Explore factors affecting the number of family members

Group04: Yiheng Yang, Yuanqing Zhang, Jie Li, Bing Gao, Ningxuan Zhang, Chao Z

1 Introduction

In order to find which household variables that influence the number of people living in a household. We use the datasets come from the Family Income and Expenditure Survey which is conducted in Philippines every three years. The response variable is the count variable, the explanatory variable Household.Head.Sex, Type.of.Household and Electricity are categorical variables, and the rest are numerical variables. All the data were based on Soccsksargen, and the variables are shown in Table 1.

Table 1: Response variable and explanatory variables

variable type	variable name	variable description
response variable	Total.Number.of.Family.members	Number of people living in the house
explanatory variables	Household.Head.Sex	Head of the households sex
	Type.of.Household	Relationship between people
	Electricity	If the house have electricity
	Total.Household.Income	Annual household income
	Total.Food.Expenditure	Annual expenditure on food
	Household.Head.Age	Head of the household age
	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

2 Data Processing

2.1 Load the data

```
data=read.csv("dataset04.csv")
```

2.2 Get packages

```
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(tidyverse)
library(ggplot2)
library(MASS)
library(knitr)
library(tidyr)
library(gt)
library(janitor)
library(skimr)
library(kableExtra)
library(gridExtra)
```

2.3 Convert some categorical variables to factors

```
# Convert some categorical variables to factors for better creating plots
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)
levels(data$Electricity)=c("No","Yes")
data$Number.of.bedrooms=as.factor(data$Number.of.bedrooms)
levels(data$Number.of.bedrooms)=c("0","1","2","3","4","5","6","7")
```

2.4 Normalize the data

Due to the different data dimensions and value ranges among variables, in order to better apply in and explain the model, we normalized the data of numerical explanatory variables using maximum and minimum scaling and reduced the data to [0,1].

```
# Normalize the data for better fitting
data2=read.csv("dataset04.csv")
data2$Household.Head.Sex=as.factor(data2$Household.Head.Sex)
data2$Type.of.Household=as.factor(data2$Type.of.Household)
data2$Electricity=as.factor(data2$Electricity)
levels(data2$Electricity)=c("No","Yes")
data.norm <-apply(data2[,c("Total.Household.Income",
                              "Total.Food.Expenditure",
                              "Household.Head.Age",
                              "House.Floor.Area",
                              "House.Age",
                              "Number.of.bedrooms")], 2, function(x)
                                (x-min(x))/(max(x)-min(x)) )

data.norm <- cbind(data.norm,data2[,c("Household.Head.Sex",
                                     "Type.of.Household",
                                     "Electricity")])
data.norm <-cbind(data2[,c("Total.Number.of.Family.members")],data.norm)
colnames(data.norm)[colnames(data.norm)=="data2[, c(\"Total.Number.of.Family.members\")]"
                    ] <- "Total.Number.of.Family.members"
```

3 Exploratory Data Analysis

3.1 Summary of response variable

```
# Create a table to summarize the characteristics of the response variables
data%>%summarize('Mean' = mean(Total.Number.of.Family.members),
'Median' = median(Total.Number.of.Family.members),
'St.Dev' = sd(Total.Number.of.Family.members),
'Variance'=var(Total.Number.of.Family.members),
'Min' = min(Total.Number.of.Family.members),
'Max' = max(Total.Number.of.Family.members),
'IQR' = quantile(Total.Number.of.Family.members,0.75)
-quantile(Total.Number.of.Family.members,0.25),
'Sample_size' = n())%>%
  gt()%>%
  fmt_number(decimals=2)%>%
  cols_label(
Mean = html("Mean"),
Median = html("Median"),
St.Dev = html("Std. Dev"),
Variance=html("Variance"),
Min = html("Minimum"),
Max = html("Maximum"),
IQR = html("Interquartile Range"),
Sample_size = html("Sample Size"))
```

Mean	Median	Std. Dev	Variance	Minimum	Maximum	Interquartile Range	Sample Size
4.53	4.00	2.22	4.91	1.00	19.00	3.00	2, 122.00

We can see from this numerical summary, the mean of number of family members is 4.53 and the variance is 4.91. If variance is bigger than mean, we can determine that we have overdispersion. We will investigate this phenomenon later.

3.2 Summary of categorical explanatory variables

```
# Select the categorical explanatory variables
data_categorical=data%>%
  dplyr::select("Household.Head.Sex","Type.of.Household","Electricity")
```

```
# Create a table to summarize the characteristics of the categorical explanatory variables
summary_table_categorical <-summary(data_categorical)
summary_table_categorical[is.na(summary_table_categorical)] <- ""
kable(summary_table_categorical,na.strings = "")
```

Household.Head.Sex	Type.of.Household	Electricity
Female: 362	Extended Family : 585	No : 363
Male :1760	Single Family :1531	Yes:1759
	Two or More Nonrelated Persons/Members: 6	

The numerical summary shows that male owners, single families and households with electricity account for a major proportion.

3.3 Summary of numerical explanatory variables

```
# Create a table to summarize the characteristics of the numerical explanatory variables
data_numerical=data[,c(1,3,5,7,8,9,10)]
data_numerical$Number.of.bedrooms=as.numeric(as.character(data_numerical$Number.of.bedroom
my_skim <- skim_with(numeric = sfl(hist = NULL),
                     base = sfl(n = length))
my_skim(data_numerical) %>%
  transmute(Variable=skim_variable, Sample_size = n, Mean=numeric.mean,
            St.Dev=numeric.sd, Min=numeric.p0, Median=numeric.p50,
            Max=numeric.p100, IQR = numeric.p75-numeric.p50) %>%
  kable(format.args = list(big.mark = ","), digits=2) %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

Variable	Sample_size	Mean	St.Dev	Min	Median	Max	IQR
Total.Household.Income	2,122	182,984.80	228,231.07	15,204	120,362.0	3,168,662	74,314.00
Total.Food.Expenditure	2,122	71,738.09	44,938.17	7,783	63,305.5	729,606	24,496.75
Household.Head.Age	2,122	49.28	14.16	9	48.0	99	11.00
Total.Number.of.Family.members	2,122	4.53	2.22	1	4.0	19	2.00
House.Floor.Area	2,122	35.74	34.67	5	26.5	450	13.50
House.Age	2,122	16.30	11.09	0	14.0	75	7.00
Number.of.bedrooms	2,122	1.77	1.00	0	2.0	7	0.00

3.4 Graphical summaries

3.4.1 Graphical summaries of response variable

As we want to plot a histogram with x axis to be number of family members, so we need to change this variable to be a factor.

```
# Convert the column "Total.Number.of.Family.members" to factor type
data$Total.Number.of.Family.members=as.factor(data$Total.Number.of.Family.members)
```

```
ggplot(data=data,aes(x=Total.Number.of.Family.members))+geom_bar()
# Plot a histogram to show the distribution of response variable
```

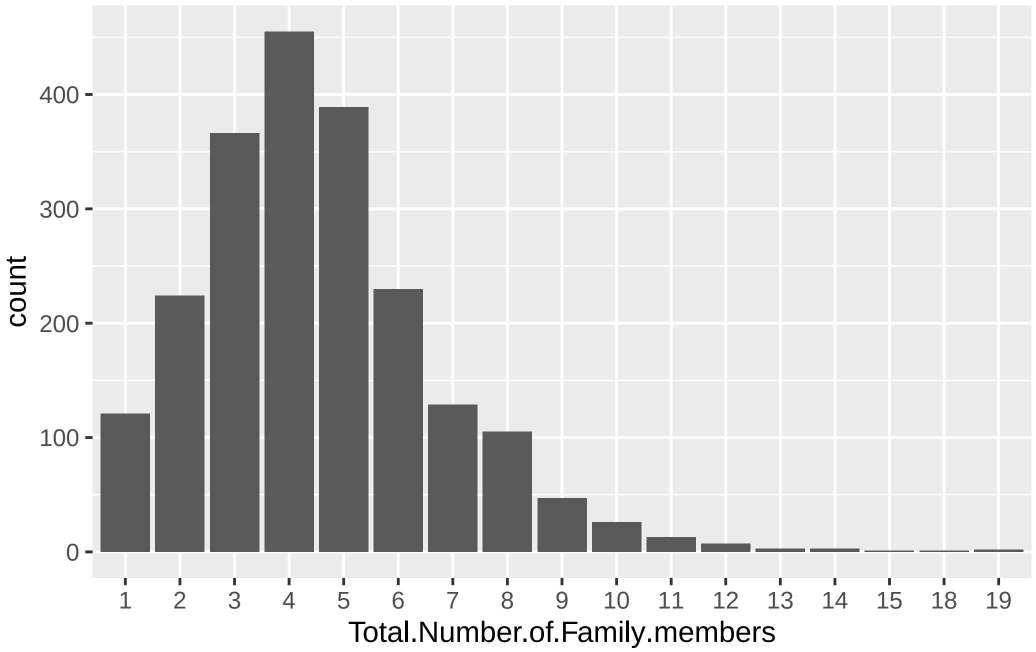


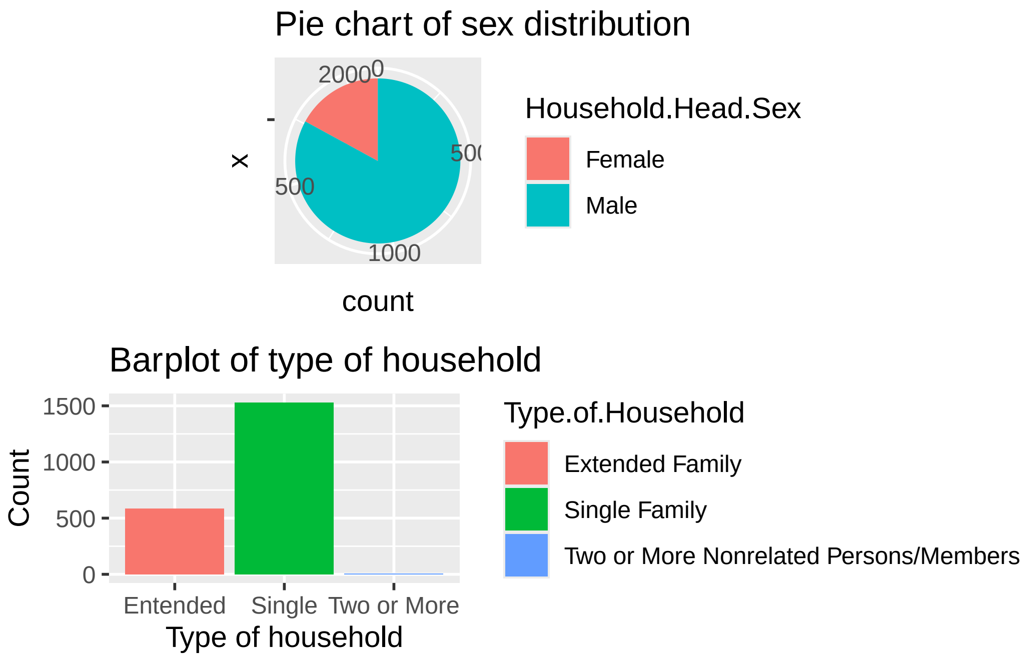
Figure 1: histogram of response variable

The Figure 1 shows that household with four family members accounts for the largest proportion. Most of the data is consisted of families with three to five family members.

3.4.2 Graphical summaries of explanatory variables

