Data Analysis Skills

Analysis of Family Income and Expenditure in Phillipines

Jianan Liu | Qijia He | Shreyansi Jain | Subhasish Behera | Tingkai Fan

Introduction

This analysis relates to the Family Income and Expenditure dataset in Philippines.

The data set contains information about the household income, food expenditure, type of household, house floor area and so on.

Aim of the Analysis

This analysis aims to identify the household related variables that influence the number of people living in a household in Philippines in the region CAR.

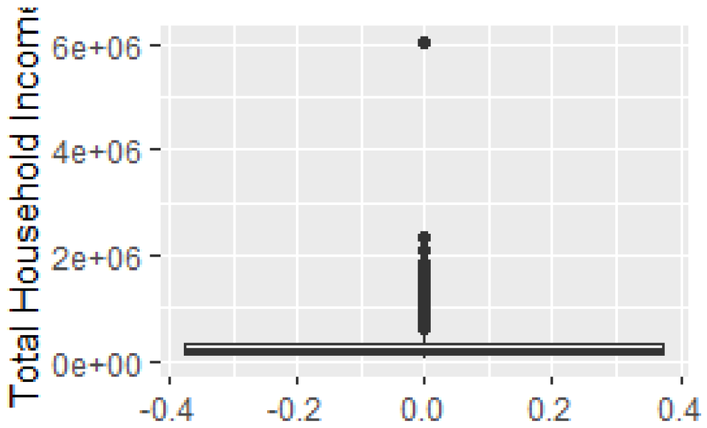
Dataset Description

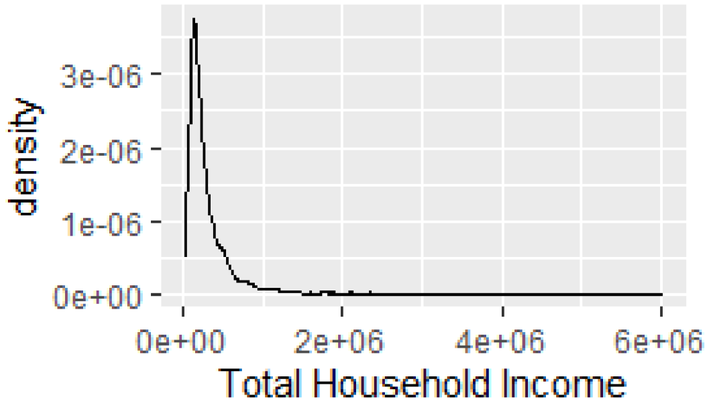
- **Total.Household.Income** – Annual household income (in Philippine peso)
- **Region** – The region under consideration is CAR in this analysis
- **Total.Food.Expenditure** – Annual expenditure by the household on food (in Philippine peso)
- **Household.Head.Sex** – Head of the households sex
- **Household.Head.Age** – Head of the households age (in years)
- **Type.of.Household** – Relationship between the group of people living in the house
- **Total.Number.of.Family.members** – Number of people living in the house
- **House.Floor.Area** – Floor area of the house (in m2)
- **House.Age** – Age of the building (in years)
- **Number.of.bedrooms** – Number of bedrooms in the house 1
- **Electricity** – Does the house have electricity? (1=Yes, 0=No)

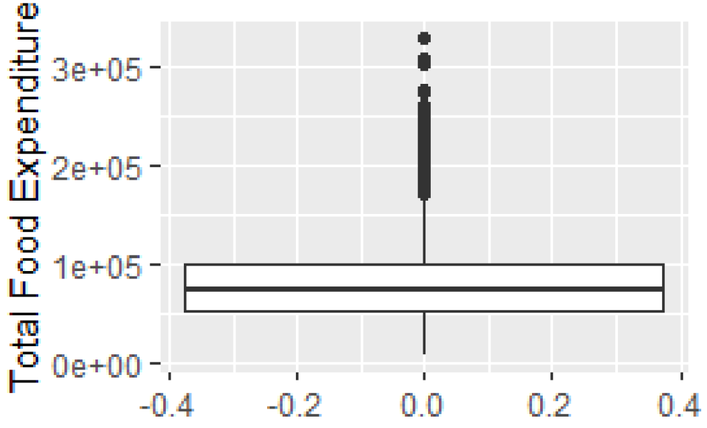
Exploratory Data Analysis

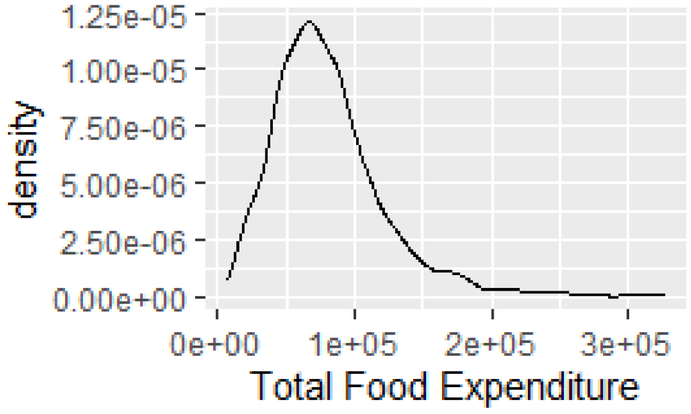
In order to explore the data,
We analyze the numerical and categorical variables separately.

