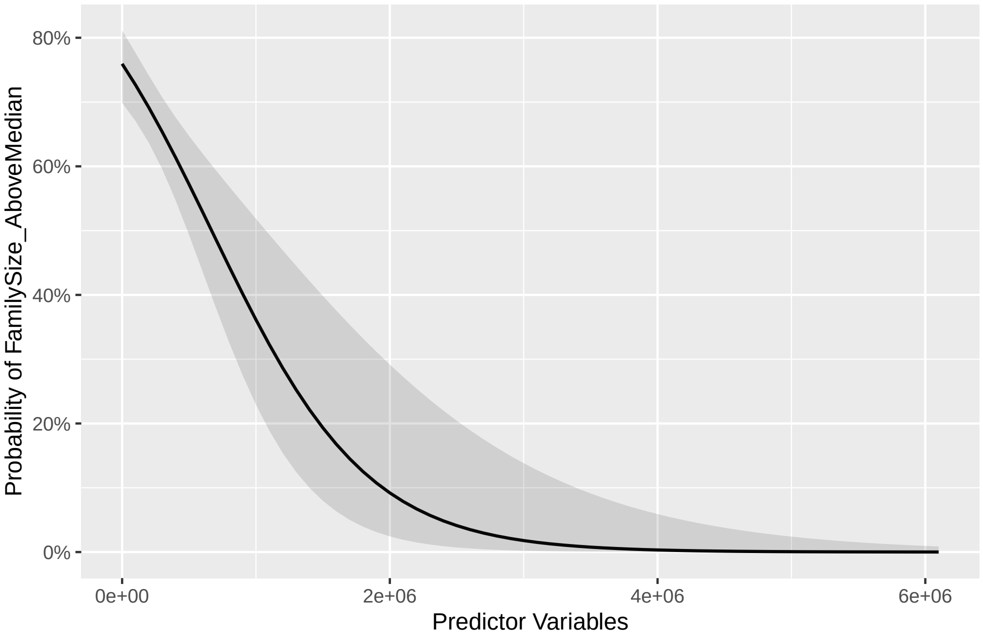


Figure 4: Plot predicted probabilities of predictor variable values

```
$Total.Food.Expenditure
```

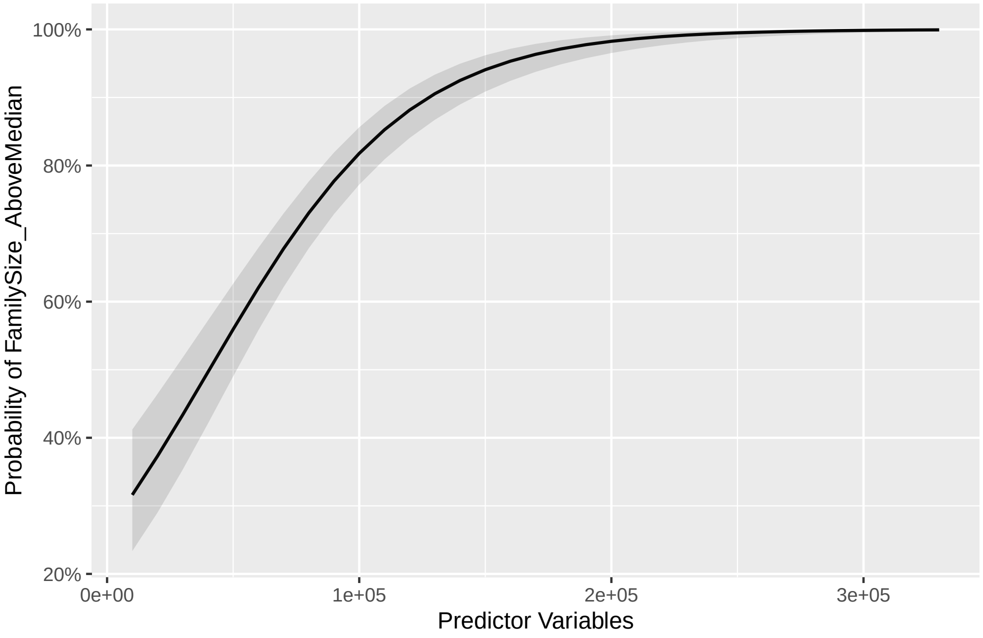


Figure 5: Plot predicted probabilities of predictor variable values

```
$Household.Head.Age
```