```
p1=ggplot(data=data,aes(y=Total.Household.Income))+geom_boxplot()+labs(y="Total household 
p2=ggplot(data=data,(aes(y=Total.Food.Expenditure)))+geom_boxplot()+labs(y="Total food exp
p3=ggplot(data=data,aes(x="",fill=Household.Head.Sex))+geom_bar(width=1)+coord_polar(theta
p4=ggplot(data=data,aes(y=Household.Head.Age))+geom_boxplot()+labs(y="Household head age",
p5=ggplot(data=data,aes(x=Type.of.Household))+geom_bar(aes(fill=Type.of.Household))+scale_
p6=ggplot(data=data,aes(y=House.Floor.Area))+geom_boxplot()+labs(y="House floor area",titl
p7=ggplot(data=data,aes(y=House.Age))+geom_boxplot()+labs(y="House age",title="Boxplot of 
p8=ggplot(data=data,aes(x=Number.of.bedrooms))+geom_bar(aes(fill=Number.of.bedrooms))+labs
p9=ggplot(data=data,aes(x=Electricity))+geom_bar(aes(fill=Electricity))+labs(y="Count",tit
```

```
# Arrange the plots in a grid layout for display
#| label: fig-piechart_sex
#| fig-cap: Pie chart of Household.Head.Sex
#| fig-align: center
#| message: false
grid.arrange(p3,p5,ncol=1)
```



```
grid.arrange(p8,p9,ncol=2)
```

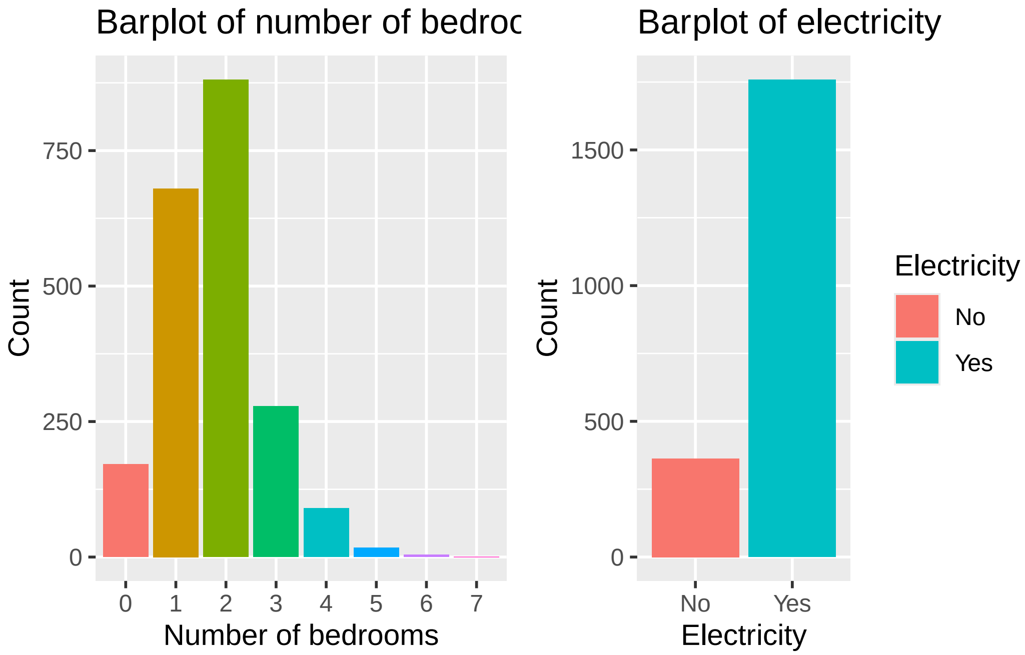


Figure 2: Barplots of some explanatory variables

```
grid.arrange(p1,p2,p4,p6,p7,ncol=3)
```

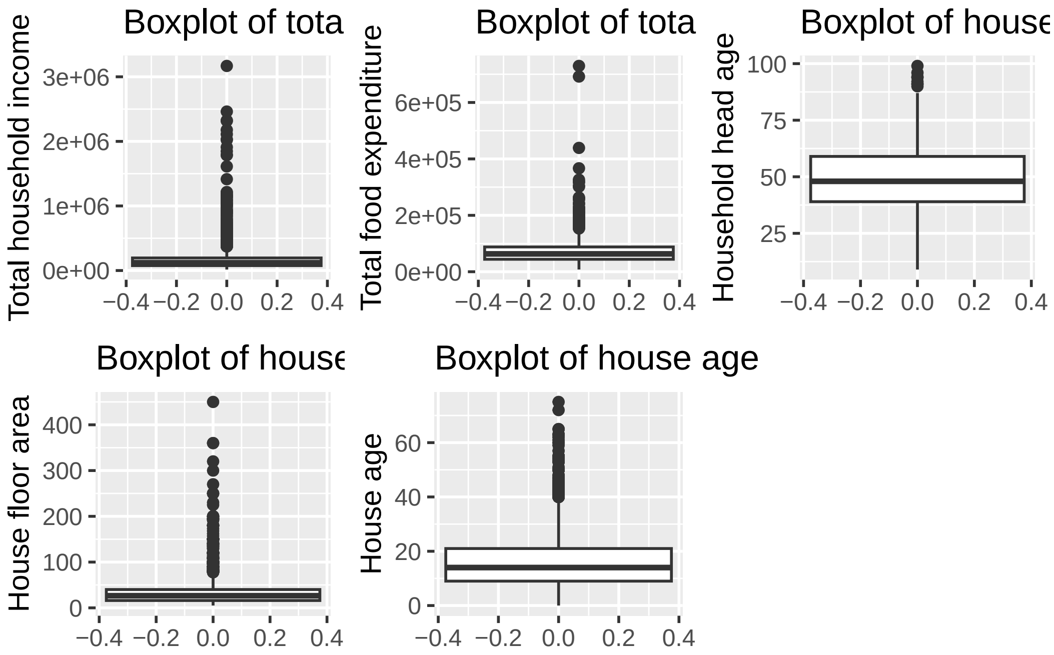


Figure 3: Boxplot of some explanatory variables

3.4.3 Relationship between explanatory variables and response variable

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=Total.Household.Income,
                     fill=Total.Number.of.Family.members))+geom_boxplot()+
                     theme(legend.position = "none")+
                     labs(x="Number of Family Members",
                                       y="Total Household Income")
```

