Group_1_Analysis

Ying Yang, Linbin Lai

1 Forward Selection

1.1 Load Package and Document

```
rm(list = ls())
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(MASS)
library(dplyr)
library(janitor)
library(tidyr)
FIES <- read.csv("~/Data Analysis Skills/Group project2/dataset01.csv")
unique(FIES$Region)#ignore the region column
```

```
[1] "CAR"
```

```
FIES <- FIES %>%
  dplyr::select(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)
FIES$Total.Number.of.Family.members <-
 ↪  as.factor(FIES$Total.Number.of.Family.members)
FIES$Household.Head.Sex <- as.factor(FIES$Household.Head.Sex)
FIES$Type.of.Household <- as.factor(FIES$Type.of.Household)
```

```r
FIES$Electricity <- as.factor(FIES$Electricity)
levels(FIES$Electricity) <- c("No", "Yes")
FIES$Number.of.bedrooms <- as.factor(FIES$Number.of.bedrooms)
levels(FIES$Number.of.bedrooms) <- c("0", "1", "2", "3", "4", "5", "6",
 ↪  "7", "8", "9")
```

1.2 Analyze each element individually

1.2.1 `Total.Number.of.Family.members` vs `Total.Household.Income`

```r
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪  Total.Household.Income, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Annual
   ↪  household income (in Philippine peso)") +
  theme(legend.position = "none")
```

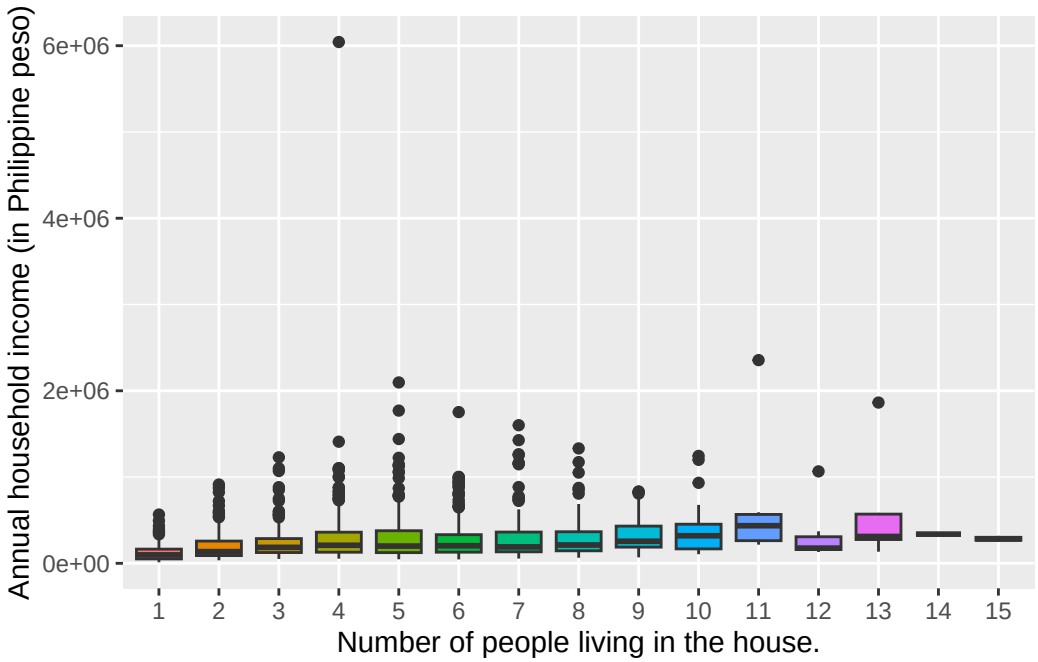


Figure 1: Annual household income by Number of family members.

From the boxplot, we can find that median annual income appears to be similar for most

household sizes (number of people), particularly from one to six person households. Moreover, as household size increases, the number of households with unusually high incomes decreases.

1.2.2 `Total.Number.of.Family.members` vs `Total.Food.Expenditure`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪  Total.Household.Income, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Annual
   ↪  expenditure by the household on food") +
  theme(legend.position = "none")
```

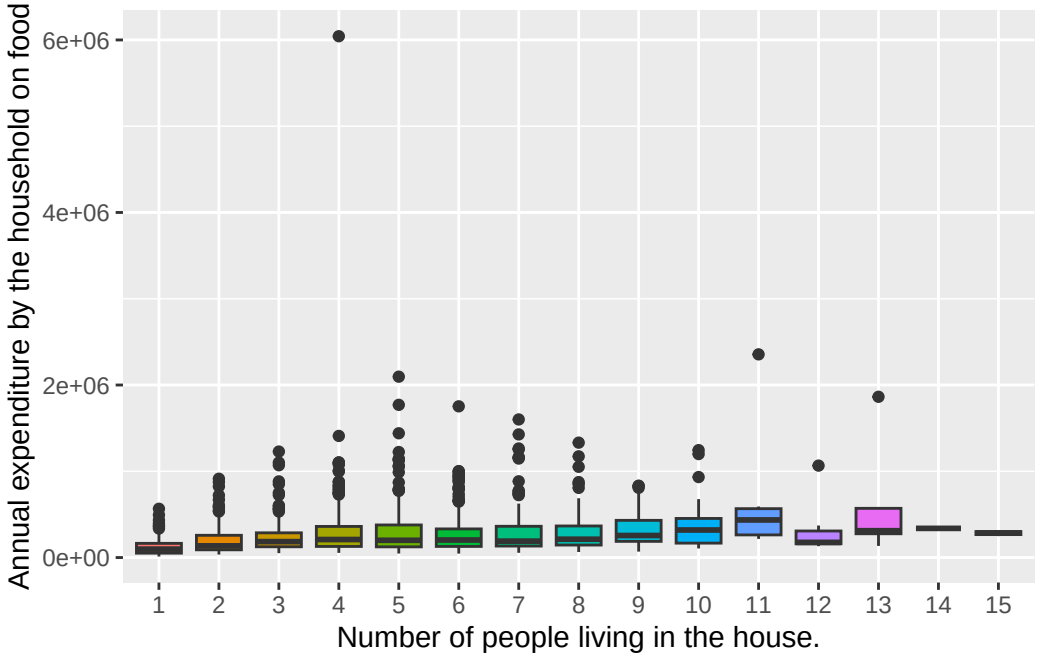


Figure 2: Annual expenditure by the household on food by Number of family members.

From this boxplot, median food expenditures appear to increase gradually as household size increases, especially from one to nine person households. There are fewer data points for larger households (e.g. more than 10 people), as can be seen from the width of the boxplot, with smaller box widths indicating smaller sample sizes for these household sizes.

1.2.3 Total.Number.of.Family.members vs Household.Head.Sex

```
FIES %>%
  tabyl(Household.Head.Sex, Total.Number.of.Family.members) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() #To show original counts
```

