DAS-Project2

Yiheng Yang

load the data

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

get packages

```r
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)
```

convert some categorical variables to factors

```
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")
```

explanatory analysis

summary of response variable

```
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.memb
'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.

Summary of categorical explanatory variables

```
data_categorical=data%>%
  dplyr::select("Household.Head.Sex","Type.of.Household","Electricity")
summary(data_categorical)
```

```
 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.

Summary of 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")
```

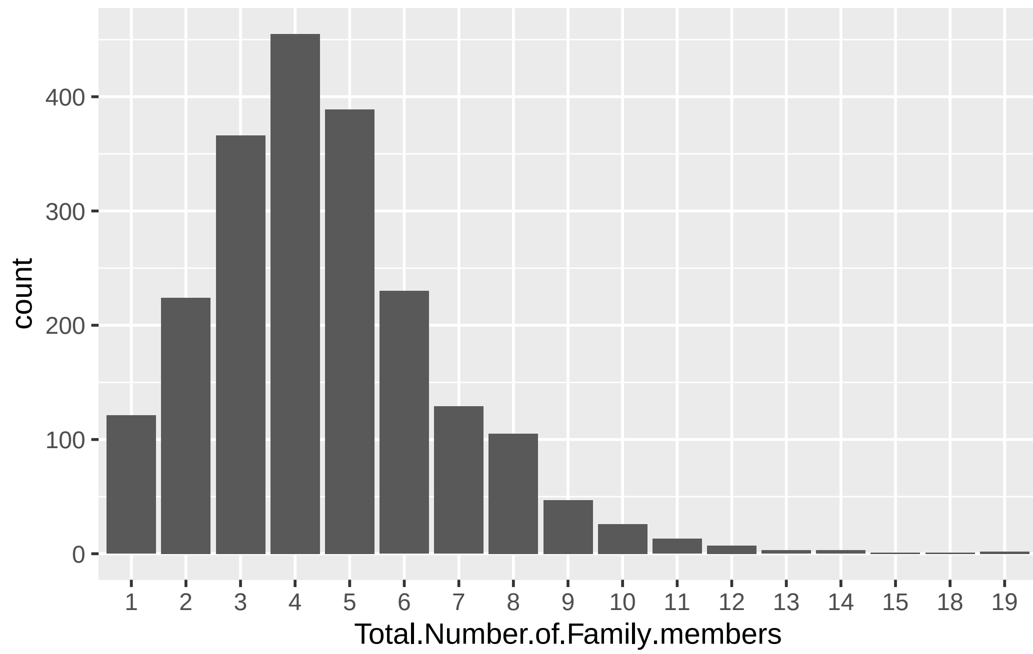
Variable	Sample_size	Mean	St.Dev	Min	Median	Max	IQR
Total.Household.Income	2,122	182,984.80	228,231.07	15,204	120,362.0	3,168,662	74,314.00
Total.Food.Expenditure	2,122	71,738.09	44,938.17	7,783	63,305.5	729,606	24,496.75
Household.Head.Age	2,122	49.28	14.16	9	48.0	99	11.00
Total.Number.of.Family.members	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

Graphical summaries

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

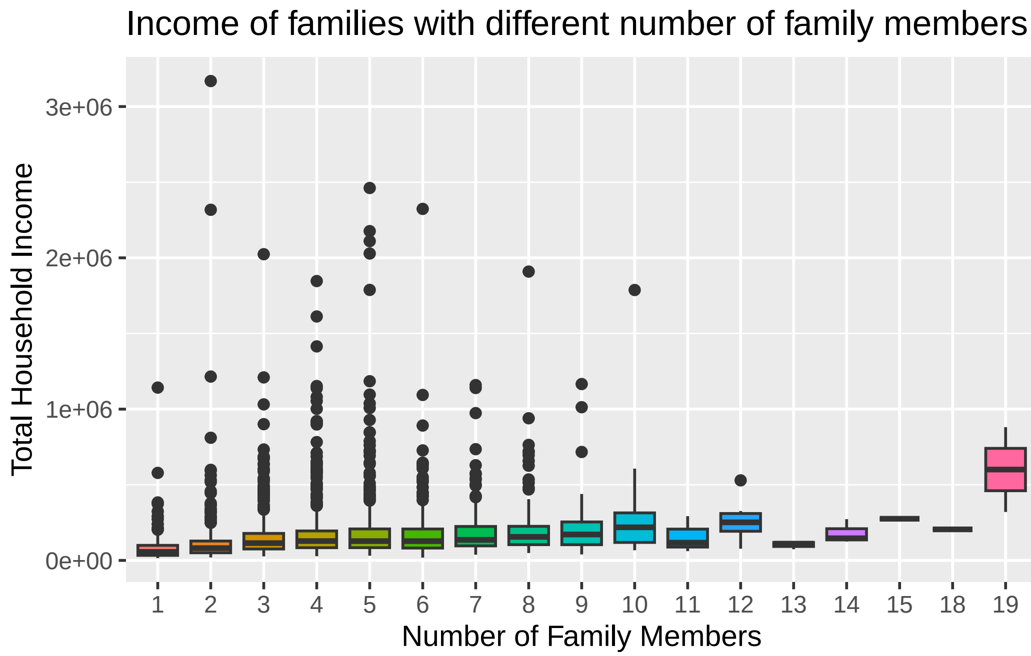
```
data$Total.Number.of.Family.members=as.factor(data$Total.Number.of.Family.members)
```