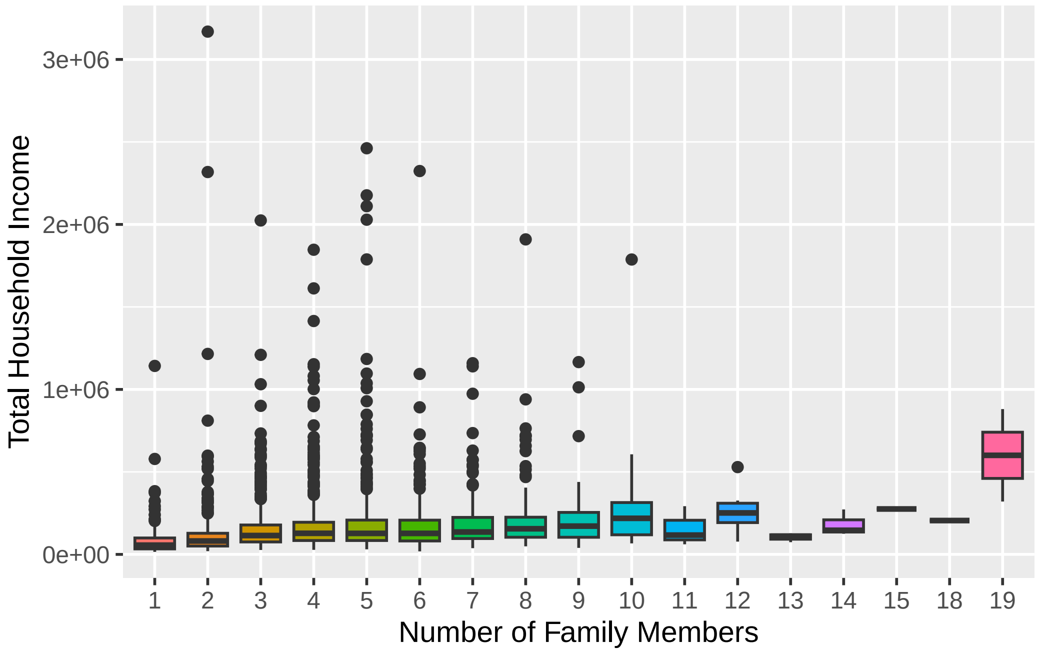


Figure 4: Income of families with different number of family members

We can see from the Figure 4 that the median of household income increase as number of family members increase.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=Total.Food.Expenditure,
                     fill=Total.Number.of.Family.members))+geom_boxplot()+
                     theme(legend.position = "none")+
                  labs(x="Number of Family Members",y="Total food expenditure")
```

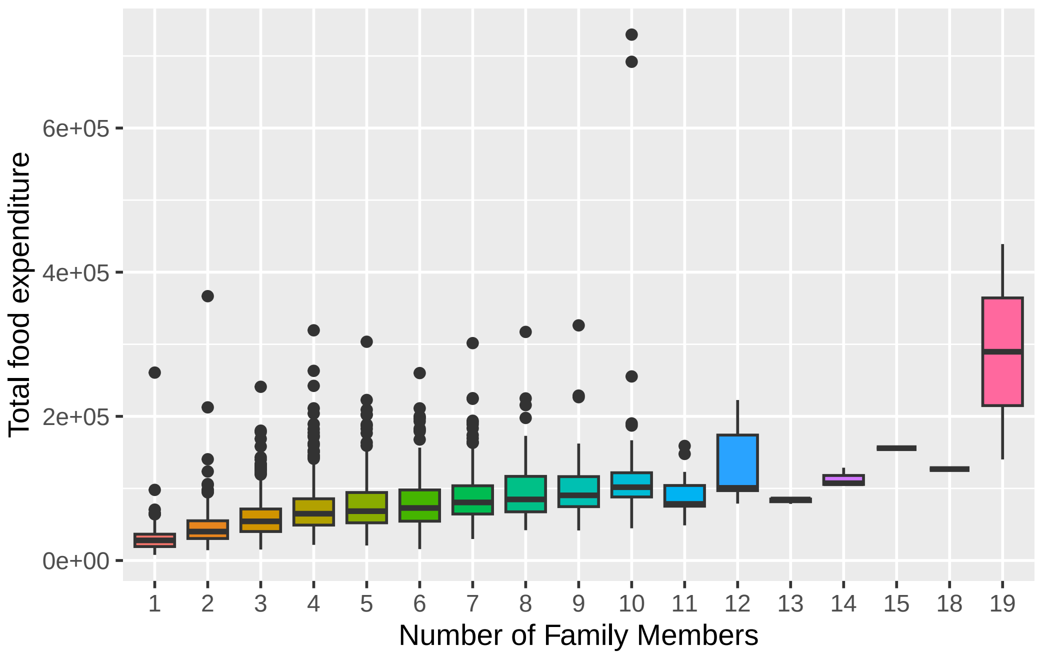


Figure 5: Food expenditure of families with different number of family members

The Figure 5 indicates that median increase significantly as the number of family members increase. Household with 19 members have the largest variance in food expenditure.

```
frequency_sex <- data%>%
  tabyl(Household.Head.Sex,Total.Number.of.Family.members)%>%
  adorn_percentages()%>%
  adorn_pct_formatting()%>%
  adorn_ns()
kable(frequency_sex)
```

Household.Head.Sex	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	18	19
Female	12.4% (45)	19.1% (69)	25.4% (92)	16.0% (58)	11.0% (40)	7.7% (28)	2.5% (9)	3.3% (12)	1.7% (6)	0.3% (1)	0.0% (0)	0.0% (0)	0.3% (1)	0.0% (0)	0.0% (0)	0.0% (0)	0.3% (1)
Male	4.3% (76)	8.8% (155)	15.6% (274)	22.6% (397)	19.8% (349)	11.5% (202)	6.8% (120)	5.3% (93)	2.3% (41)	1.4% (25)	0.7% (13)	0.4% (7)	0.1% (2)	0.2% (3)	0.1% (1)	0.1% (1)	0.1% (1)

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Household.Head.Sex))+geom_bar(aes(y=..prop..,
                    fill=Household.Head.Sex),position="dodge")+
                    labs(x="Number of Family Members",y="Proportion")
```

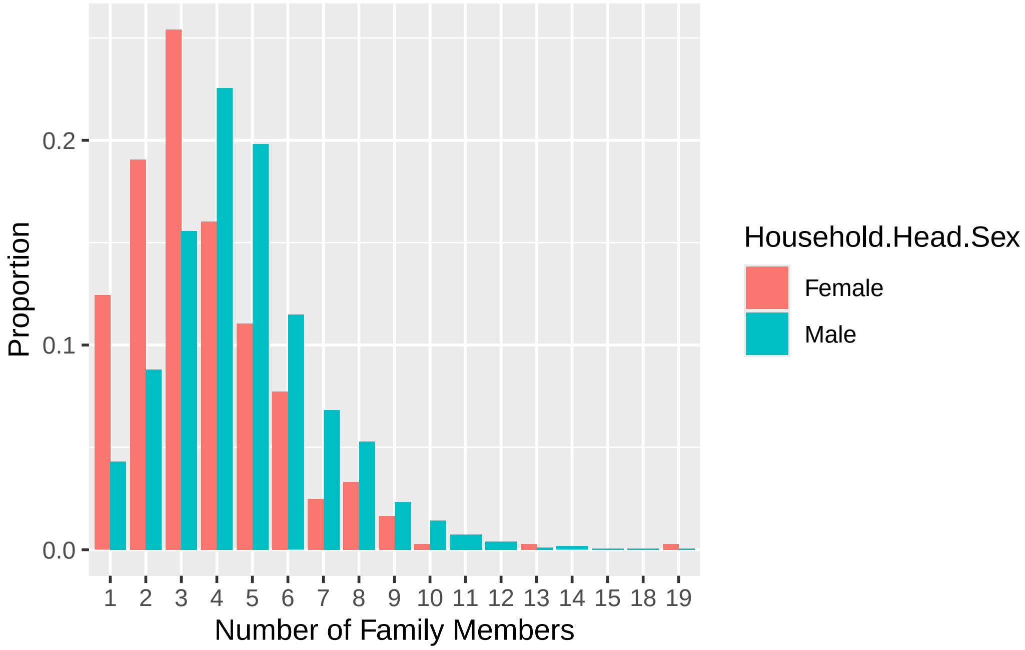


Figure 6: Head sex proportion for different size of households

We can see from the Figure 6, for those small sized households, the proportion is much higher for females than for males. However, this situation does not exist for those household with four or more family members.

```
ggplot(data=data,aes(x=Household.Head.Sex,
                     y=as.numeric(as.character(Total.Number.of.Family.members))))+
  geom_boxplot(aes(fill=Household.Head.Sex))+labs(x="Household head sex",
                                                  y="Number of family members")
```

