Group_05_Analysis

```r
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(skimr)
library(GGally)
library(MASS)
library(dplyr)
library(knitr)
library(gridExtra)
library(kableExtra)
library(ggplot2)
library(glm2)
library(gridExtra)
library(grid)
library(knitr)
```

1 Data wrangling

```r
# Load the CSV file
data5 <-  read.csv('dataset05.csv')

#Simplify variable name
colnames(data5)<-c('Income','Region','Expenditure','Sex',
                   'Household.age','Type','Members','Area','House.age',
                   'Bedrooms','Electricity')
```

```r
# Display the structure of the dataframe
str(data5)
```

```
'data.frame':   1788 obs. of  11 variables:
 $ Income       : int  192922 161843 535899 312579 154715 107244 141249 113946 115508 153087
 $ Region       : chr  "IX - Zasmboanga Peninsula" "IX - Zasmboanga Peninsula" "IX - Zasmboa
 $ Expenditure  : int  114258 78176 92464 133445 39580 58182 47960 53999 42241 54140 ...
 $ Sex          : chr  "Male" "Male" "Male" "Male" ...
 $ Household.age: int  56 66 46 46 46 40 57 61 45 32 ...
 $ Type         : chr  "Single Family" "Extended Family" "Single Family" "Extended Family" .
 $ Members      : int  3 6 4 14 3 6 3 5 8 5 ...
 $ Area         : int  32 24 12 49 32 20 35 12 12 11 ...
 $ House.age    : int  28 6 3 21 1 2 39 4 4 5 ...
 $ Bedrooms     : int  2 0 1 3 2 0 2 1 1 1 ...
 $ Electricity  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
# Convert specified columns to categorical factors
data5$Region <- as.factor(data5$Region)
data5$Household.Head.Sex <- as.factor(data5$Sex)
data5$Type.of.Household <- as.factor(data5$Type)
data5$Electricity <- as.factor(data5$Electricity)


# Provide a concise summary of the dataframe
skim(data5)
```

Table 1: Data summary

Name	data5
Number of rows	1788
Number of columns	13
Column type frequency:	
character	2
factor	4
numeric	7
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Sex	0	1	4	6	0	2	0
Type	0	1	13	38	0	3	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Region	0	1	FALSE	1	IX : 1788
Electricity	0	1	FALSE	2	1: 1445, 0: 343
Household.Head.Sex	0	1	FALSE	2	Mal: 1434, Fem: 354
Type.of.Household	0	1	FALSE	3	Sin: 1254, Ext: 522, Two: 12

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Income	0	1	191000.91	238229.27	19730	85263.50	126567	205488.8	6071030	
Expenditure	0	1	69645.32	44465.59	5408	40500.25	59256	87221.0	434881	
Household.age	0	1	50.95	13.93	16	41.00	50	61.0	98	
Members	0	1	4.55	2.19	1	3.00	4	6.0	15	
Area	0	1	38.42	37.58	7	18.00	28	45.0	638	
House.age	0	1	16.00	11.50	0	7.00	15	20.0	93	
Bedrooms	0	1	1.73	1.00	0	1.00	2	2.0	6	

The Philippine government conducts a survey every three years to obtain data on household income and expenditure. Our goal is to explore what family-related variables influence the number of people living in a household, utilizing five data sets from different parts of the Philippines.

Description:

- Income: Total.Household.Income
- Expenditure: Total.Food.Expenditure
- Sex: Household.Head.Sex
- Household.age: Household.Head.Age
- Type: Type.of.Household
- Members: Type.of.Household

- Area: House.Floor.Area
- House.age: House.Floor.Area
- Bedrooms: House.Floor.Area

2 Exploratory Data Analysis

This section mainly carries out some exploratory analysis of data and data visualization.

First, the statistical summary of the data is performed. It can be seen that data distribution of each variable from the results, such as minimum value, maximum value, quartile, etc.

```
# Display the summary statistics of the data5
summary(data5)
```

```
     Income                                    Region      Expenditure
 Min.   :  19730   IX - Zasmboanga Peninsula:1788   Min.   :  5408
 1st Qu.:  85264                                    1st Qu.: 40500
 Median : 126567                                    Median : 59256
 Mean   : 191001                                    Mean   : 69645
 3rd Qu.: 205489                                    3rd Qu.: 87221
 Max.   :6071030                                    Max.   :434881
     Sex            Household.age       Type             Members
 Length:1788        Min.   :16.00   Length:1788        Min.   : 1.000
 Class :character   1st Qu.:41.00   Class :character   1st Qu.: 3.000
 Mode  :character   Median :50.00   Mode  :character   Median : 4.000
                    Mean   :50.95                      Mean   : 4.552
                    3rd Qu.:61.00                      3rd Qu.: 6.000
                    Max.   :98.00                      Max.   :15.000
      Area           House.age      Bedrooms     Electricity Household.Head.Sex
 Min.   :  7.00   Min.   : 0   Min.   :0.000   0: 343      Female: 354
 1st Qu.: 18.00   1st Qu.: 7   1st Qu.:1.000   1:1445      Male  :1434
 Median : 28.00   Median :15   Median :2.000
 Mean   : 38.42   Mean   :16   Mean   :1.732
 3rd Qu.: 45.00   3rd Qu.:20   3rd Qu.:2.000
 Max.   :638.00   Max.   :93   Max.   :6.000
                             Type.of.Household
 Extended Family                     : 522
 Single Family                       :1254
 Two or More Nonrelated Persons/Members:  12
```

Then histograms are drawn for the continuous variables to visually identify their distribution. It can be seen from Figure 1 that Income, Expenditure, Area, House.age have relatively serious skewed distributions.

```
# Create histogram plots for continuous variables
p11 <- ggplot(data5,aes(x = Income)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
p12 <- ggplot(data5,aes(x = Expenditure)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
p13 <- ggplot(data5,aes(x = Household.age)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
p14 <- ggplot(data5,aes(x = Members)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
p15 <- ggplot(data5,aes(x = Area)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
p16 <- ggplot(data5,aes(x = House.age)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
p17 <- ggplot(data5,aes(x = Bedrooms)) +
        geom_histogram(bins = 30, color="white",fill="steelblue")
grid.arrange(p11, p12, p13, p14, p15, p16, p17, ncol=3,
             top=textGrob('Histogram plots for continuous variables'))
```

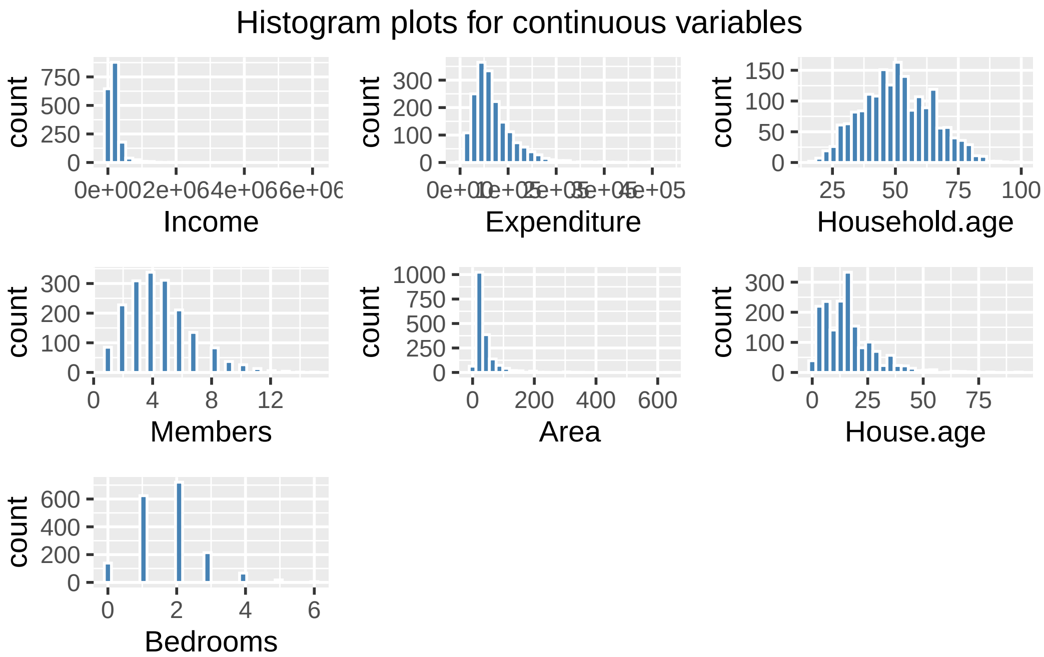


Figure 1: Histogram plots for continuous variables

Then the bar charts are drawn for the categorical variables to visually display the number of each category. It can be seen from Figure 2, male householders are three times more likely than female householders. In addition, those with electricity and single families account for the majority.

```
# Create bar plots for categorical variables
p21 <- ggplot(data5,aes(x = Sex)) +
        geom_bar(aes(fill = Sex))
p22 <- ggplot(data5,aes(x = Type)) +
        geom_bar(aes(fill = Type))+
  scale_x_discrete(labels=function(x)str_wrap(x,width= 2))
p23 <- ggplot(data5,aes(x = Electricity)) +
        geom_bar(aes(fill = Electricity))
grid.arrange(arrangeGrob(p21, p23,ncol=2) , p22, nrow=2,
             heights=c(5,8),
             top=textGrob('Bar plots for categorical variables'))
```