```
Household.Head.Sex          1           2           3           4           5
            Female 14.6% (54) 13.6%  (50) 20.3%  (75) 17.3%  (64) 11.7%  (43)
              Male  5.5% (74)  9.8% (133) 13.3% (180) 17.8% (242) 18.5(251)
          6          7         8         9        10        11       12
 9.5%  (35) 5.7%  (21) 3.0% (11) 1.9%  (7) 0.5%  (2) 0.3%  (1) 0.8     (3)
14.2% (193) 8.8% (119) 5.6% (76) 2.9% (39) 1.8% (24) 1.0% (13) 0.5     (7)
      13       14       15
0.8% (3) 0.0% (0) 0.0                                                  (0)
0.1% (2) 0.1% (1) 0.1                                                  (2)
```

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
 ↪  Household.Head.Sex)) +
  geom_bar(aes(y = ..prop.., fill = Household.Head.Sex), stat = "count",
   ↪  position = "dodge") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

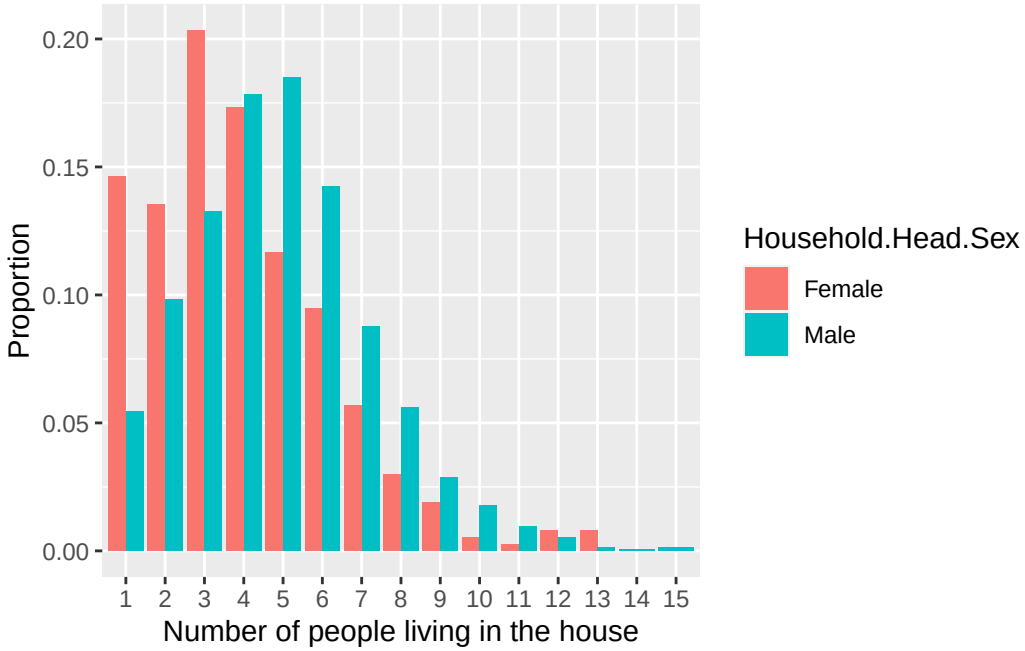


Figure 3: Barplot of Number of people living in the house by Head of the households sex.

For different family sizes, families headed by men generally account for a larger proportion, especially in families with three to seven people. Among one person, two person and three person households, the proportion of female-headed households is higher than the male-headed households.

1.2.4 `Total.Number.of.Family.members` vs `Household.Head.Age`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪  Household.Head.Age, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Head of the
   ↪  households age (in years)") +
  theme(legend.position = "none")
```

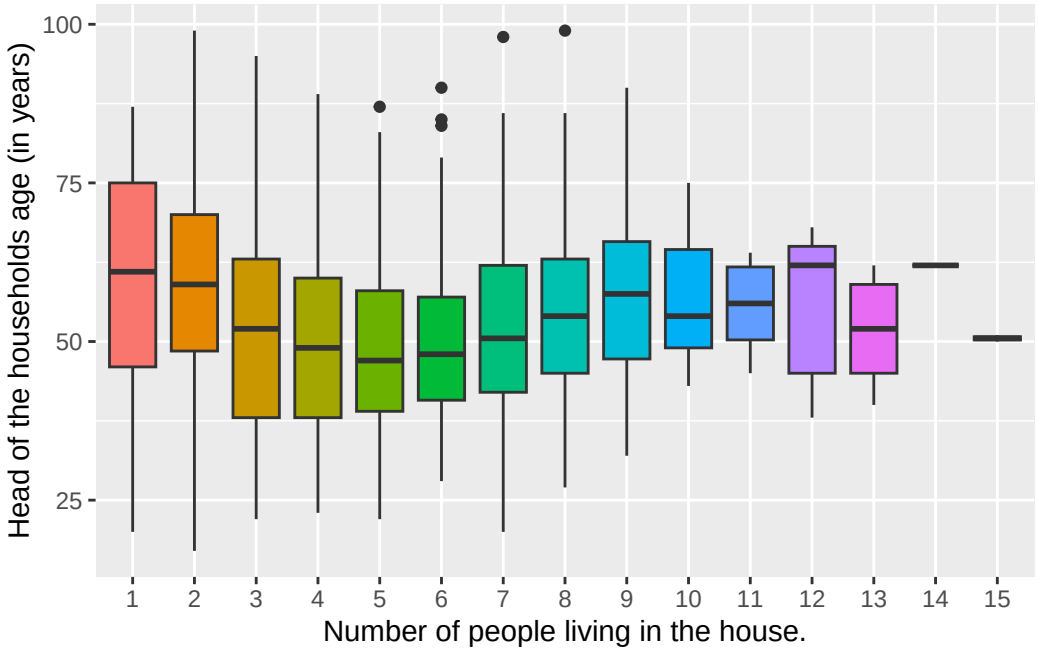


Figure 4: Head of the households age by Number of family members.

For different family sizes, the median age of the household head is relatively stable among different family sizes, mainly concentrated around 50 years old. So I think there is no obvious correlation trend between the age of the household head and the size of the family.

1.2.5 `Total.Number.of.Family.members` vs `Type.of.Household`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
 ↪  Type.of.Household)) +
  geom_bar(aes(y = ..prop.., fill = Type.of.Household), stat = "count",
   ↪  position = "dodge") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

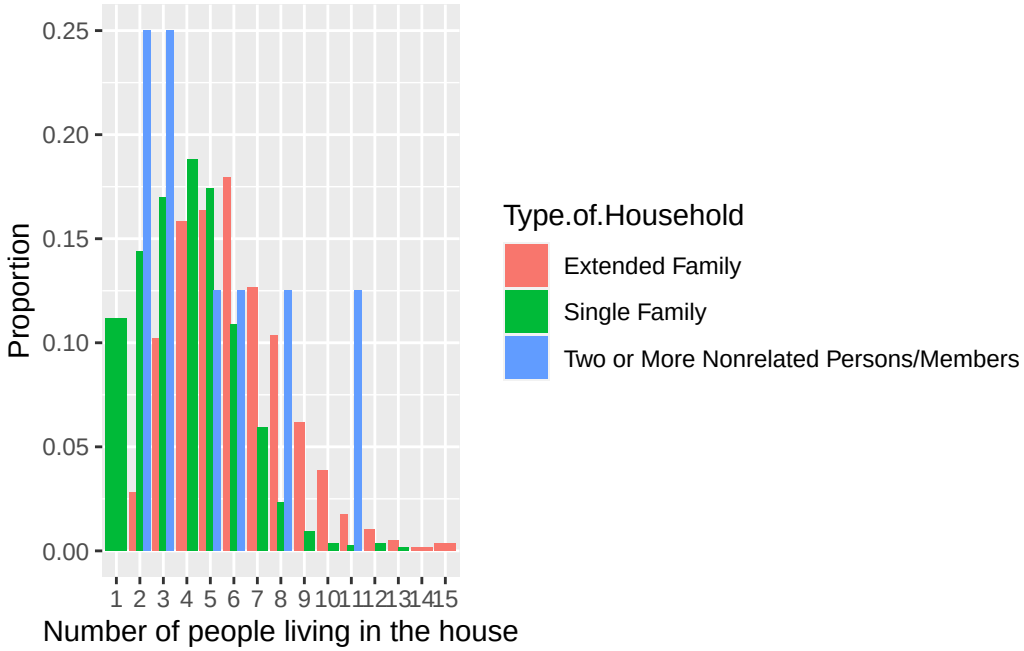


Figure 5: Barplot of Number of people living in the house by the relationship between the group of people living in the house.

There is a correlation between household type and the number of people living in the house, with extended households being more common in medium and large households, while single households predominate among small households. And for two or More Non-related Persons/Members, it has highly proportion in two person and three person household.