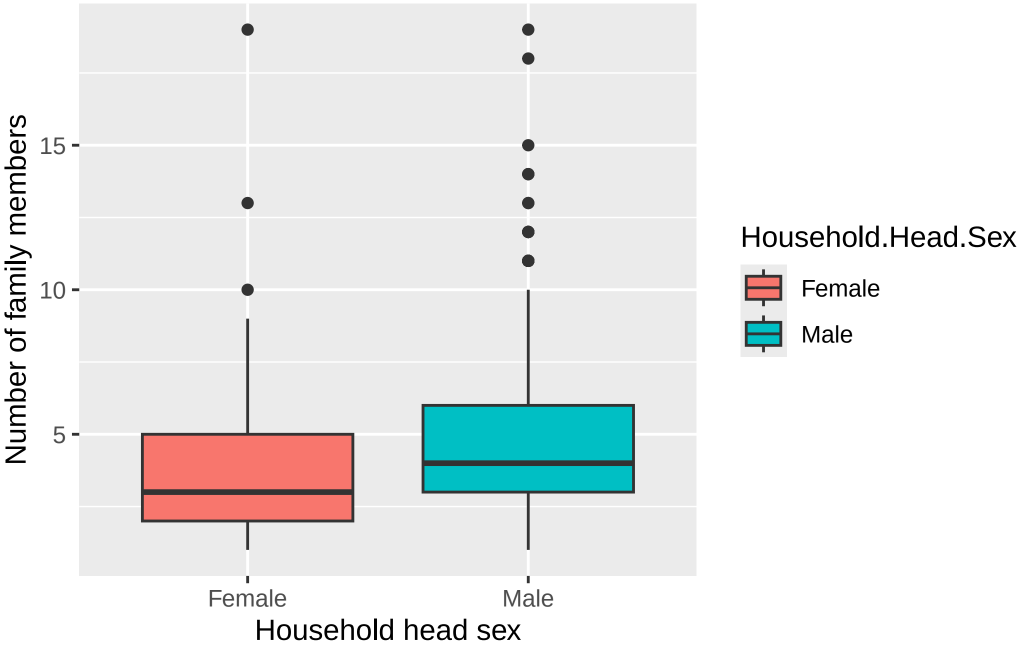


Figure 7: Number of family members by sex

We can conclude from the Figure 7 that households tend to have more family members if their owner is male.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=Household.Head.Age,fill=Total.Numb
```

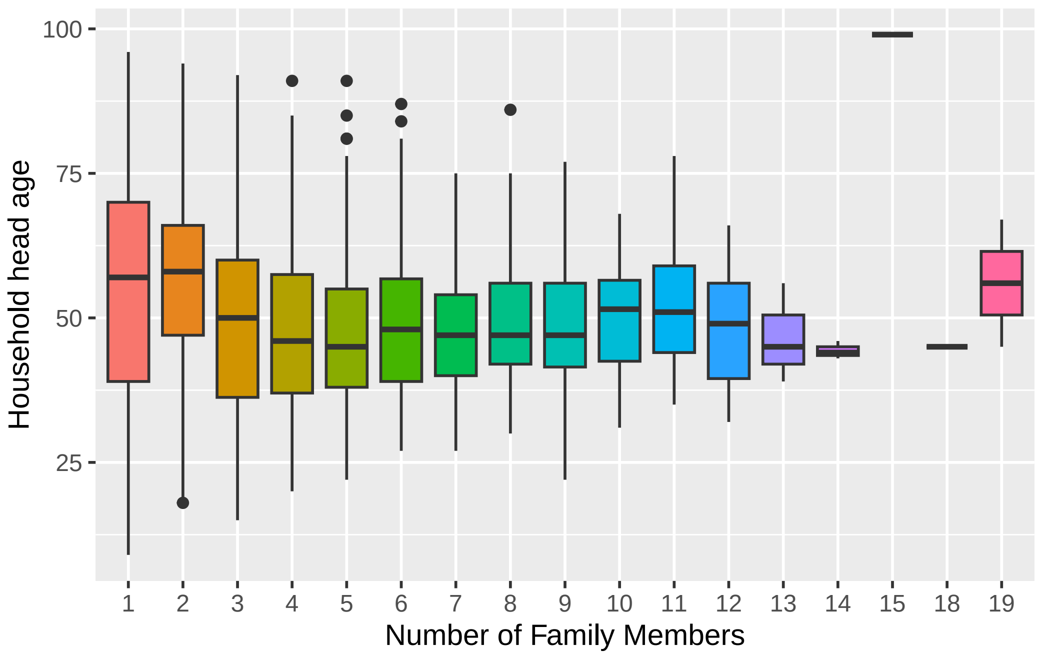


Figure 8: Head age for different size households

As shown in Figure 8, for different size of households, the median of household head age remain at a constant level around 50.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Type.of.Household))+
  geom_bar(aes(y=..prop..,fill=Type.of.Household))+
  labs(x="Number of Family Members",y="Proportion")
```

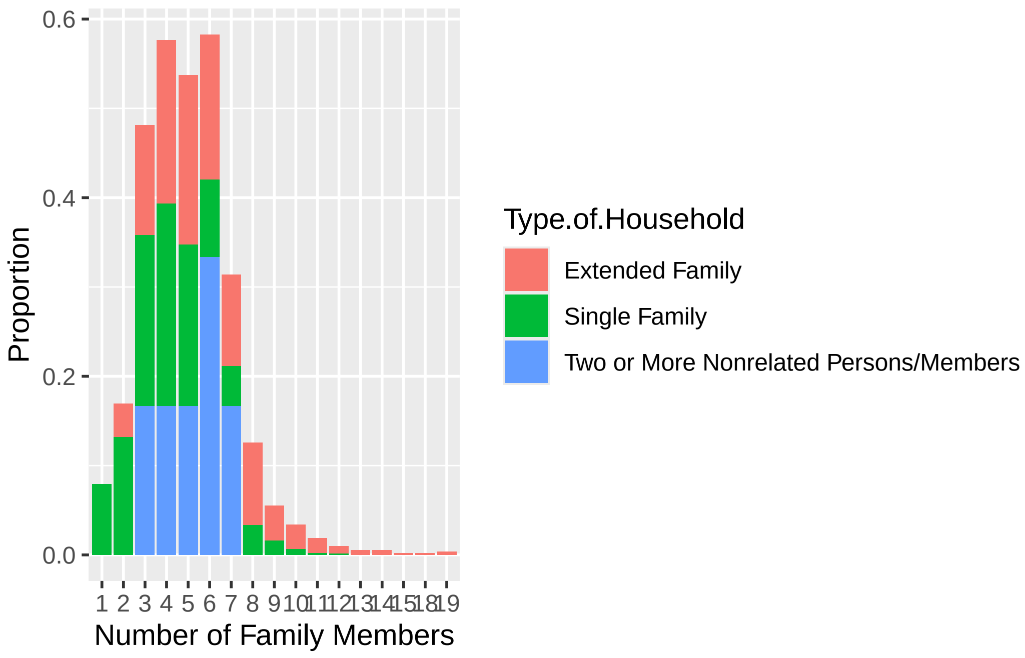


Figure 9: Type of household in different size of households

From we Figure 9 can see that these families with two or more nonrelated members only exist in medium size household. As total family members increase more than 8, single family account for a very small proportion.

```
ggplot(data=data,aes(x=Type.of.Household,
                     y=as.numeric(as.character(Total.Number.of.Family.members))))+
  geom_boxplot(aes(fill=Type.of.Household))+
  scale_x_discrete(labels=c("Extended","Single","Two or more"))+
  labs(x="Type of household",y="Number of family members")+
  theme(legend.position = "bottom")
```

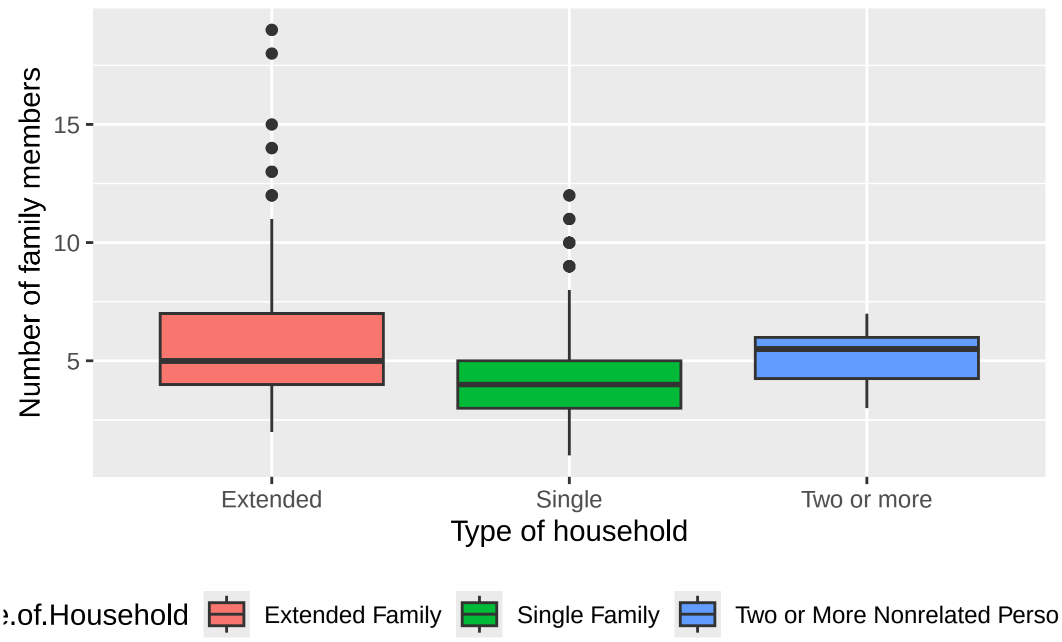


Figure 10: Number of family members by type of household

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     y=House.Floor.Area,fill=Total.Number.of.Family.members))+
  geom_boxplot()+theme(legend.position = "none")+
  labs(x="Number of Family Members",y="House floor area")
```

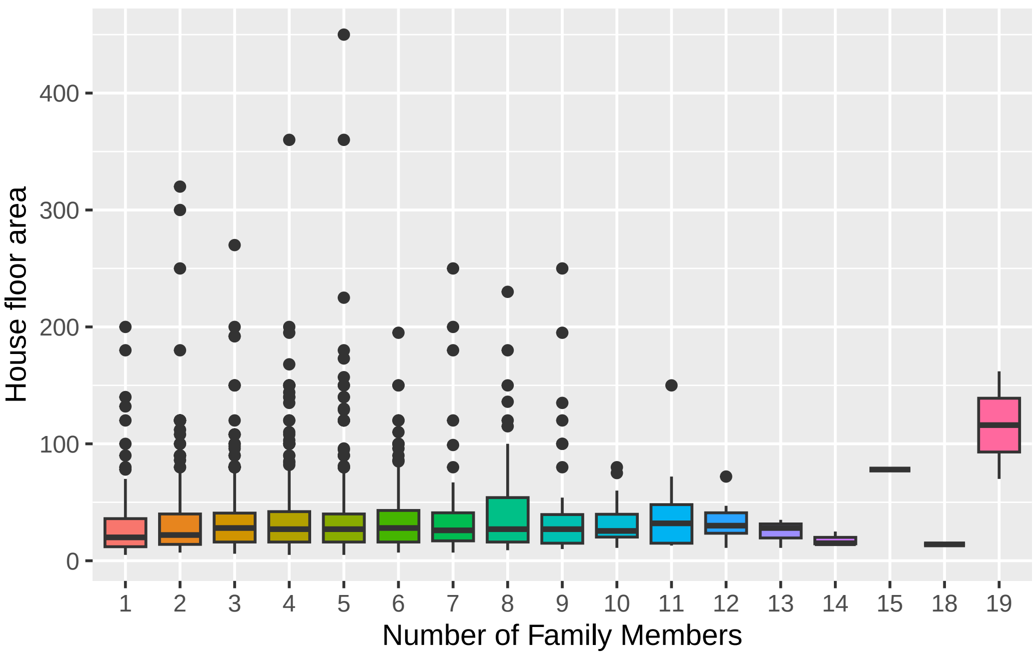


Figure 11: House floor area for different size of households

As shown in Figure 11, there are a few outliers for different sizes of households, . And the median of house floor area seems to be stable as number of family members increase.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=House.Age,
                         fill=Total.Number.of.Family.members))+
  geom_boxplot()+theme(legend.position = "none")+
  labs(x="Number of Family Members",y="House age")
```

