Analysis of Income and Expenditure Data in the Phillipines

Fanting Kai, Jianan Liu, Qijia He, Shreyansi Jain, Subasish Behera

2023-03-20

```
#Loading required packages
library(pROC)
library(relaimpo)
library(tidyverse)
library(kableExtra)
library(gridExtra)
library(skimr)
library(knitr)
library(moderndive)
library(gapminder)
library(stats)
library(GGally)
library(MASS)
```

1 Introduction

The Philippine Statistics Authority conducts a nationwide survey every 3 years which is aimed at providing the data on family income and expenditure. The data set contains information about the household income, food expenditure, type of household, house floor area and so on. 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.

The following report contains an explanatory analysis into the data including data graphics and summaries in 3. The formal analysis has been built using a Generalized Linear Model which will be seen in sections 4 and 5.

2 Dataset Description

The data set of interest contains the following variables:

- Total.Household.Income – Annual household income (in Philippine peso)
- Region – The region of the Philippines. 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 m^2)
- House.Age – Age of the building (in years)
- Number.of.bedrooms – Number of bedrooms in the house

- Electricity – Does the house have electricity? (1=Yes, 0=No)

From the above list of variables, we notice that six of the variables are numeric (namely Total.Household.Income, Total.Food.Expenditure, Household.Head.Age, House.Floor.Area, House.Age, Number.of.bedrooms) and the rest are categorical (namely Region, Household.Head.Sex, Type.of.Household, Electricity). The response variable of interest is "Total.Number.of.Family.members" (which is numeric). The following section will further analyse the relationship within these variables.

3 Exploratory Data Analysis

First, data cleaning is performed to convert some categorical variables into factors and unnecessary variables are removed. This will help in easier processing later.

```
#Data Cleaning

#Removing the variable "Region" since the information relates to only one Region: CAR
dataset <- dataset %>% select(-Region)

dataset$Electricity <- factor(dataset$Electricity)
dataset$Type.of.Household <- factor(dataset$Type.of.Household)
dataset$Household.Head.Sex <- factor(dataset$Household.Head.Sex)

df_num <- select_if(dataset, is.numeric)

#Summary of numeric variables
my_skim <- skim_with(numeric = sfl(hist = NULL),
                     base = sfl(n = length))
my_skim(df_num) %>%
  transmute(Variable=skim_variable, n = n, Mean=numeric.mean, SD=numeric.sd,
            Min=numeric.p0, Median=numeric.p50,  Max=numeric.p100,
            IQR = numeric.p75-numeric.p50) %>%
  kable(format.args = list(big.mark = ","),
  caption = '\\label{tab: Summary Statistics}
        Summary statistics for numerical variables.', digits=2) %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: Summary statistics for numerical variables.

Variable	n	Mean	SD	Min	Median	Max	IQR
Total.Household.Income	1,725	269,540.48	274,564.17	11,988	188,580	6,042,860	139,755
Total.Food.Expenditure	1,725	80,352.78	41,194.36	6,781	73,578	327,724	24,915
Household.Head.Age	1,725	52.23	14.52	17	52	99	11
Total.Number.of.Family.members	1,725	4.67	2.33	1	4	15	2
House.Floor.Area	1,725	90.92	99.20	5	54	900	48
House.Age	1,725	22.98	15.32	0	20	100	11
Number.of.bedrooms	1,725	2.26	1.44	0	2	9	1

Table 1 shows the summary of numeric information for the 1725 observations. The data shows clear signs of existence of outliers since for most of the numeric variables, the maximum values are significantly larger than the Median. The mean value and median value also seem significantly different for some variables for eg: Total.Food.Expenditure due to the existence of some extremely large values. ‘

The boxplots in figure 1 show the outliers suspected in the numeric data summary. Additionally, The density plots show that the data is skewed to the right. Due to the heavily skewed data, it is suggested to use the log scale for these variables. These have been thus converted into a log scale with base 2 to make the data more symmetrical.

```
#Boxplots and density plots of highly skewed numeric variables

ggplot(data = dataset, aes( y = Total.Household.Income)) +
  geom_boxplot() +
  ylab("Total Household Income")
ggplot(data = dataset, aes( x = Total.Household.Income)) +
  geom_density() +
  xlab("Total Household Income")

ggplot(data = dataset, aes( y = Total.Food.Expenditure)) +
  geom_boxplot() +
  ylab("Total Food Expenditure")
ggplot(data = dataset, aes( x = Total.Food.Expenditure)) +
  geom_density() +
  xlab("Total Food Expenditure")

ggplot(data = dataset, aes( y = House.Floor.Area)) +
  geom_boxplot() +
  ylab("Floor Area of House")
ggplot(data = dataset, aes( x = House.Floor.Area)) +
  geom_density() +
  xlab("Floor Area of House")
```

```
#Converting the skewed variables into log scale

#base 2 has a nice interpretation
dataset['Log.income'] <- log(dataset$Total.Household.Income, base = 2)
dataset['Log.floorarea'] <- log(dataset$House.Floor.Area, base = 2)
dataset['Log.food.exp'] <- log(dataset$Total.Food.Expenditure, base = 2)
```

```
#Removing old variables (variables before conversion)

dataset <- dataset %>%
  select(-Total.Household.Income, -House.Floor.Area, -Total.Food.Expenditure)
```

Next, a pairs plot has been created to look at the scatter plots, density plots and correlation between the transformed variables.