1.2.6 `Total.Number.of.Family.members` vs `House.Floor.Area`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪  House.Floor.Area, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Floor area of
   ↪  the house (in meter square)") +
  theme(legend.position = "none")
```

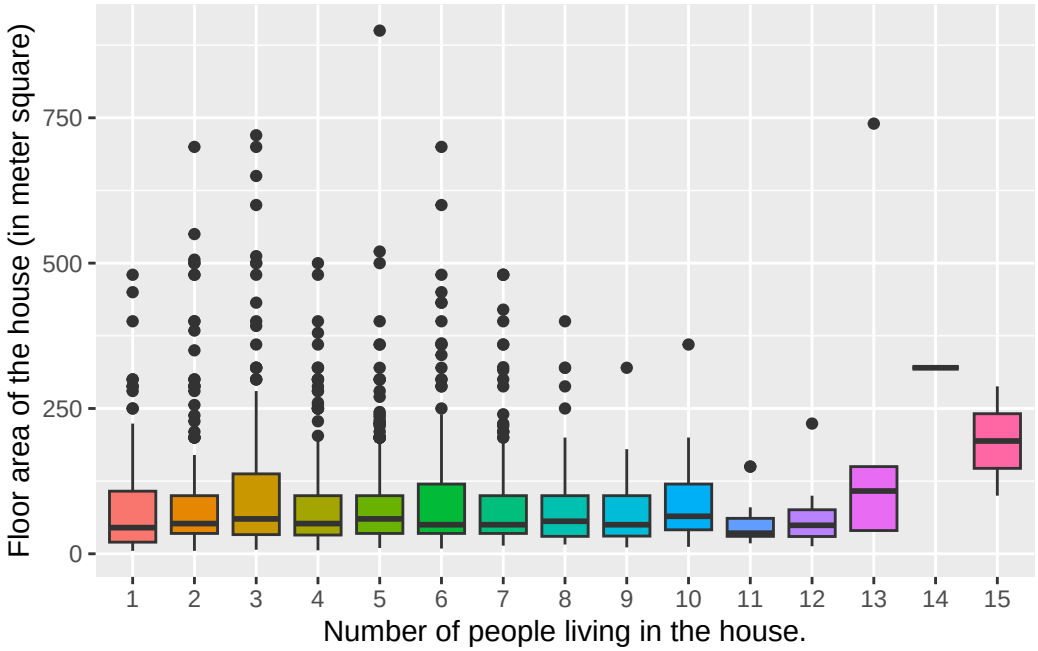


Figure 6: Floor area of the house by Number of family members.

Median house floor area appears to be relatively stable for most number of people living in the house, without increasing significantly as the number of residents increases. The graph shows a large number of outliers, indicating that some households have homes that are well outside the typical range for similar household sizes.Therefore, house floor area does not appear to increase significantly with number of people living in the house.

1.2.7 `Total.Number.of.Family.members` vs `House.Age`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪  House.Age, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Age of the
   ↪  building (in years)") +
  scale_fill_brewer(palette = "Set3") +#change to different color
  theme(legend.position = "none")
```

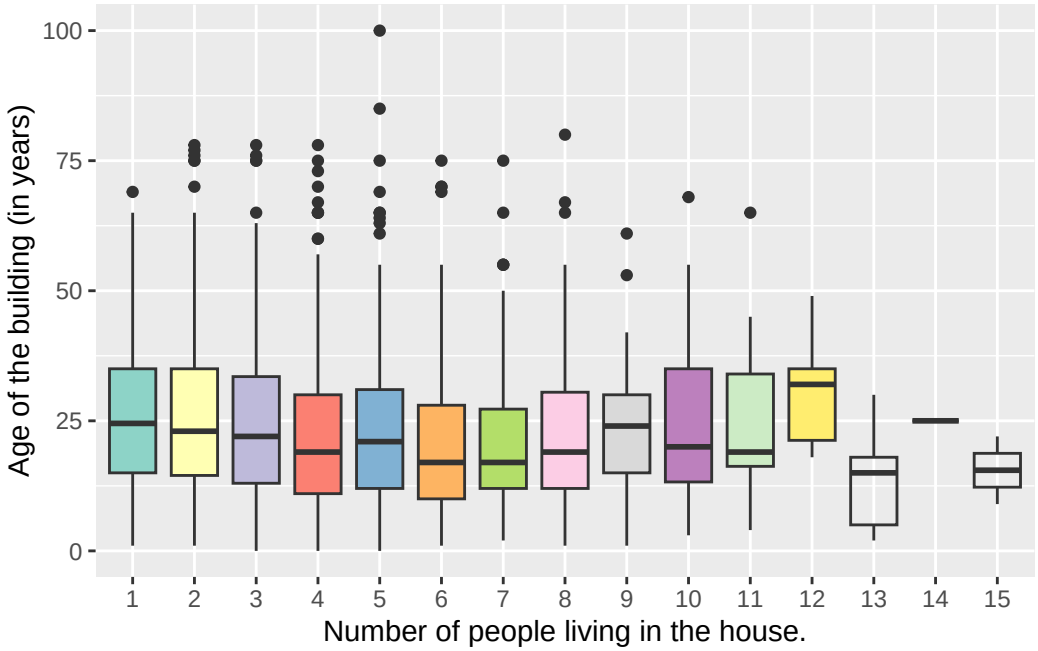


Figure 7: Age of the building by Number of family members.

The distribution range of house ages is relatively consistent, suggesting that the distribution of house ages is similar across household sizes. In conclude, there does not appear to be a direct relationship between house age and household size, with the median age remaining relatively stable across household sizes.

1.2.8 `Total.Number.of.Family.members` vs `Number.of.bedrooms`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
 ↪  Number.of.bedrooms)) +
  geom_bar(aes(y = ..prop.., fill = Number.of.bedrooms), stat = "count",
   ↪  position = "dodge") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

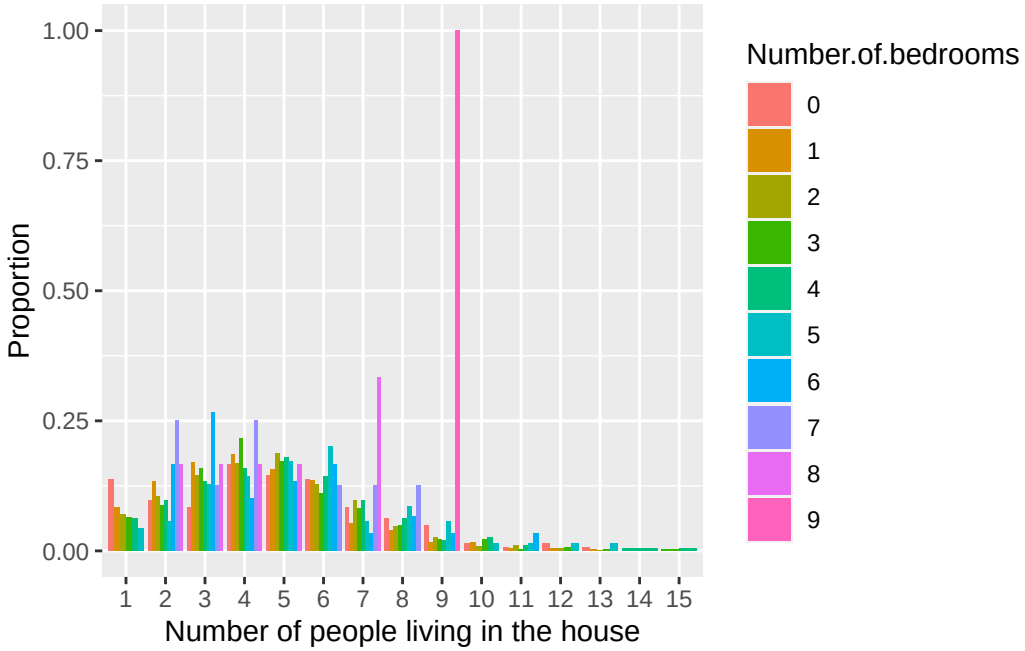


Figure 8: Barplot of Number of people living in the house by Number of bedrooms in the house.

