

# Group Project 2

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

2023-03-16

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Exploratory Data Analysis</b>	<b>1</b>
2.1	Data Mining . . . . .	1
<b>3</b>	<b>Logistic regression</b>	<b>4</b>
3.1	Data logistic regression {sec:sub} . . . . .	4
3.2	Show Logistic . . . . .	5
<b>4</b>	<b>Importance ranking of variables {sec:IRV}</b>	<b>6</b>
4.1	Re-run the analysis for 4 variables {sec:sub} . . . . .	8
4.2	Re-run the analysis for only 2 variables {sec:sub} . . . . .	9
<b>5</b>	<b>Conclusion {sec:Conc}</b>	<b>9</b>
5.1	Appendix {sec:sub} . . . . .	10

## 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 explanatory 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:

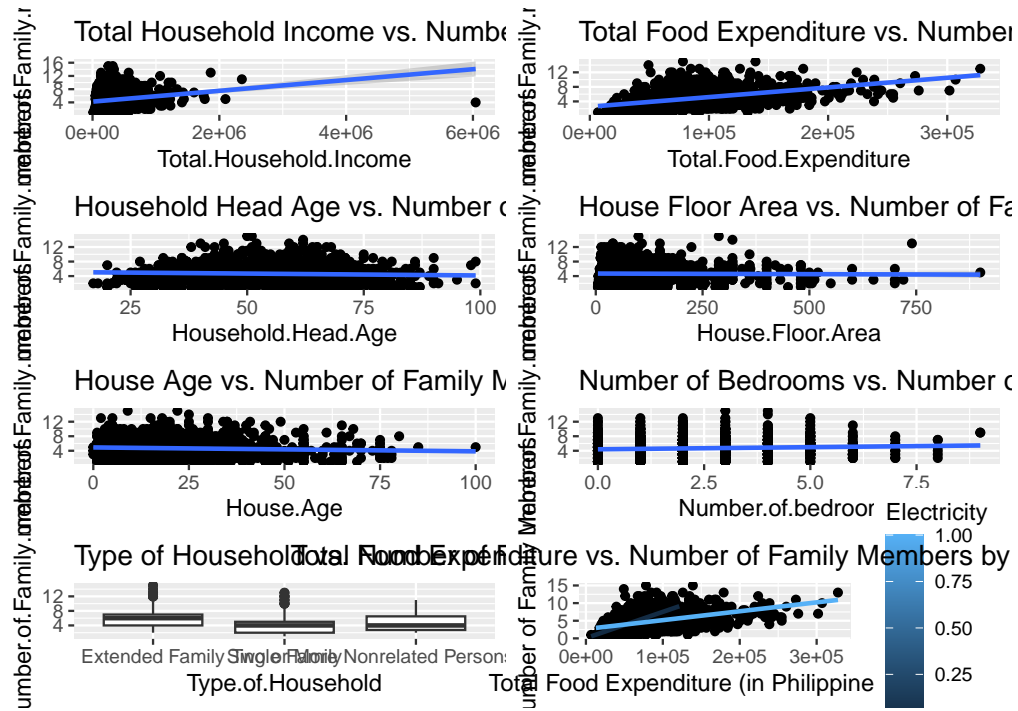


Figure 1: Scatterplots of Phillipine Family

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

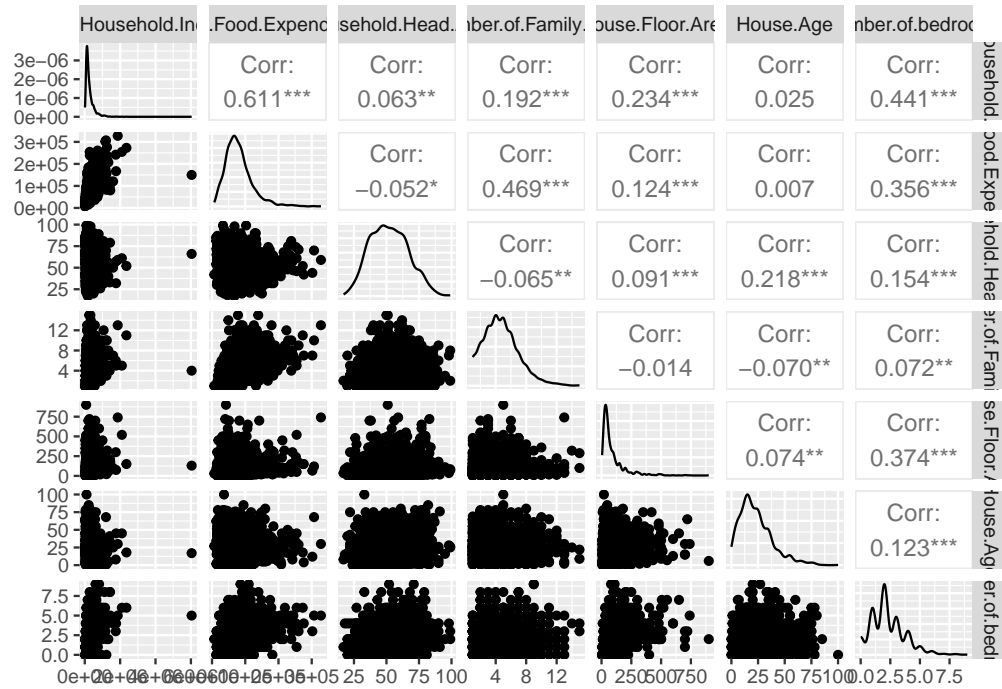