```
#Pairs plot to show the correlation, and density plots and scatter plots of the
#numeric variables
ggpairs(dataset[,-c(1, 3, 7)], axisLabels = "none",
        columnLabels = gsub('.', ' ', colnames(dataset[,-c(1,3,7)]), fixed = T),
        labeller = label_wrap_gen(10))+
   theme_bw(base_size = 9)
```

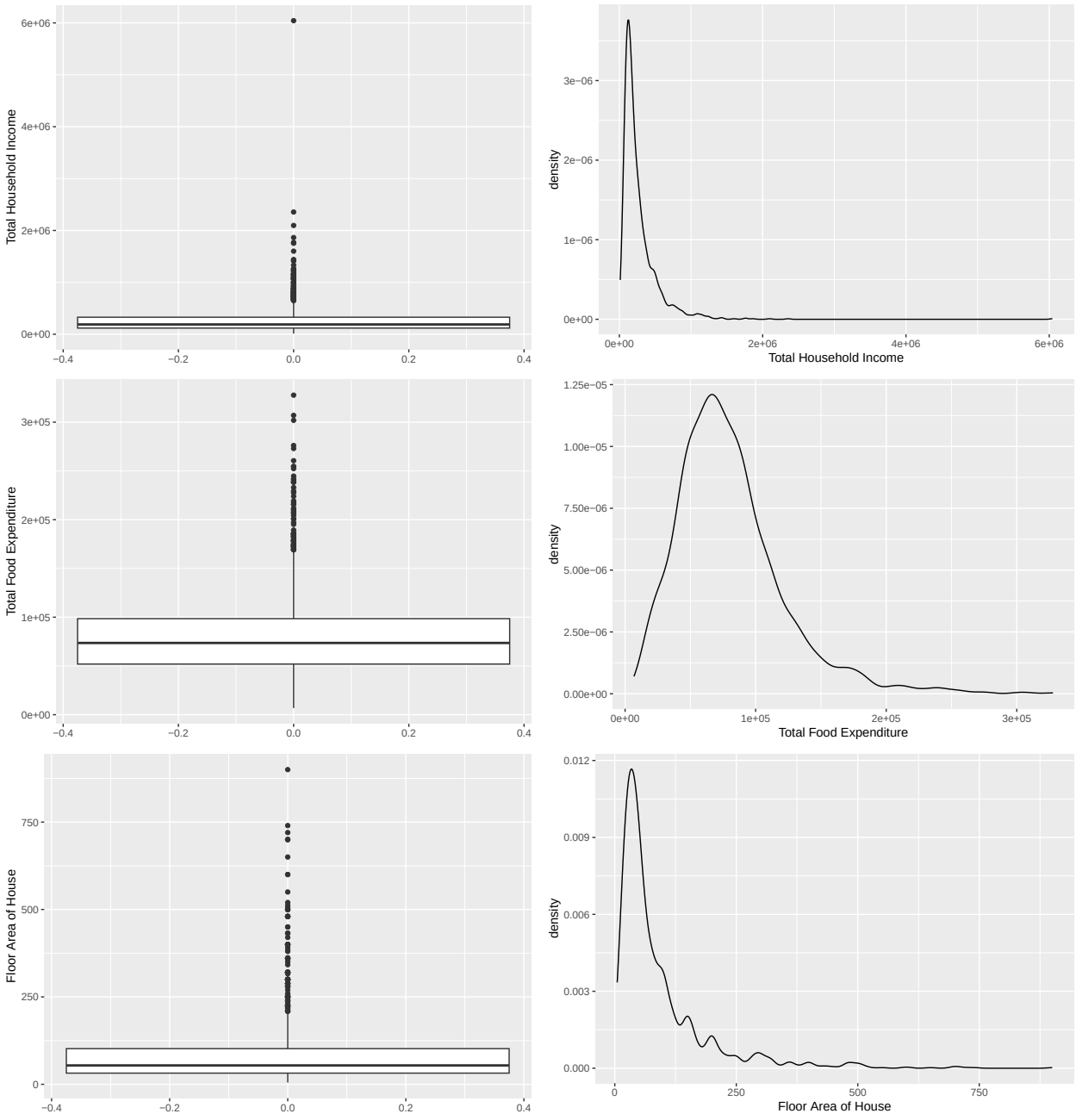

Figure 1: Boxplots and density plots of selected variables

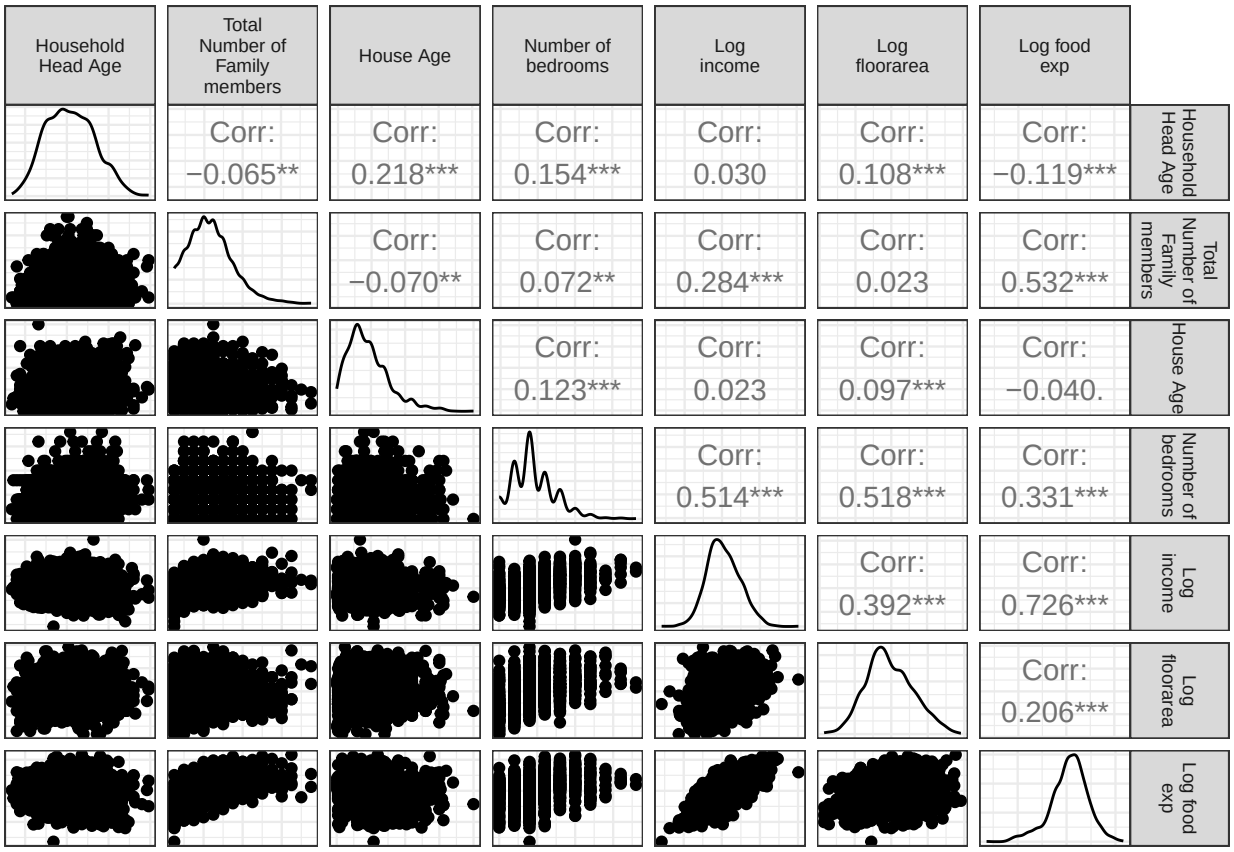

Figure 2: Correlation plots for numerical variables

Figure 2 shows the scatter plots, density plots and the correlation between the numeric variables. We can see the density plots seem relatively more symmetrical after the transformation into log. From the plot, it is also evident that multi-collinearity exists. Log of total income is highly correlated to log of food expenditure 72.63%.

Next, we move on to the analysis of the categorical variables. The graphical analysis of categorical variables can be handled with bar charts and summaries. These can be seen below.

