

# Group\_02\_Analysis

Group02: Yue Xing, Chengyang Dong, Jingyi Zhu, Qinyuan Ye

## Exploratory Data Analysis

We are using the given data set “dataset02”, which provides data on family income and expenditure. We aim to analyse which household-related variables influence the number of people living in a household.

```
# Load packages
library(MASS)
library(tidyverse)
library(psych)
library(jtools)
library(stats)
library(graphics)
library(ggplot2)
library(patchwork)
library(dplyr)
library(skimr)
library(knitr)
library(kableExtra)

# Upload the dataset02 and rename it as "data"
data <- read.csv("dataset02.csv")
glimpse(data)
```

Rows: 1,249

Columns: 11

|                           |  |
|---------------------------|--|
| \$ Total.Household.Income | <int> 144437, 56094, 215758, 159295, 140240, ~ |
| \$ Region                 | <chr> "IVB - MIMAROPA", "IVB - MIMAROPA", "IV~ |
| \$ Total.Food.Expenditure | <int> 64609, 27218, 73780, 72120, 80152, 2641~ |

```

$ Household.Head.Sex      <chr> "Male", "Male", "Male", "Female", "Fema~
$ Household.Head.Age      <int> 66, 79, 59, 60, 28, 54, 59, 58, 47, 43,~
$ Type.of.Household       <chr> "Single Family", "Single Family", "Sing~
$ Total.Number.of.Family.members <int> 3, 2, 3, 5, 3, 4, 4, 4, 4, 4, 1, 2, 2, ~
$ House.Floor.Area       <int> 45, 13, 22, 25, 13, 160, 52, 36, 9, 7, ~
$ House.Age              <int> 33, 23, 22, 18, 29, 16, 15, 19, 16, 35,~
$ Number.of.bedrooms      <int> 1, 2, 2, 0, 0, 5, 2, 2, 2, 1, 2, 0, 1, ~
$ Electricity             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~

```

Considering that “Total.Number.of.Family.members” is a set of counting variables, consisting of discrete non-negative integers. Therefore, it was chosen to first observe the characteristics of the distribution of the dependent variable.

```

# View outcome variables
ggplot(data, aes(Total.Number.of.Family.members)) +
  geom_histogram()

```

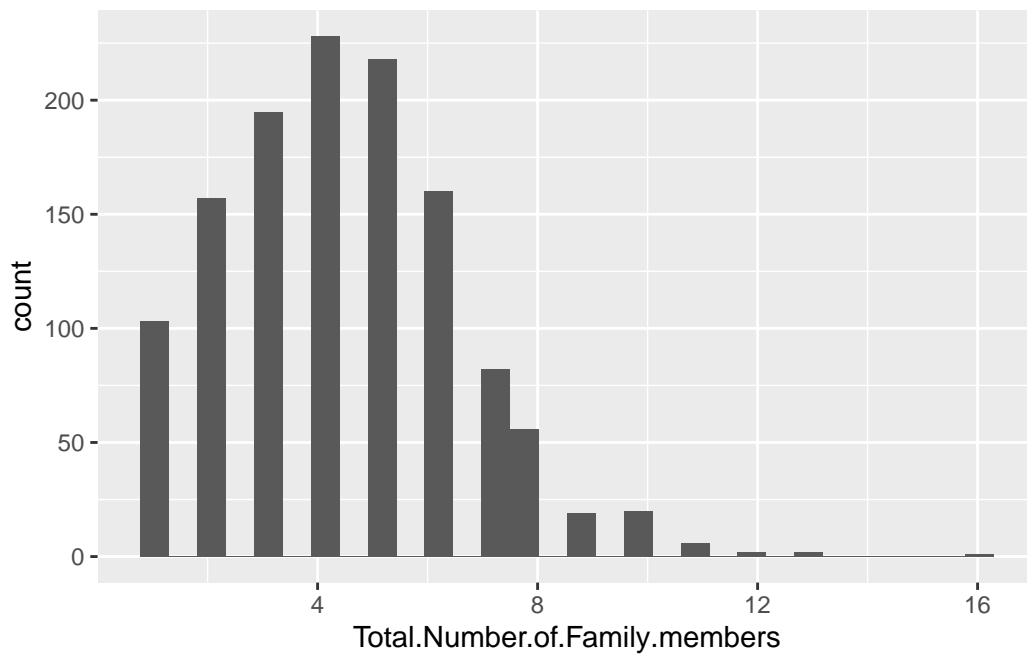


Figure 1: Dependent variable data distribution

The Figure 1 shows that the distribution of the “Total.Number.of.Family.members” variable is right skewed. Combined with the fact that the variable is a set of count-type variables, it

can be roughly tentatively assumed that the total number of family members shows a Poisson distribution.

Therefore, for the subsequent regression model selection for analysing the factors influencing family members, the Poisson regression implementation in the generalised linear model can be initially considered.

## Data Pre-processing

```
# Convert some variables into factor variables
data$Region <- as.factor(data$Region)
data$Household.Head.Sex <- as.factor(data$Household.Head.Sex)
data$Type.of.Household <- as.factor(data$Type.of.Household)
data$Electricity <- as.factor(data$Electricity)
```

```
# View levels of variables
levels(data$Region)
```

```
[1] "IVB - MIMAROPA"
```

```
levels(data$Household.Head.Sex)
```

```
[1] "Female" "Male"
```

```
levels(data$Type.of.Household)
```

```
[1] "Extended Family"
[2] "Single Family"
[3] "Two or More Nonrelated Persons/Members"
```

```
levels(data$Electricity)
```

```
[1] "0" "1"
```

Converting character variables into factor variables is convenient for analysis. We find that “Region” only have one level, so we won’t use this variables to fit model.

The level of “Household.Head.Sex” is female and male. “Type.of.Household” has three different level, which are extended Family, single family and two or more nonrelated persons/members. Because of “1” of variable “Electricity” means the house have electricity and “0” means the house doesn’t have electricity, we convert the count variable “Electricity” into a factor variable.