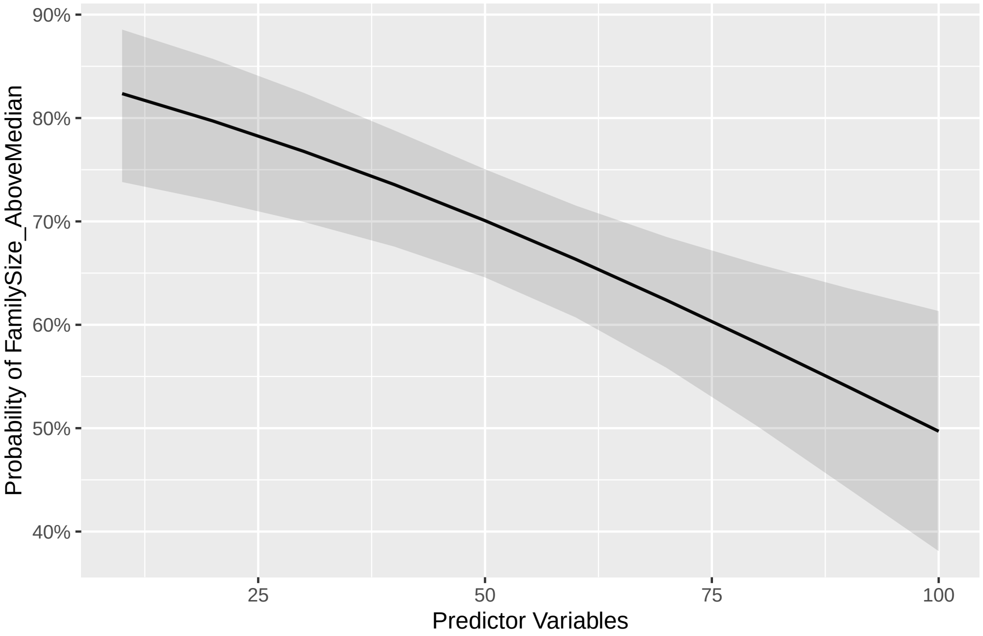


Figure 6: Plot predicted probabilities of predictor variable values

```
$Type.of.Household
```

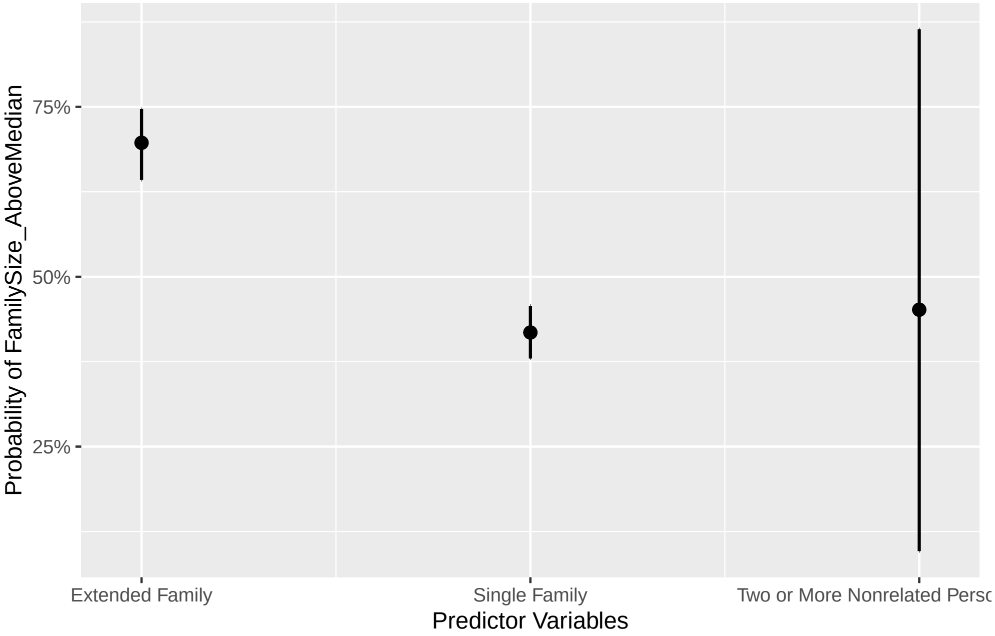


Figure 7: Plot predicted probabilities of predictor variable values

```
$House.Age
```