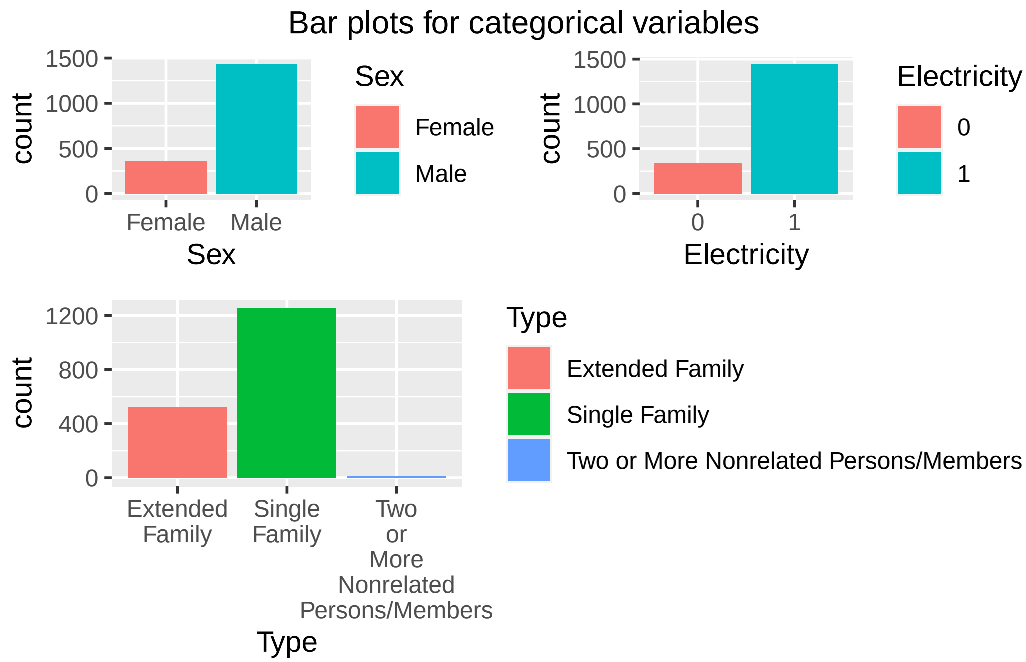


Figure 2: Bar plots for categorical variables

Furthermore, boxplots are drawn for the continuous variables, and it can also be seen from Figure 3 that Income, Expenditure, Area, House.age have relatively serious skewed distributions, And their outliers are almost all on the same side.

```
# Create boxplots for continuous variables
p31<-ggplot(data=data5,mapping=aes(y=Income))+
  geom_boxplot(fill="steelblue")+
  labs(y='Income')
p32<-ggplot(data=data5,mapping=aes(y=Expenditure))+
  geom_boxplot(fill="steelblue")+
  labs(y='Expenditure')
p33<-ggplot(data=data5,mapping=aes(y=Household.age))+
  geom_boxplot(fill="steelblue")+
  labs(y='Household age')
p34<-ggplot(data=data5,mapping=aes(y=Members))+
  geom_boxplot(fill="steelblue")+
  labs(y='Members')
p35<-ggplot(data=data5,mapping=aes(y=Area))+
  geom_boxplot(fill="steelblue")+
  labs(y='Area')
```

```r
p36<-ggplot(data=data5,mapping=aes(y=House.age))+
  geom_boxplot(fill="steelblue")+
  labs(y='House age')
p37<-ggplot(data=data5,mapping=aes(y=Bedrooms))+
  geom_boxplot(fill="steelblue")+
  labs(y='Bedrooms')
grid.arrange(p31, p32, p33, p34, p35, p36, p37, ncol=3,
             heights=c(8,8,8),
             top=textGrob('Boxplots for continuous variables'))
```

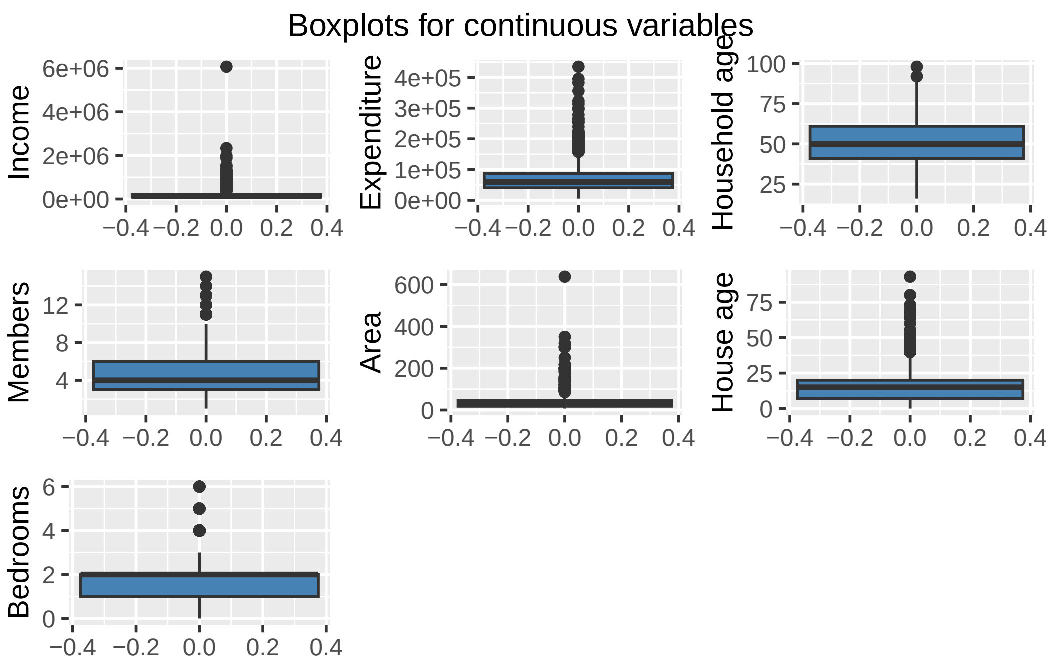


Figure 3: Boxplots for continuous variables

Finally, logarithm transformation is performed on these four skewed distributed variables, the new dataset is formed. Then, scatter plots are drawn for the predictor variables and response variables to determine their relationships. As can be seen from the Figure 4, Expenditure and Members show a relatively obvious correlation, Income, Household.age and Members have a weak correlation, and the remaining variables cannot see the obvious correlation.

```r
# Perform log transformation on selected variables
data5_log<-data5 %>%
  mutate(
```

```r
    log_Income=log(Income),
    log_Expenditure=log(Expenditure),
    log_Area=log(Area),
    log_House.age=log1p(House.age)
  )
data5_log<-data5_log[,c(-1,-3,-8,-9)]

# Create scatterplots for each predictor variable and response variable
p41<-ggplot(data=data5_log,aes(y=Members,
                               x=log_Income))+
  geom_point()+
  labs(x='Log.Income',y='Members')
p42<-ggplot(data=data5_log,aes(y=Members,
                               x=log_Expenditure))+
  geom_point()+
  labs(x='Log.Expenditure',y='Members')
p43<-ggplot(data=data5_log,aes(y=Members,
                               x=Household.age))+
  geom_point()+
  labs(x='Household age',y='Members')
p44<-ggplot(data=data5_log,aes(y=Members,
                               x=log_Area))+
  geom_point()+
  labs(x='Log.Area',y='Members')
p45<-ggplot(data=data5_log,aes(y=Members,
                               x=log_House.age))+
  geom_point()+
  labs(x='Log.House age',y='Members')
p46<-ggplot(data=data5_log,aes(y=Members,
                               x=Bedrooms))+
  geom_point()+
  labs(x='Number of bedrooms',y='Members')
grid.arrange(p41, p42, p43, p44, p45, p46, ncol=3,
             top=textGrob('Scatterplots for each predictor variable and response variable
                                                  '))
```