```
#Violin plot of total number of family members based on sex of the household head

dataset %>%
  ggplot(aes(x = Household.Head.Sex, y = Total.Number.of.Family.members,
             fill = Household.Head.Sex))+
  geom_violin(trim = FALSE)+
  geom_boxplot(width = 0.1, fill = "white")+
  theme(legend.position = "none")  +
  labs(title="Total number of family members based on Sex of Household Head",
       x="Sex of Household Head", y = "Total number of family members")
```

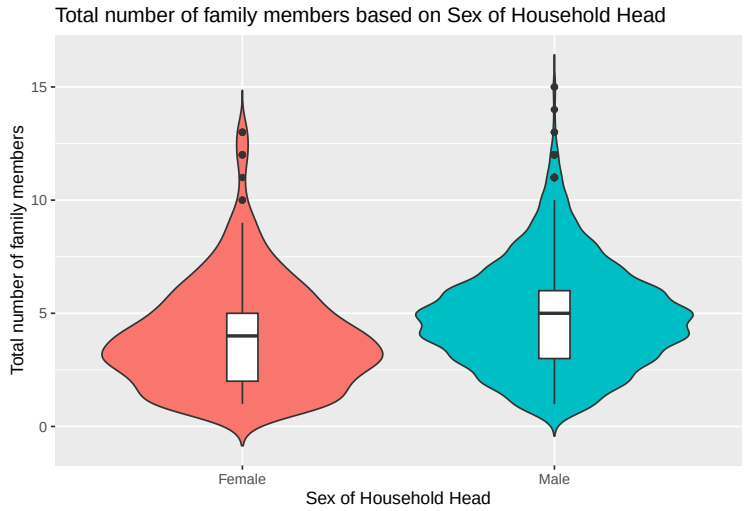

Figure 3: Violin plot of number of family members across gender of household head

```
#Barplot of total number of family members based on sex of the household head

dataset %>% ggplot(aes(x = Total.Number.of.Family.members, fill = Household.Head.Sex)) +
  geom_bar(position = 'fill')+
  labs(title="Total number of family members based on Sex of Household Head",
       x="Total number of family members", y = "Percentage of families")
```

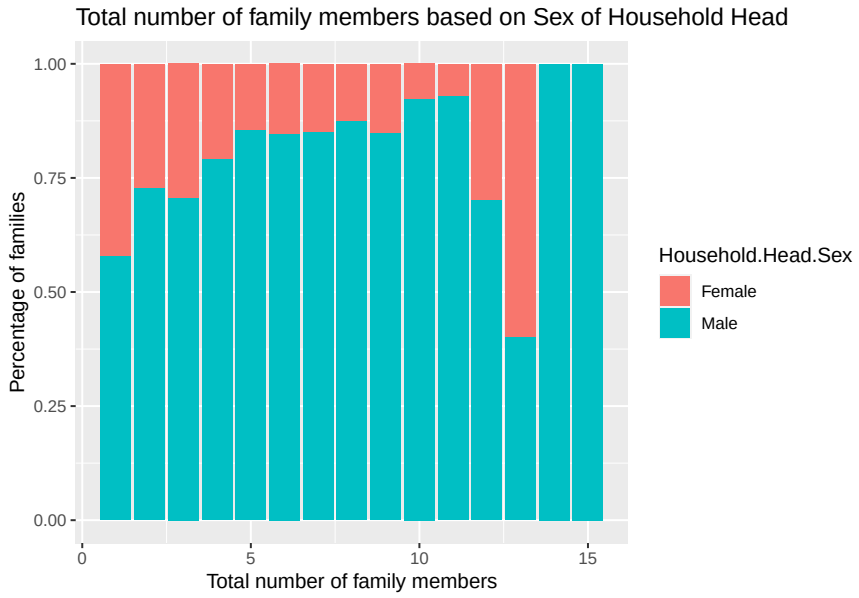

Figure 4: Distribution of number of family members across gender of household head

From figure 3, there seems to be a difference between the number of total family numbers based on the sex of the household head. The figure 4 shows that the percentage of families with male heads are more than that of female heads. It can also be seen that the percentage of families with male heads increase as the size of family members increase up til 10 members. After this point, the data is too scarce to make a conclusion.

```
#Viloin plot of total number of family members based on type of household

dataset %>%
  ggplot(aes(x = Type.of.Household, y = Total.Number.of.Family.members,
             fill = Type.of.Household))+
  geom_violin(trim = FALSE)+
  geom_boxplot(width = 0.1, fill = "white")+
  theme(legend.position = "none")  +
  labs(title="Total number of family members based on Type of household",
       x="Type of household", y = "Total number of family members")
```

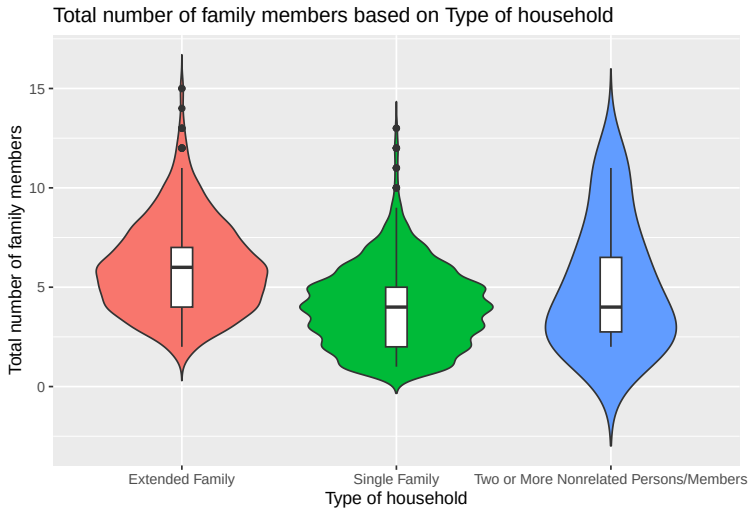

Figure 5: Violin plot of number of family members across type of household