```
ggplot(data=data,aes(x=Total.Number.of.Family.members))+geom_bar()
```



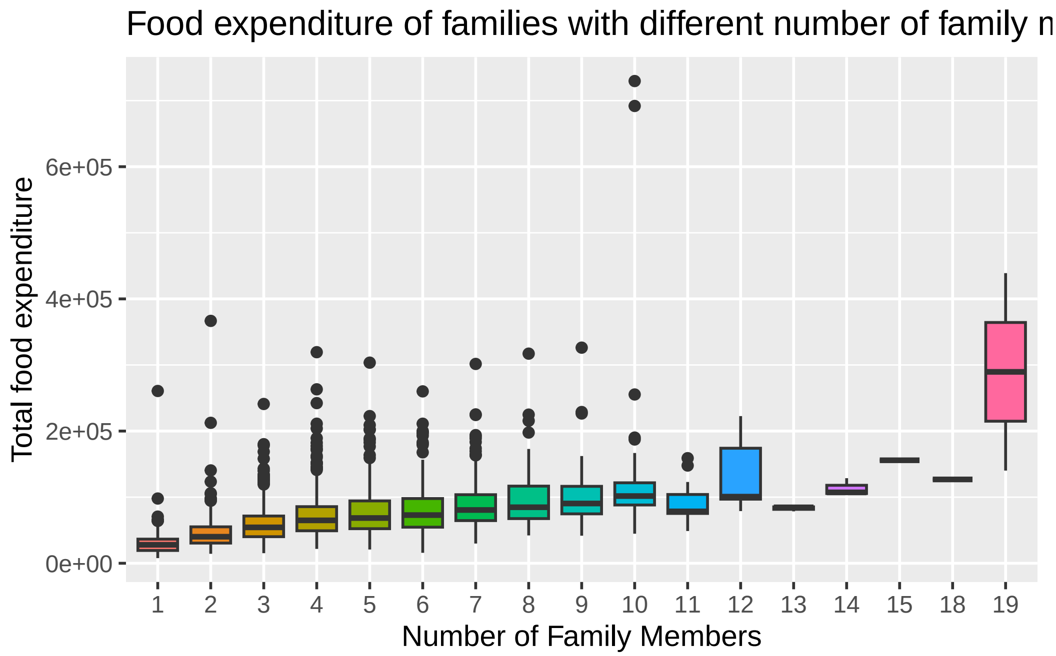
The boxplot shows that household with four family members accounts for the largest proportion. Most of the data is consisted of families with three to five family members.

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



We can see from the above boxplot 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.
```

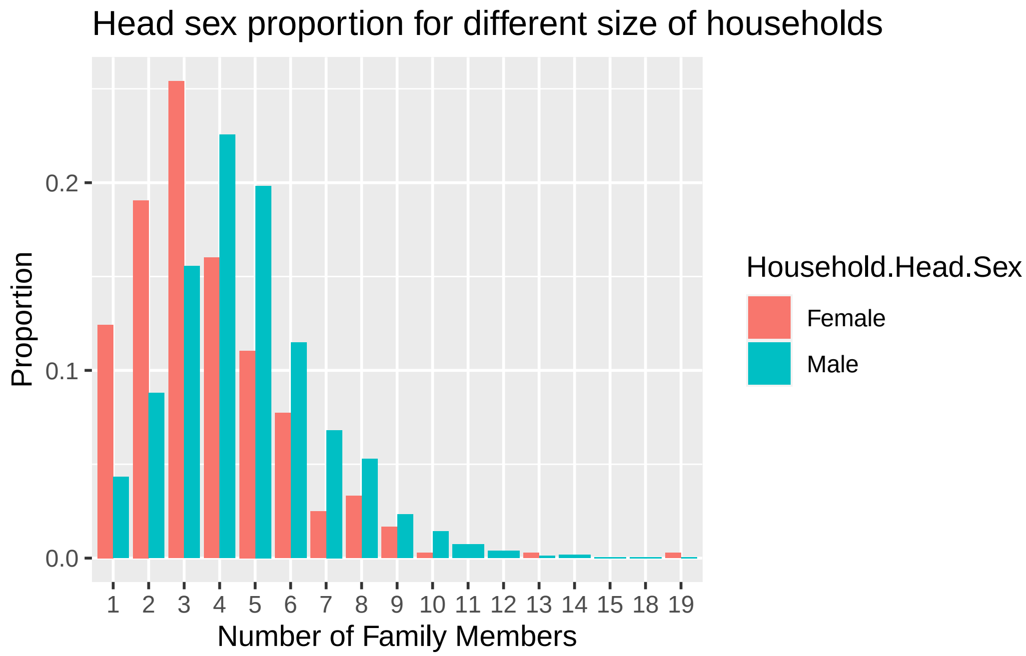


The boxplot indicates that median increase significantly as the number of family members increase. Household with 19 members have the largest variance in food expenditure.