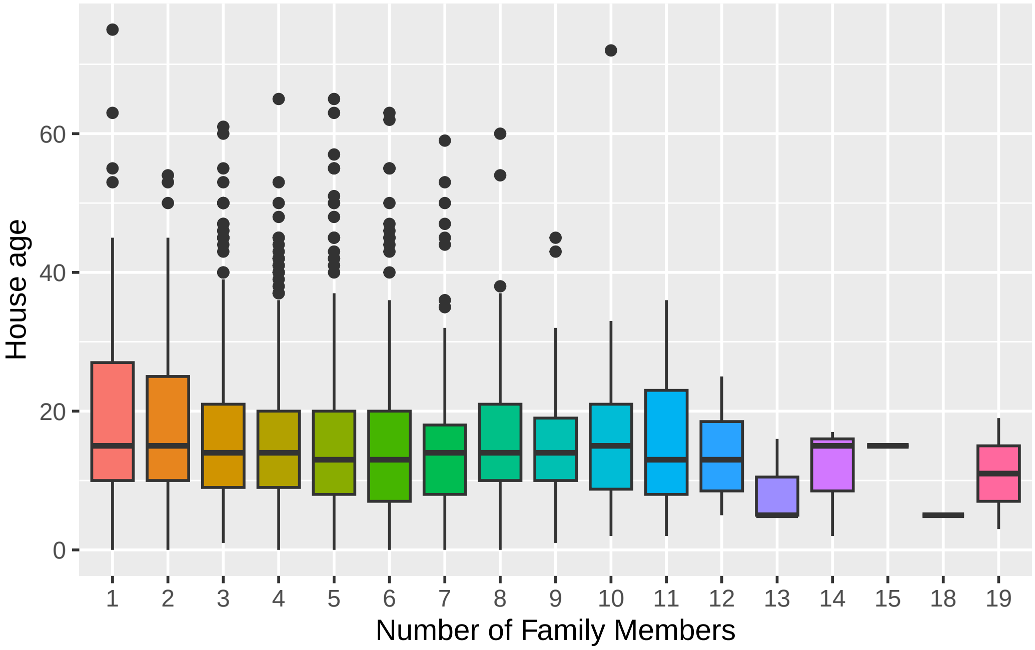


Figure 12: House age for different sizes of households

The median house age of different sizes of households are less than 20 years, which is relatively stable as number of family members increase. (Figure 12)

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Number.of.bedrooms))+
  geom_bar(aes(y=..prop..,fill=Number.of.bedrooms))+
  labs(x="Number of Family Members",y="Proportion")
```

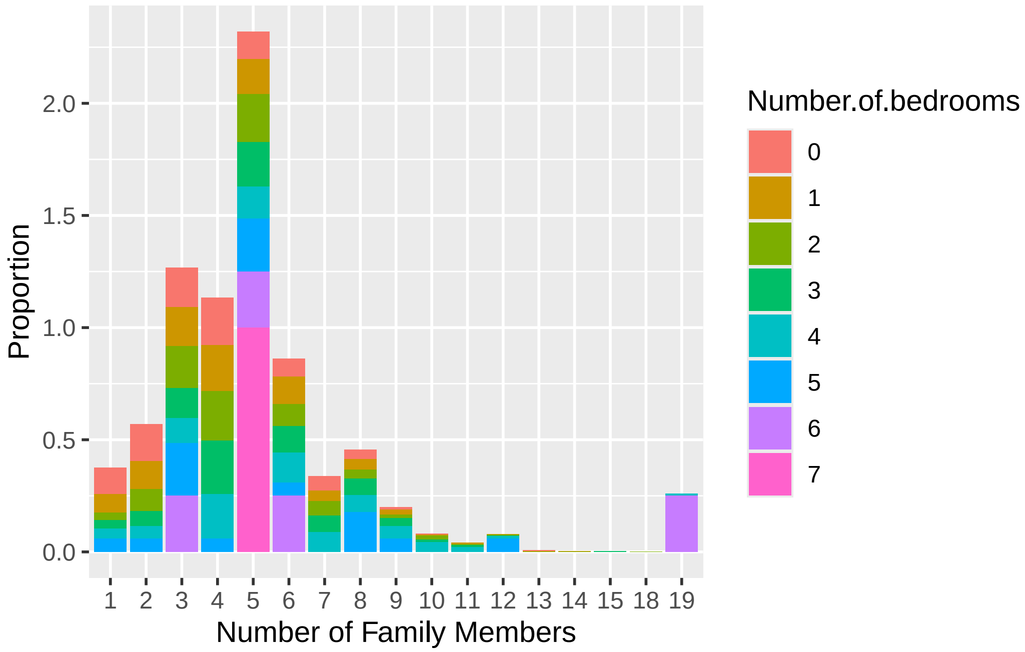


Figure 13: Number of bedrooms by different sizes of households

As the number of family members increases, number of bedrooms increase, but for household with 5 family members, proportion of 7 bedrooms is incredibly high.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Electricity))+
  geom_bar(aes(y=..prop..,fill=Electricity),position="dodge")+
  labs(x="Number of Family Members",y="Proportion")
```

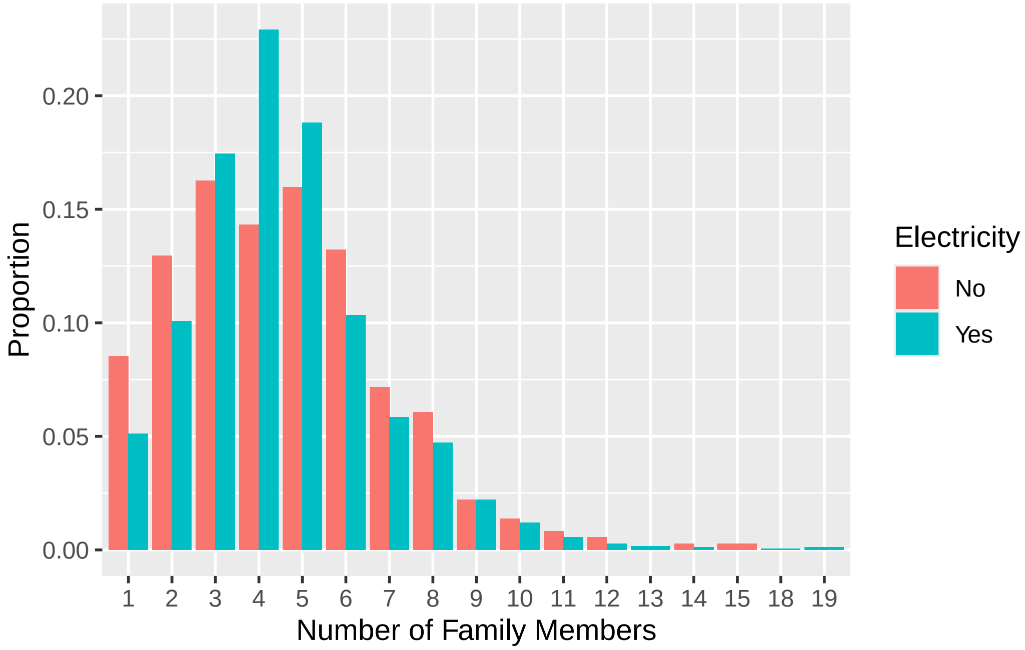


Figure 14: Electricity by different sizes of households

For those small size households, the proportion without electricity is relatively high.

```
ggplot(data=data,aes(x=Electricity,y=as.numeric(as.character(
  Total.Number.of.Family.members))))+geom_boxplot(aes(fill=Electricity))+
  labs(x="Electricity",y="Number of family members")
```

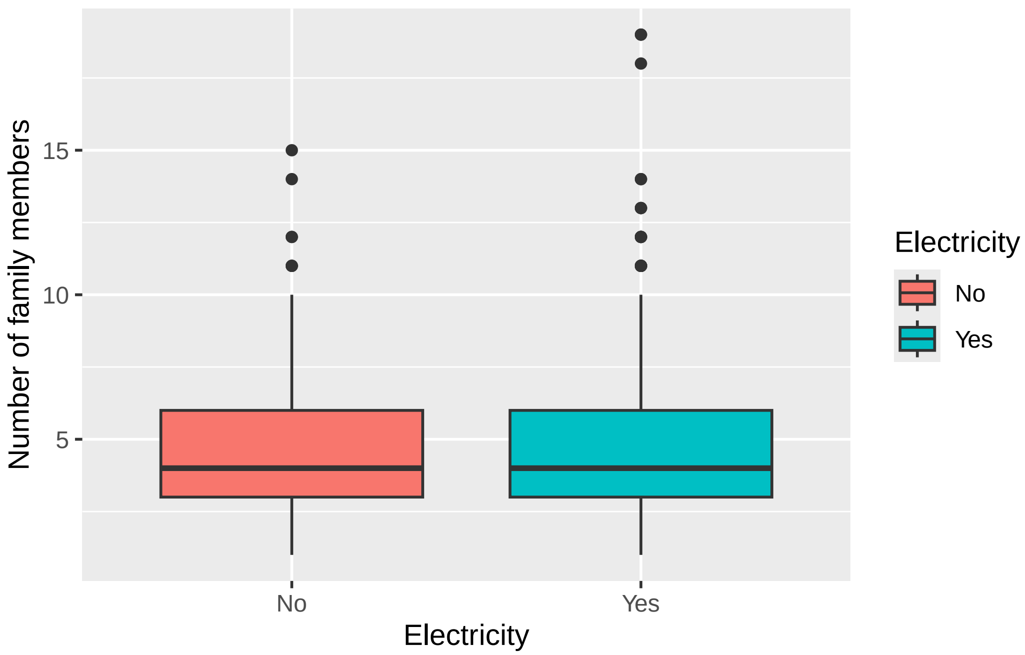


Figure 15: Number of family members by electricity

From the above Figure 15, households with electricity and without electricity have the same distribution of family members.