## Visualization

```
# Descriptive statistical analysis of selected variables
data_summary <- data %>%
  select(Total.Number.of.Family.members, Total.Household.Income,
    ↪ Total.Food.Expenditure,
    ↪ Household.Head.Age, House.Floor.Area, House.Age,
    ↪ Number.of.bedrooms) %>%
  skim() %>%
  transmute(
    Variable = case_when(
      skim_variable == "Total.Number.of.Family.members" ~ "Total Number
    ↪ of Family members",
      skim_variable == "Total.Household.Income" ~ "Total Household
    ↪ Income",
      skim_variable == "Total.Food.Expenditure" ~ "Total Food
    ↪ Expenditure",
      skim_variable == "Household.Head.Age" ~ "Household Head Age",
      skim_variable == "House.Floor.Area" ~ "House Floor Area",
      skim_variable == "House.Age" ~ "House Age",
      skim_variable == "Number.of.bedrooms" ~ "Number of bedrooms",
      TRUE ~ as.character(skim_variable) #
    ),
    Mean = numeric.mean,
    SD = numeric.sd,
    IQR = numeric.p75 - numeric.p50,
    Min = numeric.p0,
    Median = numeric.p50,
    Max = numeric.p100
  )

kable(data_summary, booktabs = TRUE, format = "latex", digits = 2) %>%
```

```
kable_styling(font_size = 12, latex_options = c('scale_down',
↪ 'hold_position'))
```

| Variable                       | Mean      | SD        | IQR   | Min   | Median | Max     |
|--------------------------------|-----------|-----------|-------|-------|--------|---------|
| Total Number of Family members | 4.39      | 2.19      | 2     | 1     | 4      | 16      |
| Total Household Income         | 216685.12 | 263207.20 | 89919 | 18784 | 140483 | 2891788 |
| Total Food Expenditure         | 70760.29  | 41638.03  | 24118 | 10488 | 62590  | 413844  |
| Household Head Age             | 51.37     | 14.24     | 10    | 15    | 51     | 87      |
| House Floor Area               | 48.95     | 49.43     | 24    | 5     | 36     | 750     |
| House Age                      | 16.49     | 12.51     | 8     | 0     | 14     | 105     |
| Number of bedrooms             | 1.78      | 0.98      | 0     | 0     | 2      | 7       |

```
# Correlation graph except categorical variables
pairs.panels(data[, -c(2,4,6)],
  smooth = TRUE,
  scale = FALSE,
  density = TRUE,
  pch = 21,
  lm = TRUE,
  cor = TRUE,
  jiggle = FALSE,
  factor = 2,
  hist.col = 4,
  ci = TRUE
)
```

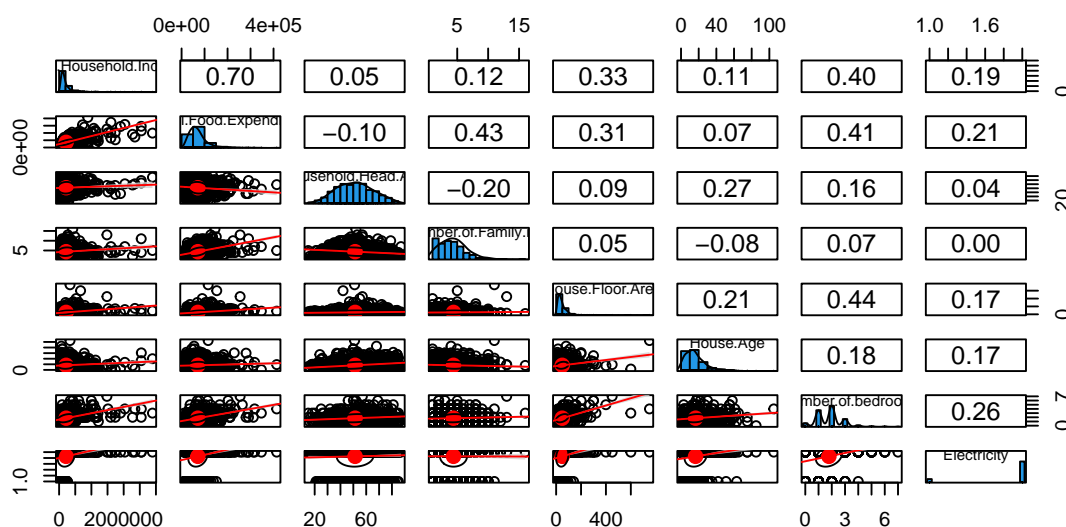


Figure 2: Correlation graph except categorical variables

From the correlation graph of Figure 2 we can see three variables most correlate with the total number of family member: 1.Total household income 2.Total food expenditure 3.Household head age.

```
# The graph with only strong correlate variable
pairs.panels(data[,c(1,3,5,7)],
  smooth = TRUE,
  scale = FALSE,
  density = TRUE,
  pch = 21,
  lm = TRUE,
  cor = TRUE,
  jiggle = FALSE,
  factor = 2,
  hist.col = 4,
  ci = TRUE
)
```

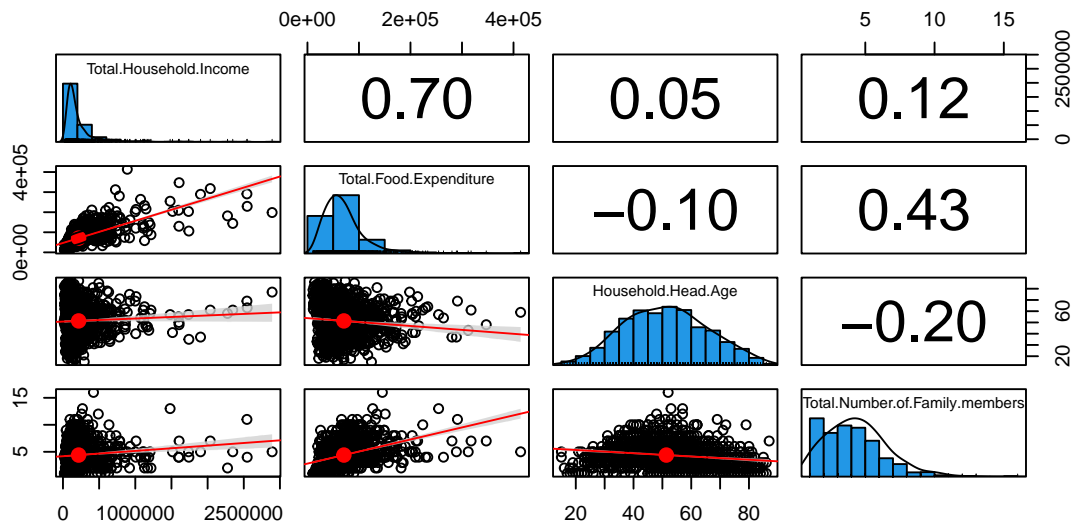


Figure 3: The pairs plot of strong correlate variable

Then we draw boxplots of these variables to analyse:

- First, analyze the data distribution of the total number of family members in the house under different categorical variables.

```
plot_num_sex <- ggplot(data = data, aes(x = Household.Head.Sex,
                                         y =
                                           Total.Number.of.Family.members,
                                         fill = Household.Head.Sex)) +
  geom_boxplot() +
  labs(x = "Household.Head.Sex", y = "Total.Number.of.Family.members")+
  theme(legend.position = "none")
print(plot_num_sex)
```