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

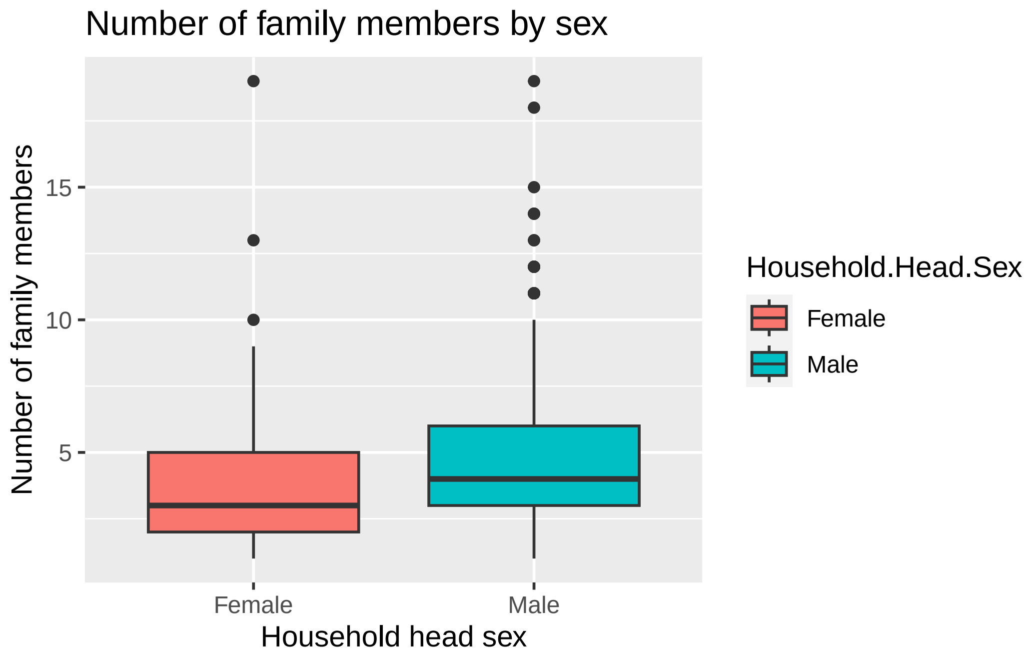
```
Household.Head.Sex           1            2            3            4            5
            Female 12.4% (45) 19.1%  (69) 25.4%  (92) 16.0%  (58) 11.0%  (40)
              Male  4.3% (76)  8.8% (155) 15.6% (274) 22.6% (397) 19.8(349)
          6          7          8         9         10     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

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,group=Household.Head.Sex))+geom_bar(
```



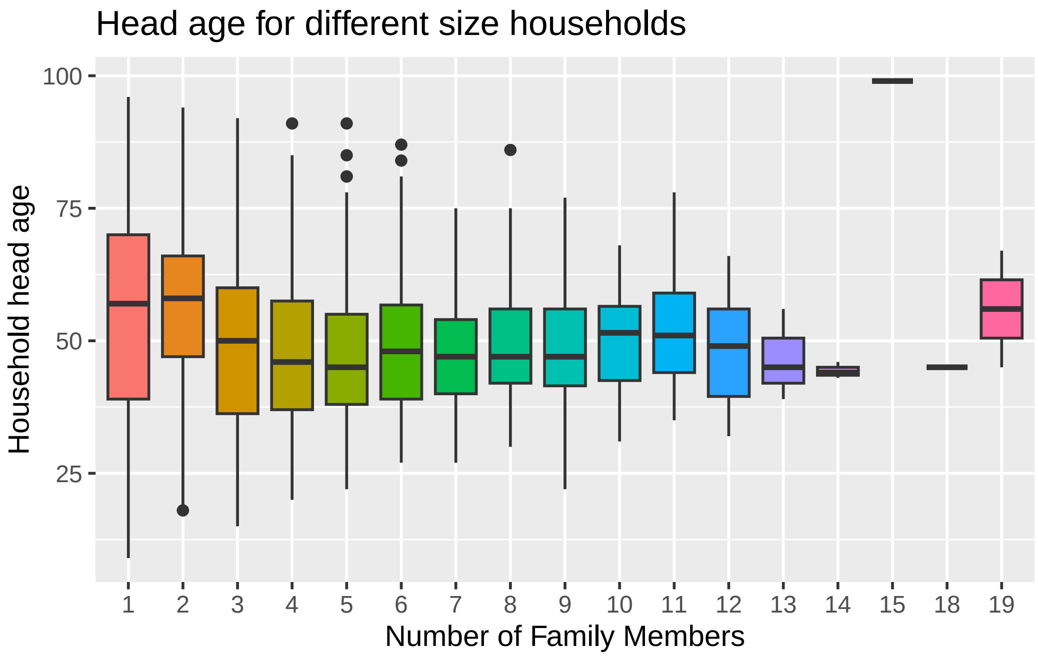
We can see from the barplot, for those small sized households, the proportion is much higher for females than for males. However, this situation does not exist for those household with four or more family members.

```
ggplot(data=data,aes(x=Household.Head.Sex,y=as.numeric(as.character(Total.Number.of.Family
```