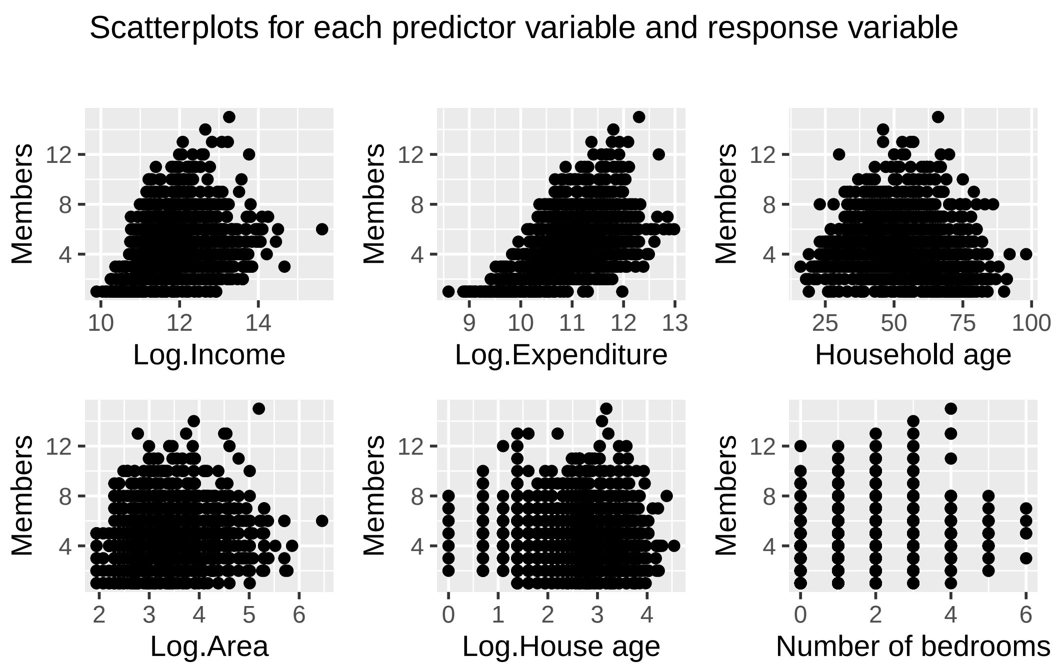


Figure 4: Scatterplots for each predictor variable and response variable

Additionally, draw a relationship diagram as shown in Figure 5.

```
ggpairs(data5_log,upper=list(continuous=wrap("points", alpha=0.4, color="#d73027")),
lower="blank", axisLabels="none")+
  ggtitle('The relationship between variables')+
  theme(plot.title = element_text(hjust = 0.5))
```

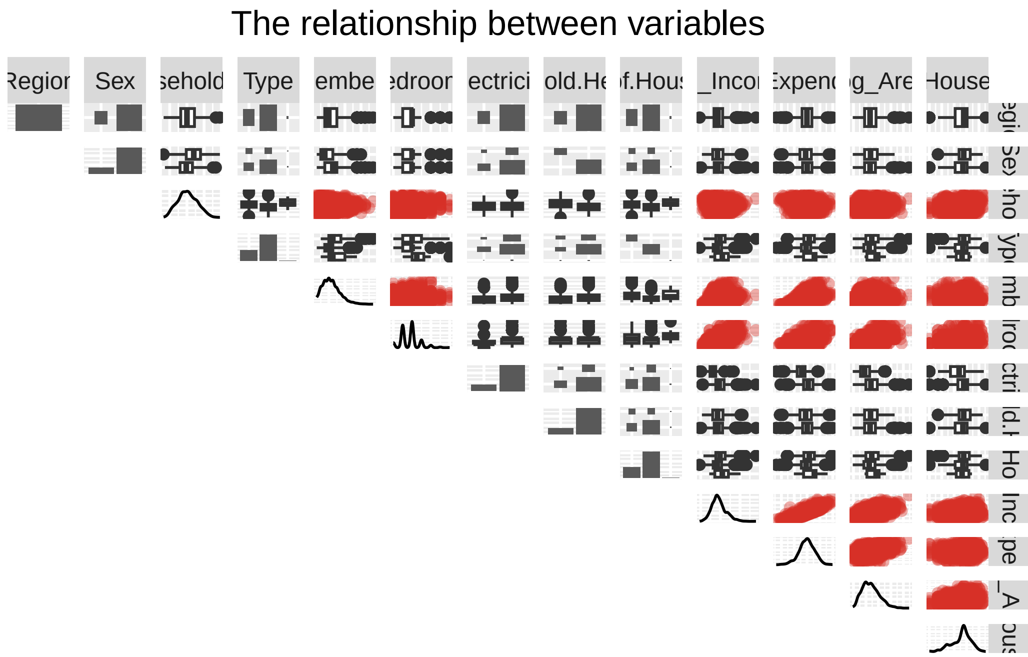


Figure 5: The relationship between variables

3 Model Construction(all variables)

By analyzing our data, we can see that the response variable is a count variable. To prevent the problem of underdispersion, we selected and compared four possible feasible models: the Poisson regression model, the generalized Poisson regression model, the negative binomial regression model, and the Quasi-Poisson regression.

3.1 Poisson regression model

```
library(car)
# Fit the Poisson regression model with the log link function
model_pois <- glm(Members ~
                      log_Income +
                      log_Expenditure +
                      Sex +
                      Household.age +
                      Type +
                      log_Area +
```

```r
                    log_House.age +
                    Bedrooms +
                    Electricity,
                    family=poisson(link="log"),
                    data = data5_log)

# Summarize the model
summary(model_pois)
```