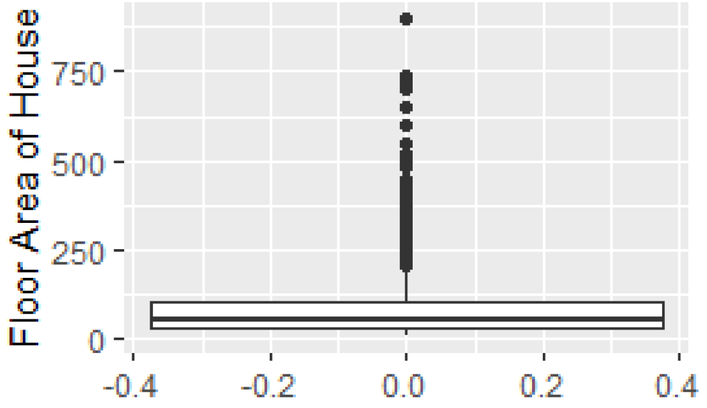
Starting off with some numeric variables>>>

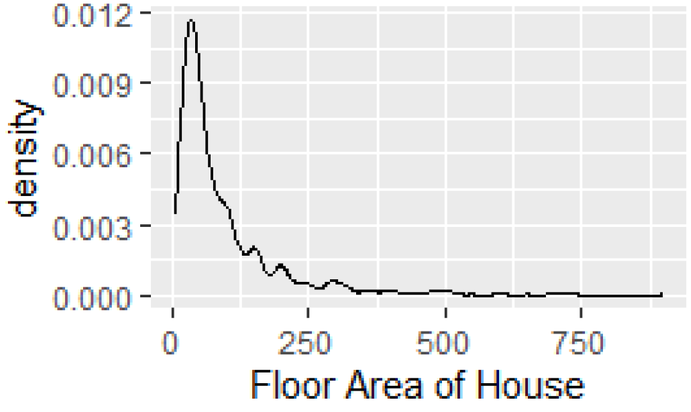

Total Household Income
6e+06
4e+06
2e+06
0e+00
-0.4
-0.2
0.0
0.2
0.4


density
3e-06
2e-06
1e-06
0e+00
0e+00
2e+06
4e+06
6e+06
Total Household Income


Total Food Expenditure
3e+05
2e+05
1e+05
0e+00
-0.4
-0.2
0.0
0.2
0.4

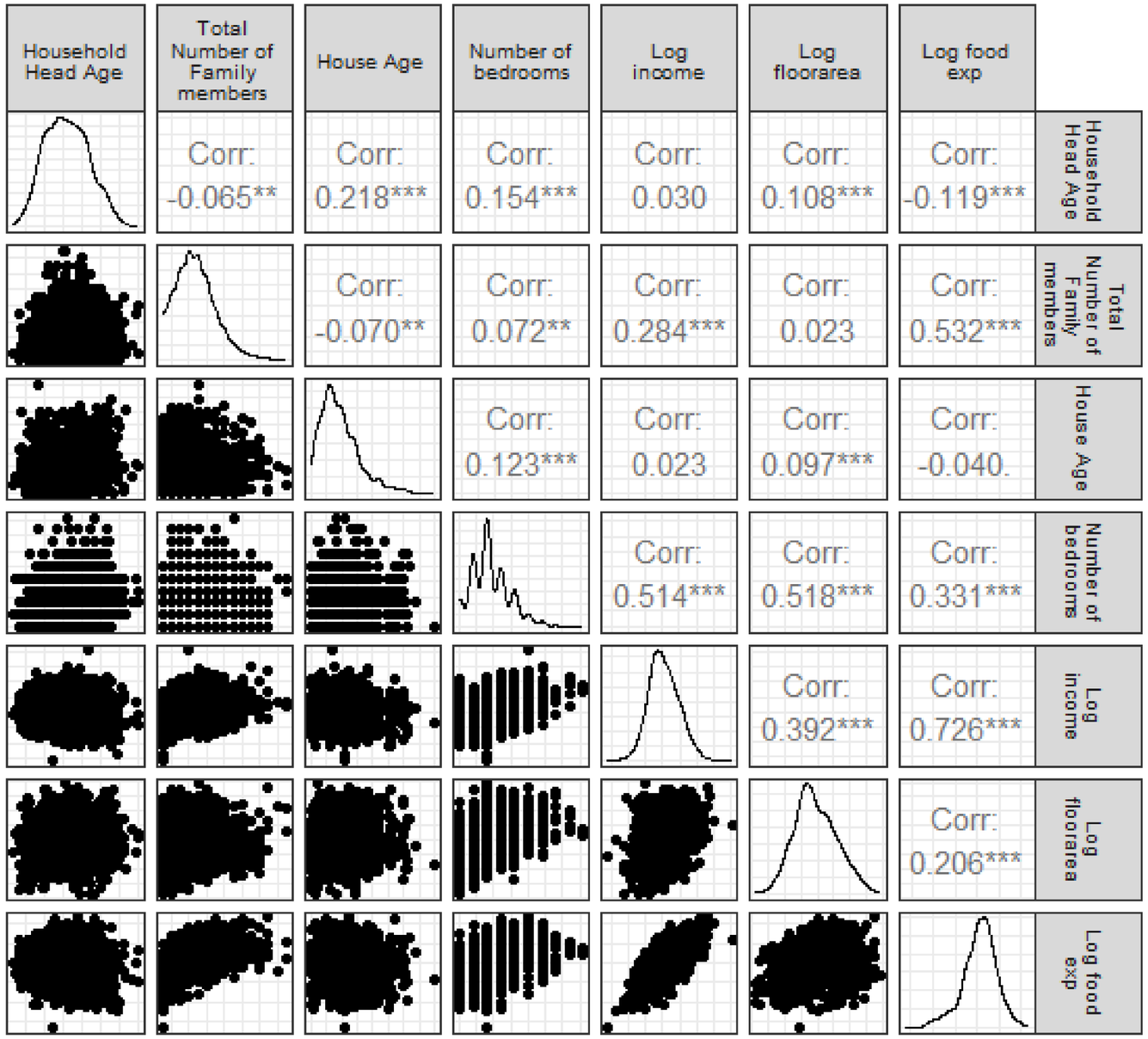

density
1.25e-05
1.00e-05
7.50e-06
5.00e-06
2.50e-06
0.00e+00
0e+00
1e+05
2e+05
3e+05
Total Food Expenditure