Figure 2: Correlation Matrix of Phillipine Households

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

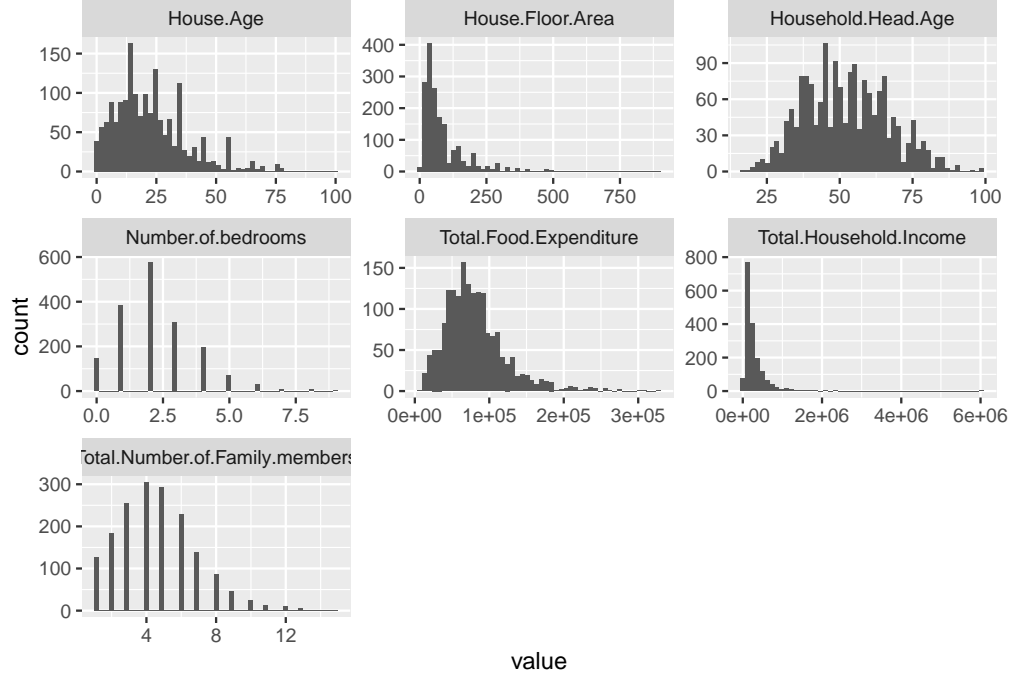


Figure 3: Plot histograms of the variables

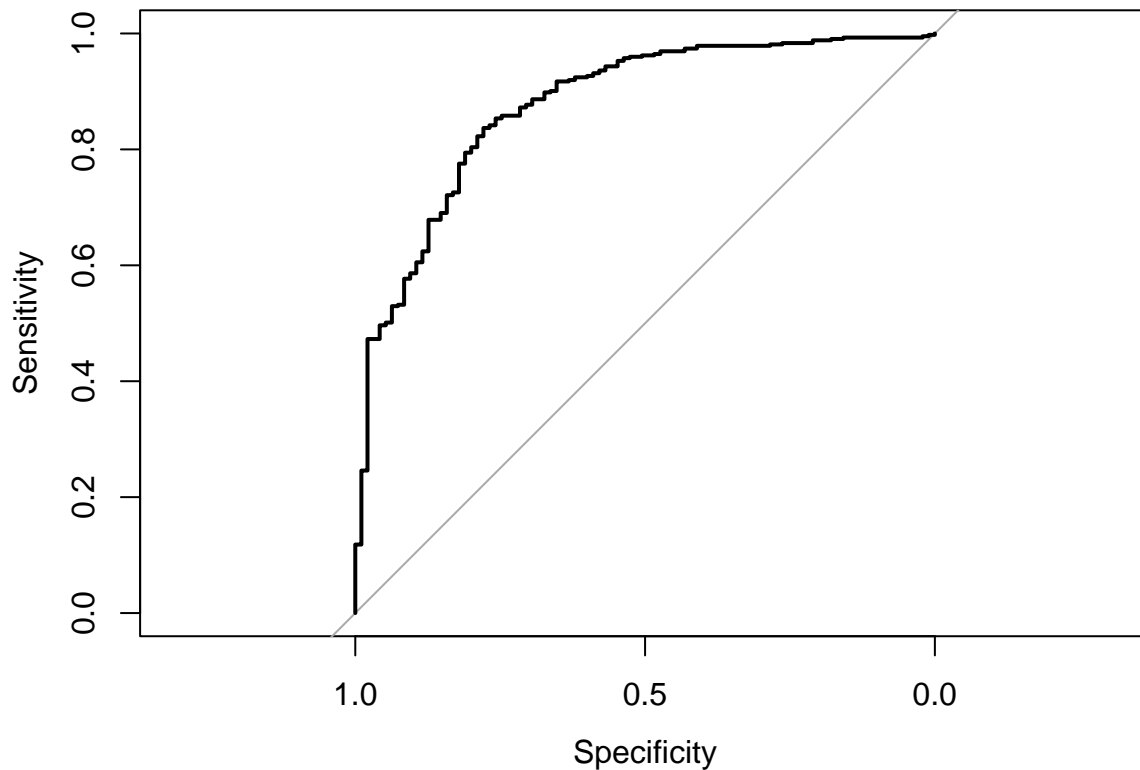
Using the trends seen in the data exploration, our group has taken this further to examine causal relationships through generalized linear models in the Logistic regression.

### 3 Logistic regression

#### 3.1 Data logistic regression {sec;sub}

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 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
-5.008	-1.480	-0.297	1.191	10.610

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.884e+00	2.587e-01	11.150	< 2e-16 ***
Total.Household.Income	-1.120e-06	2.710e-07	-4.132	3.84e-05 ***
Total.Food.Expenditure	3.303e-05	1.787e-06	18.480	< 2e-16 ***
Household.Head.Age	-5.728e-04	4.226e-03	-0.136	0.8922
House.Floor.Area	-1.194e-03	6.438e-04	-1.854	0.0639 .
House.Age	-9.702e-03	3.916e-03	-2.478	0.0134 *