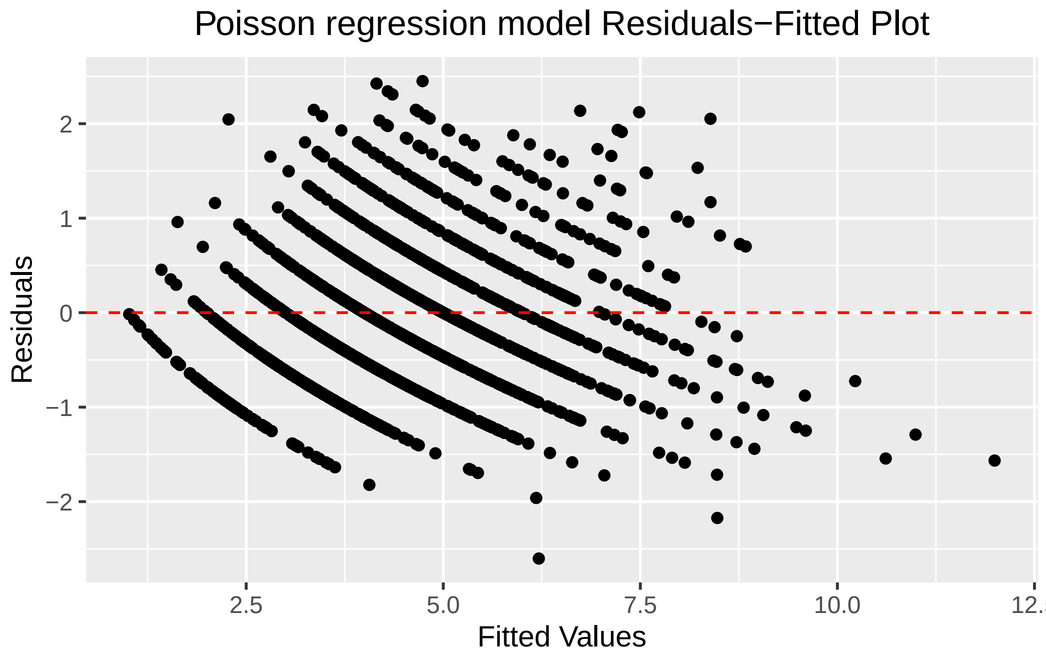
```

Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex + 
    Household.age + Type + log_Area + log_House.age + Bedrooms + 
    Electricity, family = poisson(link = "log"), data = data5_log)

Coefficients:
                                            Estimate Std. Error z value
(Intercept)                               -2.0730583  0.2569212  -8.069
log_Income                                -0.3805169  0.0368852 -10.316
log_Expenditure                            0.7611959  0.0418083  18.207
SexMale                                    0.1996045  0.0314011   6.357
Household.age                             -0.0023426  0.0009279  -2.525
TypeSingle Family                         -0.2911703  0.0254793 -11.428
TypeTwo or More Nonrelated Persons/Members  0.0125574  0.1206989   0.104
log_Area                                   0.0092809  0.0200343   0.463
log_House.age                             -0.0621789  0.0156432  -3.975
Bedrooms                                   0.0047464  0.0142395   0.333
Electricity1                              -0.0326431  0.0325815  -1.002
                                          Pr(>|z|)    
(Intercept)                               7.10e-16 ***
log_Income                                 < 2e-16 ***
log_Expenditure                            < 2e-16 ***
SexMale                                   2.06e-10 ***
Household.age                               0.0116 *  
TypeSingle Family                          < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members  0.9171    
log_Area                                    0.6432    
log_House.age                             7.04e-05 ***
Bedrooms                                    0.7389    
Electricity1                                0.3164    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 1854.6  on 1787  degrees of freedom
Residual deviance: 1033.7  on 1777  degrees of freedom
AIC: 6911

Number of Fisher Scoring iterations: 4
```

```r
summary_table11 <- as.data.frame(summary(model_pois)$coefficients)
kable(summary_table11, "html", digits = 2)%>%
  kable_styling(font_size = 12,latex_options =c('scale_down','hold_position'))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.07	0.26	-8.07	0.00
log_Income	-0.38	0.04	-10.32	0.00
log_Expenditure	0.76	0.04	18.21	0.00
SexMale	0.20	0.03	6.36	0.00
Household.age	0.00	0.00	-2.52	0.01
TypeSingle Family	-0.29	0.03	-11.43	0.00
TypeTwo or More Nonrelated Persons/Members	0.01	0.12	0.10	0.92
log_Area	0.01	0.02	0.46	0.64
log_House.age	-0.06	0.02	-3.97	0.00
Bedrooms	0.00	0.01	0.33	0.74
Electricity1	-0.03	0.03	-1.00	0.32

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_pois)
residuals_values <- residuals(model_pois)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title = 'Poisson regression model Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Calculate VIF to check for multicollinearity
vif(model_pois)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.514643  1        2.348328
log_Expenditure 4.659949  1        2.158691
Sex             1.094467  1        1.046168
Household.age   1.258154  1        1.121675
Type            1.239748  2        1.055196
log_Area        1.585886  1        1.259320
log_House.age   1.126629  1        1.061428
Bedrooms        1.698309  1        1.303192
Electricity     1.234713  1        1.111176
```

3.2 Generalized Poisson regression model

```
# Fit the generalized Poisson regression model
model_gp <- glm2(Members ~
                   log_Income +
                   log_Expenditure +
```

```r
                     Sex +
                     Household.age +
                     Type +
                     log_Area +
                     log_House.age +
                     Bedrooms +
                     Electricity,
                     family=poisson(link="log"),
                     data = data5_log)

# Summarize the model
summary(model_gp)
```

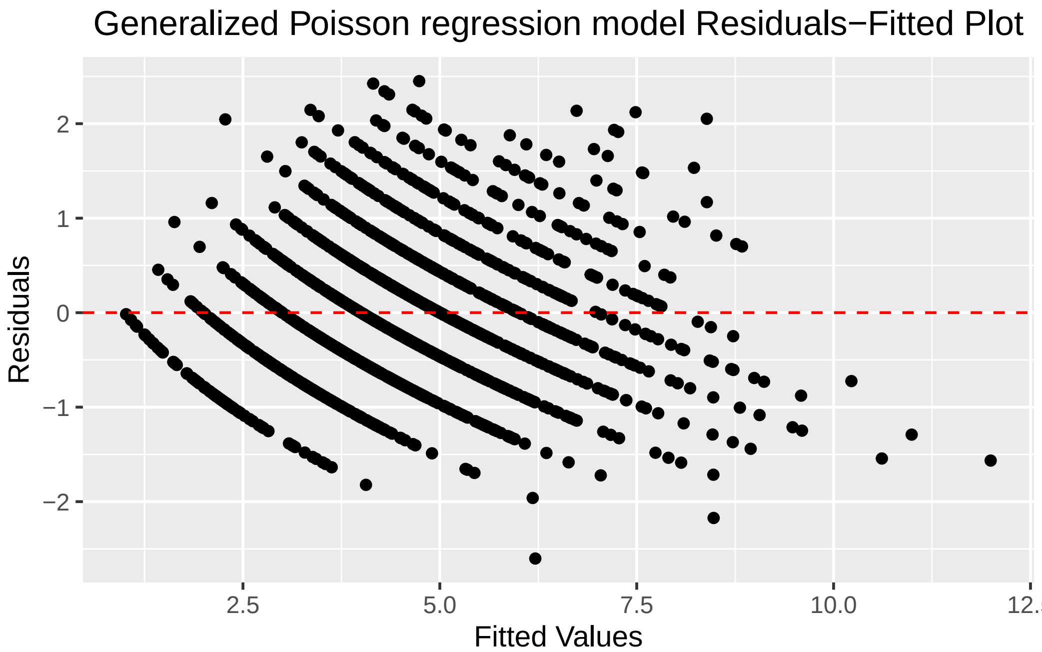
```
Call:
glm2(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_Area + log_House.age + Bedrooms +
    Electricity, family = poisson(link = "log"), data = data5_log)

Coefficients:
                                           Estimate Std. Error z value
(Intercept)                              -2.0730583  0.2569212  -8.069
log_Income                               -0.3805169  0.0368852 -10.316
log_Expenditure                           0.7611959  0.0418083  18.207
SexMale                                   0.1996045  0.0314011   6.357
Household.age                            -0.0023426  0.0009279  -2.525
TypeSingle Family                        -0.2911703  0.0254793 -11.428
TypeTwo or More Nonrelated Persons/Members  0.0125574  0.1206989   0.104
log_Area                                  0.0092809  0.0200343   0.463
log_House.age                            -0.0621789  0.0156432  -3.975
Bedrooms                                  0.0047464  0.0142395   0.333
Electricity1                             -0.0326431  0.0325815  -1.002
                                         Pr(>|z|)
(Intercept)                              7.10e-16 ***
log_Income                                < 2e-16 ***
log_Expenditure                           < 2e-16 ***
SexMale                                  2.06e-10 ***
Household.age                              0.0116 *
TypeSingle Family                         < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members   0.9171
log_Area                                   0.6432
log_House.age                            7.04e-05 ***
```

```
Bedrooms                                      0.7389
Electricity1                                  0.3164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1854.6  on 1787  degrees of freedom
Residual deviance: 1033.7  on 1777  degrees of freedom
AIC: 6911

Number of Fisher Scoring iterations: 4
```

```r
summary_table11 <- as.data.frame(summary(model_gp)$coefficients)
kable(summary_table11, "html", digits = 2)%>%
  kable_styling(font_size = 12,latex_options =c('scale_down','hold_position'))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.07	0.26	-8.07	0.00
log_Income	-0.38	0.04	-10.32	0.00
log_Expenditure	0.76	0.04	18.21	0.00
SexMale	0.20	0.03	6.36	0.00
Household.age	0.00	0.00	-2.52	0.01
TypeSingle Family	-0.29	0.03	-11.43	0.00
TypeTwo or More Nonrelated Persons/Members	0.01	0.12	0.10	0.92
log_Area	0.01	0.02	0.46	0.64
log_House.age	-0.06	0.02	-3.97	0.00
Bedrooms	0.00	0.01	0.33	0.74
Electricity1	-0.03	0.03	-1.00	0.32

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_gp)
residuals_values <- residuals(model_gp)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title = 'Generalized Poisson regression model Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Calculate VIF to check for multicollinearity
vif(model_gp)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.514643  1        2.348328
log_Expenditure 4.659949  1        2.158691
Sex             1.094467  1        1.046168
Household.age   1.258154  1        1.121675
Type            1.239748  2        1.055196
log_Area        1.585886  1        1.259320
log_House.age   1.126629  1        1.061428
Bedrooms        1.698309  1        1.303192
Electricity     1.234713  1        1.111176
```

The fitting results of the Poisson regression model and the generalized Poisson regression model are identical. The AIC value of them is 6911, relatively low, which indicates that the model explains the observed data well. Meanwhile, the null deviance is 1854.6, and the residual deviance is 1033.7. The smaller residual deviance suggests that the model has a good fit relative to the null model.