4 Formal analysis

4.1 Poisson Regression Model

4.1.1 Fit model with all variables

```
# As the response variable is the number of people living in a household,
# which is counts data, we tend to use a poisson model to fit it.
model1=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,data=data.norm,family = poisson)
model1%>%
  summary()
```

```
Call:
glm(formula = 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.norm)

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                              1.57997    0.05607
Total.Household.Income                                  -0.75227    0.17767
Total.Food.Expenditure                                   2.11528    0.13570
Household.Head.SexMale                                   0.26306    0.03053
Household.Head.Age                                      -0.34169    0.07294
Type.of.HouseholdSingle Family                          -0.34673    0.02291
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.10585    0.18088
House.Floor.Area                                        -0.21983    0.15139
House.Age                                               -0.27860    0.07727
Number.of.bedrooms                                       0.35079    0.08638
ElectricityYes                                          -0.09028    0.02850
                                                        z value Pr(>|z|)
(Intercept)                                              28.177  < 2e-16 ***
Total.Household.Income                                   -4.234 2.29e-05 ***
Total.Food.Expenditure                                   15.588  < 2e-16 ***
Household.Head.SexMale                                    8.616  < 2e-16 ***
Household.Head.Age                                       -4.684 2.81e-06 ***
Type.of.HouseholdSingle Family                          -15.135  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.585 0.558423
House.Floor.Area                                         -1.452 0.146476
House.Age                                                -3.606 0.000311 ***
Number.of.bedrooms                                        4.061 4.89e-05 ***
ElectricityYes                                           -3.168 0.001536 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1551.8  on 2111  degrees of freedom
AIC: 8511.9

Number of Fisher Scoring iterations: 5
```

```r
confint(model1)%>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	1.4697699	1.6895784
Total.Household.Income	-1.1047276	-0.4083056
Total.Food.Expenditure	1.8438582	2.3758808
Household.Head.SexMale	0.2036003	0.3232971
Household.Head.Age	-0.4847617	-0.1988282
Type.of.HouseholdSingle Family	-0.3915529	-0.3017466
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.4820181	0.2294578
House.Floor.Area	-0.5203892	0.0730630
House.Age	-0.4306764	-0.1277901
Number.of.bedrooms	0.1813764	0.5199776
ElectricityYes	-0.1458759	-0.0341516

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

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

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

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

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

```
[1] "No"  "Yes"
```

The default baseline in R being taken as the one which comes first alphabetically. So these three categorical variables adopt female, Extended Family, No as baseline.

From the above summary we can observe that one continuous explanatory variable floor area is not significant and compared to extended family, Two or More Nonrelated Persons/Members is not significant while single family is significant according to the p-value and the 95% CI of estimates of coefficients.

4.1.1.1 Rate Ratio

```
model_summary <- summary(model1)
coef <- model_summary$coefficients[,1]
std_err <- model_summary$coefficients[,2]
rate_ratio <- exp(model_summary$coef)
conf_interval <- exp(cbind(coef - 1.96 * std_err, coef + 1.96 * std_err))
result <- data.frame(coef = coef, std_err = std_err, rate_ratio = rate_ratio, conf_interva
print(result)
```

```
                                                               coef    std_err
(Intercept)                                              1.57996848 0.05607394
Total.Household.Income                                  -0.75227175 0.17766885
Total.Food.Expenditure                                   2.11527547 0.13570013
Household.Head.SexMale                                   0.26305999 0.03053305
Household.Head.Age                                      -0.34169093 0.07294341
Type.of.HouseholdSingle Family                          -0.34672880 0.02290952
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.10584735 0.18087820
House.Floor.Area                                        -0.21983329 0.15139076
House.Age                                               -0.27859648 0.07726785
Number.of.bedrooms                                       0.35078526 0.08637971
ElectricityYes                                          -0.09028251 0.02849982
                                                        rate_ratio.Estimate
(Intercept)                                                       4.8548028
Total.Household.Income                                            0.4712947
Total.Food.Expenditure                                            8.2918696
Household.Head.SexMale                                            1.3009048
Household.Head.Age                                                0.7105678
Type.of.HouseholdSingle Family                                    0.7069970
Type.of.HouseholdTwo or More Nonrelated Persons/Members           0.8995620
House.Floor.Area                                                  0.8026526
House.Age                                                         0.7568452
Number.of.bedrooms                                                1.4201823
ElectricityYes                                                    0.9136730
                                                        rate_ratio.Std..Error
(Intercept)                                                          1.057676
Total.Household.Income                                               1.194430
Total.Food.Expenditure                                               1.145338
Household.Head.SexMale                                               1.031004
Household.Head.Age                                                   1.075670
Type.of.HouseholdSingle Family                                       1.023174
Type.of.HouseholdTwo or More Nonrelated Persons/Members              1.198269
```

```
House.Floor.Area                                                     1.163451
House.Age                                                            1.080331
Number.of.bedrooms                                                   1.090220
ElectricityYes                                                       1.028910
                                                        rate_ratio.z.value
(Intercept)                                                   1.725466e+12
Total.Household.Income                                        1.449251e-02
Total.Food.Expenditure                                        5.884700e+06
Household.Head.SexMale                                        5.516954e+03
Household.Head.Age                                            9.238931e-03
Type.of.HouseholdSingle Family                                2.673506e-07
Type.of.HouseholdTwo or More Nonrelated Persons/Members       5.570024e-01
House.Floor.Area                                              2.340801e-01
House.Age                                                     2.717131e-02
Number.of.bedrooms                                            5.803047e+01
ElectricityYes                                                4.209497e-02
                                                        rate_ratio.Pr...z..
(Intercept)                                                        1.000000
Total.Household.Income                                             1.000023
Total.Food.Expenditure                                             1.000000
Household.Head.SexMale                                             1.000000
Household.Head.Age                                                 1.000003
Type.of.HouseholdSingle Family                                     1.000000
Type.of.HouseholdTwo or More Nonrelated Persons/Members            1.747914
House.Floor.Area                                                   1.157747
House.Age                                                          1.000311
Number.of.bedrooms                                                 1.000049
ElectricityYes                                                     1.001537
                                                               X1         X2
(Intercept)                                             4.3495116  5.4187947
Total.Household.Income                                  0.3327038  0.6676169
Total.Food.Expenditure                                  6.3553897 10.8183928
Household.Head.SexMale                                  1.2253361  1.3811338
Household.Head.Age                                      0.6159066  0.8197779
Type.of.HouseholdSingle Family                          0.6759532  0.7394666
Type.of.HouseholdTwo or More Nonrelated Persons/Members 0.6310510  1.2823238
House.Floor.Area                                        0.5965697  1.0799261
House.Age                                               0.6504821  0.8806003
Number.of.bedrooms                                      1.1989918  1.6821782
ElectricityYes                                          0.8640349  0.9661629
```

The result from the rate ratio agree with that from p-values and confidence intervals. We can observe that the type “Two or More Nonrelated Persons/Members” is not significantly different compared to the baseline “Extended family”. So we can firstly merge these two kinds of types of household to “Not Single”, while another is “Single Family”.