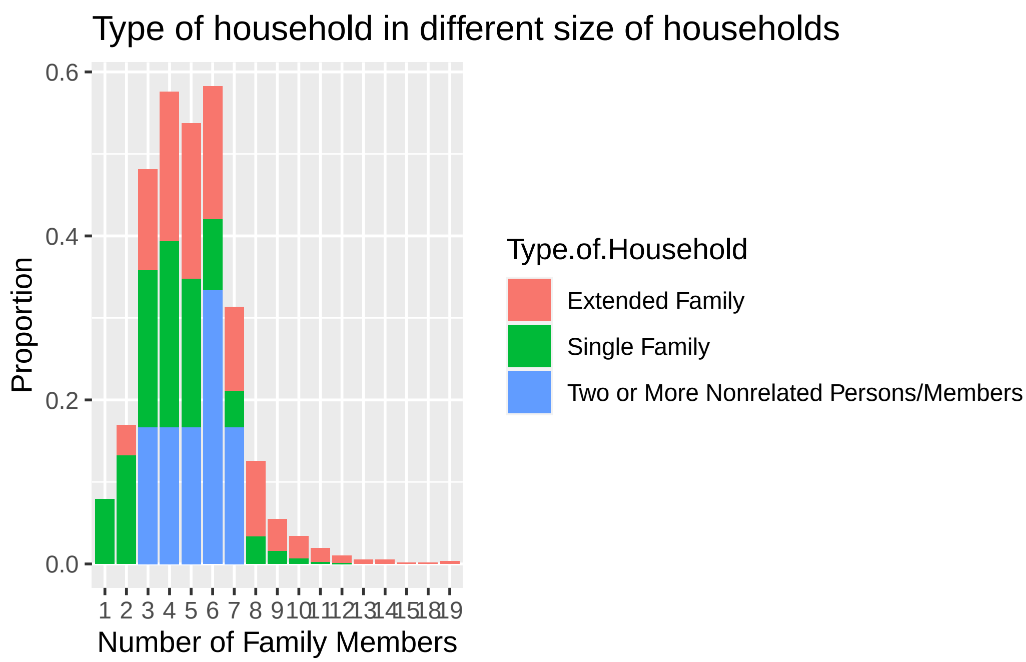
We can conclude from the boxplot that households tend to have more family members if their owner is male.

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



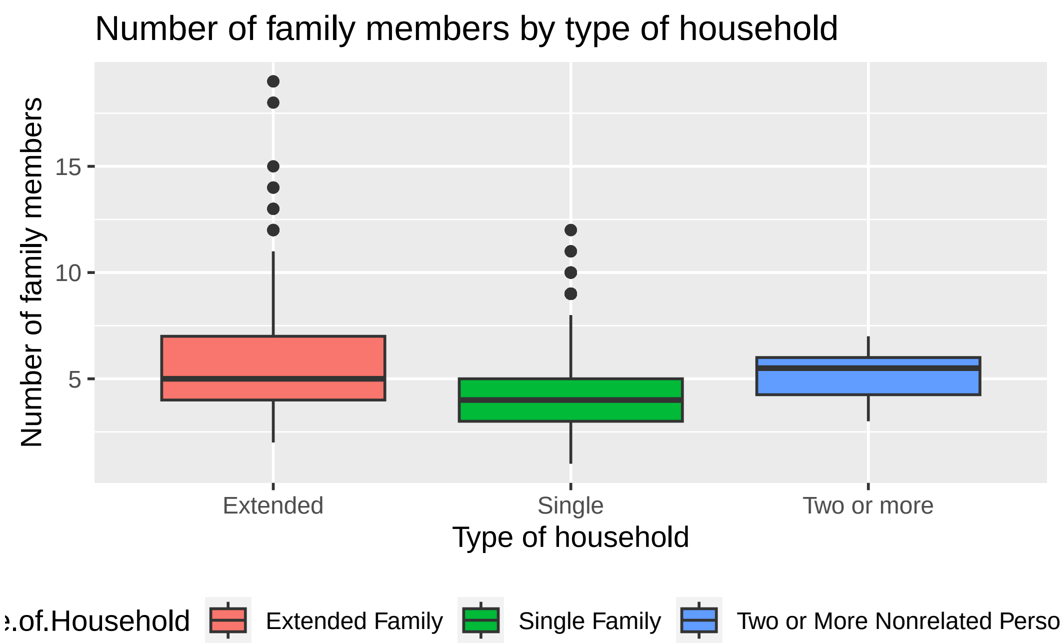
For different size of households, the median of household head age remain at a constant level around 50.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,group=Type.of.Household))+geom_bar(a
```

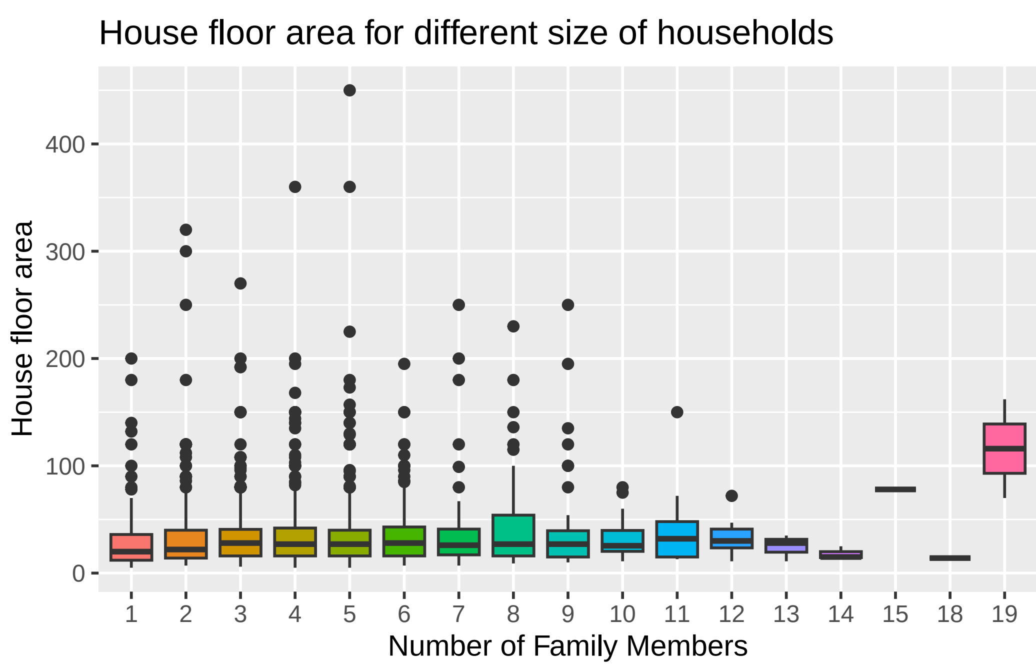


These families with two or more nonrelated members only exist in medium size household. As total family members increase more than 8, single family account for a very small proportion.