3.3 Negative binomial regression model

```r
# Fit the negative binomial regression model
  model_nb <- glm.nb(Members ~
                       log_Income +
                       log_Expenditure +
                       Sex +
                       Household.age +
                       Type +
                       log_Area +
                       log_House.age +
                       Bedrooms +
                       Electricity,
                       data = data5_log)

# Summarize the model
summary(model_nb)
```

```

Call:
glm.nb(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_Area + log_House.age + Bedrooms +
    Electricity, data = data5_log, init.theta = 120988.1938,
    link = log)

Coefficients:
                                         Estimate Std. Error z value
(Intercept)                            -2.0730716  0.2569267  -8.069
log_Income                             -0.3805173  0.0368859 -10.316
log_Expenditure                         0.7611978  0.0418092  18.206
SexMale                                 0.1996052  0.0314018   6.356
Household.age                          -0.0023426  0.0009279  -2.525
TypeSingle Family                      -0.2911711  0.0254799 -11.427
TypeTwo or More Nonrelated Persons/Members  0.0125558  0.1207021   0.104
log_Area                                0.0092809  0.0200348   0.463
log_House.age                          -0.0621793  0.0156436  -3.975
Bedrooms                                0.0047466  0.0142398   0.333
Electricity1                           -0.0326428  0.0325822  -1.002
                                       Pr(>|z|)
(Intercept)                            7.10e-16 ***
log_Income                              < 2e-16 ***
```

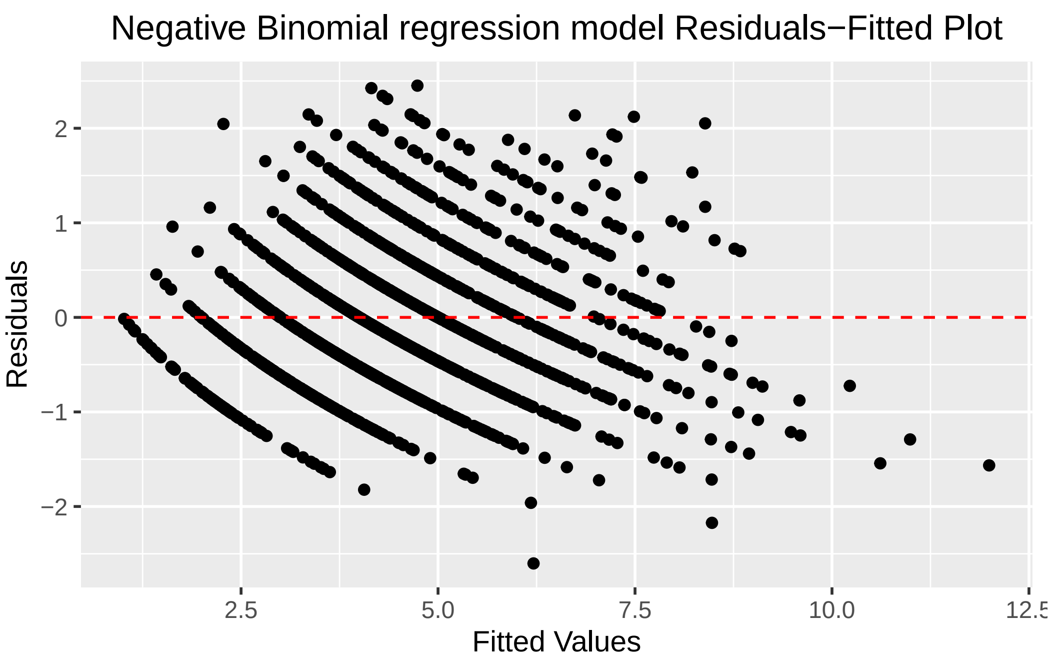
```
log_Expenditure                               < 2e-16 ***
SexMale                                     2.06e-10 ***
Household.age                                 0.0116 *
TypeSingle Family                             < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members    0.9172
log_Area                                      0.6432
log_House.age                               7.05e-05 ***
Bedrooms                                      0.7389
Electricity1                                  0.3164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(120988.2) family taken to be 1)

    Null deviance: 1854.6  on 1787  degrees of freedom
Residual deviance: 1033.6  on 1777  degrees of freedom
AIC: 6913.1

Number of Fisher Scoring iterations: 1

              Theta:  120988
          Std. Err.:  417483
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -6889.061
```

```r
summary_table11 <- as.data.frame(summary(model_nb)$coefficients)
kable(summary_table11, "html", digits = 2)%>%
  kable_styling(font_size = 12,latex_options =c('scale_down','hold_position'))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.07	0.26	-8.07	0.00
log_Income	-0.38	0.04	-10.32	0.00
log_Expenditure	0.76	0.04	18.21	0.00
SexMale	0.20	0.03	6.36	0.00
Household.age	0.00	0.00	-2.52	0.01
TypeSingle Family	-0.29	0.03	-11.43	0.00
TypeTwo or More Nonrelated Persons/Members	0.01	0.12	0.10	0.92
log_Area	0.01	0.02	0.46	0.64
log_House.age	-0.06	0.02	-3.97	0.00
Bedrooms	0.00	0.01	0.33	0.74

	Estimate	Std. Error	z value	Pr(>|z|)
Electricity1	-0.03	0.03	-1.00	0.32

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_nb)
residuals_values <- residuals(model_nb)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title = 'Negative Binomial regression model Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Calculate VIF to check for multicollinearity
vif(model_nb)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.514625  1        2.348324
log_Expenditure 4.659934  1        2.158688
```

```
Sex               1.094468  1         1.046168
Household.age     1.258155  1         1.121675
Type              1.239746  2         1.055196
log_Area          1.585885  1         1.259319
log_House.age     1.126630  1         1.061428
Bedrooms          1.698307  1         1.303191
Electricity       1.234713  1         1.111177
```

The AIC value of the negative binomial regression model is 6913.1, higher than the 6911 for the previous models. This suggests that the negative binomial regression model may not provide a better fit compared to the previous models, indicating potentially weaker explanatory power. Additionally, despite its higher AIC value, there is a slight decrease in residual deviance (1033.6) for the negative binomial regression model.

3.4 Quasi-Poisson regression model

```r
# Fit the Quasi-Poisson regression model
model_qp <- glm(Members ~
                   log_Income +
                   log_Expenditure +
                   Sex +
                   Household.age +
                   Type.of.Household +
                   log_Area +
                   log_House.age +
                   Bedrooms +
                   Electricity,
                   family=quasipoisson(link="log"),
                   data = data5_log)
# Summarize the model
summary(model_qp)
```

```
Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type.of.Household + log_Area + log_House.age +
    Bedrooms + Electricity, family = quasipoisson(link = "log"),
    data = data5_log)

Coefficients:
```

```
                                                 Estimate Std. Error
(Intercept)                                     -2.0730583  0.1992678
log_Income                                      -0.3805169  0.0286081
log_Expenditure                                  0.7611959  0.0324265
SexMale                                          0.1996045  0.0243547
Household.age                                   -0.0023426  0.0007197
Type.of.HouseholdSingle Family                  -0.2911703  0.0197617
Type.of.HouseholdTwo or More Nonrelated Persons/Members  0.0125574  0.0936139
log_Area                                         0.0092809  0.0155386
log_House.age                                   -0.0621789  0.0121329
Bedrooms                                         0.0047464  0.0110441
Electricity1                                    -0.0326431  0.0252702
                                                t value Pr(>|t|)
(Intercept)                                     -10.403  < 2e-16 ***
log_Income                                      -13.301  < 2e-16 ***
log_Expenditure                                  23.475  < 2e-16 ***
SexMale                                           8.196 4.71e-16 ***
Household.age                                    -3.255  0.00116 **
Type.of.HouseholdSingle Family                  -14.734  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members   0.134  0.89331
log_Area                                          0.597  0.55040
log_House.age                                    -5.125 3.30e-07 ***
Bedrooms                                          0.430  0.66742
Electricity1                                     -1.292  0.19661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.6015535)

    Null deviance: 1854.6  on 1787  degrees of freedom
Residual deviance: 1033.7  on 1777  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_qp)
residuals_values <- residuals(model_qp)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
```

```r
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title = 'Quasi-Possion regression model Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```


