Analysis of Factors Affecting Number of People in Household of Philippines

Siqi Wang, Yubin Lyu, Liting Wang, Han Xu, Jiahao Mao

Introduction

The Family Income and Expenditure Survey (FIES) is a significant source of data for understanding the wellbeing of households in Philippines. It provides valuable information on family income and expenditure, which can be used to investigate various research questions related to household characteristics.

In this analysis, we are interested in identifying which household-related factors influence the size of a household. Using Generalized Linear Model (GLM), we will explore the datasets obtained from the FIES survey for XII - SOCCSKSARGEN region in Philippines. The results of our analysis could help the government to make informed decisions related to household policies and other related matters.

Data Processing

```
# Load data
household <- read.csv("dataset4.csv")
# Factorize the categorical variables
household$Electricity <- as.factor(household$Electricity)
household$Household.Head.Sex <- as.factor(household$Household.Head.Sex)
household$Type.of.Household <- as.factor(household$Type.of.Household)
```

```
# Simplified column names
colnames(household)[1]<-"Income"            # original name "Total.Household.Income"
colnames(household)[3]<-"FoodExp"           # original name "Total.Food.Expenditure"
colnames(household)[4]<-"Householder_Sex"    # original name "Household.Head.Sex"
colnames(household)[5]<-"Householder_Age"    # original name "Household.Head.Age"
colnames(household)[6]<-"Household_Type"     # original name "Type.of.Household"
colnames(household)[7]<-"Number_Members"     # original name "Total.Number.of.Family.members"
colnames(household)[8]<-"Floorarea"          # original name "House.Floor.Area"
colnames(household)[10]<-"Number_bedrooms"   # original name "Number.of.bedrooms"
# change the long type name "Two or More Unrelated Persons/Members"
household$Household_Type <- ifelse(household$Household_Type == "Two or More Nonrelated Persons/Members"
```

Data Summary

```
# Summary of Categorical Variables
household_cat <- household %>%
  dplyr::select("Electricity","Householder_Sex","Household_Type")
summary(household_cat)
```

```
 Electricity Householder_Sex Household_Type
 0: 363      Female: 362     Length:2122
 1:1759      Male  :1760     Class :character
                             Mode  :character
```

```
# Summary of Numerical Variables
household_num <- household[, sapply(household, is.numeric)]
my_skim <- skim_with(base = sfl(n = length))
household_num %>%
  my_skim() %>%
  transmute(Variable=skim_variable, n=n,
            Mean = format(signif(numeric.mean, 3), scientific = TRUE, digits = 2),
            SD = format(signif(numeric.sd, 3), scientific = TRUE, digits = 2),
            Min= format(signif(numeric.p0, 3), scientific = TRUE, digits = 2),
            Median=format(signif(numeric.p50, 3), scientific = TRUE, digits = 2),
            Max=format(signif(numeric.p100, 3), scientific = TRUE, digits = 2),
            IQR = format(signif(numeric.p75-numeric.p50, 3), scientific = TRUE, digits = 2) ) %>%
   kable(caption = '\\label{tab:summarybyskim} Summary statistics of variables',
         booktabs = TRUE, linesep = "", digits = 2) %>%
   kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: Summary statistics of variables

Variable	n	Mean	SD	Min	Median	Max	IQR
Income	2122	1.8e+05	2.3e+05	1.5e+04	1.2e+05	3.2e+06	7.4e+04
FoodExp	2122	7.2e+04	4.5e+04	7.8e+03	6.3e+04	7.3e+05	2.4e+04
Householder_Age	2122	4.9e+01	1.4e+01	9.0e+00	4.8e+01	9.9e+01	1.1e+01
Number_Members	2122	4.5e+00	2.2e+00	1.0e+00	4.0e+00	1.9e+01	2.0e+00
Floorarea	2122	3.6e+01	3.5e+01	5.0e+00	2.6e+01	4.5e+02	1.4e+01
House.Age	2122	1.6e+01	1.1e+01	0.0e+00	1.4e+01	7.5e+01	7.0e+00
Number_bedrooms	2122	1.8e+00	1.0e+00	0.0e+00	2.0e+00	7.0e+00	0.0e+00

Distribution Check

Test if y follows the poisson distribution.

```
# check if the variable follows poisson distribution by testing if the variance equals to the mean
print(var(household$Number_Members))
```

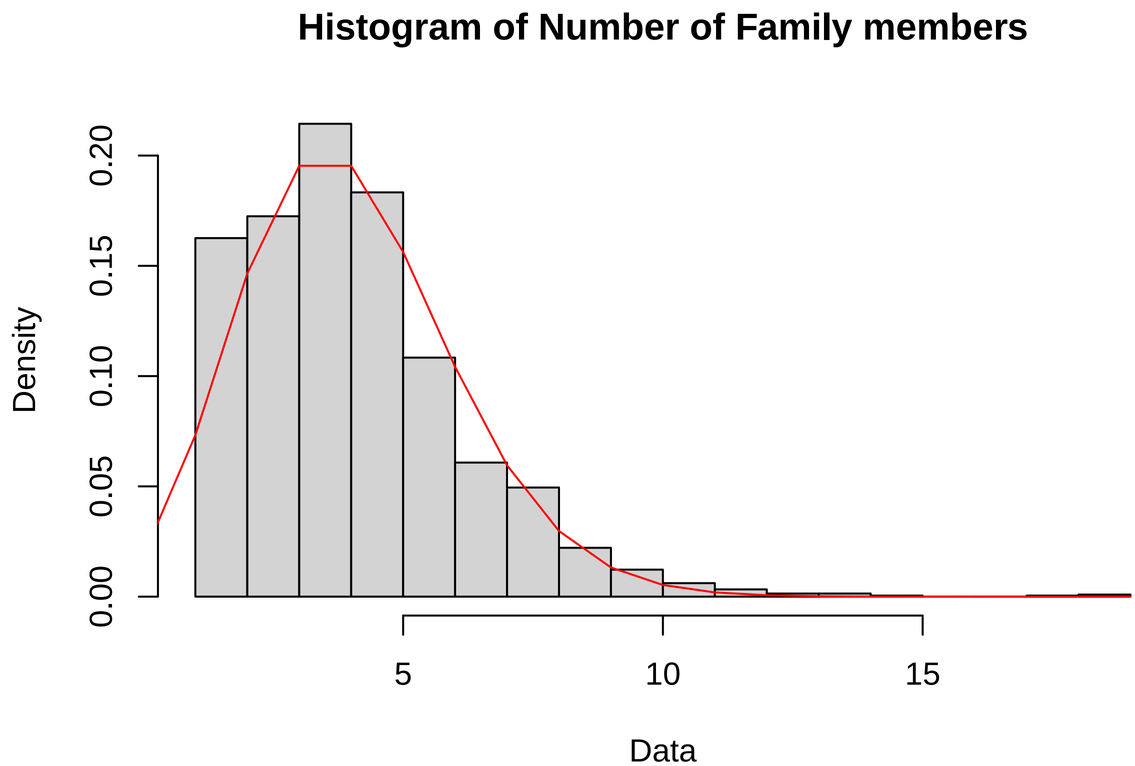
```
[1] 4.9
```