As the number of residents increases, the median number of bedrooms gradually increases, but this growth is not linear.

1.2.9 `Total.Number.of.Family.members` vs `Electricity`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
 ↪  Electricity)) +
  geom_bar(aes(y = ..prop.., fill = Electricity), stat = "count",
   ↪  position = "dodge") +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

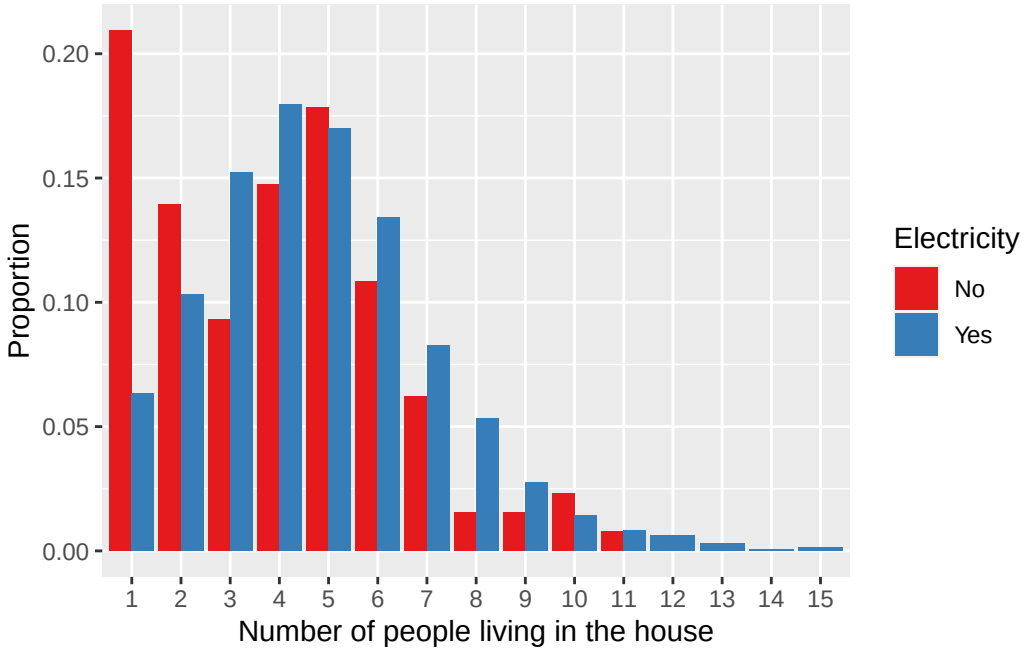


Figure 9: Barplot of Number of people living in the house by the electricity.

For most of the families, regardless of size, they have electricity. Among small scale households (especially single person households), the proportion of households without electricity is relatively high. As household size increases, the proportion of households without electricity gradually decreases. In larger households (seven people and above), almost all households have electricity supply.

1.3 Summary of Analyze each element individually

Overall, according to the separate analysis of explanatory variables, the variables that should be retained for now are:Annual household income, Annual expenditure by the household on food, Head of the households sex, Relationship between the group of people living in the house and Electricity.

1.4 Analysis of the overall part

```r
FIES <- FIES %>%
  dplyr::select(Total.Number.of.Family.members, Total.Household.Income,
 ↪  Total.Food.Expenditure, Household.Head.Sex, Type.of.Household,
 ↪  Number.of.bedrooms, Electricity)

FIES$Total.Number.of.Family.members <-
 ↪  as.numeric(as.character(FIES$Total.Number.of.Family.members))

model.FIES <- glm(Total.Number.of.Family.members ~
 ↪  Total.Household.Income + Total.Food.Expenditure + Household.Head.Sex
 ↪  + Type.of.Household + Electricity, data = FIES, family =
 ↪  poisson(link = "log"))

model.FIES %>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Type.of.Household +
    Electricity, family = poisson(link = "log"), data = FIES)

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                            1.185e+00  5.675e-02
Total.Household.Income                                -2.646e-07  5.541e-08
Total.Food.Expenditure                                 5.017e-06  3.354e-07
Household.Head.SexMale                                 2.480e-01  2.921e-02
Type.of.HouseholdSingle Family                        -3.214e-01  2.390e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.265e-01  1.594e-01
ElectricityYes                                         4.664e-03  4.693e-02
                                                       z value Pr(>|z|)
(Intercept)                                             20.885  < 2e-16 ***
Total.Household.Income                                  -4.776 1.79e-06 ***
Total.Food.Expenditure                                  14.958  < 2e-16 ***
Household.Head.SexMale                                   8.489  < 2e-16 ***
Type.of.HouseholdSingle Family                         -13.446  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.794    0.427
ElectricityYes                                            0.099    0.921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1364.9  on 1718  degrees of freedom
AIC: 7041.4

Number of Fisher Scoring iterations: 4
```

`Type.of.HouseholdTwo or More Nonrelated Persons/Members`: The p value is 0.365302, indicating that the effect of this variable is not significant.
`Number.of.bedrooms (7 to 9 bedrooms)`: The p-values of these variables range from 0.129907 to 0.371571, indicating that their impact on the number of family members is not significant.
`Electricity Yes`: The p-value is 0.462128, which means that the presence or absence of electricity has no significant impact on the number of household members.

```
plot_model(model.FIES, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.25)
```

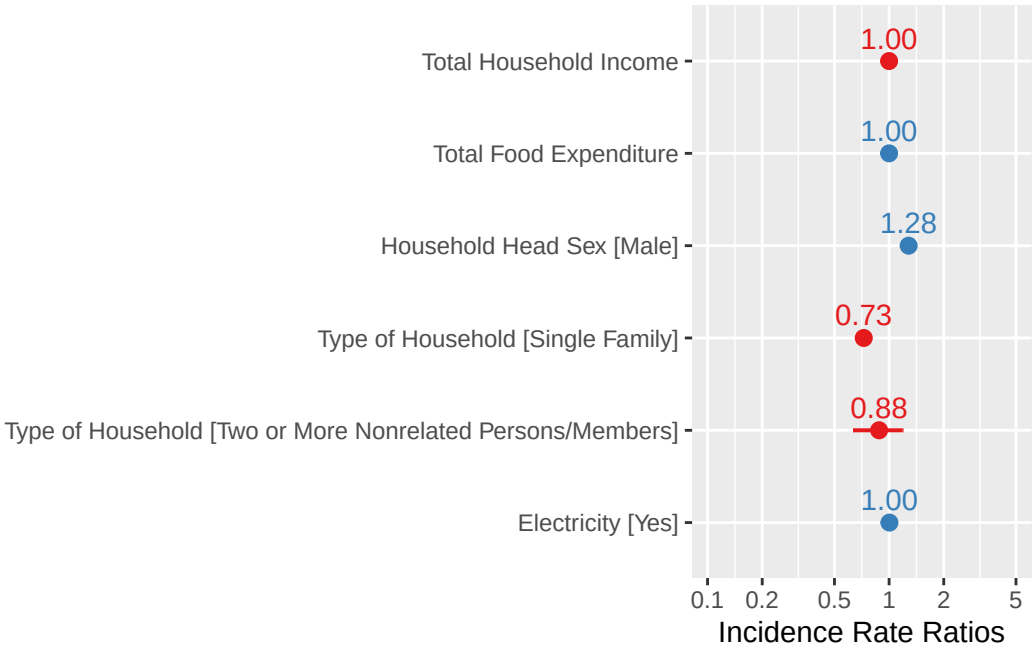


Figure 10: Rate ratios plot.