```
#Barplot of total number of family members based on Type of household

dataset %>% ggplot(aes(x = Total.Number.of.Family.members, fill = Type.of.Household)) +
  geom_bar(position = 'fill')+
  labs(title="Total number of family members based on Type of Household",
       x="Total number of family members", y = "Percentage of families")
```

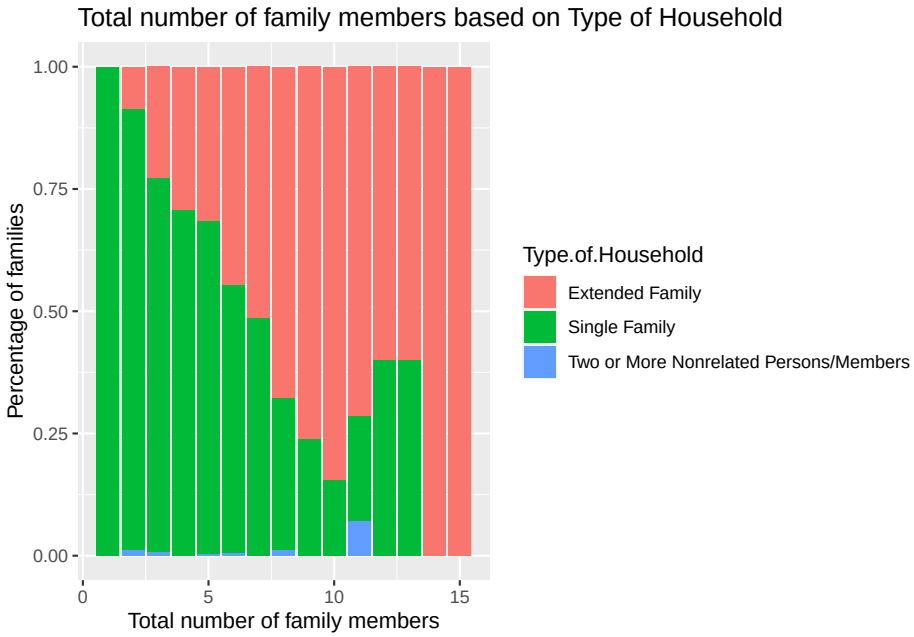

Figure 6: Distribution of number of family members across type of houshold

For 'type of household' variable, a similar trend is seen as before. With higher family members, there is higher proportion of extended family as the type. Given that there is not much data points for the category "Two or More Nonrelated Persons/Members", we can choose to remove it.

The variables 'number of bedrooms' and 'log of total floor area' tend to vary together i.e the correlation is 51.81%. Households having high floor area generally have a higher number of bedrooms. This suggests that once we control for the floor area, number of bedrooms will not be significant. This can tested empirically during the model fit.

Using the trends seen in the data exploration, the analysis can be extended further to examine the relationships using generalized linear models and we bring out 2 different ways to solve the research problem.

4 Generalised Linear Model

4.1 Model 1: Poisson Model

The first model that is used for this data is the possion model. This assumes that the response, given the covariates, follows poisson distribution and the mean and the variance are equal. The check for assumptions will be done after the model fitting.

We start with the full model, using all the variables, except food expenditure because of the high correlation with the income variable.

```
#Fitting a poison model on complete data
pois.fit <- glm(Total.Number.of.Family.members ~ ., data = dataset, family = poisson())

summary(pois.fit)
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3939  -0.6416  -0.1129   0.4502   3.9140

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                           -3.5625633  0.3093005
Household.Head.SexMale                                 0.1926618  0.0297840
Household.Head.Age                                    -0.0010655  0.0008877
Type.of.HouseholdSingle Family                        -0.3029227  0.0250045
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.1869405  0.1598921
House.Age                                             -0.0016425  0.0007755
Number.of.bedrooms                                    -0.0168027  0.0101199
Electricity1                                          -0.0217887  0.0479829
Log.income                                            -0.0773623  0.0174703
Log.floorarea                                         -0.0149389  0.0108468
Log.food.exp                                           0.4160430  0.0243434
                                                      z value Pr(>|z|)
(Intercept)                                           -11.518  < 2e-16 ***
Household.Head.SexMale                                  6.469 9.89e-11 ***
Household.Head.Age                                     -1.200   0.2300
Type.of.HouseholdSingle Family                        -12.115  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -1.169   0.2423
House.Age                                              -2.118   0.0342 *
Number.of.bedrooms                                     -1.660   0.0968 .
Electricity1                                           -0.454   0.6498
Log.income                                             -4.428 9.50e-06 ***
Log.floorarea                                          -1.377   0.1684
Log.food.exp                                           17.091  < 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: 1183.5  on 1714  degrees of freedom
AIC: 6868