Floor Area of House
750
500
250
0
-0.4
-0.2
0.0
0.2
0.4

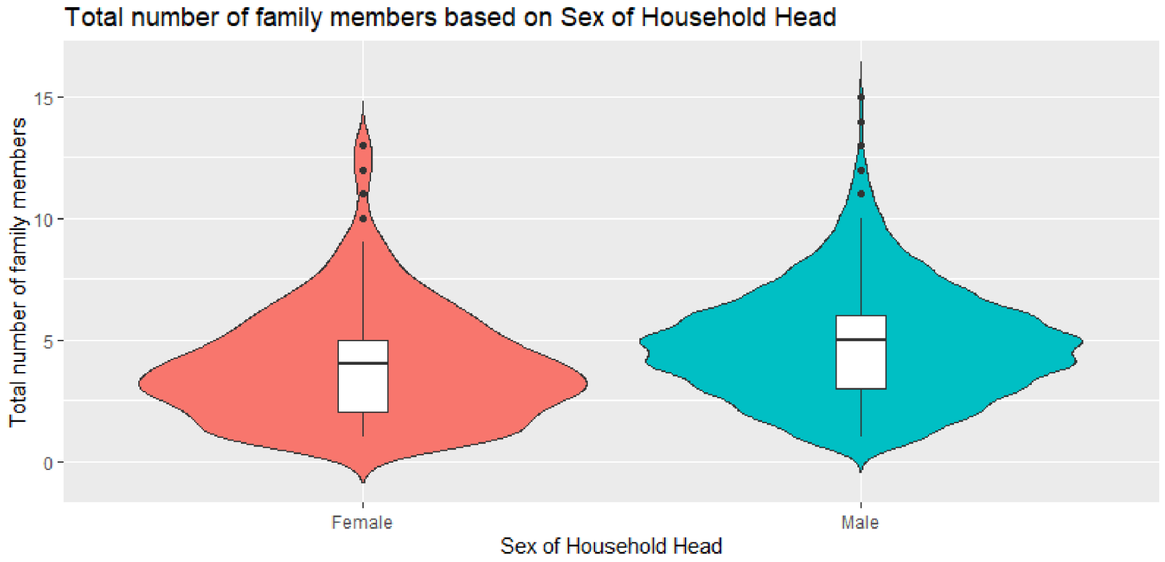

density
0.012
0.009
0.006
0.003
0.000
0
250
500
750
Floor Area of House

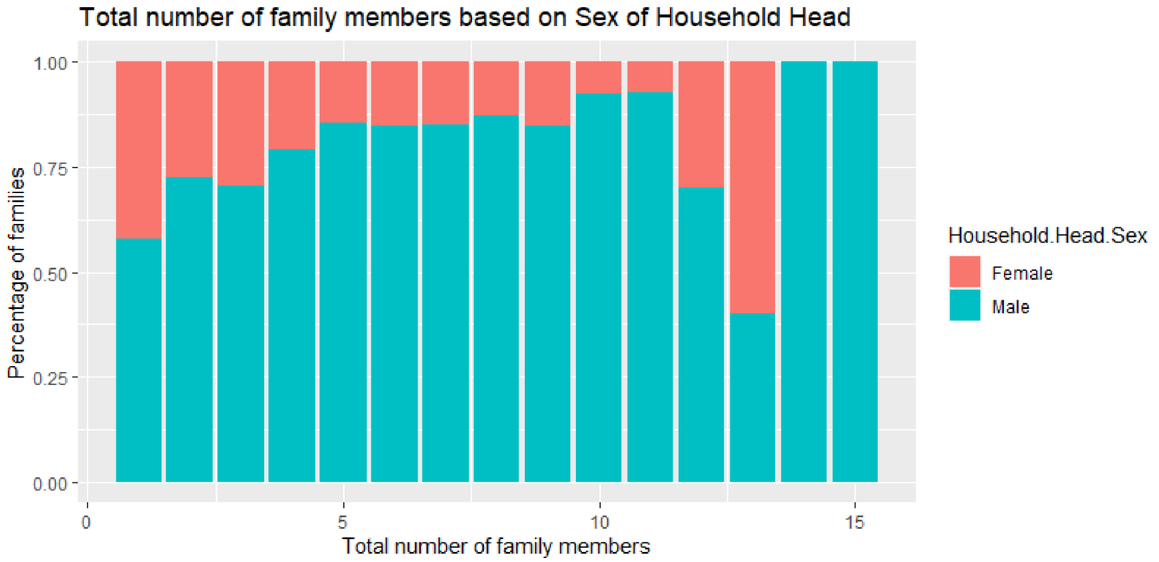
Due to high skewness, Log transform the variables using base 2

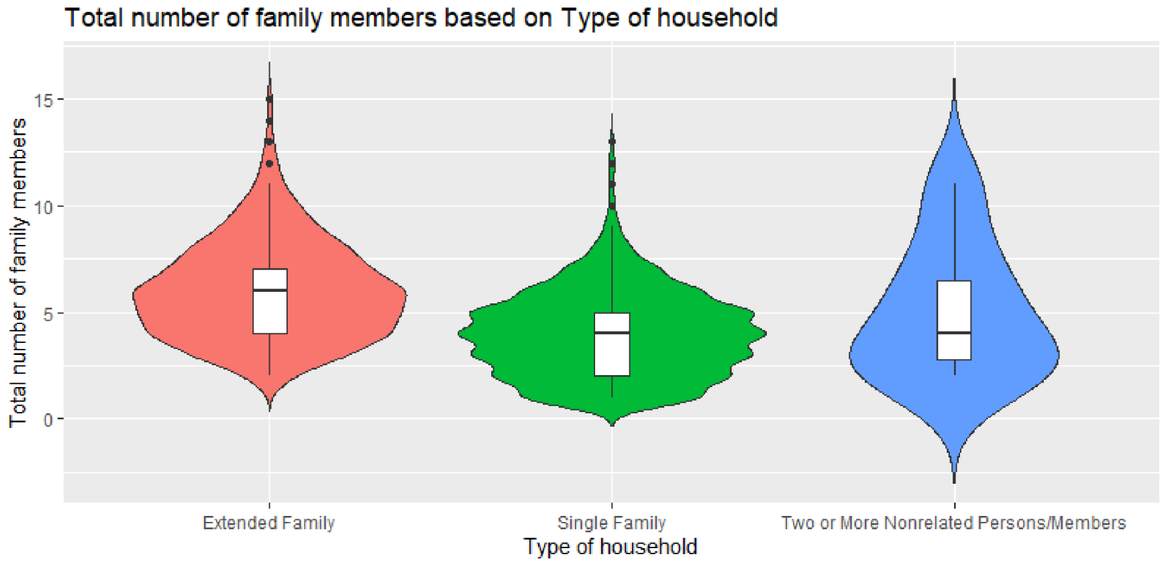
After the transformation, the pairs plot of the numeric variables looks like the following