```
# for conviniency, introduce a new dataset,
# which we can merge these two kinds of type of households in it
# without changing the original dataset.
data.norm.merged=data.frame(data.norm)
data.norm.merged$Type.of.Household <- as.character(data.norm.merged$Type.of.Household)
data.norm.merged$Type.of.Household[data.norm.merged$Type.of.Household != "Single Family"]
```

4.1.2 Fit model on the merged dataset

```
model2=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,data=data.norm.merged,family = poisson)
model2%>%
  summary()
```

```
Call:
glm(formula = 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.norm.merged)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     1.57764    0.05594  28.202  < 2e-16 ***
Total.Household.Income         -0.75201    0.17764  -4.233 2.30e-05 ***
Total.Food.Expenditure          2.11713    0.13561  15.611  < 2e-16 ***
Household.Head.SexMale          0.26293    0.03053   8.611  < 2e-16 ***
Household.Head.Age             -0.34035    0.07291  -4.668 3.04e-06 ***
Type.of.HouseholdSingle Family -0.34540    0.02280 -15.147  < 2e-16 ***
House.Floor.Area               -0.21843    0.15136  -1.443 0.148971
House.Age                      -0.27721    0.07723  -3.589 0.000331 ***
Number.of.bedrooms              0.35115    0.08639   4.065 4.81e-05 ***
ElectricityYes                 -0.09036    0.02850  -3.171 0.001521 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1552.2  on 2112  degrees of freedom
AIC: 8510.2

Number of Fisher Scoring iterations: 5
```

we can find the variable floor area is still not significant, so we remove it then.

4.1.3 Remove floor area

```r
model3=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, data=data.norm.merged,family = poisson)
```

```r
model3%>%
  summary()
```

```
Call:
glm(formula = 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.norm.merged)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     1.57856    0.05593  28.223  < 2e-16 ***
Total.Household.Income         -0.79803    0.17461  -4.570 4.87e-06 ***
Total.Food.Expenditure          2.11999    0.13571  15.622  < 2e-16 ***
Household.Head.SexMale          0.26332    0.03053   8.625  < 2e-16 ***
Household.Head.Age             -0.34533    0.07283  -4.741 2.12e-06 ***
Type.of.HouseholdSingle Family -0.34576    0.02280 -15.164  < 2e-16 ***
House.Age                      -0.28066    0.07720  -3.635 0.000278 ***
Number.of.bedrooms              0.31177    0.08201   3.802 0.000144 ***
ElectricityYes                 -0.09140    0.02849  -3.209 0.001334 **
---
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1554.3  on 2113  degrees of freedom
AIC: 8510.4

Number of Fisher Scoring iterations: 5
```

The explanatory variables in final model are all significant , with an AIC value of 8510.4.

4.2 Overdispersion

```
ggplot(model2, aes(x=log(fitted(model2)),
                   y=log((data.norm.merged$Total.Number.of.Family.members-fitted(model2))^
geom_point(col="#f46d43") +
geom_abline(slope=1, intercept=0, col="#a6d96a", linewidth=1) +
ylab(expression((y-hat(mu))^2)) + xlab(expression(hat(mu)))
```

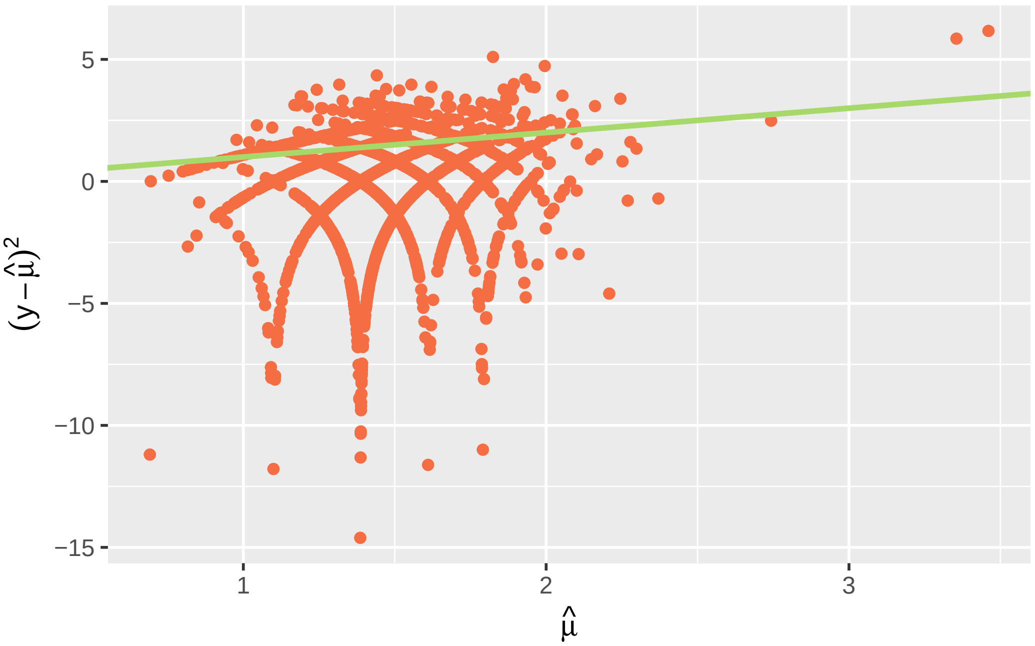


Figure 16: scatterplot of mean and variance

From Figure 16, we can find most of the points lie above the line of equality for mean and variance. In this case, we are not to able to determine which explanatory variables are significant.

4.2.1 Examine existence of overdispersion

```
library(qcc)
data$Total.Number.of.Family.members=as.numeric(as.character(
  data$Total.Number.of.Family.members))
qcc.overdispersion.test(data$Total.Number.of.Family.members)
```

```
Overdispersion test Obs.Var/Theor.Var Statistic   p-value
       poisson data          1.082586  2296.164 0.0042826
```

From the overdispersion test we know that the p-value<0.05, indicating that the overdispersion does exist in number of family members. So we should consider to fit a Quasi-Poisson model or a negative binomial model to the data.

4.2.2 Quasi-Poisson model

we can define a dispersion parameter ϕ such that $Var(Y_i) = \phi\mu_i$, we can estimate this parameter by

$$\hat{\phi} = \frac{X^2}{n-p}$$

```
X2=sum(resid(model1,type="pearson")^2)
dp=X2/model1$df.res
# With the use of the estimated dispersion parameter the Wald tests are not very reliable,
# so we turn to an F test to determine the significance of the regression coefficients:
drop1(model1,test="F")
```

```
Single term deletions

Model:
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  F value    Pr(>F)
```

```
<none>                      1551.8 8511.9
Total.Household.Income  1   1570.8 8528.8  25.7704 4.182e-07 ***
Total.Food.Expenditure  1   1737.1 8695.2 252.0856 < 2.2e-16 ***
Household.Head.Sex      1   1630.4 8588.4 106.8233 < 2.2e-16 ***
Household.Head.Age      1   1573.8 8531.9  29.9530 4.952e-08 ***
Type.of.Household       2   1774.8 8730.9 151.6907 < 2.2e-16 ***
House.Floor.Area        1   1554.0 8512.0   2.9244 0.0873964 .
House.Age               1   1565.0 8523.1  17.9624 2.350e-05 ***
Number.of.bedrooms      1   1568.3 8526.3  22.3752 2.391e-06 ***
Electricity             1   1561.7 8519.8  13.4388 0.0002526 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the model summary above, we are supposed to delete the variable House.Floor.Area.

```
model_quasi <- 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,
                   data=data.norm,
                   family = quasipoisson(link = "log"))
summary(model_quasi)
```

```
Call:
glm(formula = 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 = quasipoisson(link = "log"), data = data.norm)

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                       1.57997    0.04858
Total.Household.Income                           -0.75227    0.15391
Total.Food.Expenditure                            2.11528    0.11755
Household.Head.SexMale                            0.26306    0.02645
Household.Head.Age                               -0.34169    0.06319
Type.of.HouseholdSingle Family                   -0.34673    0.01985
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.10585    0.15669
House.Floor.Area                                 -0.21983    0.13115
House.Age                                        -0.27860    0.06693
Number.of.bedrooms                                0.35079    0.07483
```

```
ElectricityYes                                          -0.09028    0.02469
                                                        t value Pr(>|t|)
(Intercept)                                              32.526  < 2e-16 ***
Total.Household.Income                                   -4.888 1.10e-06 ***
Total.Food.Expenditure                                   17.994  < 2e-16 ***
Household.Head.SexMale                                    9.946  < 2e-16 ***
Household.Head.Age                                       -5.407 7.11e-08 ***
Type.of.HouseholdSingle Family                          -17.471  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.676 0.499417
House.Floor.Area                                         -1.676 0.093836 .
House.Age                                                -4.162 3.28e-05 ***
Number.of.bedrooms                                        4.688 2.94e-06 ***
ElectricityYes                                           -3.657 0.000262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.7504229)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1551.8  on 2111  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

In a Quasi-Poisson model, Two or More Nonrelated Persons/Members is still not significantly different compared to Extended Family. So we need to fit this model again using merged dataset.

```
model_quasi_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,
                     data=data.norm.merged,family = quasipoisson(link = "log"))
summary(model_quasi_1)
```

```
Call:
glm(formula = 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 = quasipoisson(link = "log"), data = data.norm.merged)
```

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.57764    0.04846  32.558  < 2e-16 ***
Total.Household.Income         -0.75201    0.15387  -4.887 1.10e-06 ***
Total.Food.Expenditure          2.11713    0.11747  18.022  < 2e-16 ***
Household.Head.SexMale          0.26293    0.02645   9.941  < 2e-16 ***
Household.Head.Age             -0.34035    0.06316  -5.389 7.89e-08 ***
Type.of.HouseholdSingle Family -0.34540    0.01975 -17.486  < 2e-16 ***
House.Floor.Area               -0.21843    0.13111  -1.666 0.095849 .  
House.Age                      -0.27721    0.06690  -4.144 3.55e-05 ***
Number.of.bedrooms              0.35115    0.07483   4.693 2.87e-06 ***
ElectricityYes                 -0.09036    0.02469  -3.660 0.000258 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.7503333)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1552.2  on 2112  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

We can see that the p-value of House.Floor.Area is 0.0958 which is larger than 0.05, so we need to remove the floor area variable.