```
print(mean(household$Number_Members))
```

```
[1] 4.5
```

```
# As we can see there is no significant difference between mean and variance
# Plot the histogram of response variable and compare it with poisson distribution
hist(household$Number_Members, freq = FALSE, xlab = "Data",
```

```
    main = "Histogram of Number of Family members")
# Overlay a Poisson probability mass function
x <- 0:max(household$Number_Members)
lines(x, dpois(x, lambda = 4), col = "red")
```



From the plot it can be seen that the number of members in a household follows a poisson distribution with lambda = 4.

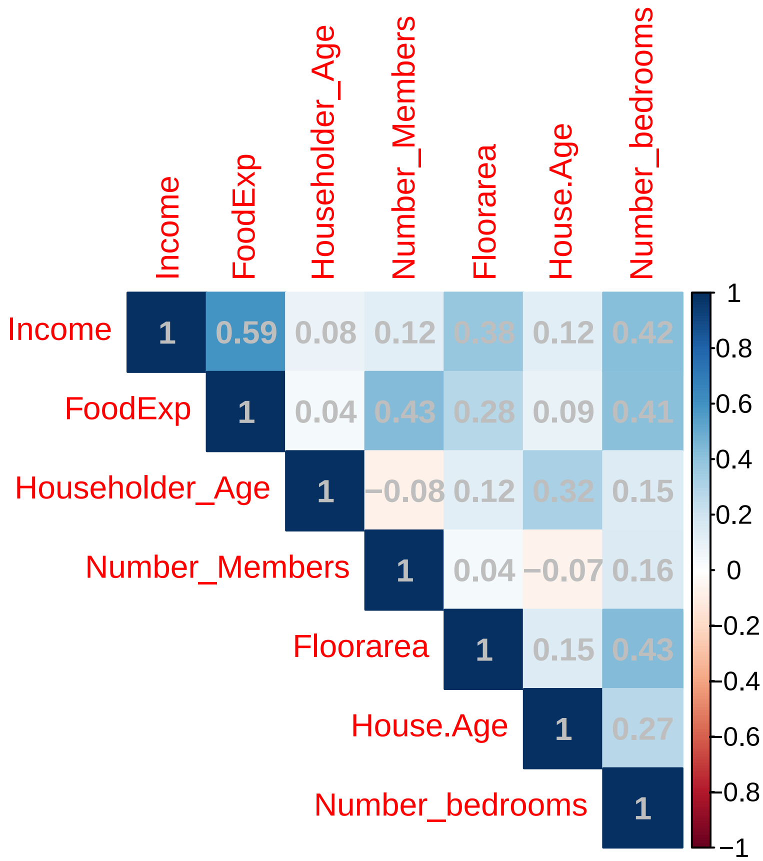
Correlation Matrix and ggpairs

```
# Create the correlation matrix of variables
cor_matrix <- cor(household_num) %>%
  round(2)
cor_matrix
```

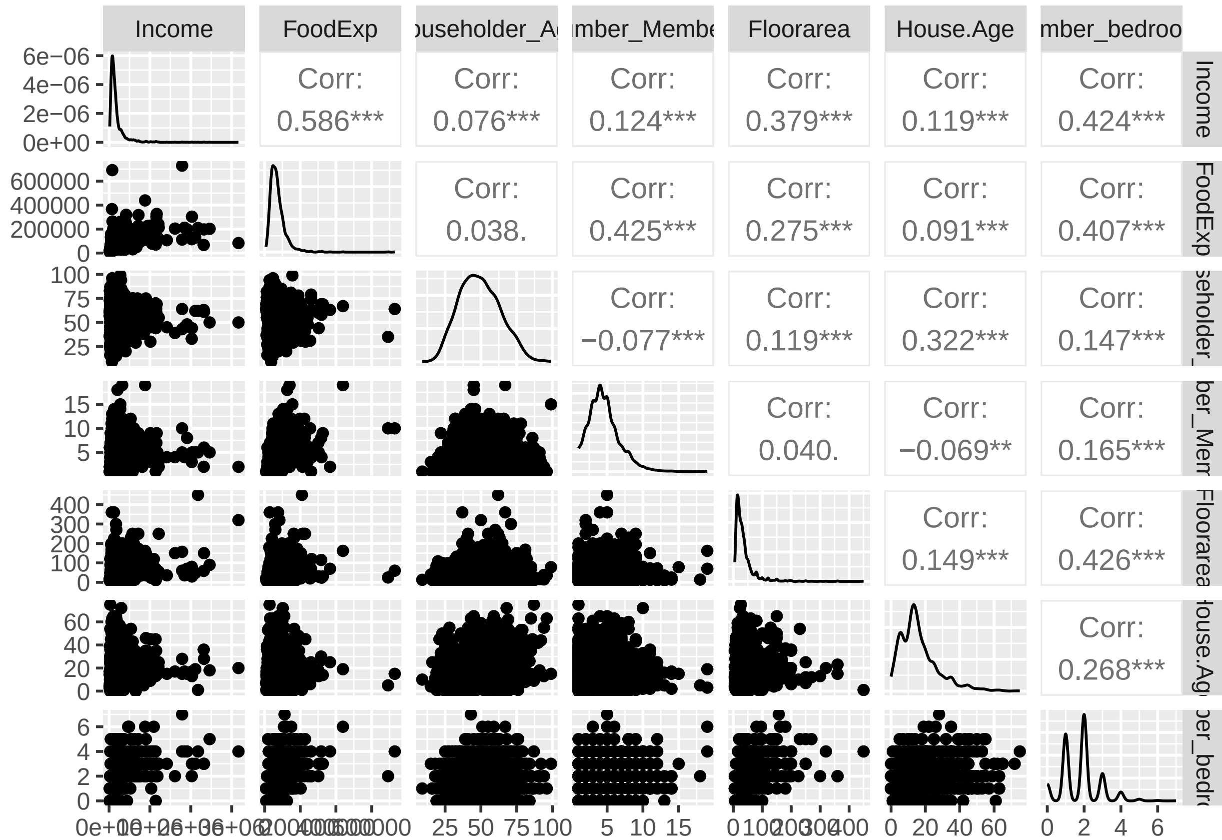
```
                Income FoodExp Householder_Age Number_Members Floorarea
Income            1.00    0.59            0.08           0.12      0.38
FoodExp           0.59    1.00            0.04           0.43      0.28
Householder_Age   0.08    0.04            1.00          -0.08      0.12
Number_Members    0.12    0.43           -0.08           1.00      0.04
Floorarea         0.38    0.28            0.12           0.04      1.00
House.Age         0.12    0.09            0.32          -0.07      0.15
Number_bedrooms   0.42    0.41            0.15           0.16      0.43
```