From total household income and total food expenditure, their rate ratios is exactly 1.00, they prove no effect on the number of family members. Household Head Sex's rate ratio is 1.28, which suggests that households headed by males have a 28% increase in the expected count of family members compared to households headed by females, assuming all other factors are held constant. Type of Household has the rate ratios about 0.72 indicates that single-family households have a 28% decrease in the expected count of family members compared to the reference household type (likely non-single family households). And for type of household With an rate ratios of 0.87, this suggests that households consisting of two or more non-related persons/members have a 13% decrease in the expected count of family members compared to the reference category. Number of Bedrooms suggesting a small decrease of 3% in the expected count of family members for each additional bedroom in the household.

1.5 Result

At last , we keep `Household.Head.Sex` and `Number.of.bedrooms` as explanatory variables:

$$Total.Number.of.Family.members = \beta_0 + \beta_1 \times \text{Household.Head.Sex}$$

```
FIES$Total.Number.of.Family.members <-
 ↪  as.numeric(as.character(FIES$Total.Number.of.Family.members))

model.FIES <- glm(Total.Number.of.Family.members ~ Household.Head.Sex,
 ↪  data = FIES, family = poisson(link = "log"))

model.FIES %>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Household.Head.Sex,
    family = poisson(link = "log"), data = FIES)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.37813    0.02613  52.732  < 2e-16 ***
Household.Head.SexMale  0.20288    0.02889   7.022 2.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1972.9  on 1723  degrees of freedom
AIC: 7639.5

Number of Fisher Scoring iterations: 4
```

```
model <- glm.nb(Total.Number.of.Family.members ~ Household.Head.Sex,data
 ↪  = FIES)
summary(model)
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Household.Head.Sex,
    data = FIES, init.theta = 36.04958898, link = log)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.37813    0.02754  50.048  < 2e-16 ***
Household.Head.SexMale 0.20288    0.03050   6.651 2.91e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(36.0496) family taken to be 1)

    Null deviance: 1796.6  on 1724  degrees of freedom
Residual deviance: 1750.9  on 1723  degrees of freedom
AIC: 7627.6

Number of Fisher Scoring iterations: 1

              Theta:  36.0
          Std. Err.:  10.6

 2 x log-likelihood:  -7621.585
```

```
model$aic
```

```
[1] 7627.585
```

2 Backward Elimination

2.1 Find the suitable generalized linear model using family=poisson(Backward regression)

```
data <- read.csv('~/Data Analysis Skills/Group project2/dataset01.csv')
house_data1 <- glm(Total.Number.of.Family.members~
                     Total.Household.Income+
                     Total.Food.Expenditure+
                     Household.Head.Sex+
                     Household.Head.Age+
                     Type.of.Household+
                     Total.Number.of.Family.members+
                     House.Floor.Area+House.Age+
                     Number.of.bedrooms+
                     Electricity,
                   family = poisson, data=data)
house_data1%>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + Total.Number.of.Family.members + House.Floor.Area +
    House.Age + Number.of.bedrooms + Electricity, family = poisson,
    data = data)

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                             1.430e+00  7.951e-02
Total.Household.Income                                 -1.881e-07  5.693e-08
Total.Food.Expenditure                                  4.893e-06  3.370e-07
Household.Head.SexMale                                  2.213e-01  2.970e-02
Household.Head.Age                                     -2.536e-03  8.704e-04
Type.of.HouseholdSingle Family                         -3.490e-01  2.479e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.429e-01  1.599e-01
House.Floor.Area                                       -2.048e-04  1.276e-04
House.Age                                              -2.309e-03  7.735e-04
Number.of.bedrooms                                     -1.569e-02  9.489e-03
Electricity                                             2.776e-02  4.755e-02
```

```
                                                        z value Pr(>|z|)
(Intercept)                                              17.982  < 2e-16 ***
Total.Household.Income                                   -3.303 0.000956 ***
Total.Food.Expenditure                                   14.518  < 2e-16 ***
Household.Head.SexMale                                    7.452 9.21e-14 ***
Household.Head.Age                                       -2.914 0.003568 **
Type.of.HouseholdSingle Family                          -14.081  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.894 0.371271
House.Floor.Area                                         -1.606 0.108320
House.Age                                                -2.986 0.002830 **
Number.of.bedrooms                                       -1.654 0.098175 .
Electricity                                               0.584 0.559313
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1331.2  on 1714  degrees of freedom
AIC: 7015.8

Number of Fisher Scoring iterations: 4
```

The variable Electricity has p-value greater than 0.05, so delete this variable and refit the model.

```
house_data2 <- glm(Total.Number.of.Family.members~
                     Total.Household.Income+
                     Total.Food.Expenditure+
                     Household.Head.Sex+
                     Household.Head.Age+
                     Type.of.Household+
                     Total.Number.of.Family.members+
                     House.Floor.Area+House.Age+
                     Number.of.bedrooms,
                   family = poisson, data=data)
house_data2%>%
  summary()
```

```
Call:
```

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

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                        1.454e+00  6.754e-02
Total.Household.Income                            -1.884e-07  5.692e-08
Total.Food.Expenditure                             4.909e-06  3.358e-07
Household.Head.SexMale                             2.212e-01  2.970e-02
Household.Head.Age                                -2.568e-03  8.688e-04
Type.of.HouseholdSingle Family                    -3.499e-01  2.474e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.422e-01  1.599e-01
House.Floor.Area                                  -2.033e-04  1.275e-04
House.Age                                         -2.279e-03  7.717e-04
Number.of.bedrooms                                -1.499e-02  9.411e-03
                                                  z value Pr(>|z|)
(Intercept)                                        21.532  < 2e-16 ***
Total.Household.Income                             -3.310 0.000934 ***
Total.Food.Expenditure                             14.619  < 2e-16 ***
Household.Head.SexMale                              7.448  9.5e-14 ***
Household.Head.Age                                 -2.955 0.003122 **
Type.of.HouseholdSingle Family                    -14.147  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.890 0.373721
House.Floor.Area                                   -1.595 0.110806
House.Age                                          -2.953 0.003148 **
Number.of.bedrooms                                 -1.593 0.111200
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1331.6  on 1715  degrees of freedom
AIC: 7014.1

Number of Fisher Scoring iterations: 4
```

The variable Number.of.bedrooms has p-value greater than 0.05, so delete this variable and refit the model.

```r
house_data3 <- glm(Total.Number.of.Family.members~
                     Total.Household.Income+
                     Total.Food.Expenditure+
                     Household.Head.Sex+
                     Household.Head.Age+
                     Type.of.Household+
                     Total.Number.of.Family.members+
                     House.Floor.Area+House.Age,
                   family = poisson, data=data)
house_data3%>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + Total.Number.of.Family.members + House.Floor.Area +
    House.Age, family = poisson, data = data)

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                         1.449e+00  6.742e-02
Total.Household.Income                             -2.122e-07  5.572e-08
Total.Food.Expenditure                              4.846e-06  3.351e-07
Household.Head.SexMale                              2.209e-01  2.969e-02
Household.Head.Age                                 -2.727e-03  8.628e-04
Type.of.HouseholdSingle Family                     -3.503e-01  2.473e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.371e-01  1.598e-01
House.Floor.Area                                   -2.601e-04  1.231e-04
House.Age                                          -2.399e-03  7.691e-04
                                                   z value Pr(>|z|)
(Intercept)                                         21.485  < 2e-16 ***
Total.Household.Income                              -3.808  0.00014 ***
Total.Food.Expenditure                              14.460  < 2e-16 ***
Household.Head.SexMale                               7.440 1.01e-13 ***
Household.Head.Age                                  -3.161  0.00157 **
Type.of.HouseholdSingle Family                     -14.164  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.857  0.39117
House.Floor.Area                                    -2.113  0.03460 *
House.Age                                           -3.119  0.00182 **
---
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1334.1  on 1716  degrees of freedom
AIC: 7014.7

Number of Fisher Scoring iterations: 4
```

The variable Type.of.Household has p-value greater than 0.05, so delete this variable and refit the model.