Number of Fisher Scoring iterations: 4
```

Using the p-values as a metric to simplify the model, it can be seen that there are variables that need to be taken out, as they are not significant. As was suspected, the variable, number of bedrooms did not turn to be significant.Along with that, whether or not the household has electricity also did not affect the number of members. During the exploratory analysis, it was seen that the category 'Two or more non related persons' had low sample size. The corresponding high standard error could be the effect of that. This also suggests that the 'Type of household' variable could be transformed into a binary variable. A new model is then fit below:

```r
#Combing 2 types of households
dataset2 <- dataset %>% mutate(
  householdtype_binary = fct_recode(Type.of.Household,
                                    'Not Extended Family' = 'Single Family',
                                    'Not Extended Family' =
                                      "Two or More Nonrelated Persons/Members"))
dataset2 <- dataset2 %>%
  select(-Type.of.Household)
```

```r
#Fitting a poison model to the edited data
pois.fit2 <- glm(Total.Number.of.Family.members ~ Household.Head.Sex  + House.Age
                 + Log.income + Log.floorarea + Household.Head.Age + householdtype_binary,
                 data = dataset2, family = poisson())

summary(pois.fit2)
```

```
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

(Intercept)
Household.Head.SexMale                  ***
House.Age                               **
Log.income                              ***
Log.floorarea                           ***
Household.Head.Age                      ***
householdtype_binaryNot Extended Family ***
---
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
```

Using p-values as a metric, it is seen that all of the variables seem to be significant. The variable of 'Type of Household' was transformed to a binary variable,with its levels being 'Extended family' and 'not extended family'; the later includes both 'single family' and 'Two or more no related persons'.

The interpretation of coefficients in this model is different from that of coefficients in OLS. The model itself is multiplicative. So, for example, one unit increase in Log.income means the number of members increase by 11.96%

The deviance for the model, as read from the output, is '1489.3 1489.28 at 1718 degrees of freedom. This value can be compared with the chi-square quantile for assessing lack of fit. The chi-square quantile is 1815.54. The deviance is less, which suggests that fit is better than the saturated model(at 5% significance level).

After the appropriate model is fitted, the assumptions are needed to be checked. Diagnostic plots can be used which involves plot of fitted values and deviance or pearson residuals.

```
#Preparng data for assumption check
diagnostic.data <- dataset2 %>% select(Total.Number.of.Family.members) %>%
  rename(actual = Total.Number.of.Family.members)

diagnostic.data['y_pois'] <- predict(pois.fit2, type = 'response')
diagnostic.data['y_link'] <- predict(pois.fit2, type = 'link') #contains log(lambda) values
diagnostic.data['pois_deviance_resid'] <- resid(pois.fit2, type = 'deviance')
```