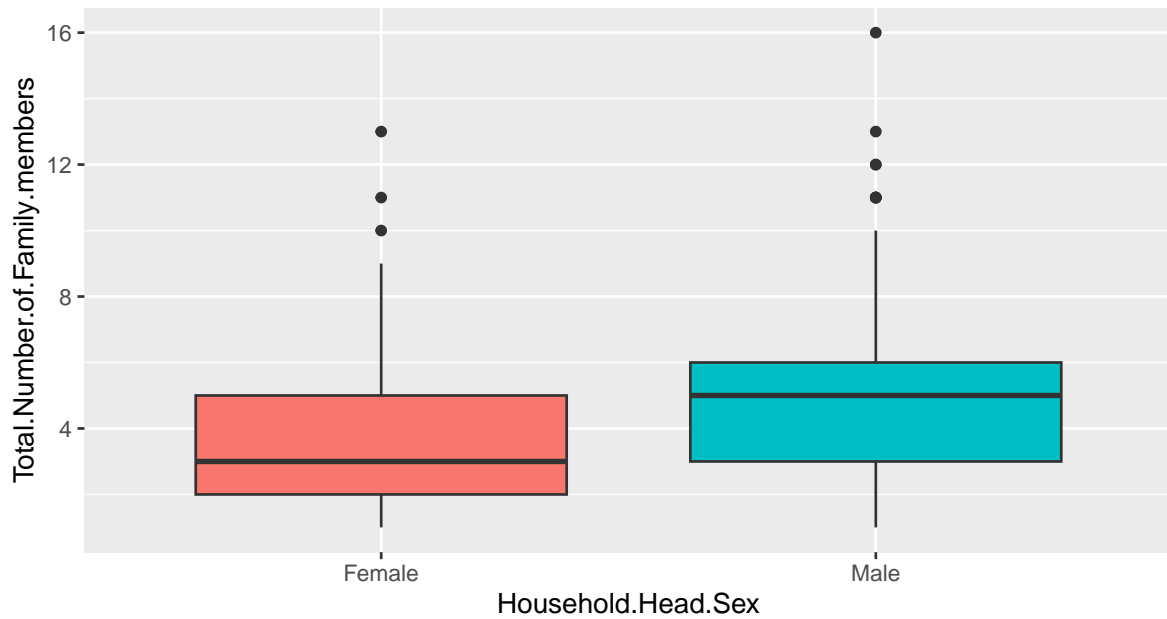


Figure 4: The boxplot of total numbers of family members and gender of head of household

```
plot_num_type <- ggplot(data = data, aes(x = Type.of.Household,
                                          y =
                                            ↪ Total.Number.of.Family.members,
                                          fill = Type.of.Household)) +
  geom_boxplot() +
  labs(x = "Type.of.Household", y = "Total.Number.of.Family.members")+
  theme(legend.position = "none")
print(plot_num_type)
```



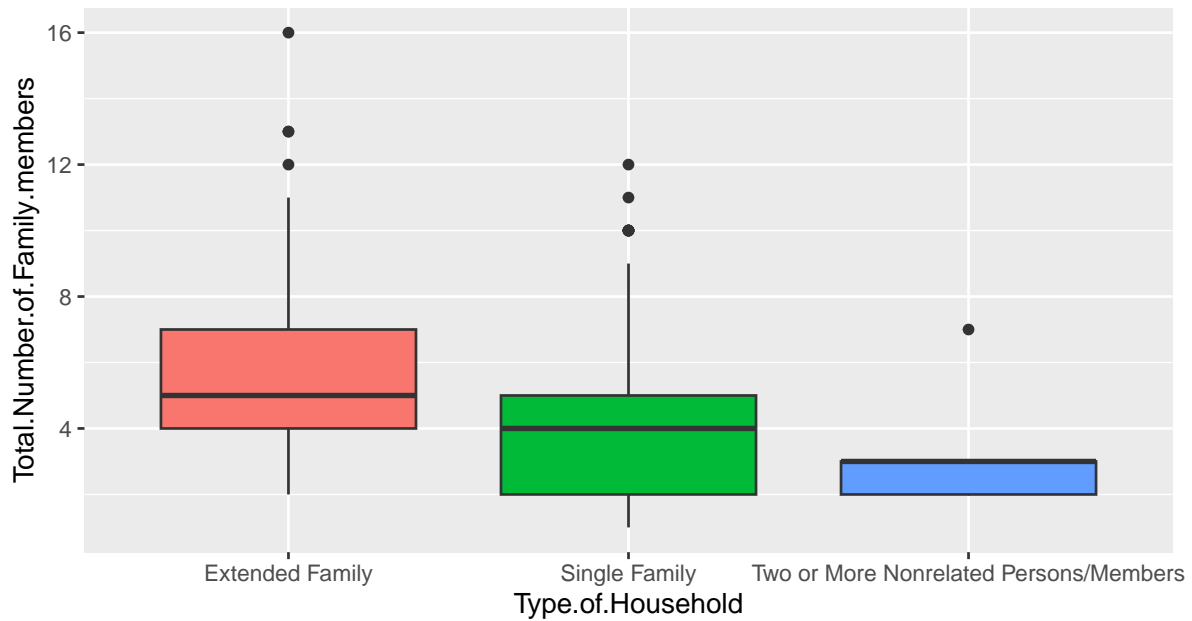


Figure 5: The boxplot of total numbers of family members and type of household

```
plot_num_ele <- ggplot(data = data, aes(x = Electricity,
                                         y =
                                           ↪ Total.Number.of.Family.members,
                                         fill = Electricity)) +
  geom_boxplot() +
  labs(x = "Electricity", y = "Total.Number.of.Family.members")+
  theme(legend.position = "none")
print(plot_num_ele)
```