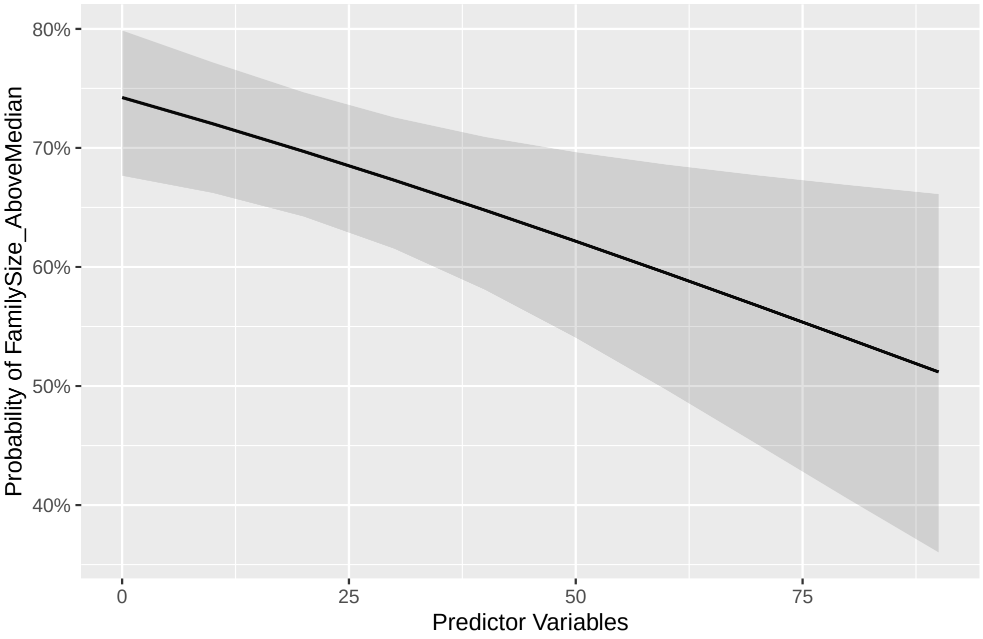


Figure 8: Plot predicted probabilities of predictor variable values

Finally we plot to Log-Odds plot in order to visualize the relationship between the predictor variables and the log-odds of the response variable. After creating a predicted probability plot, we find out that Total.HouseHold.Income is the most important perdictor.