```
model_quasi_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,
                     data=data.norm.merged,family = quasipoisson(link = "log"))
summary(model_quasi_2)
```

```

Call:
glm(formula = 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 = quasipoisson(link = "log"), data = data.norm.merged)

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.57856    0.04848  32.564  < 2e-16 ***
```

```
Total.Household.Income         -0.79803    0.15134  -5.273 1.48e-07 ***
Total.Food.Expenditure          2.11999    0.11762  18.024  < 2e-16 ***
Household.Head.SexMale          0.26332    0.02646   9.951  < 2e-16 ***
Household.Head.Age             -0.34533    0.06313  -5.470 5.02e-08 ***
Type.of.HouseholdSingle Family -0.34576    0.01976 -17.496  < 2e-16 ***
House.Age                      -0.28066    0.06691  -4.194 2.85e-05 ***
Number.of.bedrooms              0.31177    0.07108   4.386 1.21e-05 ***
ElectricityYes                 -0.09140    0.02469  -3.702 0.000219 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.7511975)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1554.3  on 2113  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Using the Quasi-Poisson model, we reach the same conclusion as what we get in the ordinary glm model, which removes only floor area variable.

4.2.3 Negative binomial models

Considering the Overdispersion, another choice is the Negative-binomial model.

```
model_nb=glm.nb(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,data=data.norm)
summary(model_nb)
```

```
Call:
glm.nb(formula = 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, data = data.norm, init.theta = 76069.2575, link = log)

Coefficients:
                                                        Estimate Std. Error
```

```
(Intercept)                                              1.57996    0.05608
Total.Household.Income                                  -0.75233    0.17768
Total.Food.Expenditure                                   2.11546    0.13571
Household.Head.SexMale                                   0.26306    0.03053
Household.Head.Age                                      -0.34169    0.07295
Type.of.HouseholdSingle Family                          -0.34673    0.02291
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.10584    0.18088
House.Floor.Area                                        -0.21985    0.15140
House.Age                                               -0.27860    0.07727
Number.of.bedrooms                                       0.35077    0.08638
ElectricityYes                                          -0.09029    0.02850
                                                        z value Pr(>|z|)
(Intercept)                                              28.176  < 2e-16 ***
Total.Household.Income                                   -4.234 2.29e-05 ***
Total.Food.Expenditure                                   15.588  < 2e-16 ***
Household.Head.SexMale                                    8.615  < 2e-16 ***
Household.Head.Age                                       -4.684 2.81e-06 ***
Type.of.HouseholdSingle Family                          -15.134  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.585 0.558455
House.Floor.Area                                         -1.452 0.146465
House.Age                                                -3.605 0.000312 ***
Number.of.bedrooms                                        4.061 4.89e-05 ***
ElectricityYes                                           -3.168 0.001536 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(76069.33) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1551.7  on 2111  degrees of freedom
AIC: 8513.9

Number of Fisher Scoring iterations: 1

              Theta:  76069
          Std. Err.:  280723
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -8489.906
```

Similarly, we can see that the categorical variable Type.of.Household(Two or More Nonrelated Persons/Members) and continuous variable House.Floor.Area seem not to be statistically significant with the response variable.

```r
model_nb1 <- glm.nb(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,data=data.norm.merged)
summary(model_nb1)
```

```

Call:
glm.nb(formula = 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, data = data.norm.merged, init.theta = 75964.08369,
    link = log)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                     1.57763    0.05594  28.201  < 2e-16 ***
Total.Household.Income         -0.75206    0.17765  -4.233 2.30e-05 ***
Total.Food.Expenditure          2.11731    0.13563  15.611  < 2e-16 ***
Household.Head.SexMale          0.26293    0.03053   8.611  < 2e-16 ***
Household.Head.Age             -0.34035    0.07292  -4.668 3.05e-06 ***
Type.of.HouseholdSingle Family -0.34540    0.02280 -15.146  < 2e-16 ***
House.Floor.Area               -0.21845    0.15136  -1.443 0.148959
House.Age                      -0.27721    0.07723  -3.589 0.000332 ***
Number.of.bedrooms              0.35114    0.08639   4.065 4.81e-05 ***
ElectricityYes                 -0.09037    0.02850  -3.171 0.001520 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(75964.11) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1552.1  on 2112  degrees of freedom
AIC: 8512.3

Number of Fisher Scoring iterations: 1


              Theta:  75964
          Std. Err.:  280536
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -8490.261
```

```
model_nb1$aic
```

```
[1] 8512.261
```

We firstly fit a negative model using the merged dataset and find the floor area is still not significant. So we need to remove it in our next model.

```
model_nb2 <- glm.nb(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,
                    data=data.norm.merged)
summary(model_nb2)
```

```
Call:
glm.nb(formula = 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, 
    data = data.norm.merged, init.theta = 76018.81357, link = log)

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     1.57856    0.05593  28.222  < 2e-16 ***
Total.Household.Income         -0.79809    0.17462  -4.570 4.87e-06 ***
Total.Food.Expenditure          2.12017    0.13572  15.622  < 2e-16 ***
Household.Head.SexMale          0.26332    0.03053   8.625  < 2e-16 ***
Household.Head.Age             -0.34533    0.07284  -4.741 2.12e-06 ***
Type.of.HouseholdSingle Family -0.34576    0.02280 -15.163  < 2e-16 ***
House.Age                      -0.28066    0.07720  -3.635 0.000278 ***
Number.of.bedrooms              0.31176    0.08201   3.801 0.000144 ***
ElectricityYes                 -0.09141    0.02849  -3.209 0.001334 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(76018.67) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1554.2  on 2113  degrees of freedom
AIC: 8512.4
```

```
Number of Fisher Scoring iterations: 1

              Theta:  76019
          Std. Err.:  280042
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -8492.384
```

```
model_nb2$aic
```

```
[1] 8512.384
```

Using the negative binomial model, all the variables except floor area are significant and the AIC value is 8512.384.

4.3 Model selection by AIC

4.3.1 GLM model

```
c(glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househo
glm(Total.Number.of.Family.members~Total.Food.Expenditure+Household.Head.Sex+Household.Hea
glm(Total.Number.of.Family.members~Total.Household.Income+Household.Head.Sex+Household.Hea
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
```

```
[1] 8510.362 8530.664 8694.374 8587.067 8530.923 8731.536 8521.788 8522.761
[9] 8518.495
```

Table 5: AIC value of different GLM models

Removed variable	AIC value
None	8510.362
Total.Household.Income	8530.664
Total.Food.Expenditure	8694.374

Removed variable	AIC value
Household.Head.Sex	8587.067
Household.Head.Age	8530.923
Type.of.Household	8731.536
House.Age	8521.788
Number.of.bedrooms	8522.761
Electricity	8518.495

As shown in Table 5, the full model with all explanatory variables except House.Floor.Area has the lowest AIC value between different Poisson regression models.

4.3.2 Negative binomial model

```
c(glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Hous
glm.nb(Total.Number.of.Family.members~Total.Food.Expenditure+Household.Head.Sex+Household.
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Household.Head.Sex+Household.
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
```

```
[1] 8512.384 8532.686 8696.396 8589.086 8532.944 8733.542 8523.809 8524.781
[9] 8520.517
```

Table 6: AIC value of different negative binomial models

Removed variable	AIC value
None	8512.384
Total.Household.Income	8532.686
Total.Food.Expenditure	8696.396
Household.Head.Sex	8589.086
Household.Head.Age	8532.944
Type.of.Household	8733.542
House.Age	8523.809
Number.of.bedrooms	8524.781
Electricity	8520.517

As shown in Table 6, the full model with all explanatory variables except House.Floor.Area has the lowest AIC value between different Negative binomial models.

5 Final model

We find that GLM model with only floor area variable removed has the lowest AIC value.

The final model is:

$$log(Total.Number.of.Family.members) = \beta_0 + \beta_1 \cdot Total.Household.Income + \beta_2 \cdot Total.Food.Expenditure + \beta_3$$

$$\mathbb{1}_{\text{Male}}(x) = \begin{cases} 1 & \text{If the head of household is male,} \\ 0 & \text{If the head of household is female.} \end{cases}$$

$$\mathbb{1}_{\text{Family}}(x) = \begin{cases} 1 & \text{Single family,} \\ 0 & \text{Not Single family.} \end{cases}$$

$$\mathbb{1}_{\text{Electricity}}(x) = \begin{cases} 1 & \text{If the house has electricity,} \\ 0 & \text{Otherwise.} \end{cases}$$

For extended family and two or more nonrelated persons/members, the final model is (all the data of numerical explanatory variables need to be normalized by maximum and minimum scaling):

$$log(Total.Number.of.Family.members) = 1.579 - 0.798 \cdot Total.Household.Income + 2.120 \cdot Total.Food.Expendi$$

For single family, the final model is:

$$log(Total.Number.of.Family.members) = 1.2328 - 0.798 \cdot Total.Household.Income + 2.120 \cdot Total.Food.Expend$$

6 Conclusion and future work

6.1 Conclusions

After removing the insignificant variables and comparing the AIC values of different models, it is found that the variables Total.Household.Income, Total.Food.Expenditure, Household.Head.Sex, Household.Head.Age, Type.of.Household, House.Age, Number.of.bedrooms and Electricity could influence response variable Total.Number.of.Family.members (the number of people living in a household).

As for the numerical explanatory variables, we can conclude that the total income of household, the age of householder and the age of house has a positive effect on the number of family members. However, the total expenditure on food and the number of bedrooms has a negative effect on the family size.

For categorical variables, we find that when other variables are constant, the head of household is female, the household type is single family, and the house has electricity, it is more likely to have a smaller household size.

6.2 Future work

Firstly, our study is based on only one region in the Philippines, and can then combine multiple regions and even countries to compare whether there are geographical differences in factors affecting family size.

Secondly, we could also track these factors over time through longitudinal studies to see how they affect family size changes. This would give us insights into how things change over time and could be really helpful for making decisions about families in society.