```r
# Calculate VIF to check for multicollinearity
vif(model_qp)
```

```
                      GVIF Df GVIF^(1/(2*Df))
log_Income        5.514643  1        2.348328
log_Expenditure   4.659949  1        2.158691
Sex               1.094467  1        1.046168
Household.age     1.258154  1        1.121675
Type.of.Household 1.239748  2        1.055196
log_Area          1.585886  1        1.259320
log_House.age     1.126629  1        1.061428
Bedrooms          1.698309  1        1.303192
Electricity       1.234713  1        1.111176
```

```r
# Perform an analysis of variance (ANOVA) to compare the different models fitted
anova(model_pois, model_gp, model_nb, model_qp, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
Model 2: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
Model 3: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
Model 4: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type.of.Household + log_Area + log_House.age + Bedrooms +
    Electricity
  Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1      1777     1033.7
2      1777     1033.7  0  0.000000
3      1777     1033.6  0  0.042386
4      1777     1033.7  0 -0.042386
```

From the summary of the Quasi-Possion regression, we can see that residual deviance is 1033.7, similar to the Poisson regression model.

Additionally, by analyzing the VIF values for models, the GVIF values for log_Total.Household.Income and log_Total.Food.Expenditure are 5.51 and 4.66, respectively, suggesting some degree of multicollinearity between these variables. The GVIF values for other variables range from 1.09 to 1.70, indicating relatively low correlations among them, which are unlikely to lead to multicollinearity problems. In summary, while some multicollinearity exists in the model, it does not appear to be severe enough to significantly impact the stability or interpretation of the model.

```r
model_compare_allvariables <- data.frame(
  Model = character(),
  AIC = numeric(),
  Deviance = numeric(),
  stringsAsFactors = FALSE)

model_compare_allvariables <- rbind(model_compare_allvariables,
                          c("Poisson", AIC(model_pois), deviance(model_pois)),
                          c("Generalized Poisson", AIC(model_gp), deviance(model_gp)),
                          c("Negative Binomial", AIC(model_nb), deviance(model_nb)),
```

```
                              c("Quasi-Poisson", AIC(model_qp), deviance(model_qp)))

names(model_compare_allvariables) <- c("Model", "AIC","deviance")
print(model_compare_allvariables)
```

```
               Model              AIC         deviance
1            Poisson 6911.03638159648 1033.68150395747
2 Generalized Poisson 6911.03638159648 1033.68150395747
3   Negative Binomial 6913.06126519013 1033.63911822879
4       Quasi-Poisson             <NA> 1033.68150395747
```

By establishing a table to compare the AIC values and Residual deviance of the previous models, it can be observed that the Poisson regression model with the full set of variables has the smallest AIC value and relatively lower Residual deviance.

4 Model Selection

In order to prevent overfitting, we split the data set into a training set and a test set.

```
set.seed(123)
train_index <- sample(seq_len(nrow(data5_log)),size = floor(0.8*nrow(data5_log)))
train_set <- data5_log[train_index, ]
test_set <- data5_log[train_index, ]
```

```
# Fit the Poisson regression model with the log link function
model_pois <- glm(Members ~
                   log_Income +
                   log_Expenditure +
                   Sex +
                   Household.age +
                   Type +
                   log_Area +
                   log_House.age +
                   Bedrooms +
                   Electricity,
                   family=poisson(link="log"),
                   data = train_set)
# Summarize the model
summary(model_pois)
```

```
Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_Area + log_House.age + Bedrooms +
    Electricity, family = poisson(link = "log"), data = train_set)

Coefficients:
                                          Estimate Std. Error z value
(Intercept)                              -2.028774   0.287998  -7.044
log_Income                               -0.358536   0.042051  -8.526
log_Expenditure                           0.732099   0.047364  15.457
SexMale                                   0.227061   0.035847   6.334
Household.age                            -0.002609   0.001043  -2.500
TypeSingle Family                        -0.297818   0.028608 -10.410
TypeTwo or More Nonrelated Persons/Members  0.022943   0.125482   0.183
log_Area                                  0.010683   0.022547   0.474
log_House.age                            -0.058620   0.017415  -3.366
Bedrooms                                  0.003827   0.015917   0.240
Electricity1                             -0.036914   0.036373  -1.015
                                         Pr(>|z|)
(Intercept)                              1.86e-12 ***
log_Income                                < 2e-16 ***
log_Expenditure                           < 2e-16 ***
SexMale                                  2.39e-10 ***
Household.age                            0.012416 *
TypeSingle Family                         < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members 0.854922
log_Area                                 0.635616
log_House.age                            0.000763 ***
Bedrooms                                 0.809992
Electricity1                             0.310164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1455.50  on 1429  degrees of freedom
Residual deviance:  813.03  on 1419  degrees of freedom
AIC: 5514.8

Number of Fisher Scoring iterations: 4
```

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_pois)
residuals_values <- residuals(model_pois)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title = 'Train set Poisson regression model Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```


```r
# Calculate VIF to check for multicollinearity
vif(model_pois)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.695973  1        2.386624
log_Expenditure 4.722522  1        2.173136
Sex             1.096541  1        1.047159
Household.age   1.273282  1        1.128398
Type            1.239572  2        1.055159
log_Area        1.650023  1        1.284532
```

```
log_House.age   1.134872  1        1.065304
Bedrooms        1.751089  1        1.323287
Electricity     1.222289  1        1.105572
```

```r
# Fit the generalized Poisson regression model
model_gp <- glm2(Members ~
                   log_Income +
                   log_Expenditure +
                   Sex +
                   Household.age +
                   Type +
                   log_Area +
                   log_House.age +
                   Bedrooms +
                   Electricity,
                   family=poisson(link="log"),
                   data = train_set)
# Summarize the model
summary(model_gp)
```

```
Call:
glm2(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_Area + log_House.age + Bedrooms +
    Electricity, family = poisson(link = "log"), data = train_set)

Coefficients:
                                            Estimate Std. Error z value
(Intercept)                                -2.028774   0.287998  -7.044
log_Income                                 -0.358536   0.042051  -8.526
log_Expenditure                             0.732099   0.047364  15.457
SexMale                                     0.227061   0.035847   6.334
Household.age                              -0.002609   0.001043  -2.500
TypeSingle Family                          -0.297818   0.028608 -10.410
TypeTwo or More Nonrelated Persons/Members  0.022943   0.125482   0.183
log_Area                                    0.010683   0.022547   0.474
log_House.age                              -0.058620   0.017415  -3.366
Bedrooms                                    0.003827   0.015917   0.240
Electricity1                               -0.036914   0.036373  -1.015
                                           Pr(>|z|)
(Intercept)                                1.86e-12 ***
```

```
log_Income                                     < 2e-16 ***
log_Expenditure                                < 2e-16 ***
SexMale                                      2.39e-10 ***
Household.age                                0.012416 *
TypeSingle Family                              < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members 0.854922
log_Area                                     0.635616
log_House.age                                0.000763 ***
Bedrooms                                     0.809992
Electricity1                                 0.310164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1455.50  on 1429  degrees of freedom
Residual deviance:  813.03  on 1419  degrees of freedom
AIC: 5514.8

Number of Fisher Scoring iterations: 4
```

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_gp)
residuals_values <- residuals(model_gp)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
  title = 'Train set Generalized Poisson regression model
       Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```