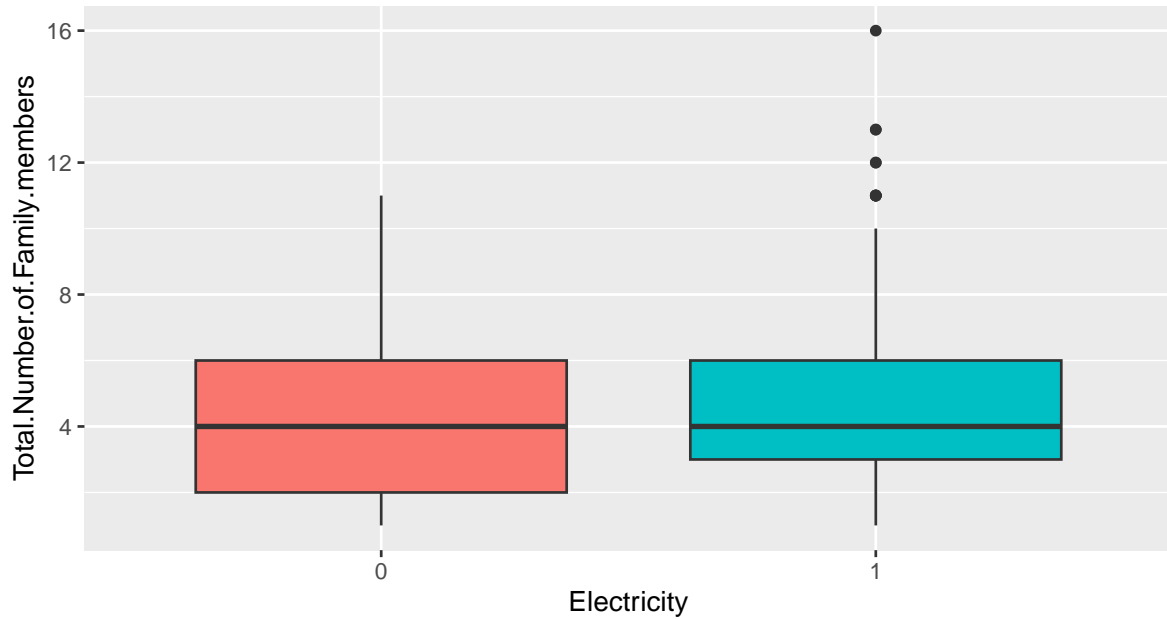


Figure 6: The boxplot of total numbers of family members and electricity

As can be seen from the Figure 4 to Figure 6, there are significant differences between sex of household head and type of household, but the distribution of the total of number of family members is basically the same under different electricity types. Therefore, it is inferred that this variable has no significant impact on the number of family members.

- Second, we are also interested in whether food expenditures differed by type of household.

```
plot_food_type <- ggplot(data = data, aes(x = Type.of.Household,
                                           y = Total.Food.Expenditure,
                                           fill = Type.of.Household)) +
  geom_boxplot() +
  labs(x = "Type.of.Household", y = "Total.Food.Expenditure")+
  theme(legend.position = "none")
print(plot_food_type)
```

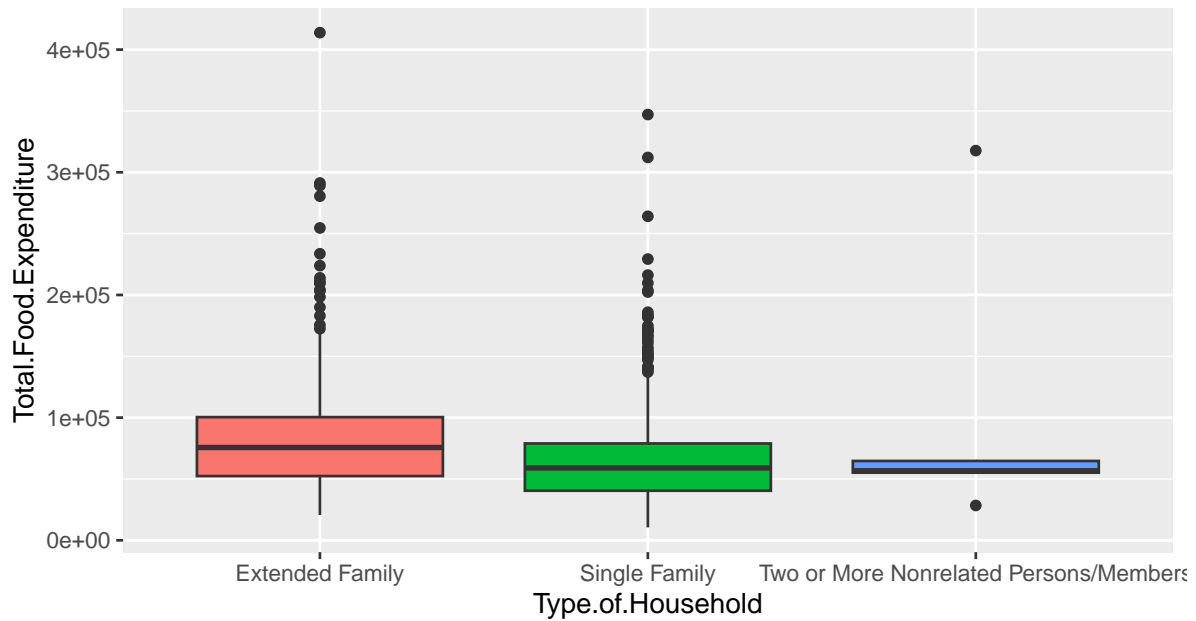


Figure 7: The boxplot of total food expenditure and type of household

Through the Figure 7, it is found that the difference in food expenditures among different house types is smaller than the difference in number of family members.

```
p1 <- ggplot(data = data, aes(x = "", y = Total.Household.Income)) +
  geom_boxplot() +
  labs(x = "Total.Household.Income", y = "")
p2 <- ggplot(data = data, aes(x = "", y = Total.Food.Expenditure)) +
  geom_boxplot() +
  labs(x = "Total.Food.Expenditure", y = "")
p3 <- ggplot(data = data, aes(x = "", y = House.Floor.Area)) +
  geom_boxplot() +
  labs(x = "House.Floor.Area", y = "")
p1+p2+p3+plot_layout(nrow=1)
```

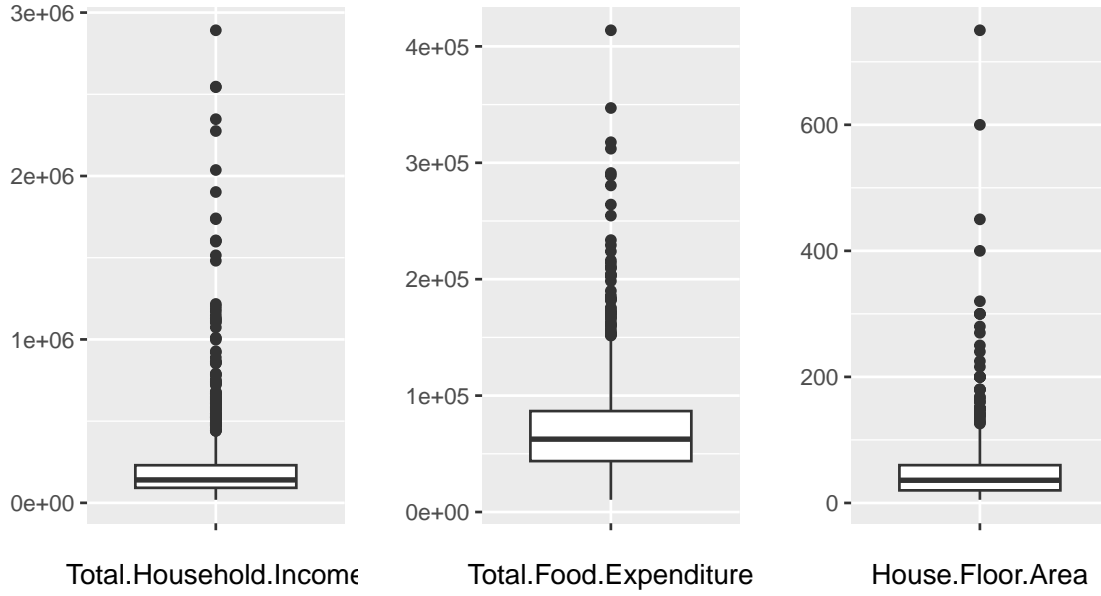


Figure 8: The boxplot of number of family member, food expenditure and house area

Figure 8 is a box plot of number of family member, food expenditure and house area. The data distribution of these variables is analyzed through the box plot.