```
                 House.Age Number_bedrooms
Income                0.12            0.42
FoodExp               0.09            0.41
Householder_Age       0.32            0.15
Number_Members       -0.07            0.16
Floorarea             0.15            0.43
House.Age             1.00            0.27
Number_bedrooms       0.27            1.00
```

```
corrplot <- corrplot(cor_matrix, method = "color", addCoef.col = "gray",type = "upper",)
```

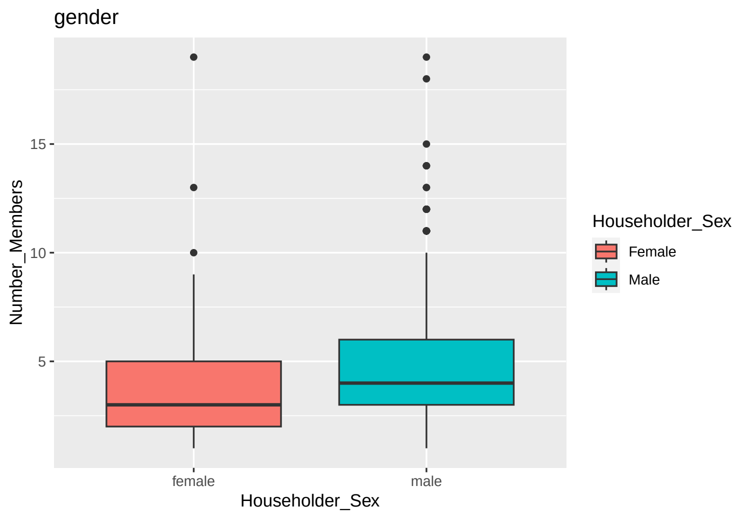



```
# Create the ggpairs of variables
ggpairs(household_num)
```

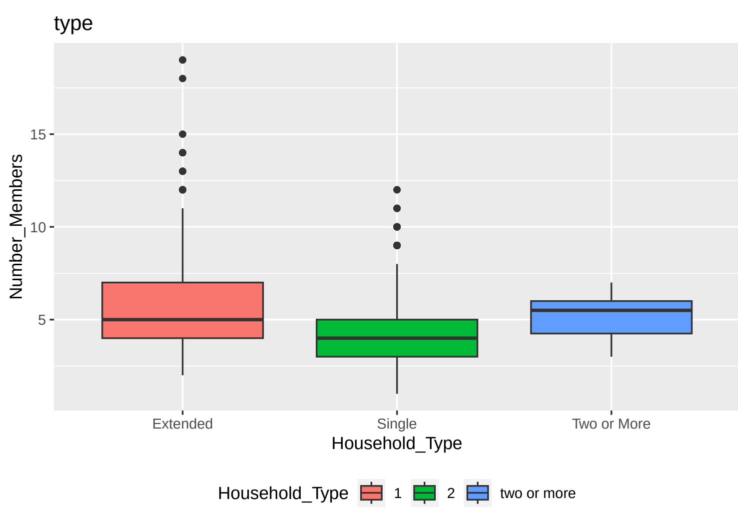


Boxplots of Variables

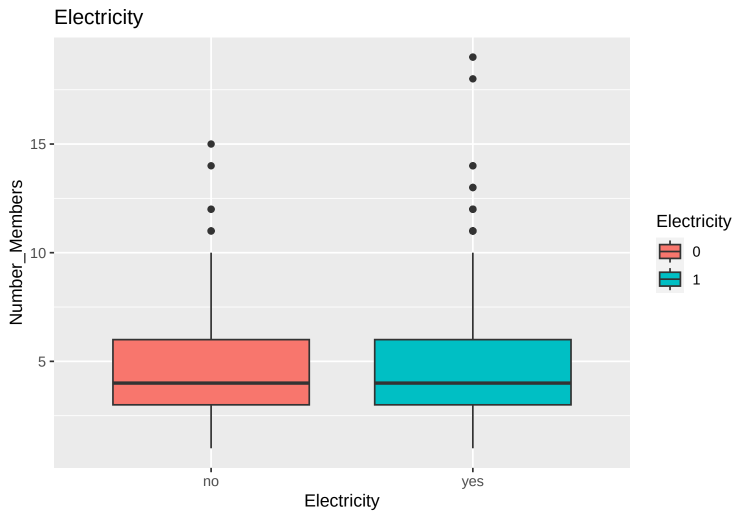
```
# Explanatory analysis on numeric variables
ggplot(data = household, mapping = aes(x = Householder_Sex, y = Number_Members)) +
  geom_boxplot(aes(fill =Householder_Sex ) )+
  labs(x = "Householder_Sex", y = "Number_Members",
      title = "gender")  +
  scale_x_discrete(labels = c("female","male"))
```