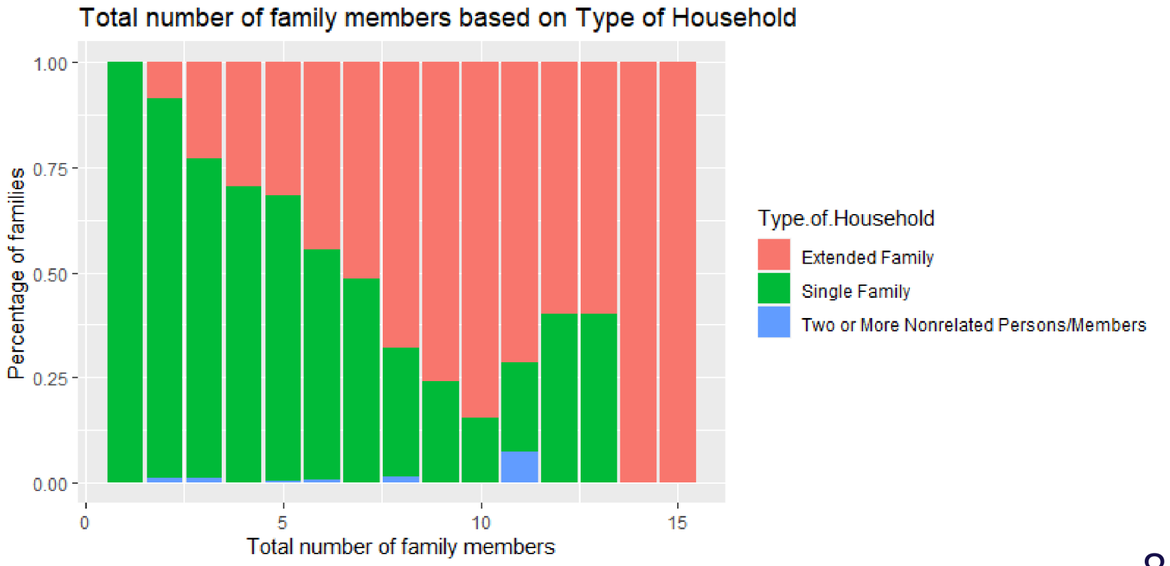

Household Head Age
Total Number of Family members
House Age
Number of bedrooms
Log income
Log floorarea
Log food exp
Corr: -0.065**
Corr: 0.218***
Corr: 0.154***
Corr: 0.030
Corr: 0.108***
Corr: -0.119***
Corr: -0.070**
Corr: 0.072**
Corr: 0.284***
Corr: 0.023
Corr: 0.532***
Corr: 0.123***
Corr: 0.023
Corr: 0.097***
Corr: -0.040.
Corr: 0.514***
Corr: 0.518***
Corr: 0.331***
Corr: 0.392***