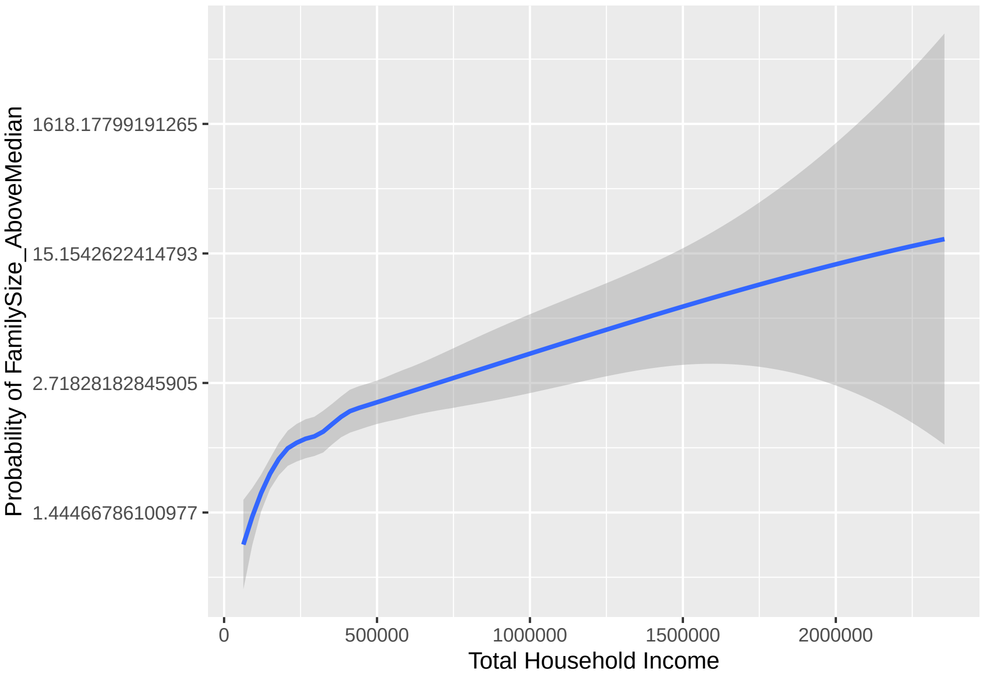


Figure 9: Log-Odds plot

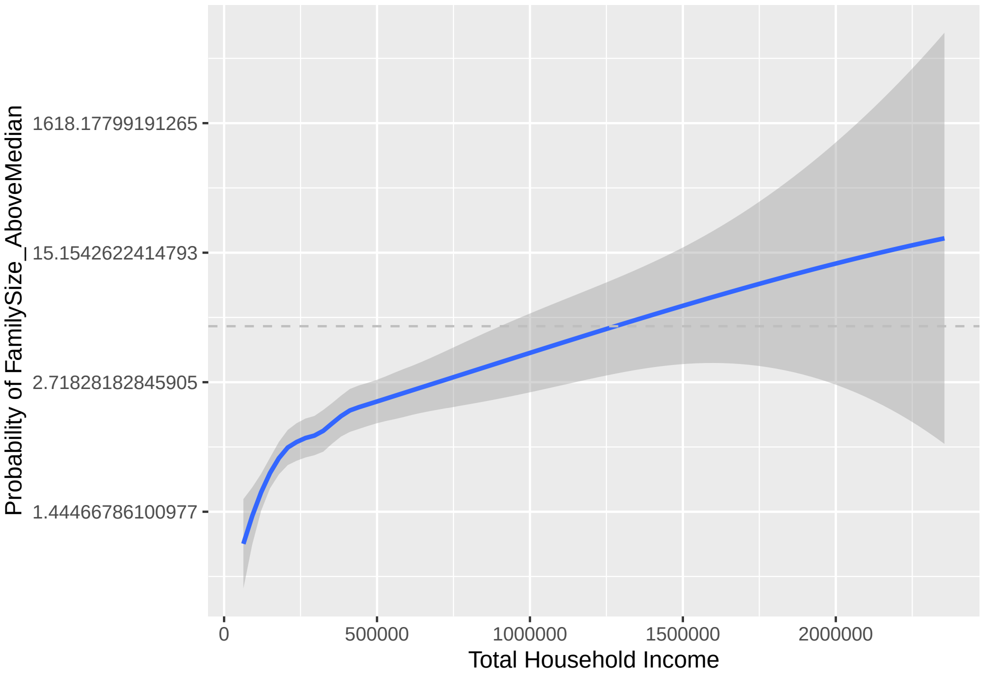


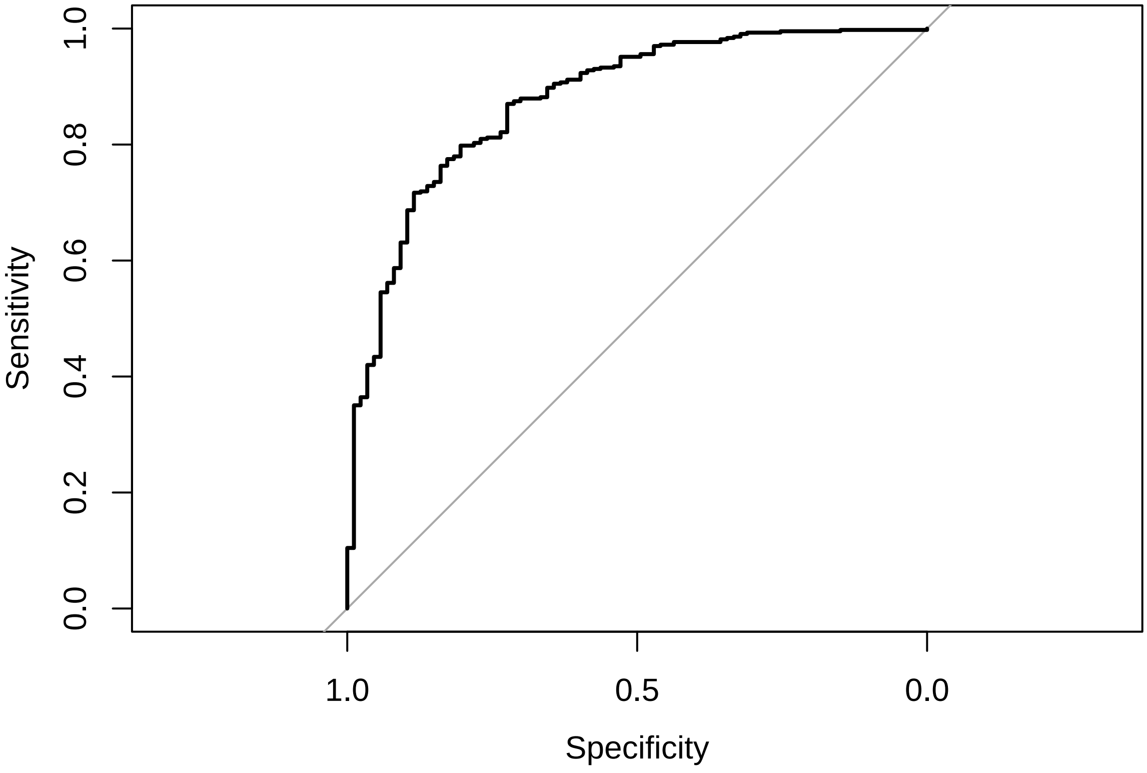
Figure 10: Log-Odds plot

3.2 Model 2: Logistic regression Poission Distribution

3.2.1 Data logistic regression

After drawing the scatter plot and data correlation, we trying to do a logistic regression from data. we dividing the dataset into training and testing sets, logistic regression models are fitted for each predictor variable and a full model using the glm function in R. The response variable in all the models is Total.Number.of.Family.members, and the predictor variables are Total.Household.Income, Total.Food.Expenditure, Household.Head.Age, House.Floor.Area, House.Age, and Number.of.bedrooms. Multiple logistic regression models are also fitted for each predictor variable separately.

The predict function is then used to generate predicted probabilities for the test data using the full model and the type = "response" argument. And the table function is used to compare the predicted probabilities with the actual values in the test data. And a binary response variable is created by collapsing some levels of the original response variable, and the pROC package is used to calculate the ROC curve using the binary response variable and predicted probabilities. Finally, a Poisson GLM model is fit using all the predictor variables in the dataset with the glm function and family = poisson() argument, which specifies that the response variable has a Poisson distribution. The resulting model is stored in the Our_glm_model object for further examination