Number.of.bedrooms	-8.417e-02	4.964e-02	-1.696	0.0902 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.153247)

Null deviance: 6659.8 on 1206 degrees of freedom  
 Residual deviance: 4983.9 on 1200 degrees of freedom  
 AIC: 5152.9

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.

## 4 Importance ranking of variables {sec:IRV}

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.b  
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: 5.522236  
Analysis based on 1207 observations

6 Regressors:

Total.Household.Income Total.Food.Expenditure Household.Head.Age House.Floor.Area House.Age Number.of.b  
Proportion of variance explained by model: 25.16%  
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

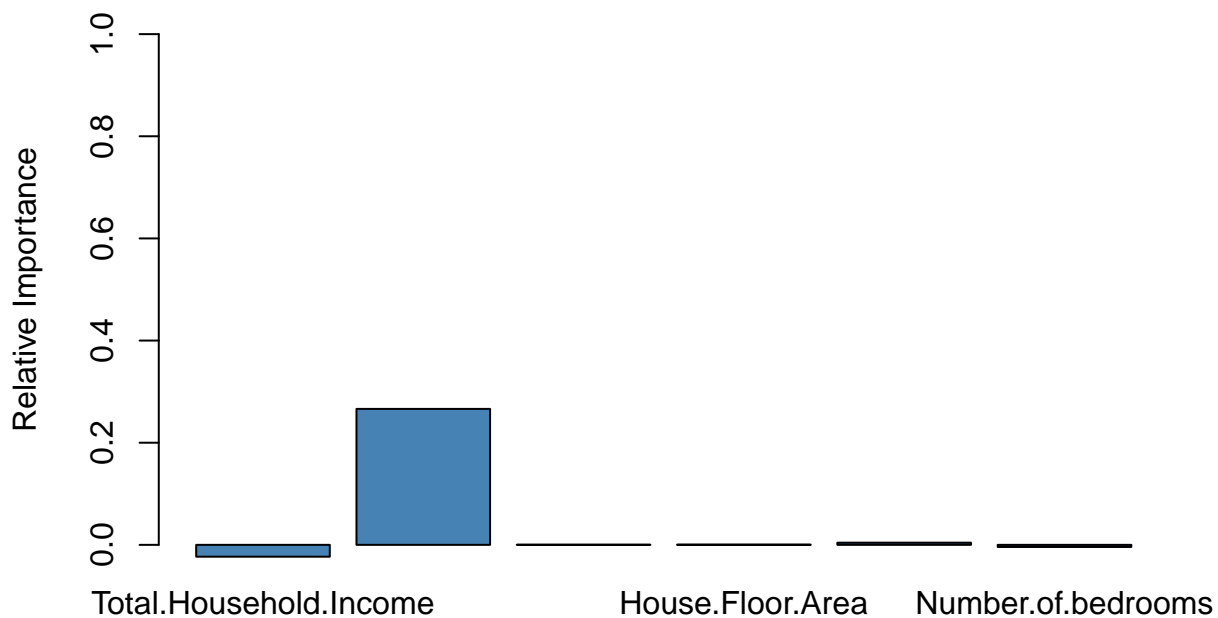
	lmg
Total.Household.Income	0.019474099
Total.Food.Expenditure	0.219431706
Household.Head.Age	0.001998996
House.Floor.Area	0.002294971
House.Age	0.003936260