Corr: 0.726***
Corr: 0.206***

Categorical Variables









Statistical Modelling

Modelling Count Data- Number of family members in a household

Poisson Model

Negative Binomial Model

Model Selection

Steps to choose the best model include

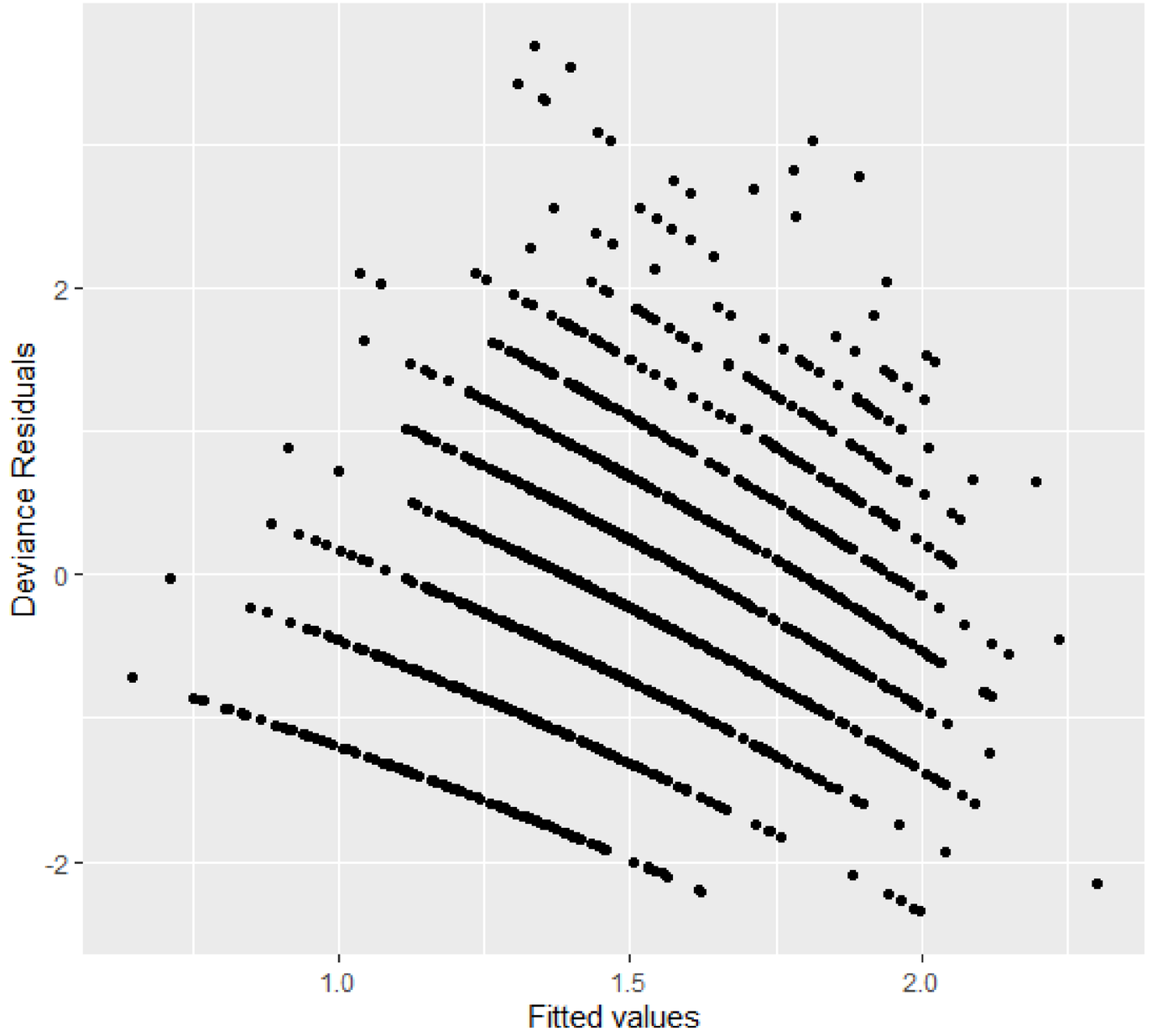
1. Fitting a full model
2. Removing variables with pvalue > 5%
3. Check assumptions
4. Choose the best model based on a particular criteria- AIC or Deviance