```
Call:  glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms, family = poisson(), data = dataset)

Coefficients:
           (Intercept)  Total.Household.Income  Total.Food.Expenditure
           1.224e+00              -2.294e-07               6.099e-06
    Household.Head.Age        House.Floor.Area               House.Age
          -6.002e-04              -1.948e-04              -2.282e-03
    Number.of.bedrooms
          -1.552e-02

Degrees of Freedom: 1724 Total (i.e. Null);  1718 Residual
Null Deviance:      2024
Residual Deviance: 1574     AIC: 7251
```

3.2.2 Show Logistic

We get results listed below:

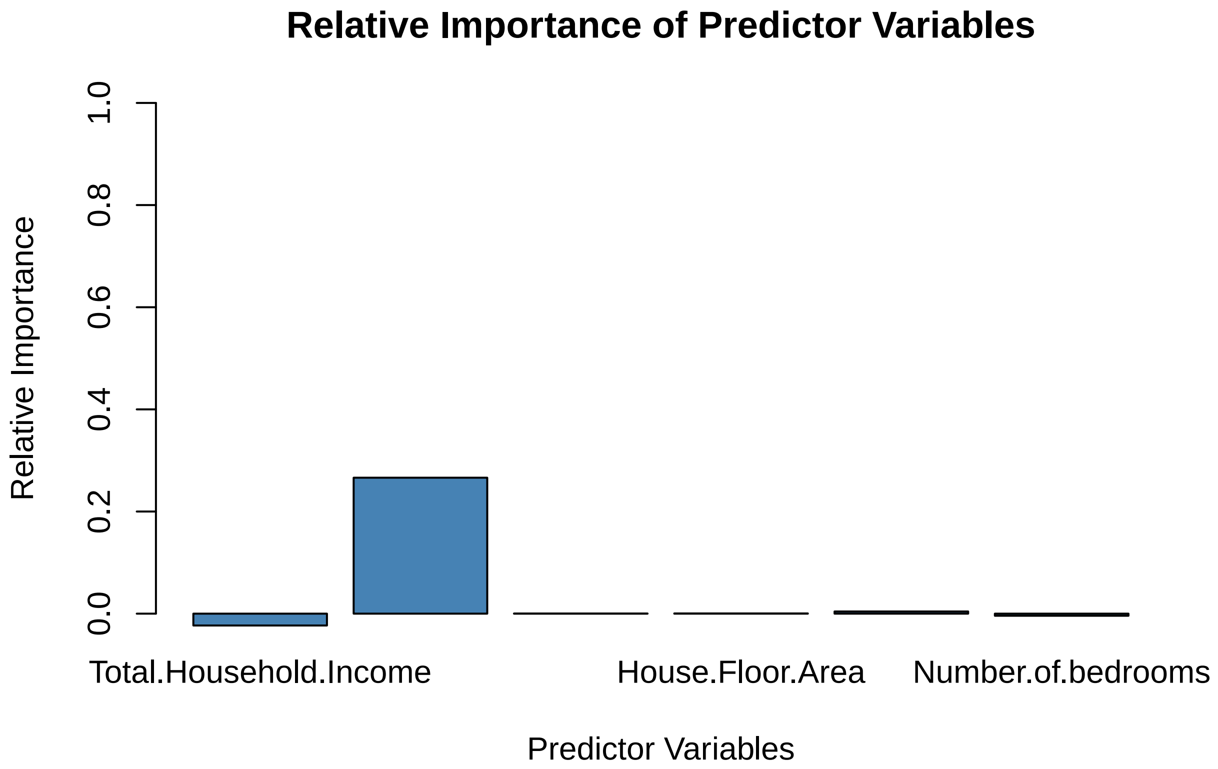
```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms, data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5374  -1.4909  -0.2615   1.2290  10.7184

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.782e+00  2.625e-01  10.596  < 2e-16 ***
Total.Household.Income -9.065e-07  2.621e-07  -3.459 0.000562 ***
Total.Food.Expenditure  3.176e-05  1.796e-06  17.679  < 2e-16 ***
Household.Head.Age     -1.047e-03  4.212e-03  -0.249 0.803657
House.Floor.Area       -7.722e-04  6.404e-04  -1.206 0.228120
House.Age              -7.350e-03  3.943e-03  -1.864 0.062575 .
Number.of.bedrooms     -6.559e-02  5.043e-02  -1.301 0.193640
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for gaussian family taken to be 4.177119)

    Null deviance: 6635.0  on 1206  degrees of freedom
Residual deviance: 5012.5  on 1200  degrees of freedom
AIC: 5159.9

Number of Fisher Scoring iterations: 2
```