```
house_data4 <- glm(Total.Number.of.Family.members~
                     Total.Household.Income+
                     Total.Food.Expenditure+
                     Household.Head.Sex+
                     Household.Head.Age+
                     Total.Number.of.Family.members+
                     House.Floor.Area+House.Age,
                   family = poisson, data=data)
house_data4%>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Total.Number.of.Family.members + House.Floor.Area + House.Age,
    family = poisson, data = data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              9.938e-01  5.935e-02  16.746  < 2e-16 ***
Total.Household.Income -2.442e-07  5.756e-08  -4.242 2.22e-05 ***
Total.Food.Expenditure  6.049e-06  3.246e-07  18.634  < 2e-16 ***
Household.Head.SexMale  1.960e-01  2.958e-02   6.625 3.46e-11 ***
Household.Head.Age      2.442e-04  8.331e-04   0.293  0.76944
House.Floor.Area       -2.369e-04  1.226e-04  -1.932  0.05340 .
House.Age              -2.053e-03  7.667e-04  -2.678  0.00741 **
---
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1531.5  on 1718  degrees of freedom
AIC: 7208

Number of Fisher Scoring iterations: 4
```

The variable Household.Head.Age has p-value greater than 0.05, so delete this variable and refit the model.

```
house_data5 <- glm(Total.Number.of.Family.members~
                   Total.Household.Income+
                   Total.Food.Expenditure+
                   Household.Head.Sex+
                   Total.Number.of.Family.members+
                   House.Floor.Area+
                   House.Age,
                 family = poisson, data=data)
house_data5%>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Total.Number.of.Family.members +
    House.Floor.Area + House.Age, family = poisson, data = data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.007e+00  3.962e-02  25.409  < 2e-16 ***
Total.Household.Income -2.425e-07  5.725e-08  -4.236 2.27e-05 ***
Total.Food.Expenditure  6.040e-06  3.234e-07  18.678  < 2e-16 ***
Household.Head.SexMale  1.945e-01  2.913e-02   6.676 2.46e-11 ***
House.Floor.Area       -2.348e-04  1.224e-04  -1.918  0.05508 .
House.Age              -2.009e-03  7.517e-04  -2.673  0.00753 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1531.6  on 1719  degrees of freedom
AIC: 7206.1

Number of Fisher Scoring iterations: 4
```

The variable House.Floor.Area has p-value greater than 0.05, so delete this variable and refit the model.

```
house_data6 <- glm(Total.Number.of.Family.members~
                     Total.Household.Income+
                     Total.Food.Expenditure+
                     Household.Head.Sex+
                     Total.Number.of.Family.members+
                     House.Age,
                   family = poisson, data=data)
house_data6%>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Total.Number.of.Family.members +
    House.Age, family = poisson, data = data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             9.914e-01  3.885e-02  25.517  < 2e-16 ***
Total.Household.Income -2.711e-07  5.616e-08  -4.828 1.38e-06 ***
Total.Food.Expenditure  6.076e-06  3.249e-07  18.703  < 2e-16 ***
Household.Head.SexMale  1.962e-01  2.913e-02   6.737 1.62e-11 ***
House.Age              -2.109e-03  7.504e-04  -2.810  0.00496 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1535.4  on 1720  degrees of freedom
```

```
AIC: 7207.9

Number of Fisher Scoring iterations: 4
```

The model Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household.He has the lowest AIC which is 7014.1.

After reducing variables which with p-value large than 0.05, the model becomes as follows:

Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household.Head.Sex+Ho

Therefore, these five variables(Total.Household.Income, Total.Food.Expenditure, Household.Head.Sex, Household.Head.Age and House.Age) are significant. So the household related variables(Total.Household.Income, Total.Food.Expenditure, Household.Head.Sex, Household.Head.Age and House.Age)influence the number of people living in a household.

We see that the coefficient for Total.Household.Income is negative, indicating that if a household has more total household income, there will be less number of people living in the household. Similarly, the coefficient for Total.Food.Expenditure is positive, which means as the total food expenditure increase, there will be more people living in the household. The Household.Head.SexMale coefficient is positive, indicating a higher number of people living in a household for male house head. Besides, the coefficient for Household.Head.Age is negative, that is, as household head age increases there will be less member living in the house. Finally, the House.Age coefficient is negative, which means if the house get older, there will be less number of people living in the household.

2.1.1 Odds ratios

```r
plot_model(house_data2, show.values = TRUE,
title = "", show.p = FALSE, value.offset = 0.25)
```

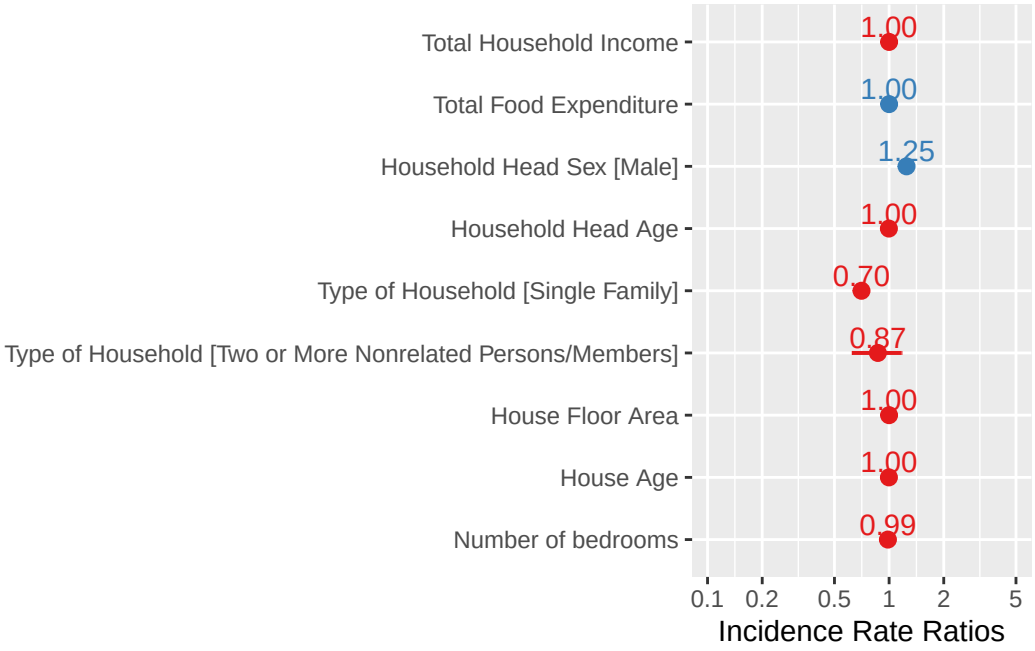


Figure 11: The odds ratios plot.

We interpret the odds ratios as follows: women's odds of number of family members were 0.8 times those of man. For each unit change for Total.Household.Income, Total.Food.Expenditure,Household.Head.Age and House.Age the odds of number of family members do not change, which means they prove no effect on the number of family members.

Therefore, the best model is:

$$\begin{aligned} Total.Number.of.Family.members &= \beta_0 + \beta_1 * \cdot \mathbb{I}_{\text{Household.Head.SexFemale}} \\ &= (1.533e+00) - (2.212e-01) \cdot \mathbb{I}_{\text{Household.Head.SexFemale}} \end{aligned}$$

$\mathbb{I}_i$ is is an indicator function such that representing each gender such that

$$\mathbb{I}_i = \begin{cases} 1 & \text{if the household head is female,} \\ 0 & \text{Otherwise.} \end{cases}$$

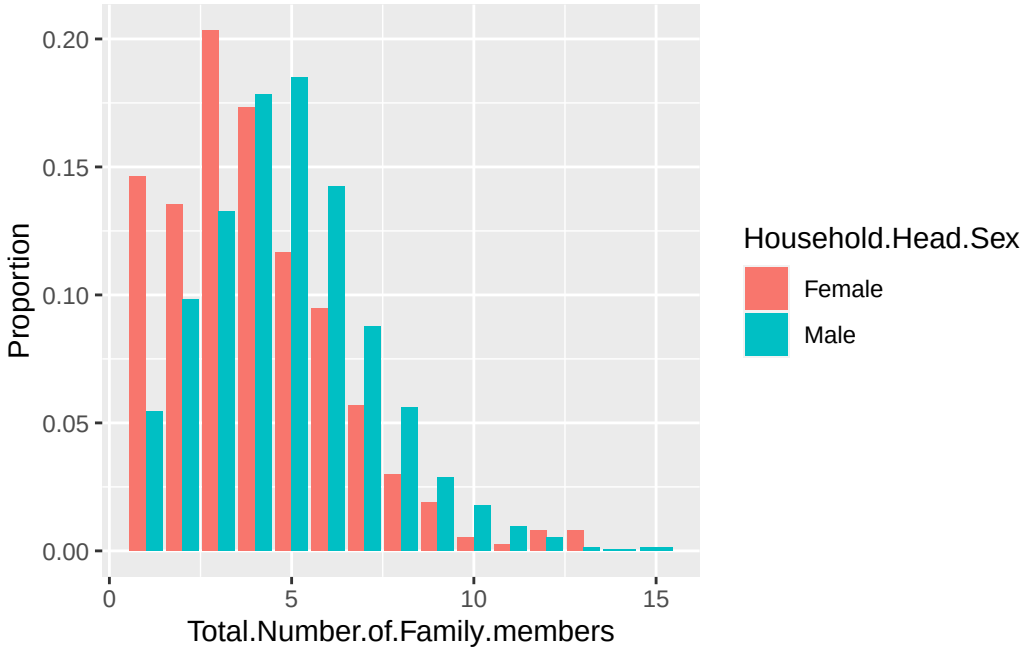
2.2 Data visualization: Analysis for significant explanatory variable(Household.Head.Sex) with response variable(Total.Number.of.Family.members)

2.2.1 Household.Head.Sex~Total.Number.of.Family.members