```
ggplot(data=data,aes(x=Type.of.Household,y=as.numeric(as.character(Total.Number.of.Family.
```

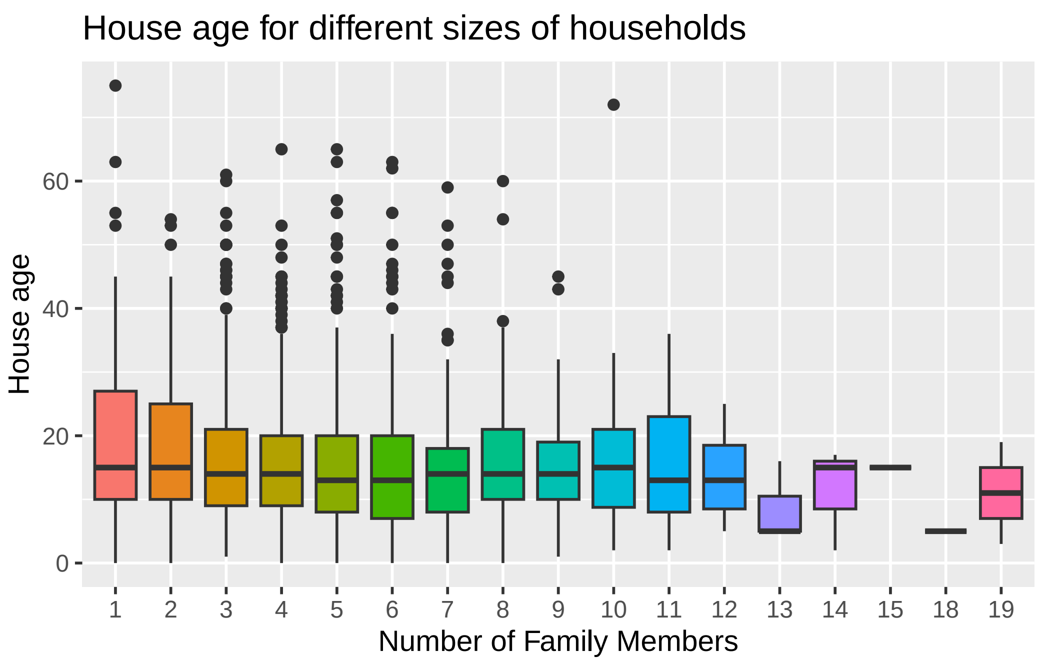



```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=House.Floor.Area,fill=Total.Number
```



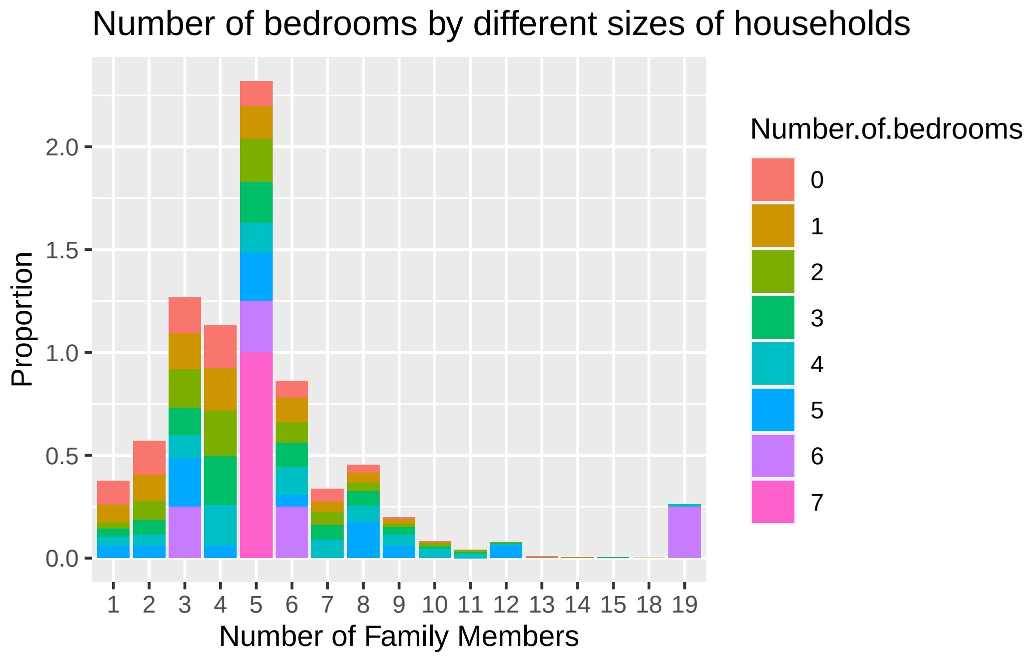
For different sizes of households, there are a few outliers. And the median of house floor area seems to be stable as number of family members increase.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=House.Age,fill=Total.Number.of.Fam
```