```r
# Calculate VIF to check for multicollinearity
vif(model_gp)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.695973  1        2.386624
log_Expenditure 4.722522  1        2.173136
Sex             1.096541  1        1.047159
Household.age   1.273282  1        1.128398
Type            1.239572  2        1.055159
log_Area        1.650023  1        1.284532
log_House.age   1.134872  1        1.065304
Bedrooms        1.751089  1        1.323287
Electricity     1.222289  1        1.105572
```

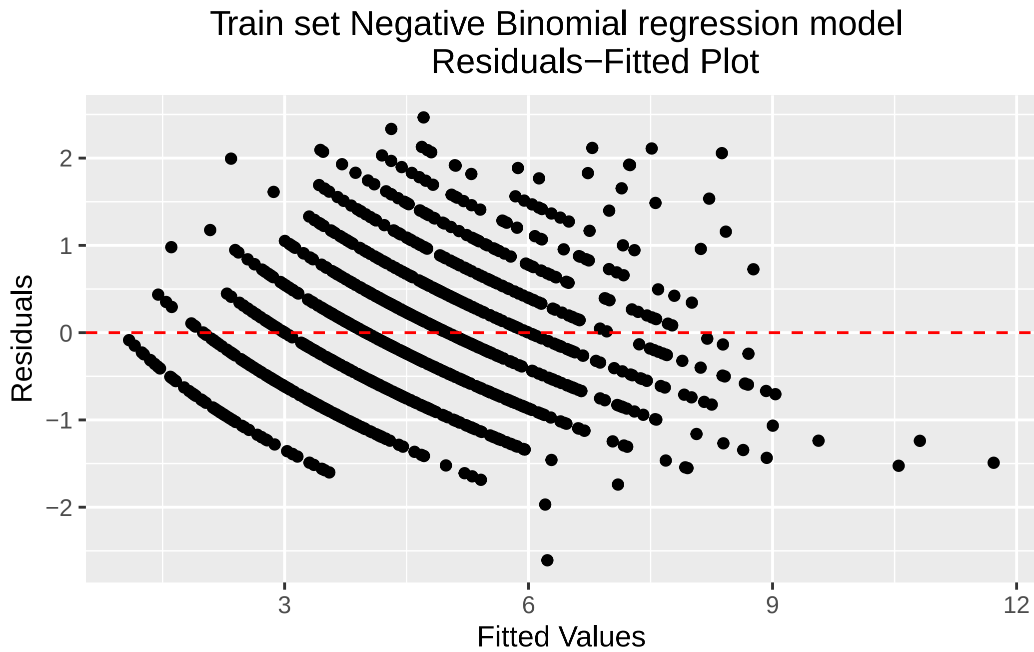
```r
# Fit the negative binomial regression model
model_nb <- glm.nb(Members ~
                     log_Income +
                     log_Expenditure +
                     Sex +
                     Household.age +
                     Type +
```

```r
                    log_Area +
                    log_House.age +
                    Bedrooms +
                    Electricity,
                    data = train_set)
# Summarize the model
summary(model_nb)
```

```
Call:
glm.nb(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_Area + log_House.age + Bedrooms +
    Electricity, data = train_set, init.theta = 124368.8299,
    link = log)

Coefficients:
                                            Estimate Std. Error z value
(Intercept)                                -2.028789   0.288004  -7.044
log_Income                                 -0.358536   0.042052  -8.526
log_Expenditure                             0.732101   0.047365  15.457
SexMale                                     0.227061   0.035848   6.334
Household.age                              -0.002609   0.001043  -2.500
TypeSingle Family                          -0.297819   0.028609 -10.410
TypeTwo or More Nonrelated Persons/Members  0.022942   0.125485   0.183
log_Area                                    0.010684   0.022547   0.474
log_House.age                              -0.058620   0.017415  -3.366
Bedrooms                                    0.003827   0.015917   0.240
Electricity1                               -0.036914   0.036374  -1.015
                                           Pr(>|z|)
(Intercept)                                1.86e-12 ***
log_Income                                  < 2e-16 ***
log_Expenditure                             < 2e-16 ***
SexMale                                    2.39e-10 ***
Household.age                              0.012417 *
TypeSingle Family                           < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members 0.854937
log_Area                                   0.635621
log_House.age                              0.000763 ***
Bedrooms                                   0.809992
Electricity1                               0.310179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for Negative Binomial(124368.8) family taken to be 1)

    Null deviance: 1455.4  on 1429  degrees of freedom
Residual deviance:  813.0  on 1419  degrees of freedom
AIC: 5516.9

Number of Fisher Scoring iterations: 1

              Theta:  124369
          Std. Err.:  480371
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -5492.864
```

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_nb)
residuals_values <- residuals(model_nb)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title =
         'Train set Negative Binomial regression model
       Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Calculate VIF to check for multicollinearity
vif(model_nb)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.695957  1        2.386620
log_Expenditure 4.722508  1        2.173133
Sex             1.096542  1        1.047159
Household.age   1.273282  1        1.128398
Type            1.239571  2        1.055159
log_Area        1.650022  1        1.284532
log_House.age   1.134872  1        1.065304
Bedrooms        1.751086  1        1.323286
Electricity     1.222289  1        1.105572
```

```r
# Fit the Quasi-Poisson regression model
model_qp <- glm(Members ~
                  log_Income +
                  log_Expenditure +
                  Sex +
                  Household.age +
                  Type +
```

```r
                     log_Area +
                     log_House.age +
                     Bedrooms +
                     Electricity,
                     family=quasipoisson(link="log"),
                     data = train_set)
summary(model_qp)
```

```
Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_Area + log_House.age + Bedrooms +
    Electricity, family = quasipoisson(link = "log"), data = train_set)

Coefficients:
                                         Estimate Std. Error t value
(Intercept)                            -2.0287735  0.2212836  -9.168
log_Income                             -0.3585357  0.0323101 -11.097
log_Expenditure                         0.7320985  0.0363919  20.117
SexMale                                 0.2270605  0.0275433   8.244
Household.age                          -0.0026085  0.0008017  -3.254
TypeSingle Family                      -0.2978181  0.0219811 -13.549
TypeTwo or More Nonrelated Persons/Members  0.0229432  0.0964138   0.238
log_Area                                0.0106834  0.0173237   0.617
log_House.age                          -0.0586195  0.0133807  -4.381
Bedrooms                                0.0038270  0.0122299   0.313
Electricity1                           -0.0369145  0.0279475  -1.321
                                       Pr(>|t|)
(Intercept)                             < 2e-16 ***
log_Income                              < 2e-16 ***
log_Expenditure                         < 2e-16 ***
SexMale                                3.76e-16 ***
Household.age                           0.00117 **
TypeSingle Family                       < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members  0.81194
log_Area                                0.53754
log_House.age                          1.27e-05 ***
Bedrooms                                0.75438
Electricity1                            0.18676
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for quasipoisson family taken to be 0.590362)

    Null deviance: 1455.50  on 1429  degrees of freedom
Residual deviance:  813.03  on 1419  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

```r
# Calculate and plot fitted values vs residuals for diagnostic checking
fitted_values <- fitted(model_qp)
residuals_values <- residuals(model_qp)
ggplot(data.frame(Fitted=fitted_values, Residuals=residuals_values),
       aes(x=Fitted, y=Residuals))+
  geom_point()+
  geom_hline(yintercept = 0,linetype = "dashed",color="red")+
  labs(x="Fitted Values",y="Residuals",
       title = 'Train set Quasi-Possion regression model
       Residuals-Fitted Plot')+
  theme(plot.title = element_text(hjust = 0.5))
```

```r
# Calculate VIF to check for multicollinearity
vif(model_qp)
```

```
                    GVIF Df GVIF^(1/(2*Df))
log_Income      5.695973  1        2.386624
log_Expenditure 4.722522  1        2.173136
Sex             1.096541  1        1.047159
Household.age   1.273282  1        1.128398
Type            1.239572  2        1.055159
log_Area        1.650023  1        1.284532
log_House.age   1.134872  1        1.065304
Bedrooms        1.751089  1        1.323287
Electricity     1.222289  1        1.105572
```

```r
# Perform an analysis of variance (ANOVA) to compare the different models fitted
anova(model_pois, model_gp, model_nb, model_qp, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
Model 2: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
Model 3: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
Model 4: Members ~ log_Income + log_Expenditure + Sex + Household.age +
    Type + log_Area + log_House.age + Bedrooms + Electricity
  Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1      1419     813.03
2      1419     813.03  0  0.000000
3      1419     813.00  0  0.032258
4      1419     813.03  0 -0.032258
```

In the comparison of the above four models with the full set of variables, the Poisson regression model has the smallest AIC(5514.8).

Meanwhile, four of the explanatory variables in the Poisson regression model do not appear to be statistically significant, as their p-values are greater than 0.05.

Therefore, we first remove three of the non-significant explanatory variables: House.Floor.Area, Number.of.bedrooms and Electricity.