The deviance residuals measure how much probabilities estimated from our model differ from the observed proportions of successes. Bigger values indicate a bigger difference. Smaller values mean a smaller difference. The summary of the logistic regression model shows the coefficient estimates, standard errors, z-values, and p-values for each predictor variable in the model. The deviance residuals measure the difference between the observed and predicted values, with larger values indicating a larger difference. Although the deviance residuals suggest a generally good fit of the model, the presence of a large maximum value suggests that there may be non-relevant variables included in the model. Hence, it is advisable to rank the importance of variables and remove any irrelevant variables from the model.

3.2.3 Importance ranking of variables

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms, data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5814  -1.4710  -0.3067   1.1883  10.7487

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.882e+00  2.177e-01  13.236  < 2e-16 ***
Total.Household.Income -1.024e-06  2.383e-07  -4.298 1.82e-05 ***
Total.Food.Expenditure  3.213e-05  1.529e-06  21.007  < 2e-16 ***
Household.Head.Age     -5.947e-04  3.516e-03  -0.169  0.86569
House.Floor.Area       -7.110e-04  5.354e-04  -1.328  0.18437
House.Age              -9.258e-03  3.292e-03  -2.813  0.00497 **
Number.of.bedrooms     -9.277e-02  4.085e-02  -2.271  0.02326 *
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for gaussian family taken to be 4.130421)

    Null deviance: 9384.0  on 1724  degrees of freedom
Residual deviance: 7096.1  on 1718  degrees of freedom
AIC: 7351

Number of Fisher Scoring iterations: 2

Response variable: Total.Number.of.Family.members
Total response variance: 5.44315
Analysis based on 1725 observations

6 Regressors:
Total.Household.Income Total.Food.Expenditure Household.Head.Age House.Floor.Area House.Age Number.of.be
Proportion of variance explained by model: 24.38%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                               pratt
Total.Household.Income -0.0231816731
Total.Food.Expenditure  0.2661854246
Household.Head.Age      0.0002421048
House.Floor.Area        0.0004279960
House.Age               0.0042567456
Number.of.bedrooms     -0.0041188170

Total.Household.Income Total.Food.Expenditure     Household.Head.Age
         -0.0231816731           0.2661854246           0.0002421048
      House.Floor.Area              House.Age     Number.of.bedrooms
```

```
          0.0004279960           0.0042567456          -0.0041188170

Response variable: Total.Number.of.Family.members
Total response variance: 5.501699
Analysis based on 1207 observations

6 Regressors:
Total.Household.Income Total.Food.Expenditure Household.Head.Age House.Floor.Area House.Age Number.of.be
Proportion of variance explained by model: 24.45%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                               lmg
Total.Household.Income 0.021561216
Total.Food.Expenditure 0.211453783
Household.Head.Age     0.002184888
House.Floor.Area       0.001254435
House.Age              0.002534027
Number.of.bedrooms     0.005547262

Average coefficients for different model sizes:

                                  1X           2Xs           3Xs           4Xs
Total.Household.Income  1.547316e-06  1.026446e-06  5.174466e-07  2.442630e-08
Total.Food.Expenditure  2.684151e-05  2.822185e-05  2.933925e-05  3.026679e-05
Household.Head.Age     -9.711088e-03 -9.589826e-03 -8.153701e-03 -6.030505e-03
House.Floor.Area        1.333608e-04 -5.960191e-04 -9.514071e-04 -1.035862e-03
House.Age              -7.955306e-03 -8.341368e-03 -8.060118e-03 -7.651949e-03
Number.of.bedrooms      1.753170e-01  9.490145e-02  3.170543e-02 -1.541926e-02
                                 5Xs           6Xs
Total.Household.Income -4.504572e-07 -9.065212e-07
Total.Food.Expenditure  3.105775e-05  3.175535e-05
Household.Head.Age     -3.593733e-03 -1.047287e-03
House.Floor.Area       -9.450803e-04 -7.721986e-04
House.Age              -7.379106e-03 -7.349658e-03
Number.of.bedrooms     -4.751489e-02 -6.558918e-02
```



function calc.relimp() calculates several relative importance metrics for the linear model. From the results displayed above, we witness two sets of important index. Consistent to the logistic above, Total.Household.Income and Total.Food.Expenditure show greater vitality than other four indexes.House age and number of bedroom have a slightly impact. Now we first drop2 variables and then drop 2 slightly influenced variables and re-run de logistic and importance ranking to see the best we can do.

3.2.4 Re-run the analysis for 4 variables

```
Response variable: Total.Number.of.Family.members
Total response variance: 5.501699
Analysis based on 1207 observations

4 Regressors:
Total.Household.Income Total.Food.Expenditure House.Age Number.of.bedrooms
Proportion of variance explained by model: 24.36%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                               lmg
Total.Household.Income 0.021629500
Total.Food.Expenditure 0.213384562
House.Age              0.002888822
Number.of.bedrooms     0.005674576

Average coefficients for different model sizes:
```

```
                                  1X           2Xs           3Xs           4Xs
Total.Household.Income  1.547316e-06  6.440462e-07 -1.857277e-07 -9.356729e-07
Total.Food.Expenditure  2.684151e-05  2.899556e-05  3.065648e-05  3.189216e-05
House.Age              -7.955306e-03 -9.070002e-03 -8.366715e-03 -7.664098e-03
Number.of.bedrooms      1.753170e-01  2.663222e-02 -6.051560e-02 -8.625704e-02
```

Based on the output of the calc.relimp function, we can see that the total response variance for this variable is 5.363603, and the proportion of variance explained by this model is 26.42%.

3.2.5 Re-run the analysis for only 2 variables {sec:sub}

```
Response variable: Total.Number.of.Family.members
Total response variance: 5.501699
Analysis based on 1207 observations

2 Regressors:
Total.Household.Income Total.Food.Expenditure
Proportion of variance explained by model: 23.85%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

                              lmg
Total.Household.Income 0.02473607
Total.Food.Expenditure 0.21379215

Average coefficients for different model sizes:

                                 1X           2Xs
Total.Household.Income 1.547316e-06 -1.088828e-06
Total.Food.Expenditure 2.684151e-05  3.138976e-05
```

Based on the output of the calc.relimp function, we can see that the total response variance for this variable is 5.363603, and the proportion of variance explained by this model is 25.77%. Total.Food.Expenditure appears to be more important in explaining the variance in the Total.Number.of.Family.members, as its relative importance (lmg) is higher than that of Total.Household.Income.And the coefficients indicate the relationship between each predictor and the response variable (Total.Number.of.Family.members). In the 1X model, a unit increase in Total.Household.Income is associated with a 2.276022e-06 increase in Total.Number.of.Family.members, whereas a unit increase in Total.Food.Expenditure is associated with a 2.539690e-05 increase in Total.Number.of.Family.members. In the 2Xs model, the relationships are -1.510802e-06 and 3.125984e-05, respectively.

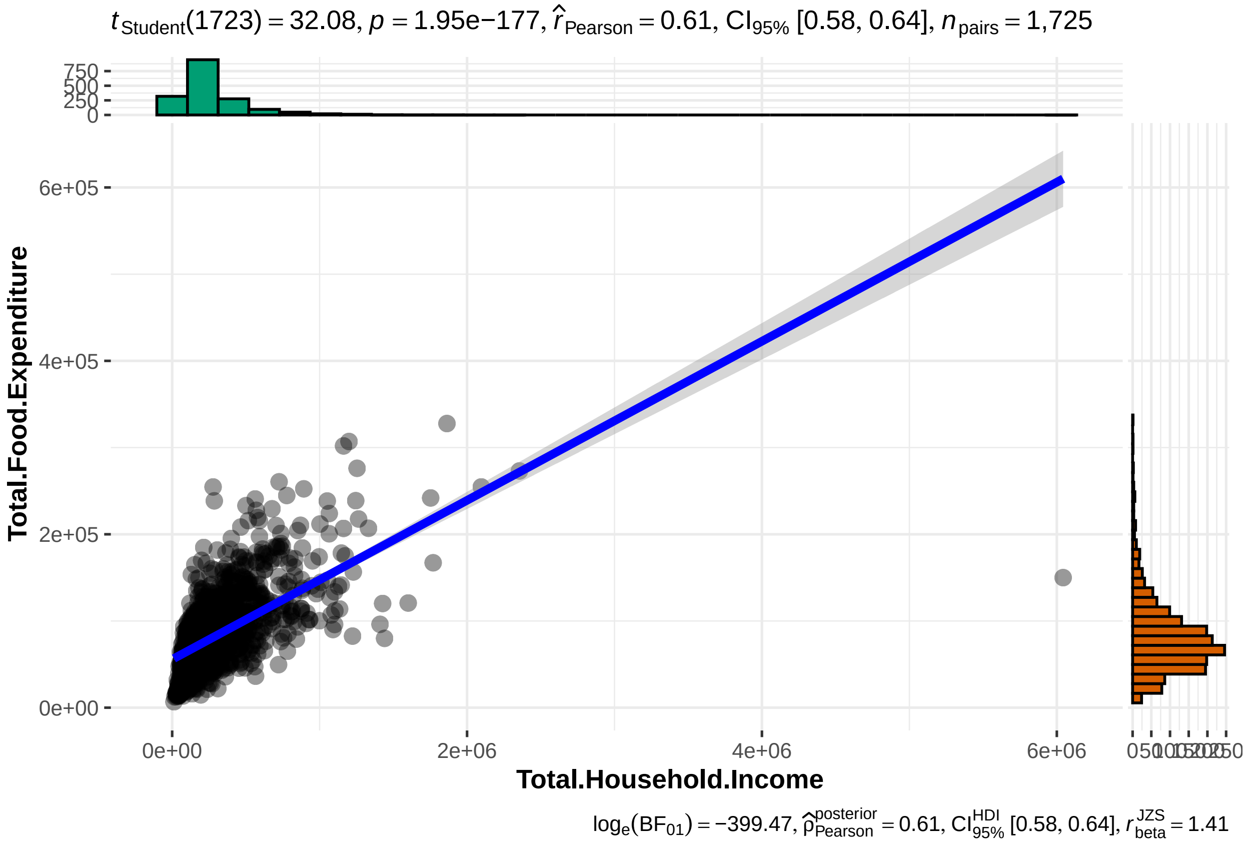
4 Conclusion {sec:Conc}

From the results we get above, we have concluded that Total.Household.Income and Total.Food.Expenditure are the two most important variables. However, the average coefficients are still relatively small. This can be attributed to several reasons. The most significant reason is that Income and Expenditure are not on the same scale as the number of family members. Actually, we can see from the output above that roughly each additional unit of income, the number of household members increase by 2.276022e-06, namely every additional 407947 units of household income could attribute to one member plus. So high income families tend to have more members. As for the food expenditure, it does not against our common sense that with the number of family members increase, the food expenditure would rise with a 2.539690e-05 coefficient and relatively high correlation.

4.1 Appendix {sec:sub}

Here, we discover the correlations between variables, in other words the columns of the dataset.

The results from above conducted on the dataset show that both p-values are extremely small, leading to the rejection of the null hypothesis (data is normally distributed). Consequently, neither Total.Household.Income nor Total.Food.Expenditure is normally distributed.The extremely small p-value in Pearson's product-moment correlation test reveals a statistically significant correlation between Total.Household.Income and Total.Food.Expenditure. The 95% confidence interval for this correlation is [0.5810680, 0.6402141], suggesting that the true correlation between these two variables lies within this range. And the sample estimate for the correlation (cor) is 0.6114945, indicating a positive and moderately strong relationship between Total.Household.Income and Total.Food.Expenditure.


$t_{\text{Student}}(1723) = 32.08$, $p = 1.95\text{e}{-177}$, $\hat{r}_{\text{Pearson}} = 0.61$, $\text{CI}_{95\%}$ [0.58, 0.64], $n_{\text{pairs}} = 1{,}725$
6e+05
4e+05
2e+05
0e+00
Total.Food.Expenditure
0e+00
2e+06
4e+06
6e+06
Total.Household.Income
$\log_e(\text{BF}_{01}) = -399.47$, $\hat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.61$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.58, 0.64], $r^{\text{JZS}}_{\text{beta}} = 1.41$