## Methodology

The response variable here is the number of people living in the household (Total.Number.of.Family.members), which is count data. And combined with the previous visualization analysis, we consider using Poisson regression.

### Model Fitted

We plan to select model by looking at the p-value of variable and using the stepwise model selection based on AIC. Model 1 covers all explanatory variables.

#### 1. P-value

First, fit all variables into a generalized linear regression model.

```
poisson_model_1 <- glm(Total.Number.of.Family.members ~
  ↪ Total.Household.Income +
    Total.Food.Expenditure +
    Household.Head.Sex +
    Household.Head.Age +
    Type.of.Household +
    House.Floor.Area +
    House.Age +
    Number.of.bedrooms +
    Electricity,
  family = "poisson", data = data)

summ(poisson_model_1)
```

|                    |                                |
|--------------------|--------------------------------|
| Observations       | 1249                           |
| Dependent variable | Total.Number.of.Family.members |
| Type               | Generalized linear model       |
| Family             | poisson                        |
| Link               | log                            |

|                                     |         |
|-------------------------------------|---------|
| $\chi^2(10)$                        | 492.62  |
| Pseudo-R <sup>2</sup> (Cragg-Uhler) | 0.33    |
| Pseudo-R <sup>2</sup> (McFadden)    | 0.09    |
| AIC                                 | 4931.87 |
| BIC                                 | 4988.30 |

From the result, we notice that P-value of house floor area is 0.77, indicating that it is not statistically significant. So we decide to eliminate this variable.

```
poisson_model_2 <- glm(Total.Number.of.Family.members ~
  ↪ Total.Household.Income +
    Total.Food.Expenditure +
    Household.Head.Sex +
    Household.Head.Age +
    Type.of.Household +
    House.Age +
    Number.of.bedrooms +
    Electricity,
  family = "poisson", data = data)

summ(poisson_model_2)
```

|   | Est.  | S.E. | z val. | p    |
|---|-------|------|--------|------|
| (Intercept)   | 1.67  | 0.08 | 20.30  | 0.00 |
| Total.Household.Income                                  | -0.00 | 0.00 | -5.62  | 0.00 |
| Total.Food.Expenditure                                  | 0.00  | 0.00 | 12.89  | 0.00 |
| Household.Head.SexMale                                  | 0.24  | 0.04 | 6.47   | 0.00 |
| Household.Head.Age                                      | -0.01 | 0.00 | -5.39  | 0.00 |
| Type.of.HouseholdSingle Family                          | -0.37 | 0.03 | -12.25 | 0.00 |
| Type.of.HouseholdTwo or More Nonrelated Persons/Members | -0.50 | 0.24 | -2.06  | 0.04 |
| House.Floor.Area  | -0.00 | 0.00 | -0.30  | 0.77 |
| House.Age   | -0.00 | 0.00 | -2.08  | 0.04 |
| Number.of.bedrooms                                      | -0.02 | 0.02 | -1.41  | 0.16 |
| Electricity1  | -0.05 | 0.04 | -1.29  | 0.20 |

Standard errors: MLE

|                    |                                |
|--------------------|--------------------------------|
| Observations       | 1249                           |
| Dependent variable | Total.Number.of.Family.members |
| Type               | Generalized linear model       |
| Family             | poisson                        |
| Link               | log                            |

|                                     |         |
|-------------------------------------|---------|
| $\chi^2(9)$                         | 492.53  |
| Pseudo-R <sup>2</sup> (Cragg-Uhler) | 0.33    |
| Pseudo-R <sup>2</sup> (McFadden)    | 0.09    |
| AIC                                 | 4929.96 |
| BIC                                 | 4981.26 |

The next step is remove the insignificant electricity variables.

```
poisson_model_3 <- glm(Total.Number.of.Family.members ~
  ↪ Total.Household.Income +
  Total.Food.Expenditure +
  Household.Head.Sex +
  Household.Head.Age +
  Type.of.Household +
  House.Age +
  Number.of.bedrooms,
  family = "poisson", data = data)
summ(poisson_model_3)
```

|   | Est.  | S.E. | z val. | p    |
|---|-------|------|--------|------|
| (Intercept)   | 1.67  | 0.08 | 20.30  | 0.00 |
| Total.Household.Income                                  | -0.00 | 0.00 | -5.70  | 0.00 |
| Total.Food.Expenditure                                  | 0.00  | 0.00 | 12.88  | 0.00 |
| Household.Head.SexMale                                  | 0.24  | 0.04 | 6.47   | 0.00 |
| Household.Head.Age                                      | -0.01 | 0.00 | -5.38  | 0.00 |
| Type.of.HouseholdSingle Family                          | -0.37 | 0.03 | -12.25 | 0.00 |
| Type.of.HouseholdTwo or More Nonrelated Persons/Members | -0.50 | 0.24 | -2.06  | 0.04 |
| House.Age   | -0.00 | 0.00 | -2.15  | 0.03 |
| Number.of.bedrooms                                      | -0.03 | 0.02 | -1.58  | 0.12 |
| Electricity1  | -0.05 | 0.04 | -1.30  | 0.19 |

Standard errors: MLE

|                    |                                |
|--------------------|--------------------------------|
| Observations       | 1249                           |
| Dependent variable | Total.Number.of.Family.members |
| Type               | Generalized linear model       |
| Family             | poisson                        |
| Link               | log                            |

|                                     |         |
|-------------------------------------|---------|
| $\chi^2(8)$                         | 490.85  |
| Pseudo-R <sup>2</sup> (Cragg-Uhler) | 0.33    |
| Pseudo-R <sup>2</sup> (McFadden)    | 0.09    |
| AIC                                 | 4929.64 |
| BIC                                 | 4975.81 |

|   | Est.  | S.E. | z val. | p    |
|---|-------|------|--------|------|
| (Intercept)   | 1.64  | 0.08 | 20.95  | 0.00 |
| Total.Household.Income                                  | -0.00 | 0.00 | -5.74  | 0.00 |
| Total.Food.Expenditure                                  | 0.00  | 0.00 | 12.81  | 0.00 |
| Household.Head.SexMale                                  | 0.24  | 0.04 | 6.54   | 0.00 |
| Household.Head.Age                                      | -0.01 | 0.00 | -5.37  | 0.00 |
| Type.of.HouseholdSingle Family                          | -0.37 | 0.03 | -12.28 | 0.00 |
| Type.of.HouseholdTwo or More Nonrelated Persons/Members | -0.50 | 0.24 | -2.06  | 0.04 |
| House.Age   | -0.00 | 0.00 | -2.35  | 0.02 |
| Number.of.bedrooms                                      | -0.03 | 0.02 | -1.81  | 0.07 |