Number.of.bedrooms 0.004510449

Average coefficients for different model sizes:

	1X	2Xs	3Xs	4Xs
Total.Household.Income	1.412141e-06	8.814137e-07	3.609831e-07	-1.464444e-07
Total.Food.Expenditure	2.705582e-05	2.873314e-05	3.009049e-05	3.121789e-05
Household.Head.Age	-9.728831e-03	-9.410229e-03	-7.882120e-03	-5.708107e-03
House.Floor.Area	-2.648738e-04	-9.494774e-04	-1.294449e-03	-1.392683e-03
House.Age	-9.562669e-03	-9.928653e-03	-9.736746e-03	-9.496361e-03
Number.of.bedrooms	1.311060e-01	5.643347e-02	-1.382831e-03	-4.346237e-02
	5Xs	6Xs		
Total.Household.Income	-6.399184e-07	-1.119917e-06		
Total.Food.Expenditure	3.218038e-05	3.303102e-05		
Household.Head.Age	-3.217397e-03	-5.728175e-04		
House.Floor.Area	-1.330768e-03	-1.193819e-03		
House.Age	-9.448196e-03	-9.702230e-03		
Number.of.bedrooms	-7.076801e-02	-8.417441e-02		

### Relative Importance of Predictor Variables



### Predictor Variables

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 drop 2 variables and then drop 2 slightly influenced variables and re-run de logistic and importance ranking to see the best we can do.

#### 4.1 Re-run the analysis for 4 variables {sec:sub}

Response variable: Total.Number.of.Family.members  
Total response variance: 5.522236



Analysis based on 1207 observations

4 Regressors:

Total.Household.Income Total.Food.Expenditure House.Age Number.of.bedrooms

Proportion of variance explained by model: 24.95%

Metrics are not normalized (rela=FALSE).

Relative importance metrics:

	lmg
Total.Household.Income	0.019623235
Total.Food.Expenditure	0.220349853
House.Age	0.004453013
Number.of.bedrooms	0.005056513

Average coefficients for different model sizes:

	1X	2Xs	3Xs	4Xs
Total.Household.Income	1.412141e-06	4.845635e-07	-3.745909e-07	-1.158882e-06
Total.Food.Expenditure	2.705582e-05	2.962036e-05	3.161088e-05	3.310575e-05
House.Age	-9.562669e-03	-1.075730e-02	-1.033496e-02	-9.967131e-03
Number.of.bedrooms	1.311060e-01	-8.185859e-03	-8.963216e-02	-1.129449e-01

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 22.68%.

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

Response variable: Total.Number.of.Family.members

Total response variance: 5.522236

Analysis based on 1207 observations

2 Regressors:

Total.Household.Income Total.Food.Expenditure

Proportion of variance explained by model: 24.07%

Metrics are not normalized (rela=FALSE).

Relative importance metrics:

	lmg
Total.Household.Income	0.02253813
Total.Food.Expenditure	0.21817957

Average coefficients for different model sizes:

	1X	2Xs
Total.Household.Income	1.412141e-06	-1.353299e-06
Total.Food.Expenditure	2.705582e-05	3.236819e-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 24.11%. 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.

## 5 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.

### 5.1 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.