Poisson Model 1

```
Call:
glm(formula = Total.Number.of.Family.members ~ ., family = poisson(),
    data = dataset[, -c(10)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3692  -0.7292  -0.1354   0.5286   3.6089

Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -0.0322993  0.2258494  -0.143  0.88628
Household.Head.SexMale                             0.2335947  0.0296685   7.873 3.45e-15 ***
Household.Head.Age                                -0.0038907  0.0008554  -4.548 5.41e-06 ***
Type.of.HouseholdSingle Family                    -0.4147942  0.0241591 -17.169  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.1246197  0.1598410  -0.780  0.43560
House.Age                                         -0.0020011  0.0007731  -2.589  0.00964 **
Number.of.bedrooms                                -0.0141210  0.0099169  -1.424  0.15447
Electricity1                                       0.0376406  0.0478872   0.786  0.43185
Log.income                                         0.1182066  0.0127336   9.283  < 2e-16 ***
Log.floorarea                                     -0.0342849  0.0105677  -3.244  0.00118 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1483.6  on 1715  degrees of freedom
AIC: 7166.1

Number of Fisher Scoring iterations: 4
```

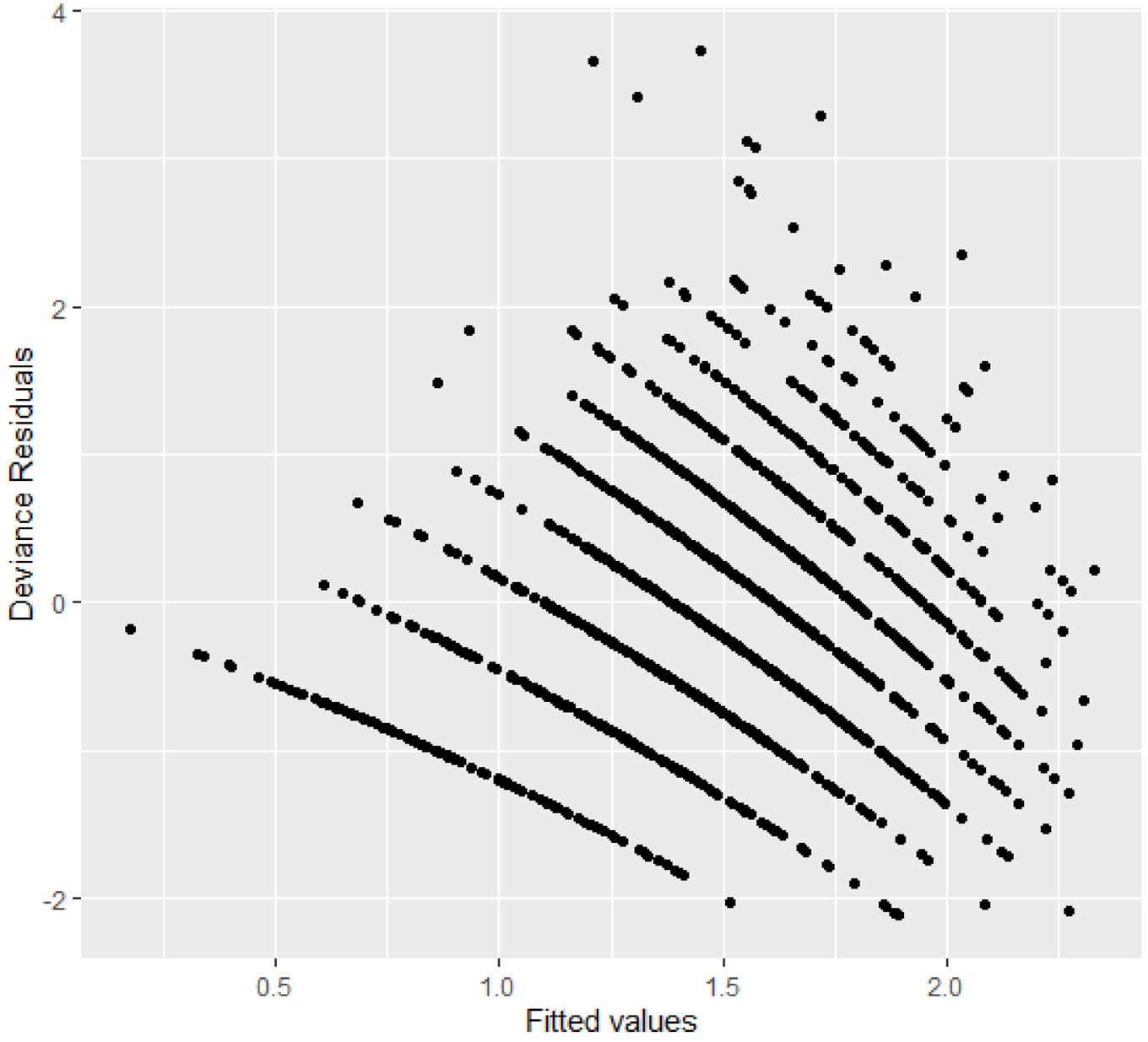
Poisson Model 2

```
Call:
glm(formula = Total.Number.of.Family.members ~ Household.Head.Sex +
    House.Age + Log.income + Log.floorarea + Household.Head.Age +
    householdtype_binary, family = poisson(), data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3503  -0.7354  -0.1360   0.5214   3.6785

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              0.1094018  0.2022987   0.541  0.58865
Household.Head.SexMale                   0.2294369  0.0295836   7.756 8.80e-15 ***
House.Age                               -0.0019835  0.0007682  -2.582  0.00982 **
Log.income                               0.1129868  0.0114992   9.826  < 2e-16 ***
Log.floorarea                           -0.0396252  0.0097853  -4.049 5.13e-05 ***
Household.Head.Age                      -0.0040981  0.0008486  -4.830 1.37e-06 ***
householdtype_binaryNot Extended Family -0.4145507  0.0240528 -17.235  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1489.3  on 1718  degrees of freedom
AIC: 7165.8

Number of Fisher Scoring iterations: 4
```