Standard errors: MLE

The p-value of number of bedrooms is 0.07. We try to delete this variable to check the model result.

```
poisson_model_4 <- glm(Total.Number.of.Family.members ~
  ↪ Total.Household.Income +
    Total.Food.Expenditure +
    Household.Head.Sex +
    Household.Head.Age +
    Type.of.Household +
    House.Age,
  family = "poisson", data = data)
summ(poisson_model_4)
```

|                    |                                |
|--------------------|--------------------------------|
| Observations       | 1249                           |
| Dependent variable | Total.Number.of.Family.members |
| Type               | Generalized linear model       |
| Family             | poisson                        |
| Link               | log                            |

|                                     |         |
|-------------------------------------|---------|
| $\chi^2(7)$                         | 487.57  |
| Pseudo-R <sup>2</sup> (Cragg-Uhler) | 0.33    |
| Pseudo-R <sup>2</sup> (McFadden)    | 0.09    |
| AIC                                 | 4930.92 |
| BIC                                 | 4971.96 |

|   | Est.  | S.E. | z val. | p    |
|---|-------|------|--------|------|
| (Intercept)   | 1.62  | 0.08 | 20.90  | 0.00 |
| Total.Household.Income                                  | -0.00 | 0.00 | -6.11  | 0.00 |
| Total.Food.Expenditure                                  | 0.00  | 0.00 | 12.64  | 0.00 |
| Household.Head.SexMale                                  | 0.24  | 0.04 | 6.51   | 0.00 |
| Household.Head.Age                                      | -0.01 | 0.00 | -5.70  | 0.00 |
| Type.of.HouseholdSingle Family                          | -0.37 | 0.03 | -12.19 | 0.00 |
| Type.of.HouseholdTwo or More Nonrelated Persons/Members | -0.51 | 0.24 | -2.07  | 0.04 |
| House.Age   | -0.00 | 0.00 | -2.53  | 0.01 |

Standard errors: MLE

```
#plot(poisson_model_4)
```

All variables are significant in Model 4.

## 2. Step-wise AIC



```
library(MASS)
poisson_model_step <- stepAIC(poisson_model_1, direction = "both")
```

Start: AIC=4931.87

Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +  
Household.Head.Sex + Household.Head.Age + Type.of.Household +  
House.Floor.Area + House.Age + Number.of.bedrooms + Electricity

|                          | Df | Deviance | AIC    |
|--------------------------|----|----------|--------|
| - House.Floor.Area       | 1  | 881.10   | 4930.0 |
| - Electricity            | 1  | 882.67   | 4931.5 |
| - Number.of.bedrooms     | 1  | 883.00   | 4931.9 |
| <none>                   |    | 881.01   | 4931.9 |
| - House.Age              | 1  | 885.41   | 4934.3 |
| - Household.Head.Age     | 1  | 910.02   | 4958.9 |
| - Total.Household.Income | 1  | 916.63   | 4965.5 |
| - Household.Head.Sex     | 1  | 924.80   | 4973.7 |
| - Type.of.Household      | 2  | 1028.11  | 5075.0 |
| - Total.Food.Expenditure | 1  | 1033.71  | 5082.6 |

Step: AIC=4929.96

Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +  
Household.Head.Sex + Household.Head.Age + Type.of.Household +  
House.Age + Number.of.bedrooms + Electricity

|                          | Df | Deviance | AIC    |
|--------------------------|----|----------|--------|
| - Electricity            | 1  | 882.78   | 4929.6 |
| <none>                   |    | 881.10   | 4930.0 |
| - Number.of.bedrooms     | 1  | 883.59   | 4930.4 |
| + House.Floor.Area       | 1  | 881.01   | 4931.9 |
| - House.Age              | 1  | 885.80   | 4932.7 |
| - Household.Head.Age     | 1  | 910.07   | 4956.9 |
| - Total.Household.Income | 1  | 917.76   | 4964.6 |
| - Household.Head.Sex     | 1  | 924.93   | 4971.8 |
| - Type.of.Household      | 2  | 1028.11  | 5073.0 |
| - Total.Food.Expenditure | 1  | 1033.71  | 5080.6 |

Step: AIC=4929.64

Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +  
Household.Head.Sex + Household.Head.Age + Type.of.Household +  
House.Age + Number.of.bedrooms

|                          | Df | Deviance | AIC    |
|--------------------------|----|----------|--------|
| <none>                   |    | 882.78   | 4929.6 |
| + Electricity            | 1  | 881.10   | 4930.0 |
| - Number.of.bedrooms     | 1  | 886.06   | 4930.9 |
| + House.Floor.Area       | 1  | 882.67   | 4931.5 |
| - House.Age              | 1  | 888.38   | 4933.2 |
| - Household.Head.Age     | 1  | 911.64   | 4956.5 |
| - Total.Household.Income | 1  | 919.96   | 4964.8 |
| - Household.Head.Sex     | 1  | 927.56   | 4972.4 |
| - Type.of.Household      | 2  | 1030.52  | 5073.4 |
| - Total.Food.Expenditure | 1  | 1033.99  | 5078.8 |

```
summ(poisson_model_step)
```

|                    |                                |
|--------------------|--------------------------------|
| Observations       | 1249                           |
| Dependent variable | Total.Number.of.Family.members |
| Type               | Generalized linear model       |
| Family             | poisson                        |
| Link               | log                            |

|                                     |         |
|-------------------------------------|---------|
| $\chi^2(8)$                         | 490.85  |
| Pseudo-R <sup>2</sup> (Cragg-Uhler) | 0.33    |
| Pseudo-R <sup>2</sup> (McFadden)    | 0.09    |
| AIC                                 | 4929.64 |
| BIC                                 | 4975.81 |

|   | Est.  | S.E. | z val. | p    |
|---|-------|------|--------|------|
| (Intercept)   | 1.64  | 0.08 | 20.95  | 0.00 |
| Total.Household.Income                                  | -0.00 | 0.00 | -5.74  | 0.00 |
| Total.Food.Expenditure                                  | 0.00  | 0.00 | 12.81  | 0.00 |
| Household.Head.SexMale                                  | 0.24  | 0.04 | 6.54   | 0.00 |
| Household.Head.Age                                      | -0.01 | 0.00 | -5.37  | 0.00 |
| Type.of.HouseholdSingle Family                          | -0.37 | 0.03 | -12.28 | 0.00 |
| Type.of.HouseholdTwo or More Nonrelated Persons/Members | -0.50 | 0.24 | -2.06  | 0.04 |
| House.Age   | -0.00 | 0.00 | -2.35  | 0.02 |
| Number.of.bedrooms                                      | -0.03 | 0.02 | -1.81  | 0.07 |

Standard errors: MLE

The selected model is the same as Model 3.