```
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 14.6% (54) 13.6%  (50) 20.3%  (75) 17.3%  (64) 11.7%  (43)
              Male  5.5% (74)  9.8% (133) 13.3% (180) 17.8% (242) 18.5(251)
          6           7         8         9        10        11       12
 9.5%  (35) 5.7%  (21) 3.0% (11) 1.9%  (7) 0.5%  (2) 0.3%  (1) 0.8     (3)
14.2% (193) 8.8% (119) 5.6% (76) 2.9% (39) 1.8% (24) 1.0% (13) 0.5     (7)
      13       14       15
0.8% (3) 0.0% (0) 0.0                                                  (0)
0.1% (2) 0.1% (1) 0.1                                                  (2)
```

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



There is a clear pattern that female has higher proportion in smaller family size than male.

2.3 Find the suitable Negative Binomial generalized linear model(Backward regression)

```
library(MASS)
data$Total.Number.of.Family.members <-
 ↪  as.numeric(as.character(data$Total.Number.of.Family.members))
housenb_data1 <- glm.nb(Total.Number.of.Family.members~
                  Total.Household.Income+
                  Total.Food.Expenditure+
                  Household.Head.Sex+
                  Household.Head.Age+
                  Type.of.Household+
                  Total.Number.of.Family.members+
                  House.Floor.Area+House.Age+
                  Number.of.bedrooms+
                  Electricity,
                  data=data)
housenb_data1%>%
```

```
summary()
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + Total.Number.of.Family.members + House.Floor.Area +
    House.Age + Number.of.bedrooms + Electricity, data = data,
    init.theta = 77580.79403, link = log)

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                            1.430e+00  7.952e-02
Total.Household.Income                                -1.881e-07  5.694e-08
Total.Food.Expenditure                                 4.893e-06  3.371e-07
Household.Head.SexMale                                 2.213e-01  2.970e-02
Household.Head.Age                                    -2.536e-03  8.704e-04
Type.of.HouseholdSingle Family                        -3.490e-01  2.479e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.430e-01  1.599e-01
House.Floor.Area                                      -2.048e-04  1.276e-04
House.Age                                             -2.309e-03  7.735e-04
Number.of.bedrooms                                    -1.569e-02  9.489e-03
Electricity                                            2.776e-02  4.755e-02
                                                      z value Pr(>|z|)
(Intercept)                                            17.981  < 2e-16 ***
Total.Household.Income                                 -3.303 0.000956 ***
Total.Food.Expenditure                                 14.517  < 2e-16 ***
Household.Head.SexMale                                  7.452 9.23e-14 ***
Household.Head.Age                                     -2.914 0.003568 **
Type.of.HouseholdSingle Family                        -14.080  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.894 0.371273
House.Floor.Area                                       -1.606 0.108333
House.Age                                              -2.985 0.002831 **
Number.of.bedrooms                                     -1.654 0.098179 .
Electricity                                             0.584 0.559319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(77580.79) family taken to be 1)

    Null deviance: 2024.3  on 1724  degrees of freedom
```

```
Residual deviance: 1331.2  on 1714  degrees of freedom
AIC: 7017.8

Number of Fisher Scoring iterations: 1

              Theta:  77581
          Std. Err.:  291920
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -6993.787
```

The variable Electricity has big p-value, so we reduce this variable and refit the model.

```
housenb_data2 <- glm.nb(Total.Number.of.Family.members~
                  Total.Household.Income+
                  Total.Food.Expenditure+
                  Household.Head.Sex+
                  Household.Head.Age+
                  Type.of.Household+
                  Total.Number.of.Family.members+
                  House.Floor.Area+House.Age+
                  Number.of.bedrooms,
                  data=data)
housenb_data2%>%
  summary()
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + Total.Number.of.Family.members + House.Floor.Area +
    House.Age + Number.of.bedrooms, data = data, init.theta = 77568.64866,
    link = log)

Coefficients:
                                                 Estimate Std. Error
(Intercept)                                     1.454e+00  6.754e-02
Total.Household.Income                         -1.884e-07  5.692e-08
Total.Food.Expenditure                          4.909e-06  3.358e-07
Household.Head.SexMale                          2.212e-01  2.970e-02
Household.Head.Age                             -2.568e-03  8.688e-04
```

```
Type.of.HouseholdSingle Family                              -3.499e-01  2.474e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.422e-01  1.599e-01
House.Floor.Area                                            -2.033e-04  1.275e-04
House.Age                                                   -2.279e-03  7.717e-04
Number.of.bedrooms                                          -1.499e-02  9.412e-03
                                                            z value Pr(>|z|)
(Intercept)                                                  21.531  < 2e-16 ***
Total.Household.Income                                       -3.310 0.000934 ***
Total.Food.Expenditure                                       14.618  < 2e-16 ***
Household.Head.SexMale                                        7.447 9.52e-14 ***
Household.Head.Age                                           -2.955 0.003122 **
Type.of.HouseholdSingle Family                              -14.146  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.890 0.373724
House.Floor.Area                                             -1.595 0.110820
House.Age                                                    -2.953 0.003149 **
Number.of.bedrooms                                           -1.593 0.111204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(77568.65) family taken to be 1)

    Null deviance: 2024.3  on 1724  degrees of freedom
Residual deviance: 1331.5  on 1715  degrees of freedom
AIC: 7016.1

Number of Fisher Scoring iterations: 1

              Theta:  77569
          Std. Err.:  291876
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -6994.131
```

The variable Number.of.bedrooms has big p-value, so we reduce this variable and refit the model.