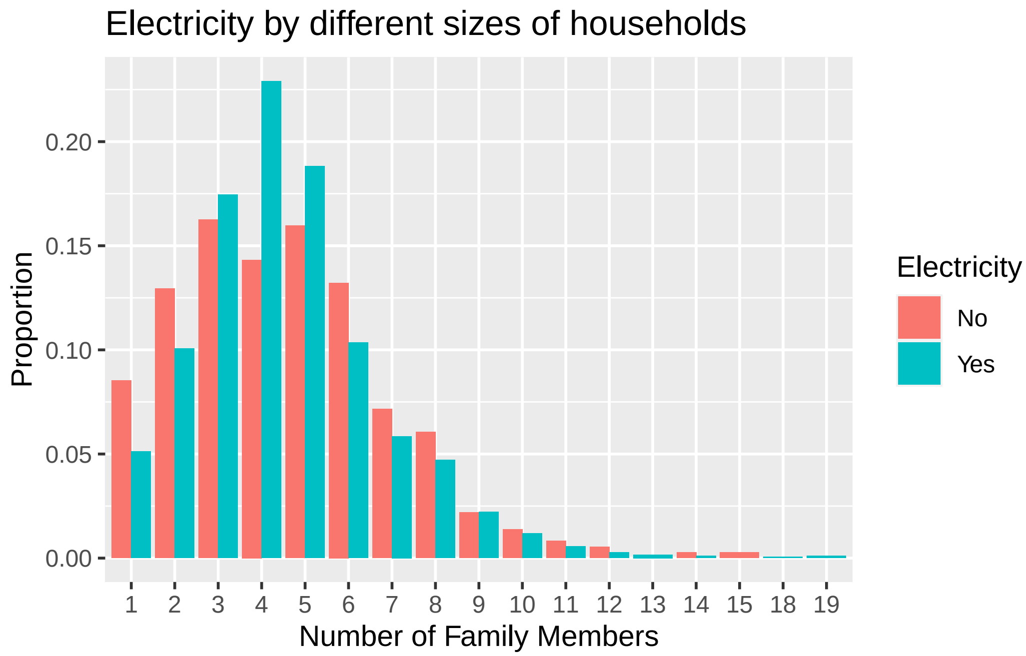
The median house age of different sizes of households are less than 20 years, which is relatively stable as number of family members increase.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,group=Number.of.bedrooms))+geom_bar(
```



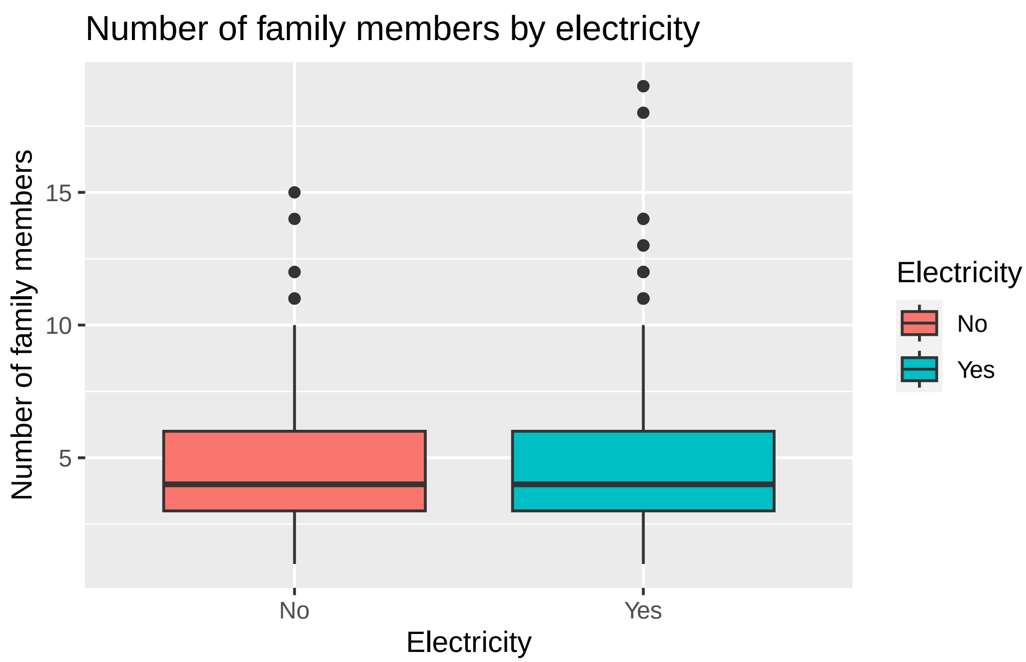
As the number of family members increases, number of bedrooms increase, but for household with 5 family members, proportion of 7 bedrooms is incredibly high.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,group=Electricity))+geom_bar(aes(y=.
```