```
#Assumption Check
diagnostic.data %>% ggplot(aes(y_link, pois_deviance_resid)) +
  geom_point() + labs(x = 'Fitted values', y = 'Deviance Residuals')
```

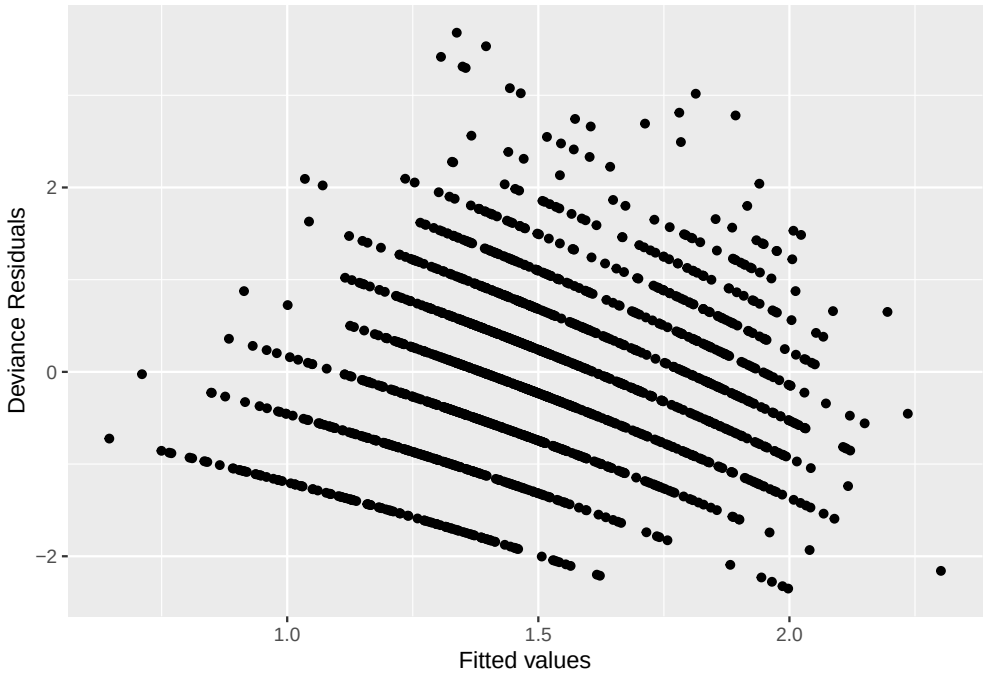

Figure 7: Residual plot for Poisson model.

The point of focus is the range of y-axis. Ideally, if the points are between +-2, it suggests a good fitting model. It can be seen in the figure 7 that most of the points are within this range.The plot exhibits curvature,

suggesting there might be some non-linearity in the relationship between the fitted values and residuals. The model may be improved with inclusion of non-linear terms in the model.

The dispersion parameter is then calculated to check for overdispersion. The estimate of the dispersion parameter for the model is 0.89. As the estimated parameter is < 1, overdispersion might not be an issue for this model. This gives assurance to the standard errors calculated for the parameters. A formal test could be done to check for the opposite case i.e., underdispersion, but this situation is unlikely in practice.

The variable 'Total food expenditure' was found to be correlated with 'Total income' and was arbitrarily excluded. The poisson model could be refitted using that variable and keeping out the income variable. The model improvement or deterioration can then judged from metrics like AIC and deviance values. Note that, the variable is used in a log base-2 scale.

```
#Fitting another poisson model using Log.food.exp instead of Log.income this time
pois.fit3 <-  glm(Total.Number.of.Family.members ~ Log.food.exp + Household.Head.Sex
                  +Household.Head.Age + householdtype_binary + Log.floorarea +
                    House.Age, data = dataset2, family = poisson())

summary(pois.fit3)
```

```
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

(Intercept)                         ***
Log.food.exp                        ***
Household.Head.SexMale              ***
Household.Head.Age                  *
householdtype_binaryNot Extended Family ***
Log.floorarea                       ***
House.Age                           *
---
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
```

There seems to good improvement in the model fitting when AIC and deviance values are considered. The variables again seem to be significant at 5% significance level. The diagnostic plot, similar to the previous model, could be graphed to assess the assumptions.

```
#Assumption check plot
diagnostic.data %>% ggplot(aes(x = predict(pois.fit3, type = 'link'),
                               y = resid(pois.fit3, type = 'deviance'))) +
  geom_point() + labs(x = 'Fitted values', y = 'Deviance Residuals')
```

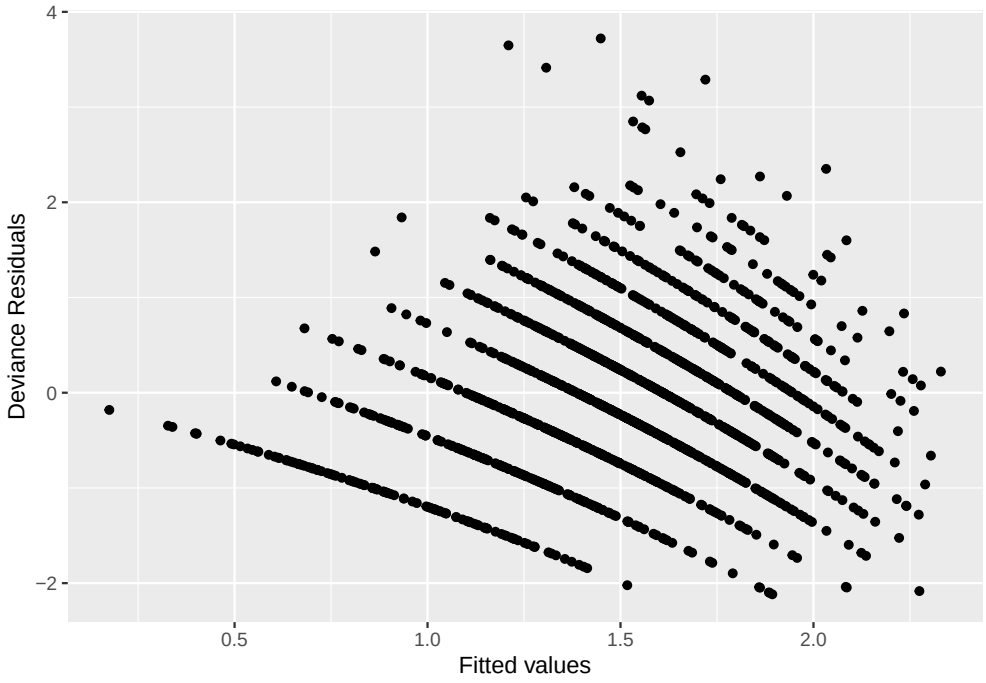

Figure 8: Residual plot for Poisson model with Food Expenditure variable.

In the figure 8, The scale of y-axis has increased, going upto a value of 4, but the quantity of points lying outside the preferred range of +-2 has decreased. The model is relatively superior to the one previously considered that used income instead of food expenditure.

Next we consider one last model that uses the Negative binomial distribution.

4.2 Model 2: Negative Binomial model

In this model, the response is assumed to be distributed according to negative binomial distribution. This has an added benefit in the sense that it does not restrict the mean to be equal to the variance and thus could fit better to the data. The link function in this case is also log link and hence the interpretation of regression coefficients is similar to that of poisson model.