```
ggplot(data = household, mapping = aes(x = Household_Type, y = Number_Members)) +
  geom_boxplot(aes(fill = Household_Type))+
  labs(x = "Household_Type", y = "Number_Members",title = "type") +
  scale_x_discrete(labels = c("Extended", "Single", "Two or More"))+
  theme(legend.position = "bottom")
```




```
ggplot(data = household, mapping = aes(x = Electricity, y = Number_Members)) +
  geom_boxplot(aes(fill =Electricity ) )+
  labs(x = "Electricity", y = "Number_Members",
       title = "Electricity")  +
  scale_x_discrete(labels = c("no","yes"))
```

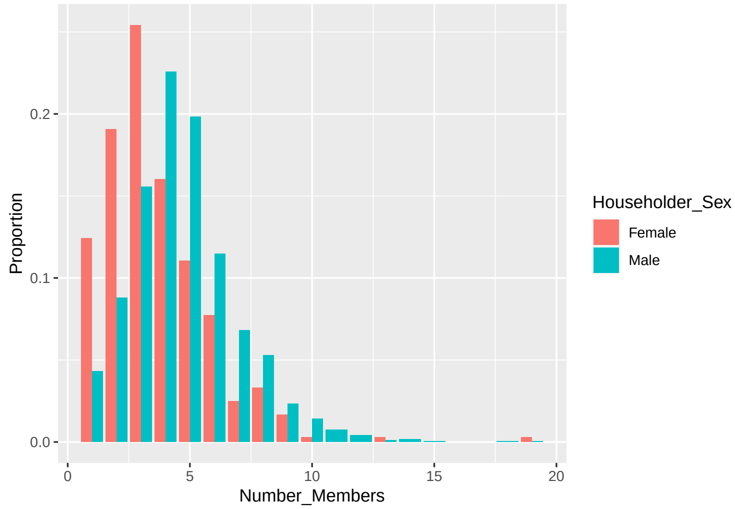


Histogram of Variables

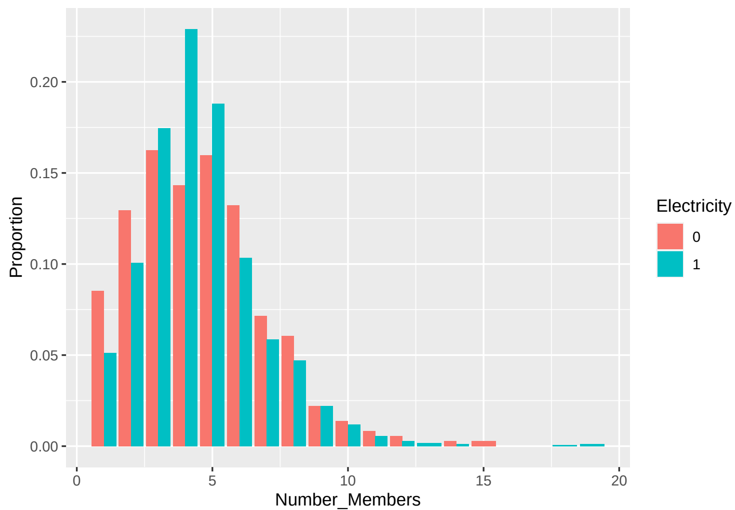
```
# Explanatory analysis on categorical variables
household %>%
  tabyl(Number_Members, Householder_Sex) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() # To show original counts
```

```
 Number_Members      Female          Male
              1 37.2% (45)  62.8%  (76)
              2 30.8% (69)  69.2% (155)
              3 25.1% (92)  74.9% (274)
              4 12.7% (58)  87.3% (397)
              5 10.3% (40)  89.7% (349)
              6 12.2% (28)  87.8% (202)
              7  7.0%  (9)  93.0% (120)
              8 11.4% (12)  88.6%  (93)
              9 12.8%  (6)  87.2%  (41)
             10  3.8%  (1)  96.2%  (25)
             11  0.0%  (0) 100.0%  (13)
             12  0.0%  (0) 100.0%   (7)
             13 33.3%  (1)  66.7%   (2)
             14  0.0%  (0) 100.0%   (3)
             15  0.0%  (0) 100.0%   (1)
             18  0.0%  (0) 100.0%   (1)
             19 50.0%  (1)  50.0%   (1)
```

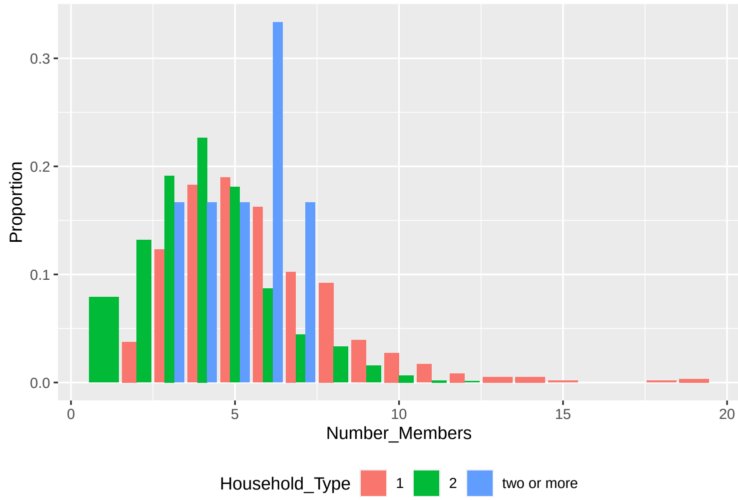
```
# barplot of Number_Members and Householder_Sex
ggplot(household,
       aes(x= Number_Members,  y = ..prop.., group=Householder_Sex, fill=Householder_Sex)) +
  geom_bar(position="dodge", stat="count") +
  labs(y = "Proportion")
```