For those small size households, the proportion wothout electricity is relatively high.

```
ggplot(data=data,aes(x=Electricity,y=as.numeric(as.character(Total.Number.of.Family.member
```



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

format analysis

```
# As the reponse variable is the number of people living in a household, which is counts d
data$Total.Number.of.Family.members=as.numeric(as.character(data$Total.Number.of.Family.me
data$Number.of.bedrooms=as.numeric(as.character(data$Number.of.bedrooms))
model1=glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Ho
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)

Coefficients:
```

```
                                                       Estimate Std. Error
(Intercept)                                           1.597e+00  6.095e-02
Total.Household.Income                               -2.385e-07  5.634e-08
Total.Food.Expenditure                                2.930e-06  1.880e-07
Household.Head.SexMale                                2.631e-01  3.053e-02
Household.Head.Age                                   -3.797e-03  8.105e-04
Type.of.HouseholdSingle Family                       -3.467e-01  2.291e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.058e-01  1.809e-01
House.Floor.Area                                     -4.940e-04  3.402e-04
House.Age                                            -3.715e-03  1.030e-03
Number.of.bedrooms                                    5.011e-02  1.234e-02
ElectricityYes                                       -9.028e-02  2.850e-02
                                                     z value Pr(>|z|)
(Intercept)                                           26.210  < 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
```

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

	2.5 %	97.5 %
(Intercept)	1.4777012	1.7166106
Total.Household.Income	-0.0000004	-0.0000001

	2.5 %	97.5 %
Total.Food.Expenditure	0.0000026	0.0000033
Household.Head.SexMale	0.2036003	0.3232971
Household.Head.Age	-0.0053862	-0.0022092
Type.of.HouseholdSingle Family	-0.3915529	-0.3017466
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.4820181	0.2294578
House.Floor.Area	-0.0011694	0.0001642
House.Age	-0.0057424	-0.0017039
Number.of.bedrooms	0.0259109	0.0742825
ElectricityYes	-0.1458759	-0.0341516

```
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] "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, 0 as baseline.

From the above summary we can observe that one continous explanaroty 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.

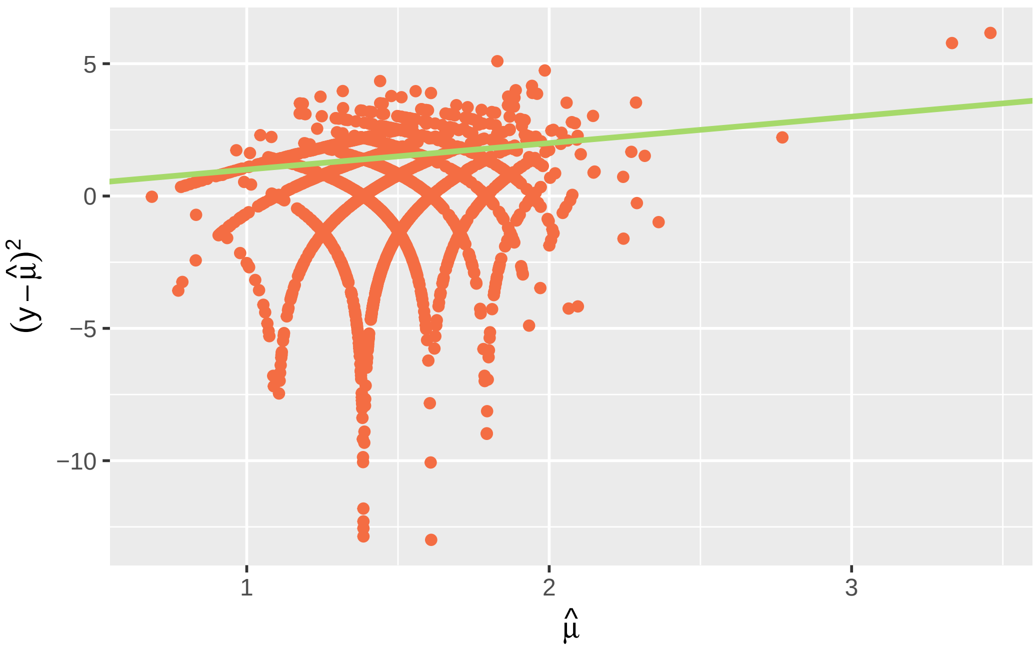
So we can remove the house floor area variable firstly.