```
#Fitting a negative binomial model
negBin.fit <- glm.nb(Total.Number.of.Family.members ~ Log.income +
```

```
                      Household.Head.Sex +Household.Head.Age + householdtype_binary +
                      Log.floorarea + House.Age, data = dataset2)

summary(negBin.fit)
```

```
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

(Intercept)
Log.income                            ***
Household.Head.SexMale                ***
Household.Head.Age                    ***
householdtype_binaryNot Extended Family ***
Log.floorarea                         ***
House.Age                             **
---
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

              Theta:  59513
          Std. Err.:  247760
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -7151.818
```

At 5% significance, the coefficients are again significant as was previously seen with poisson model. Not much difference in AIC and deviance is seen across the two models.

The diagnostic plot can also be graphed for the negative binomial model.

```
#Assumptions Check
diagnostic.data %>% ggplot(aes(y_link_negBin, negBin_deviance_resid)) +
  geom_point() + labs(x = 'Fitted values', y = 'Deviance Residuals')
```

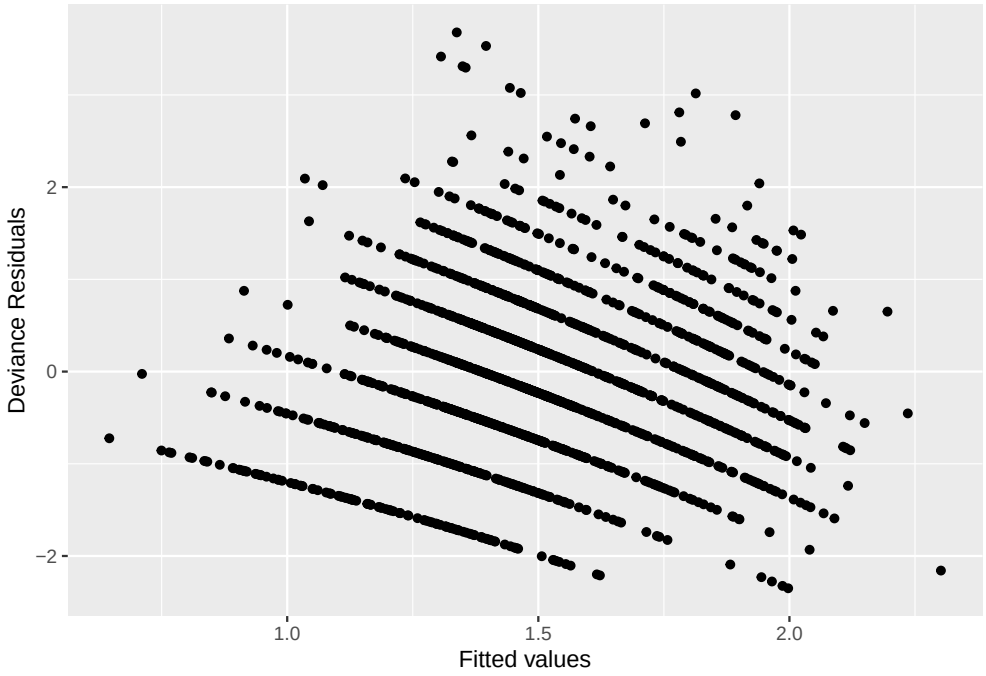

Figure 9: Residual plot for Negative binomial model.

The plot is very similar to the residual plot seen with the poisson model.

5 Conclusion

The Best model seems to be the third poisson model that was fit using the following variables: Log of expenditure on food, Sex of the household head, age of the household head, the type of household, the log of floor area and the age of house. This model has been chosen as the best one based on the AIC of the model and the accuracy of the assumptions.

Based on the coefficients calculated from the model, 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.