```r
# barplot of Number_Members and Electricity
ggplot(household,
       aes(x= Number_Members,  y = ..prop.., group=Electricity, fill=Electricity)) +
  geom_bar(position="dodge", stat="count") +
  labs(y = "Proportion")
```




```r
# barplot of Number_Members and Household_Type
ggplot(household,
       aes(x= Number_Members,  y = ..prop.., group=Household_Type, fill=Household_Type)) +
  geom_bar(position="dodge", stat="count") +
  labs(y = "Proportion") +
  theme(legend.position = "bottom")
```



Model Fitting

Since our response variable follows poisson distribution, it seems reasonable to use GLM with poisson distribution to fit the model. However, the regression coefficients for "household income" and "food expenditure" were found to be smaller than expected, possibly due to the large scale of these variables. To address this issue, a log transformation was taken on these variables, which effectively normalized their scale and improved the accuracy of the regression coefficients.

```
#model1 with all variables
m1 <- glm(formula = Number_Members ~ Income + FoodExp + Householder_Sex +
            Householder_Age + Household_Type + Floorarea +
            House.Age + Number_bedrooms + Electricity,
          family = poisson(link = "log"), data = household)


#model2 with log transformation for Income and FoodExp
m2 <- glm(formula = Number_Members ~ log(Income) + log(FoodExp) +Householder_Sex +
            Householder_Age + Household_Type + Floorarea +
            House.Age + Number_bedrooms + Electricity,
          family = poisson(link = "log"), data = household)

summary(m1)
```

```
Call:
glm(formula = Number_Members ~ Income + FoodExp + Householder_Sex +
    Householder_Age + Household_Type + Floorarea + House.Age +
    Number_bedrooms + Electricity, family = poisson(link = "log"),
    data = household)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-4.523   -0.615   -0.113    0.423    4.115

Coefficients:
```

```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.60e+00   6.09e-02   26.21  < 2e-16 ***
Income                   -2.39e-07   5.63e-08   -4.23  2.3e-05 ***
FoodExp                   2.93e-06   1.88e-07   15.59  < 2e-16 ***
Householder_SexMale       2.63e-01   3.05e-02    8.62  < 2e-16 ***
Householder_Age          -3.80e-03   8.10e-04   -4.68  2.8e-06 ***
Household_Type2          -3.47e-01   2.29e-02  -15.13  < 2e-16 ***
Household_Typetwo or more -1.06e-01   1.81e-01   -0.59  0.55842
Floorarea                -4.94e-04   3.40e-04   -1.45  0.14648
House.Age                -3.71e-03   1.03e-03   -3.61  0.00031 ***
Number_bedrooms           5.01e-02   1.23e-02    4.06  4.9e-05 ***
Electricity1             -9.03e-02   2.85e-02   -3.17  0.00154 **
---
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: 8512

Number of Fisher Scoring iterations: 5
```

```
summary(m2)
```

```
Call:
glm(formula = Number_Members ~ log(Income) + log(FoodExp) + Householder_Sex +
    Householder_Age + Household_Type + Floorarea + House.Age +
    Number_bedrooms + Electricity, family = poisson(link = "log"),
    data = household)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.960  -0.557  -0.110   0.422   3.859

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.951300   0.248609  -11.87  < 2e-16 ***
log(Income)              -0.137026   0.021684   -6.32  2.6e-10 ***
log(FoodExp)              0.577842   0.029121   19.84  < 2e-16 ***
Householder_SexMale       0.203725   0.030685    6.64  3.2e-11 ***
Householder_Age          -0.002625   0.000823   -3.19  0.00142 **
Household_Type2          -0.288165   0.023185  -12.43  < 2e-16 ***
Household_Typetwo or more -0.035410   0.180946   -0.20  0.84485
Floorarea                -0.000904   0.000341   -2.65  0.00804 **
House.Age                -0.003815   0.001032   -3.70  0.00022 ***
Number_bedrooms           0.024816   0.012572    1.97  0.04840 *
Electricity1             -0.159250   0.029844   -5.34  9.5e-08 ***
---
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: 1299.4  on 2111  degrees of freedom
AIC: 8260

Number of Fisher Scoring iterations: 4
```

Use BIC to do variable selection

BIC is implemented to found the best fitting model. In the process of model selection, the posterior probability can be utilized to evaluate the impact of each explanatory variable on the response variable and to facilitate the selection of the best model. The results show that model 1 has the highest posterior probability of 0.61, suggesting that it is the most suitable model for explaining the response variable.

```r
output <- bic.glm(Number_Members ~ log(Income) + log(FoodExp) + Householder_Sex +
                    Householder_Age + Household_Type + Floorarea +
                    House.Age + Number_bedrooms + Electricity,
                   glm.family = "poisson" , data = household)

summary(output)
```

```

Call:
bic.glm.formula(f = Number_Members ~ log(Income) + log(FoodExp) +     Householder_Sex + Householder_Age


  5  models were selected
 Best  5  models (cumulative posterior probability =  1 ):

                             p!=0    EV         SD        model 1    model 2
Intercept                    100    -3.025085  0.246656  -2.97e+00  -3.10e+00
log(Income).x                100.0  -0.136207  0.021240  -1.37e-01  -1.28e-01
log(FoodExp).x               100.0   0.582453  0.029086   5.80e-01   5.84e-01
Householder_Sex.x            100.0
                 .Male               0.205620  0.031117   2.03e-01   2.01e-01
Householder_Age.x             79.8  -0.002132  0.001308  -2.66e-03  -2.53e-03
Household_Type2.x            100.0  -0.286849  0.023801  -2.90e-01  -2.89e-01
Household_Typetwo or more.x    0.0   0.000000  0.000000      .          .
Floorarea.x                   21.2  -0.000153  0.000331      .      -7.04e-04
House.Age.x                   96.5  -0.003704  0.001268  -3.69e-03  -3.53e-03
Number_bedrooms.x              0.0   0.000000  0.000000      .          .
Electricity.x                100.0
             .1                     -0.156645  0.029881  -1.57e-01  -1.57e-01

nVar                                                        7          8
BIC                                                   -1.49e+04  -1.49e+04
post prob                                               0.610      0.152
                             model 3    model 4    model 5
Intercept                    -3.12e+00  -3.25e+00  -2.93e+00
log(Income).x                -1.42e-01  -1.32e-01  -1.43e-01
log(FoodExp).x                5.87e-01   5.90e-01   5.82e-01
Householder_Sex.x
                 .Male        2.18e-01   2.15e-01   2.07e-01
Householder_Age.x                .          .      -3.46e-03
```