Deviance Residuals
2
0
-2
1.0
1.5
2.0
Fitted values

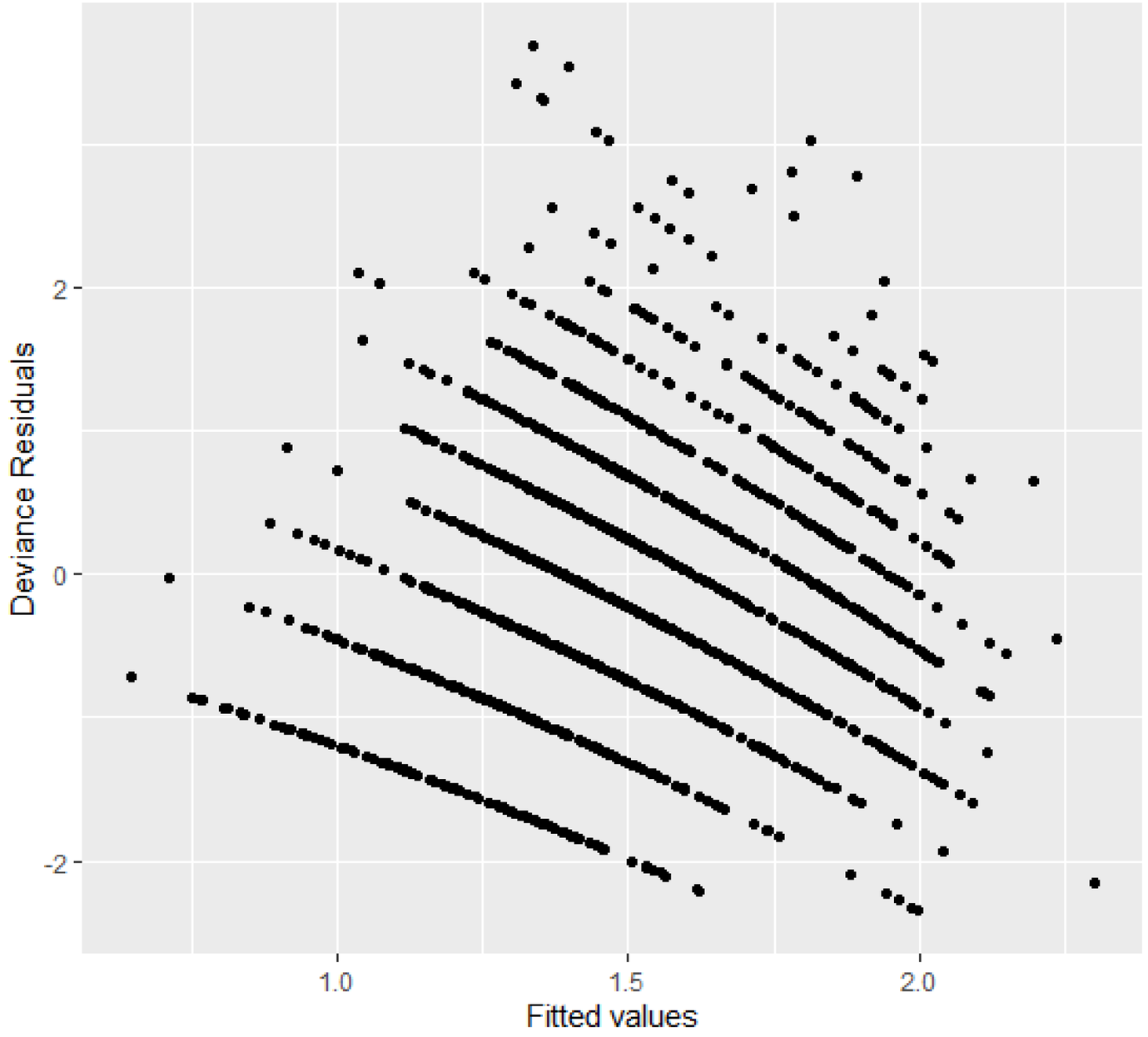
Poisson Model 3

```
Call:
glm(formula = Total.Number.of.Family.members ~ Log.food.exp +
    Household.Head.Sex + Household.Head.Age + householdtype_binary +
    Log.floorarea + House.Age, family = poisson(), data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1173  -0.6524  -0.1191   0.4602   3.7207

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -3.3491844  0.2888214 -11.596  < 2e-16 ***
Log.food.exp                             0.3265437  0.0171625  19.027  < 2e-16 ***
Household.Head.SexMale                   0.2025974  0.0295793   6.849 7.42e-12 ***
Household.Head.Age                      -0.0017581  0.0008733  -2.013   0.0441 *
householdtype_binaryNot Extended Family -0.3081896  0.0248629 -12.396  < 2e-16 ***
Log.floorarea                           -0.0410523  0.0094133  -4.361 1.29e-05 ***
House.Age                               -0.0017739  0.0007704  -2.303   0.0213 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1214.7  on 1718  degrees of freedom
AIC: 6891.3

Number of Fisher Scoring iterations: 4
```


4
2
0
-2
Deviance Residuals
0.5
1.0
1.5
2.0
Fitted values

Negative Binomial Model

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Log.income +
    Household.Head.Sex + Household.Head.Age + householdtype_binary +
    Log.floorarea + House.Age, data = dataset2, init.theta = 59513.02653,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3502  -0.7354  -0.1360   0.5214   3.6783

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            0.1093850  0.2023086   0.541  0.58873
Log.income                             0.1129883  0.0114997   9.825  < 2e-16 ***
Household.Head.SexMale                 0.2294380  0.0295850   7.755 8.82e-15 ***