```
model2=glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Ho
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.Age + Number.of.bedrooms + Electricity,
    family = poisson, data = data)

Coefficients:
                                                         Estimate Std. Error
(Intercept)                                             1.596e+00  6.094e-02
Total.Household.Income                                 -2.532e-07  5.538e-08
Total.Food.Expenditure                                  2.935e-06  1.881e-07
Household.Head.SexMale                                  2.634e-01  3.053e-02
Household.Head.Age                                     -3.852e-03  8.096e-04
Type.of.HouseholdSingle Family                         -3.470e-01  2.291e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.019e-01  1.809e-01
House.Age                                              -3.760e-03  1.030e-03
Number.of.bedrooms                                      4.445e-02  1.172e-02
ElectricityYes                                         -9.133e-02  2.849e-02
                                                       z value Pr(>|z|)
(Intercept)                                             26.199  < 2e-16 ***
Total.Household.Income                                  -4.572 4.82e-06 ***
Total.Food.Expenditure                                  15.599  < 2e-16 ***
Household.Head.SexMale                                   8.629  < 2e-16 ***
Household.Head.Age                                      -4.757 1.96e-06 ***
Type.of.HouseholdSingle Family                         -15.150  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.563 0.573307
House.Age                                               -3.651 0.000261 ***
Number.of.bedrooms                                       3.795 0.000148 ***
ElectricityYes                                          -3.206 0.001346 **
---
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: 1554.0  on 2112  degrees of freedom
AIC: 8512

Number of Fisher Scoring iterations: 5
```

Overdispersion

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



From the above scatterplot of mean and variance, 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.

Quasi-Poisson model

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

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

```r
X2=sum(resid(model2,type="pearson")^2)
dp=X2/model2$df.res
#With the use of the estimated dispersion parameter the Wald tests are not very reliable,
drop1(model2,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.Age + Number.of.bedrooms + Electricity
                       Df Deviance    AIC F value    Pr(>F)
<none>                      1554.0 8512.0
Total.Household.Income  1   1576.3 8532.4  30.340 4.068e-08 ***
Total.Food.Expenditure  1   1739.5 8695.5 252.114 < 2.2e-16 ***
Household.Head.Sex      1   1632.8 8588.8 107.072 < 2.2e-16 ***
Household.Head.Age      1   1576.7 8532.7  30.870 3.107e-08 ***
Type.of.Household       2   1777.5 8731.5 151.879 < 2.2e-16 ***
House.Age               1   1567.5 8523.6  18.406 1.865e-05 ***
Number.of.bedrooms      1   1568.3 8524.4  19.497 1.058e-05 ***
Electricity             1   1564.1 8520.2  13.750 0.0002143 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After we try to fit a quasi-poisson model, the summary still shows all the remaining variables are significant.

Negative binomial models

```r
model3=glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure
summary(model3)
```

```
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, init.theta = 76118.91171, link = log)
```

```
Coefficients:
                                                         Estimate Std. Error
(Intercept)                                             1.596e+00  6.094e-02
Total.Household.Income                                 -2.533e-07  5.538e-08
Total.Food.Expenditure                                  2.935e-06  1.881e-07
Household.Head.SexMale                                  2.634e-01  3.053e-02
Household.Head.Age                                     -3.852e-03  8.096e-04
Type.of.HouseholdSingle Family                         -3.470e-01  2.291e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.019e-01  1.809e-01
House.Age                                              -3.760e-03  1.030e-03
Number.of.bedrooms                                      4.445e-02  1.172e-02
ElectricityYes                                         -9.134e-02  2.849e-02
                                                       z value Pr(>|z|)
(Intercept)                                             26.198  < 2e-16 ***
Total.Household.Income                                  -4.573 4.82e-06 ***
Total.Food.Expenditure                                  15.599  < 2e-16 ***
Household.Head.SexMale                                   8.629  < 2e-16 ***
Household.Head.Age                                      -4.757 1.96e-06 ***
Type.of.HouseholdSingle Family                         -15.149  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.563 0.573340
House.Age                                               -3.651 0.000261 ***
Number.of.bedrooms                                       3.794 0.000148 ***
ElectricityYes                                          -3.206 0.001346 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(76118.86) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1553.9  on 2112  degrees of freedom
AIC: 8514.1

Number of Fisher Scoring iterations: 1

              Theta:  76119
          Std. Err.:  280215
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -8492.056
```

```r
c(model2$deviance,model2$aic)
```

```
[1] 1553.980 8512.034
```

```r
c(model3$deviance,model3$aic)
```

```
[1] 1553.876 8514.056
```

final model

The final model is:

$Total.Number.of.Family.members = \beta_0 + \beta_1 \cdot Total.Household.Income + \beta_2 \cdot Total.Food.Expenditure + \beta_3 \cdot \mathbb{I}_{\text{Mal}}$

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

$$\mathbb{I}_{\text{Family}}(x) = \begin{cases} 1 & \text{Single family,} \\ 0 & \text{Otherwise.} \end{cases}$$

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

For extended family and two or more nonrelated persons/members, the final model is:

$Total.Number.of.Family.members = 1.596 - 2.532 \times 10^{-7} \cdot Total.Household.Income + 2.953 \times 10^{-6} \cdot Total.Food.E$

For single family, the final model is:

$Total.Number.of.Family.members = 1.596 - 2.532 \times 10^{-7} \cdot Total.Household.Income + 2.953 \times 10^{-6} \cdot Total.Food.E$