```
Household_Type2.x            -2.74e-01  -2.74e-01  -2.87e-01
Household_Typetwo or more.x      .          .          .
Floorarea.x                      .      -7.72e-04      .
House.Age.x                  -4.58e-03  -4.35e-03      .
Number_bedrooms.x                .          .          .
Electricity.x
             .1              -1.54e-01  -1.55e-01  -1.68e-01

nVar                            6          7          6
BIC                          -1.49e+04  -1.49e+04  -1.49e+04
post prob                     0.143      0.059      0.035

  1  observations deleted due to missingness.
```

```r
# Name the best model selected by BIC m3
m3 <- glm(Number_Members ~ log(Income) + log(FoodExp) + Householder_Sex +
            Householder_Age + Household_Type +
            House.Age + Electricity,
          family = "poisson" , data = household)
summary(m3)
```

```
Call:
glm(formula = Number_Members ~ log(Income) + log(FoodExp) + Householder_Sex +
    Householder_Age + Household_Type + House.Age + Electricity,
    family = "poisson", data = household)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.912  -0.569  -0.108   0.420   3.924

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.96574    0.23096  -12.84  < 2e-16 ***
log(Income)               -0.13692    0.02069   -6.62  3.7e-11 ***
log(FoodExp)               0.58016    0.02894   20.04  < 2e-16 ***
Householder_SexMale        0.20301    0.03065    6.62  3.5e-11 ***
Householder_Age           -0.00266    0.00082   -3.25  0.00115 **
Household_Type2           -0.29073    0.02316  -12.55  < 2e-16 ***
Household_Typetwo or more -0.03024    0.18093   -0.17  0.86726
House.Age                 -0.00370    0.00102   -3.63  0.00028 ***
Electricity1              -0.15666    0.02980   -5.26  1.5e-07 ***
---
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: 1308.2  on 2113  degrees of freedom
AIC: 8264

Number of Fisher Scoring iterations: 4
```

Negative Binomial Distribution

The variance(4.9) of y is slightly larger than the mean(4.5) of y, therefore a Negative Binomial Distribution model is fitted to reduce the issue of overdispersion.

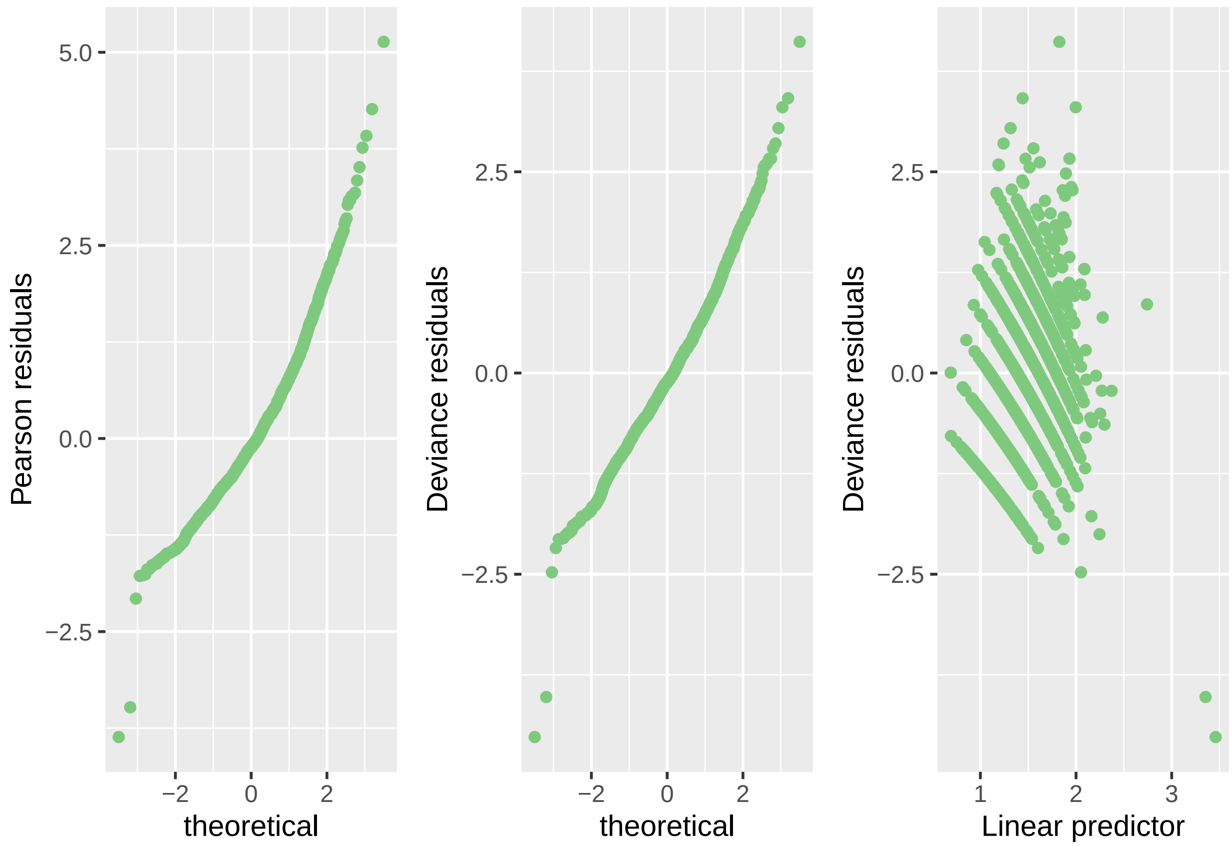
```
m4 <- glm.nb(formula = Number_Members ~ log(Income) + log(FoodExp) + Householder_Sex +
             Householder_Age + Household_Type + Floorarea +
             House.Age + Number_bedrooms + Electricity, data = household)
summary(m4)
```