Household.Head.Age                    -0.0040983  0.0008486  -4.829 1.37e-06 ***
householdtype_binaryNot Extended Family -0.4145526  0.0240540 -17.234  < 2e-16 ***
Log.floorarea                         -0.0396255  0.0097858  -4.049 5.14e-05 ***
House.Age                             -0.0019835  0.0007682  -2.582  0.00982 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(59513.03) family taken to be 1)

    Null deviance: 2024.2  on 1724  degrees of freedom
Residual deviance: 1489.2  on 1718  degrees of freedom
AIC: 7167.8

Number of Fisher Scoring iterations: 1
```


Deviance Residuals
2
0
-2
1.0
1.5
2.0
Fitted values

Results

From the models it can be seen:

1. All the variables in the Poisson model 3 are significant.
2. Poisson Model 3 has the best AIC
3. The assumptions for Poisson Model 3 seem reasonable

The Best model is the **Poisson Model 3**. It was fit using the following variables:

- Log of expenditure on food
- Sex of the household head
- age of the household head
- type of household
- log of floor area
- age of house

Conclusion

Based on the coefficients calculated from the selected model, which is the Poisson model 3, a one unit increase in an explanatory variable is associated with a multiplicative effect and results in an effect of e^β . Thus, the most influential variable is the **total expenditure on food**.

Future Work

The following work can be undertaken to expand on this analysis:

- Additional data covering more regions in the Philippines can be used to create a more accurate model.
- The Family Income and Expenditure data collected by the Philippines Statistical Association covers more variables which can be included in the analysis to build better models.
- More complex models can be built on the data.

Questions?