## Model Selection

### 1. AIC value

```
# AIC and BIC for each module
models <- list(poisson_model_1, poisson_model_2, poisson_model_3,
  ↪ poisson_model_4)
model_names <- c("Model 1", "Model 2", "Model 3", "Model 4")

aic_bic_df <- tibble(
  Model = model_names,
  AIC = sapply(models, AIC),
  BIC = sapply(models, BIC)
)
aic_bic_df
```

```
# A tibble: 4 x 3
  Model      AIC    BIC
  <chr>    <dbl> <dbl>
1 Model 1 4932. 4988.
2 Model 2 4930. 4981.
3 Model 3 4930. 4976.
4 Model 4 4931. 4972.
```

### 2. cross validation

```
library(caret)
# set cross validation
train_control <- trainControl(method = "cv", number = 10)

model_with_bed <- train(Total.Number.of.Family.members ~
  ↪ Total.Household.Income + Total.Food.Expenditure + Household.Head.Sex
  ↪ + Household.Head.Age + Type.of.Household + House.Age +
  ↪ Number.of.bedrooms,
  data = data,
  method = "glm",
  family = poisson(link = "log"),
  trControl = train_control)
```

```

model_without_bed<- train(Total.Number.of.Family.members ~
  ↪ Total.Household.Income + Total.Food.Expenditure + Household.Head.Sex
  ↪ + Household.Head.Age + Type.of.Household + House.Age ,
      data = data,
      method = "glm",
      family = poisson(link = "log"),
      trControl = train_control)

```

```

residual_deviance_with_bed <- sum(resid(model_with_bed, type =
  ↪ "deviance")^2)
residual_deviance_without_bed<- sum(resid(model_without_bed, type =
  ↪ "deviance")^2)

print(residual_deviance_with_bed)

```

```
[1] 882.7785
```

```
print(residual_deviance_without_bed)
```

```
[1] 886.0647
```

From both AIC and cross validation result, the model3 is the best model. It has lower AIC and lower residual\_deviance.

## Assumption Check

1. The assumption that the mean equals the variance

Testing for overdispersion. If the dispersion index is significantly less than 1, this indicates that the model is not overdispersed.

```

dispersion_index <- sum(residuals(poisson_model_3, type = "pearson")^2)
  ↪ / df.residual(poisson_model_3)
print(dispersion_index)

```

```
[1] 0.7079346
```

The dispersion index is below 1.

2. Assumption of linear relationships Residual Analysis: Plot the residuals of the model. A scatter plot of the residuals against the fitted values should show randomly distributed points.

```
# Residuals plot of Model 3
residuals <- residuals(poisson_model_3)
plot(residuals ~ fitted.values(poisson_model_3),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residual Plot for model 3",
     cex=1,
     pch=20)+
abline(h=0, lty=2, col="red",lwd=3)
```

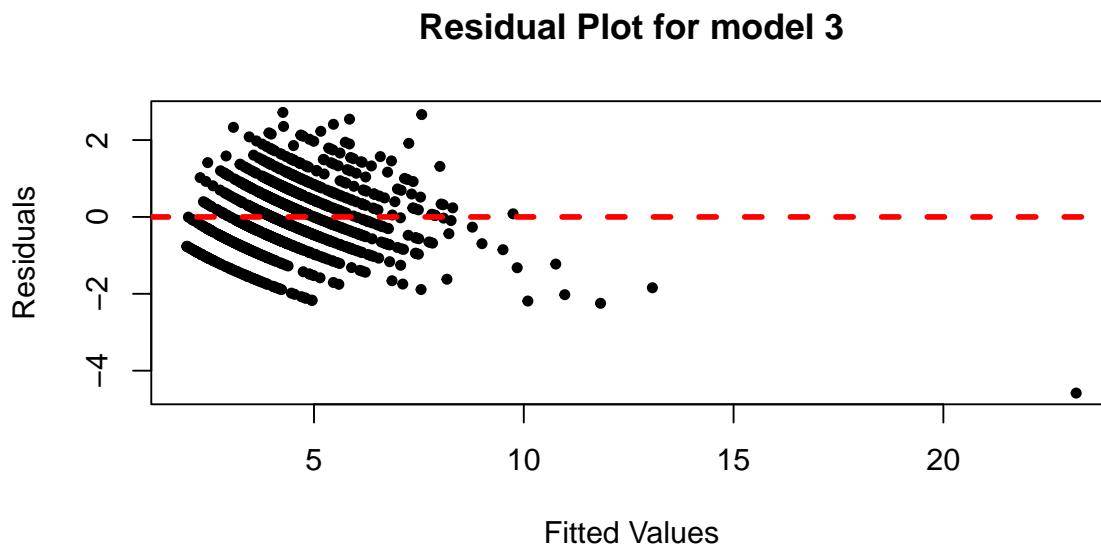


Figure 9: Residuals plot of Model 3

`integer(0)`

```
# Residuals for different variables.
par(mfrow=c(3, 2))
```

```

# Residual Plot for total household income
poisson_THI <- glm(Total.Number.of.Family.members ~
  ↪ Total.Household.Income,family = "poisson", data = data)
plot(residuals(poisson_THI) ~ fitted.values(poisson_THI),
  xlab = "Fitted Values", ylab = "Residuals",
  main = "Residual Plot for Total Household Income",
  cex=1, pch=20)
abline(h=0, lty=2, col="red", lwd=3)

# Residual Plot for total food expenditure
poisson_TFE <- glm(Total.Number.of.Family.members ~
  ↪ Total.Food.Expenditure,family = "poisson", data = data)
plot(residuals(poisson_TFE) ~ fitted.values(poisson_TFE),
  xlab = "Fitted Values", ylab = "Residuals",
  main = "Residual Plot for Total Food Expenditure",
  cex=1, pch=20)
abline(h=0, lty=2, col="red", lwd=3)

# Residual Plot for household head sex
poisson_HHS <- glm(Total.Number.of.Family.members ~
  ↪ Household.Head.Sex,family = "poisson", data = data)
plot(residuals(poisson_HHS) ~ fitted.values(poisson_HHS),
  xlab = "Fitted Values", ylab = "Residuals",
  main = "Residual Plot for Household Head Sex",
  cex=1, pch=20)
abline(h=0, lty=2, col="red", lwd=3)

# Residual Plot for household head age
poisson_HHA <- glm(Total.Number.of.Family.members ~
  ↪ Household.Head.Age,family = "poisson", data = data)
plot(residuals(poisson_HHA) ~ fitted.values(poisson_HHA),
  xlab = "Fitted Values", ylab = "Residuals",
  main = "Residual Plot for Household Head Age",
  cex=1, pch=20)
abline(h=0, lty=2, col="red", lwd=3)

# Residual Plot for type of household
poisson_TH <- glm(Total.Number.of.Family.members ~
  ↪ Type.of.Household,family = "poisson", data = data)
plot(residuals(poisson_TH) ~ fitted.values(poisson_TH),