```
Call:
glm.nb(formula = Number_Members ~ log(Income) + log(FoodExp) +
    Householder_Sex + Householder_Age + Household_Type + Floorarea +
    House.Age + Number_bedrooms + Electricity, data = household,
    init.theta = 109689.3008, link = log)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.960  -0.557  -0.110   0.422   3.859

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.951350   0.248617  -11.87  < 2e-16 ***
log(Income)              -0.137029   0.021684   -6.32  2.6e-10 ***
log(FoodExp)              0.577850   0.029122   19.84  < 2e-16 ***
Householder_SexMale       0.203724   0.030685    6.64  3.2e-11 ***
Householder_Age          -0.002625   0.000823   -3.19  0.00143 **
Household_Type2          -0.288164   0.023186  -12.43  < 2e-16 ***
Household_Typetwo or more -0.035409   0.180951   -0.20  0.84486
Floorarea                -0.000904   0.000341   -2.65  0.00804 **
House.Age                -0.003815   0.001032   -3.70  0.00022 ***
Number_bedrooms           0.024815   0.012572    1.97  0.04841 *
Electricity1             -0.159252   0.029845   -5.34  9.5e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(109689) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1299.4  on 2111  degrees of freedom
AIC: 8262

Number of Fisher Scoring iterations: 1

              Theta:  109689
          Std. Err.:  356154
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -8238
```

Deviance plots

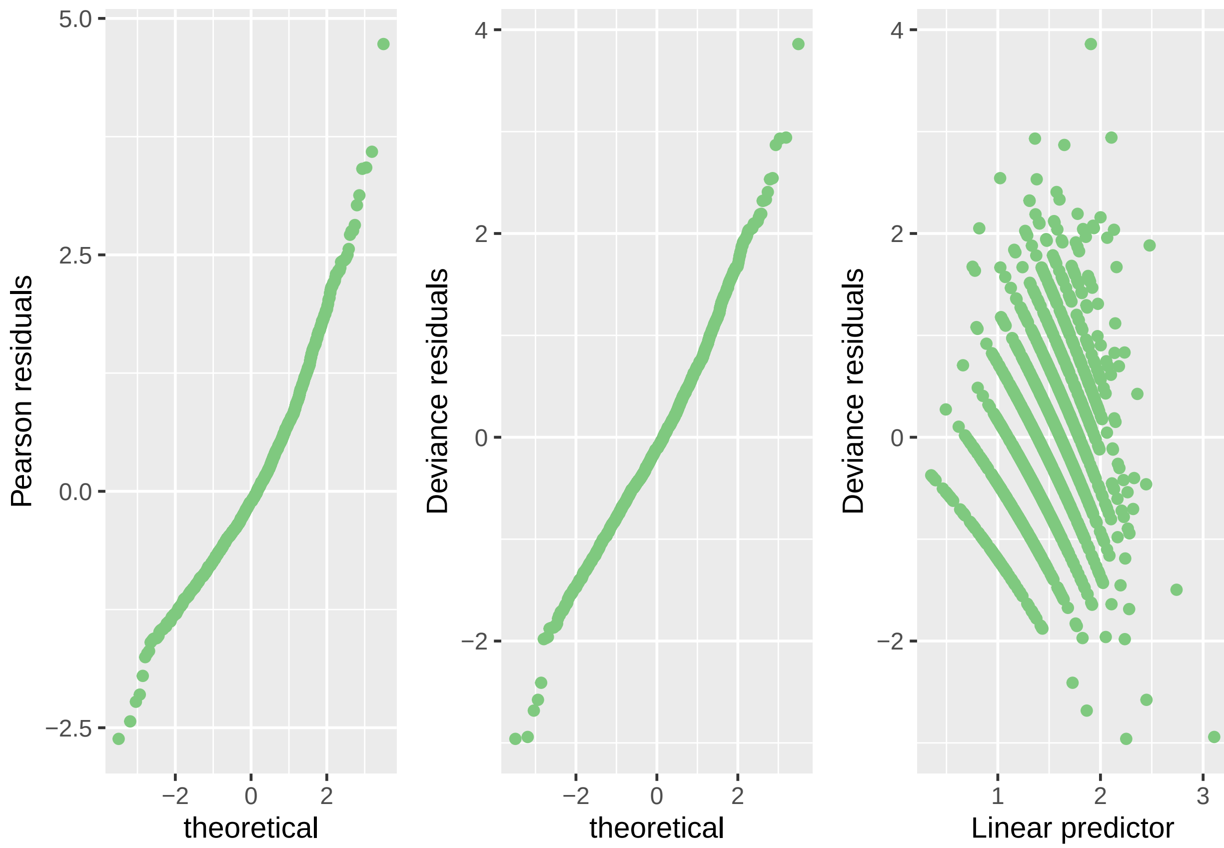
Deviance plots is shown below.

```
resp <- resid(m1, type = "pearson")
resd <- resid(m1, type = "deviance")
r1<- ggplot(m1, aes(sample = resp)) +
  geom_point(stat = "qq", color = "#7fc97f") + ylab("Pearson residuals")
r2<- ggplot(m1, aes(sample = resd)) +
  geom_point(stat = "qq", color = "#7fc97f") + ylab("Deviance residuals")
r3<- ggplot(m1, aes(x = predict(m1, type="link"), y =resd))+
  geom_point(col = "#7fc97f") +
  ylab("Deviance residuals") + xlab("Linear predictor")
grid.arrange(r1, r2, r3, nrow = 1)
```




```
resp2 <- resid(m2, type = "pearson")
resd2 <- resid(m2, type = "deviance")
r4<- ggplot(m2, aes(sample = resp2)) +
  geom_point(stat = "qq", color = "#7fc97f") + ylab("Pearson residuals")
r5<- ggplot(m2, aes(sample = resd2)) +
  geom_point(stat = "qq", color = "#7fc97f") + ylab("Deviance residuals")
r6<- ggplot(m2, aes(x = predict(m2, type="link"), y =resd2)) +
  geom_point(col = "#7fc97f") +
  ylab("Deviance residuals") + xlab("Linear predictor")
```