```
housenb_data3 <- glm.nb(Total.Number.of.Family.members~
                  Total.Household.Income+
                  Total.Food.Expenditure+
                  Household.Head.Sex+
                  Household.Head.Age+
                  Type.of.Household+
```

```
                  Total.Number.of.Family.members+
                  House.Floor.Area+
                  House.Age,
                  data=data)
housenb_data3%>%
  summary()
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + Total.Number.of.Family.members + House.Floor.Area +
    House.Age, data = data, init.theta = 77123.38422, link = log)

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              1.449e+00  6.743e-02
Total.Household.Income                                  -2.122e-07  5.572e-08
Total.Food.Expenditure                                   4.846e-06  3.352e-07
Household.Head.SexMale                                   2.209e-01  2.969e-02
Household.Head.Age                                      -2.727e-03  8.628e-04
Type.of.HouseholdSingle Family                          -3.503e-01  2.474e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.371e-01  1.598e-01
House.Floor.Area                                        -2.601e-04  1.231e-04
House.Age                                               -2.399e-03  7.691e-04
                                                        z value Pr(>|z|)
(Intercept)                                              21.484  < 2e-16 ***
Total.Household.Income                                   -3.808  0.00014 ***
Total.Food.Expenditure                                   14.459  < 2e-16 ***
Household.Head.SexMale                                    7.440 1.01e-13 ***
Household.Head.Age                                       -3.161  0.00157 **
Type.of.HouseholdSingle Family                          -14.163  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.857  0.39118
House.Floor.Area                                         -2.113  0.03461 *
House.Age                                                -3.119  0.00182 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(77123.38) family taken to be 1)

    Null deviance: 2024.3  on 1724  degrees of freedom
```

```
Residual deviance: 1334.0  on 1716  degrees of freedom
AIC: 7016.7

Number of Fisher Scoring iterations: 1

              Theta:  77123
          Std. Err.:  289568
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -6996.677
```

The variable Type.of.Household has big p-value, so we reduce this variable and refit the model.

```
housenb_data4 <- glm.nb(Total.Number.of.Family.members~
                  Total.Household.Income+
                  Total.Food.Expenditure+
                  Household.Head.Sex+
                  Household.Head.Age+
                  Total.Number.of.Family.members+
                  House.Floor.Area+
                  House.Age,
                  data=data)
housenb_data4%>%
  summary()
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Total.Number.of.Family.members + House.Floor.Area + House.Age,
    data = data, init.theta = 50502.13303, link = log)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             9.938e-01  5.935e-02  16.745  < 2e-16 ***
Total.Household.Income -2.442e-07  5.756e-08  -4.242 2.22e-05 ***
Total.Food.Expenditure  6.049e-06  3.246e-07  18.633  < 2e-16 ***
Household.Head.SexMale  1.960e-01  2.958e-02   6.625 3.47e-11 ***
Household.Head.Age      2.441e-04  8.331e-04   0.293  0.76950
House.Floor.Area       -2.369e-04  1.226e-04  -1.932  0.05342 .
```

```
House.Age              -2.053e-03  7.667e-04  -2.678  0.00741 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(50502.13) family taken to be 1)

    Null deviance: 2024.2  on 1724  degrees of freedom
Residual deviance: 1531.4  on 1718  degrees of freedom
AIC: 7210.1

Number of Fisher Scoring iterations: 1

              Theta:  50502
          Std. Err.:  233387
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -7194.058
```

The variable Household.Head.Age has big p-value, so we reduce this variable and refit the model.

```
housenb_data5 <- glm.nb(Total.Number.of.Family.members~
                  Total.Household.Income+
                  Total.Food.Expenditure+
                  Household.Head.Sex+
                  Total.Number.of.Family.members+
                  House.Floor.Area+
                  House.Age,
                  data=data)
housenb_data5%>%
  summary()
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Total.Number.of.Family.members +
    House.Floor.Area + House.Age, data = data, init.theta = 50396.01386,
    link = log)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)             1.007e+00  3.962e-02  25.408  < 2e-16 ***
Total.Household.Income -2.425e-07  5.726e-08  -4.236 2.28e-05 ***
Total.Food.Expenditure  6.041e-06  3.234e-07  18.677  < 2e-16 ***
Household.Head.SexMale  1.945e-01  2.914e-02   6.675 2.47e-11 ***
House.Floor.Area       -2.348e-04  1.224e-04  -1.918  0.05510 .
House.Age              -2.009e-03  7.518e-04  -2.673  0.00753 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(50396.01) family taken to be 1)

    Null deviance: 2024.2  on 1724  degrees of freedom
Residual deviance: 1531.5  on 1719  degrees of freedom
AIC: 7208.1

Number of Fisher Scoring iterations: 1

              Theta:  50396
          Std. Err.:  233012
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -7194.144
```

The variable House.Floor.Area has big p-value, so we reduce this variable and refit the model.

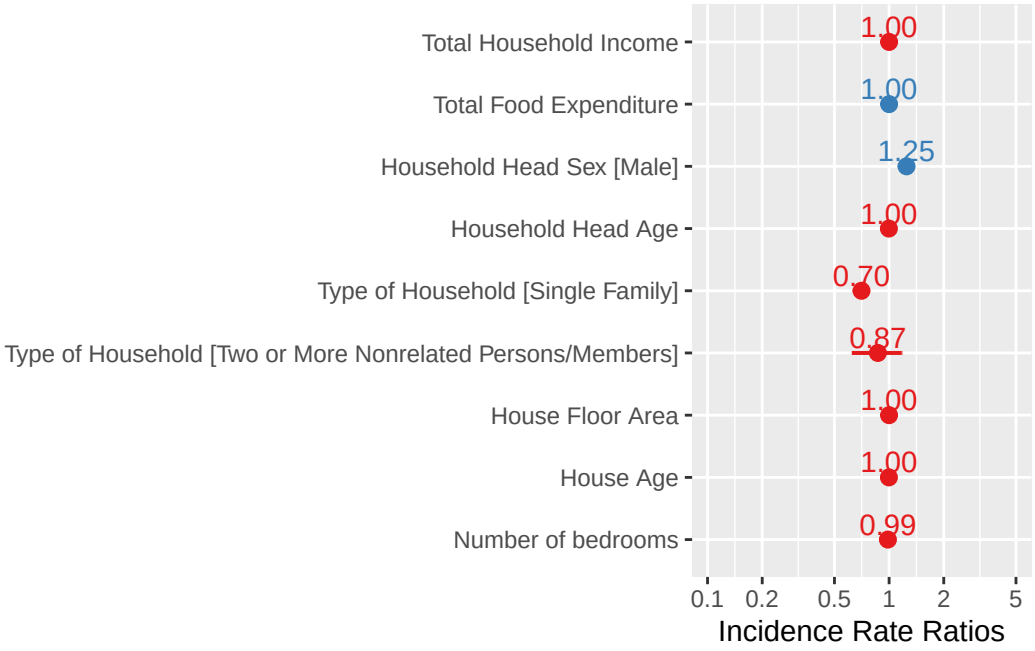
```
housenb_data6 <- glm.nb(Total.Number.of.Family.members~
                  Total.Household.Income+
                  Total.Food.Expenditure+
                  Household.Head.Sex+
                  Total.Number.of.Family.members+
                  House.Age,
                  data=data)
housenb_data6%>%
  summary()
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Total.Number.of.Family.members +
    House.Age, data = data, init.theta = 49935.62984, link = log)
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              9.914e-01  3.886e-02  25.515  < 2e-16 ***
Total.Household.Income -2.711e-07  5.617e-08  -4.827 1.38e-06 ***
Total.Food.Expenditure  6.076e-06  3.249e-07  18.702  < 2e-16 ***
Household.Head.SexMale  1.962e-01  2.913e-02   6.736 1.62e-11 ***
House.Age              -2.109e-03  7.505e-04  -2.810  0.00496 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(49935.63) family taken to be 1)

    Null deviance: 2024.2  on 1724  degrees of freedom
Residual deviance: 1535.2  on 1720  degrees of freedom
AIC: 7209.9

Number of Fisher Scoring iterations: 1

              Theta:  49936
          Std. Err.:  232615
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -7197.897
```

2.3.1 Odds ratios

```r
plot_model(housenb_data2, show.values = TRUE,
title = "", show.p = FALSE, value.offset = 0.25)
```



So the negative binomial regression method's result is the same as poisson.