```

```

      xlab = "Fitted Values", ylab = "Residuals",
      main = "Residual Plot for Type of Household",
      cex=1, pch=20)
abline(h=0, lty=2, col="red", lwd=3)

# Residual Plot for house age
poisson_HA <- glm(Total.Number.of.Family.members ~ House.Age,family =
  ↪ "poisson", data = data)
plot(residuals(poisson_HA) ~ fitted.values(poisson_HA),
      xlab = "Fitted Values", ylab = "Residuals",
      main = "Residual Plot for House Age",
      cex=1, pch=20)
abline(h=0, lty=2, col="red", lwd=3)

```

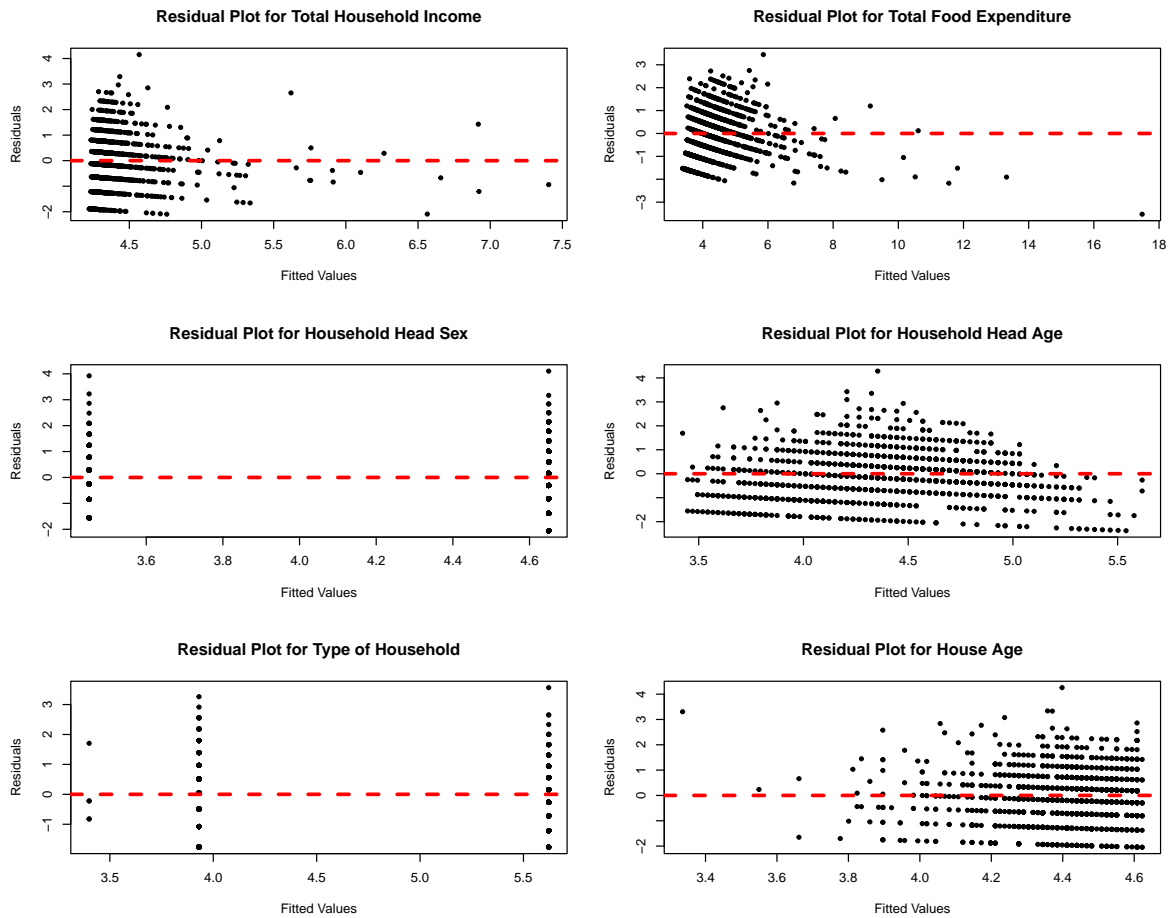


Figure 10: Residuals for different variables

```
par(mfrow=c(1, 1))
```

3. The residuals are normally distributed. Check this by Q-Q plot.

```
qqnorm(residuals)
qqline(residuals, col="red",lwd=2)
```



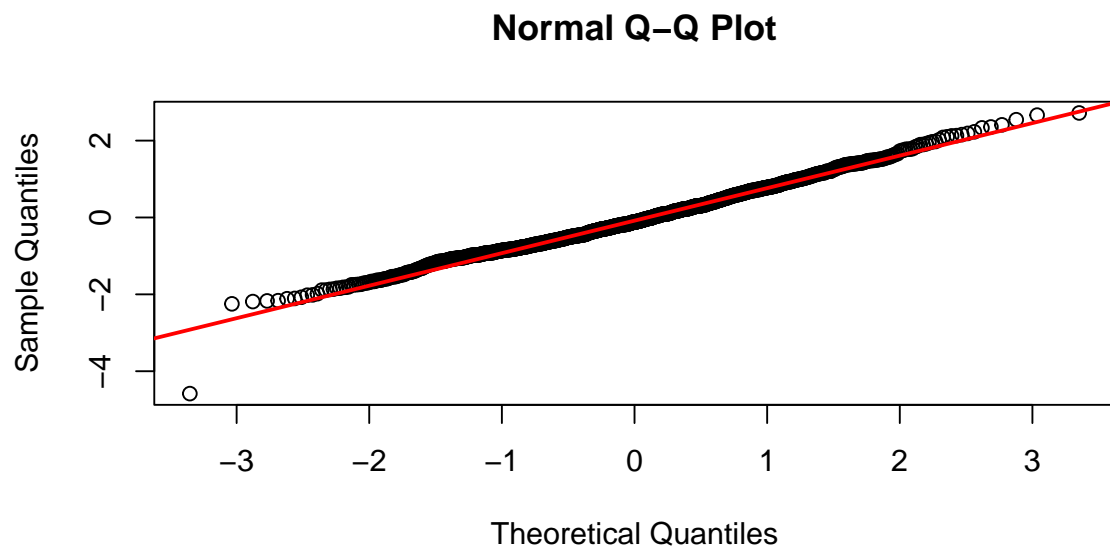


Figure 11: Q-Q plot of Model 3