```r
grid.arrange(r4, r5, r6, nrow = 1)
```



Model Evaluation

```r
# Poisson model
c(m1$deviance, m1$aic)
```

```
[1] 1552 8512
```

```r
# poission model with log transformation
c(m2$deviance, m2$aic)
```

```
[1] 1299 8260
```

```r
# BIC model
c(m3$deviance, m3$aic)
```

```
[1] 1308 8264
```

```r
# Negative binomial model
c(m4$deviance, m4$aic)
```

```
[1] 1299 8262
```

Goodness-of-fit test

```r
chisq <- with(m2, sum((household$Number_Members- fitted.values)^2/fitted.values))
df <- with(m2, df.residual)
p <- with(m2, pchisq(chisq, df, lower.tail = FALSE))
cat("Chi-square test statistic = ", chisq, "\n")
```

```
Chi-square test statistic =  1335
```

```r
cat("df = ", df, "\n")
```

```
df =  2111
```

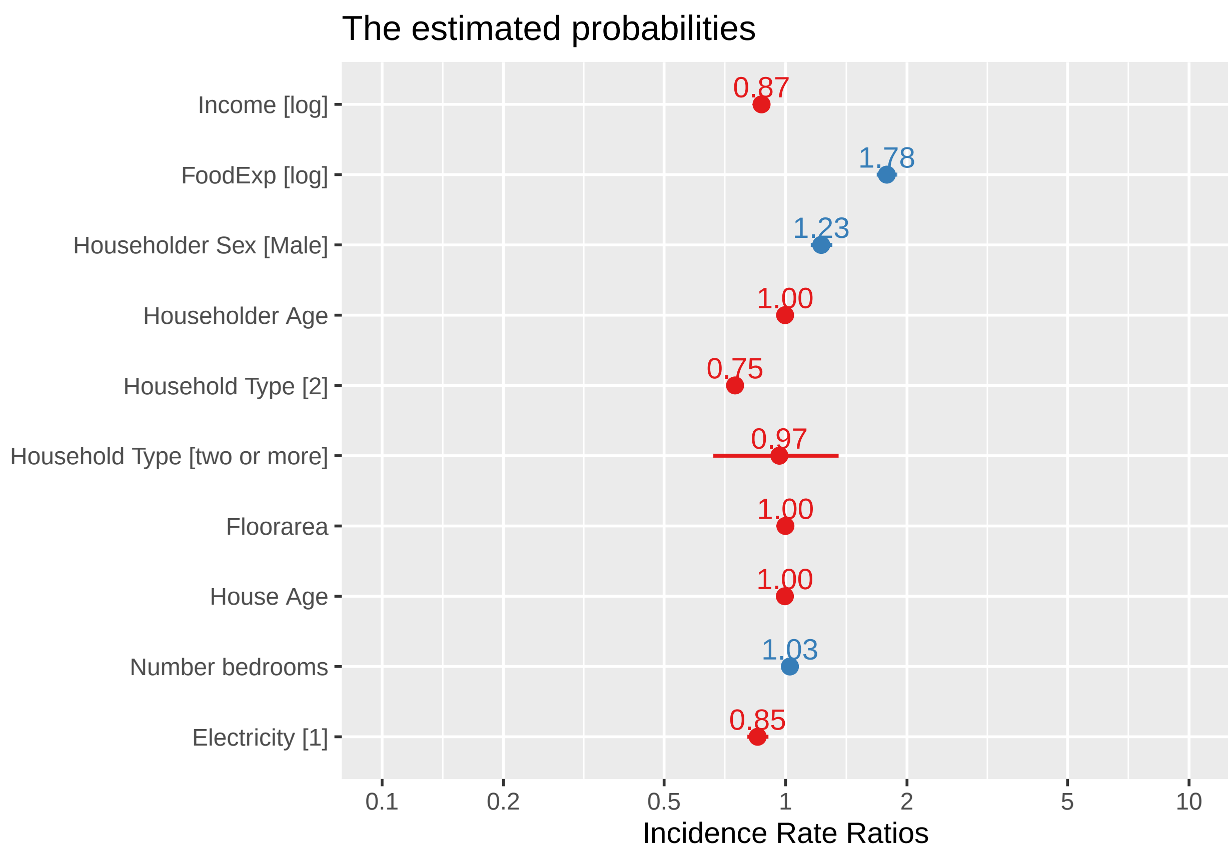
```r
cat("p-value = ", p, "\n")
```

```
p-value =  1
```

```r
# The coef() function obtains the coefficients of the model.
# The confint() function obtains the confidence interval of the model coefficients.

exp(cbind(OR = coef(m2), confint(m2)))
```

```
                            OR 2.5 % 97.5 %
(Intercept)              0.052 0.032  0.085
log(Income)              0.872 0.836  0.910
log(FoodExp)             1.782 1.683  1.886
Householder_SexMale      1.226 1.155  1.302
Householder_Age          0.997 0.996  0.999
Household_Type2          0.750 0.716  0.785
Household_Typetwo or more 0.965 0.663  1.350
Floorarea                0.999 0.998  1.000
House.Age                0.996 0.994  0.998
Number_bedrooms          1.025 1.000  1.051
Electricity1             0.853 0.805  0.904
```

```r
plot_model(m2,show.values=TRUE,title="The estimated probabilities", show.p=FALSE,value.offset=0.25)
```



In Poisson regression, OR (odds ratio) represents the probability ratio (probability ratio) of a set of variables, which is the ratio of the probability of a dependent variable between the levels of two different independent variables. Usually, a larger value of OR means that a variable has a greater effect on the dependent variable. In exp(cbind(OR = coef(model), confint(model))), coef(model) gives the coefficients of all the variables and exp converts them to OR values.

The OR value is equal to the regression coefficient of the indexed variable. The coefficient of an explanatory variable (log(foodexp)) in the regression model is 1.78, its corresponding OR value is exp(coefficient).