```r
# For the Poisson regression model
pois_modified_1 <- glm(Members ~
                       log_Income +
                       log_Expenditure +
                       Sex +
                       Household.age +
                       Type +
                       log_House.age,
                       family = poisson(link="log"),
                       data = train_set)

summary(pois_modified_1)
```

```

Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + Type + log_House.age, family = poisson(link = "log"),
    data = train_set)

Coefficients:
                                           Estimate Std. Error z value
(Intercept)                               -2.022490   0.268014  -7.546
log_Income                                -0.354219   0.037539  -9.436
log_Expenditure                            0.727619   0.046755  15.562
SexMale                                    0.229203   0.035795   6.403
Household.age                             -0.002496   0.001030  -2.424
TypeSingle Family                         -0.298080   0.028414 -10.491
TypeTwo or More Nonrelated Persons/Members  0.026154   0.125049   0.209
log_House.age                             -0.059625   0.017187  -3.469
                                           Pr(>|z|)
(Intercept)                                4.48e-14 ***
log_Income                                  < 2e-16 ***
log_Expenditure                             < 2e-16 ***
SexMale                                    1.52e-10 ***
Household.age                              0.015345 *
TypeSingle Family                           < 2e-16 ***
TypeTwo or More Nonrelated Persons/Members 0.834330
log_House.age                              0.000522 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 1455.50  on 1429  degrees of freedom
Residual deviance:  814.33  on 1422  degrees of freedom
AIC: 5510.1

Number of Fisher Scoring iterations: 4
```

```
summary_table12 <- as.data.frame(summary(pois_modified_1)$coefficients)
kable(summary_table12, "html", digits = 2)%>%
  kable_styling(font_size = 12,latex_options =c('scale_down','hold_position'))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.02	0.27	-7.55	0.00
log_Income	-0.35	0.04	-9.44	0.00
log_Expenditure	0.73	0.05	15.56	0.00
SexMale	0.23	0.04	6.40	0.00
Household.age	0.00	0.00	-2.42	0.02
TypeSingle Family	-0.30	0.03	-10.49	0.00
TypeTwo or More Nonrelated Persons/Members	0.03	0.13	0.21	0.83
log_House.age	-0.06	0.02	-3.47	0.00

The AIC of this model is 5510.1, lower than the previous Poisson regression model with all variables. This indicates that this model explains the observed data better than before.

Then, we also attempted to remove the explanatory variable, type of household, for one of the categories of this variable is not significant compared with the reference level.

```
# remove type of household
# For the Poisson regression model
pois_modified_2 <- glm(Members ~
                   log_Income +
                   log_Expenditure +
                   Sex +
                   Household.age +
                   log_House.age,
                   family = poisson(link="log"),
                   data = train_set)
summary(pois_modified_2)
```

```
Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex +
    Household.age + log_House.age, family = poisson(link = "log"),
    data = train_set)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.9743901  0.2525377 -11.778  < 2e-16 ***
log_Income      -0.3809542  0.0372168 -10.236  < 2e-16 ***
log_Expenditure  0.8158851  0.0458723  17.786  < 2e-16 ***
SexMale          0.1753831  0.0354523   4.947 7.54e-07 ***
Household.age    0.0008129  0.0009801   0.829    0.407
log_House.age   -0.0700689  0.0171782  -4.079 4.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1455.5  on 1429  degrees of freedom
Residual deviance:  923.8  on 1424  degrees of freedom
AIC: 5615.6

Number of Fisher Scoring iterations: 4
```

The AIC value of this model is 5615.6, evidently higher than before. Therefore, it's no appropriate to remove it.

As for the type of household, two or more nonrelated persons/members shows no significant difference compared to the reference level (Extended family), so is_single_family is used to replaced the Type of household to make the model more concise and precise.

```r
train_set$is_single_family <- ifelse(train_set$Type == "Single Family", 1, 0)
# For the Poisson regression model
pois_modified_3 <- glm(Members ~
                         log_Income +
                         log_Expenditure +
                         Sex +
                         Household.age +
                         is_single_family +
                         log_House.age,
                       family=poisson(link="log"),
                       data = train_set)
```

```r
summary(pois_modified_3)
```

```

Call:
glm(formula = Members ~ log_Income + log_Expenditure + Sex + 
    Household.age + is_single_family + log_House.age, family = poisson(link = "log"), 
    data = train_set)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -2.024154   0.267908  -7.555 4.18e-14 ***
log_Income       -0.354038   0.037524  -9.435  < 2e-16 ***
log_Expenditure   0.727649   0.046750  15.565  < 2e-16 ***
SexMale           0.229083   0.035791   6.401 1.55e-10 ***
Household.age    -0.002492   0.001030  -2.420 0.015507 *  
is_single_family -0.298690   0.028258 -10.570  < 2e-16 ***
log_House.age    -0.059746   0.017177  -3.478 0.000504 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1455.50  on 1429  degrees of freedom
Residual deviance:  814.37  on 1423  degrees of freedom
AIC: 5508.2

Number of Fisher Scoring iterations: 4
```

From the summary of this model, the AIC is 5508.2, which is the smallest among all the models built. This suggests that the model explains the observed data well while maintaining higher predictive ability and model simplicity. Meanwhile, the residual deviance of this model is 814.37, slightly higher than the residual deviance of the first modified model (pois_modified_1) above. However, considering the degrees of freedom of this model is also higher than the previous model, the slight increase of the residual deviance is reasonable and acceptable.

The table of AIC and Residual deviances is presented below.

```r
model_comparison <- data.frame(
  Model = character(),
  AIC = numeric(),
  Deviance = numeric(),
```

```r
  stringsAsFactors = FALSE)

model_comparison <- rbind(model_comparison,
                          c("Poisson", AIC(model_pois),
                            deviance(model_pois)),
                          c("Negative Binomial", AIC(model_nb),
                            deviance(model_nb)),
                          c("Poisson_modified_1", AIC(pois_modified_1),
                            deviance(pois_modified_1)),
                          c("Poisson_modified_2", AIC(pois_modified_2),
                            deviance(pois_modified_2)),
                          c("Poisson_modified_3", AIC(pois_modified_3),
                            deviance(pois_modified_3)))

names(model_comparison) <- c("Model", "AIC","deviance")
print(model_comparison)
```

```
               Model              AIC         deviance
1            Poisson 5514.84398974086 813.029057570012
2  Negative Binomial 5516.86384990858 812.996799560048
3 Poisson_modified_1 5510.14151070816 814.326578537313
4 Poisson_modified_2 5615.61761451534 923.802682344499
5 Poisson_modified_3 5508.18489930866 814.369967137816
```

Therefore, taking into account the model's accuracy, interpretability, and simplicity, we regard the "Poisson_modified_3" model which has the smallest AIC 5508.18 as the best fit.

5 Model prediction performance

```r
# For the Poisson regression model contains all variables
predictions <- predict(model_pois,newdata = test_set, type = "response")
actuals <- test_set$Members
rmse <- sqrt(mean((predictions - actuals)^2))
print(rmse)
```

```
[1] 1.675006
```

```r
# For the Poisson regression model after stepwise removal
test_set$is_single_family <- ifelse(train_set$Type== "Single Family", 1, 0)
predictions <- predict(pois_modified_3, newdata = test_set, type = "response")
actuals <- test_set$Members
rmse <- sqrt(mean((predictions - actuals)^2))
print(rmse)
```

```
[1] 1.676832
```

According to the Root Mean Square Error value(RMSE), it can be seen that the difference between the predicted value and the actual value of the model is small, and the prediction performance of the model is better. However, It can be seen that the RMSE of the model containing all variables is smaller than that of the model after stepwise removal, 1.675<1.676. Therefore, there may be a slight overfitting issue for the model after stepwise removal.

6 Conclusion

In summary, among the household related variables, we think total household income, total food expenditure, household head sex, household head age, type of household (whether it is single family or not) and house age, these six variables will influence the number of people living in a household significantly. The specific model is as follows.

$$Members = -2.024 - 0.354 * log\_Income + 0.728 * log\_Expenditure + 0.229 * Sex -$$

$$0.002 * Household.age - 0.299 * is\_single\_family - 0.060 * log\_House.age$$

From the model results, the coefficients of log_Income, Household.age, is_single_family, and log_House.age are negative, while the coefficients of log_Expenditure and Sex are positive, indicating that as Annual household income, Head of the households age or Age of the building increases, the expected Number of people living in the house decreases. As Annual expenditure by the household on food increases, the expected Number of people living in the house also increases. In addition, the coefficient for SexMale is 0.229083, which implies that the expected Number of people living in the house for households headed by a male is higher compared to a female-headed household, all else being equal. Finally, the variable is_single_family has a negative coefficient (-0.298690), indicating that Single-family households are expected to have a lower Number of people living in their house compared to other household types.