Analysis of Factors Influencing Philippine Family Population Based on GLM

Group_03

Introduction

In Philippine, **FIES**(Family Income and Expenditure Survey), which is undertaken every three years, is aimed at providing data on family income and expenditure. This dataset, comes from the FIES recorded in the Philippines, is analysed in this report.

In particular, this report presents numerical and graphical summaries of FIES and fits a **Generalized Linear Model(GLM)** to analyze which household related variables influence the number of people living in a household.

Research Question

Which household related variables influence the number of people living in a household?

Data Cleaning

First read the data and tidy the data using tidyverse:

```
# Read the data set
Data_FIES <- read.csv("dataset03.csv")
# Tidy the data
FIES <- Data_FIES %>%
  # Place the dependent variable in the first column and delete the unique value
  select(Total.Number.of.Family.members, everything(), -Region) %>%
  # Convert categorical variables into factors
  mutate(
    Household.Head.Sex = as.factor(Household.Head.Sex),
```

```
    Type.of.Household = as.factor(Type.of.Household),
    Electricity = as.factor(Electricity)) %>%
  # Remove Missing Values
  drop_na()
FIES_saved <- FIES # Used for model fitting
```

We moved the implicit variable to the first row and deleted the region which has no difference in the data. 'Household.Head.Sex', 'Type.of.Household', and 'Electricity' were converted to factors to ensure that the statistical model correctly handled the categorical variables. Also we removed missing values to ensure data integrity.

The dependent variable and independent variables are shown as below:

Dependent Variable:

- **Total.Number.of.Family.members**: Number of people living in the house.

Independent Variables:

- **Total.Household.Income**: Annual household income (in Philippine peso)
- **Total.Food.Expenditure**: Annual expenditure by the household on food (in Philippine peso)
- **Household.Head.Sex**: Head of the households sex
- **Household.Head.Age**: Head of the households age (in years)
- **Type.of.Household**: Relationship between the group of people living in the house
- **House.Floor.Area**: Floor area of the house (in m^2)
- **House.Age**: Age of the building (in years)
- **Number.of.bedrooms**: Number of bedrooms in the house
- **Electricity**: Does the house have electricity? (1=Yes, 0=No)

Exploratory Data Analysis

Then we check the data structure and get summary statistics of all variables:

```
str(FIES) # Check data structure
```

```
'data.frame':   1887 obs. of  10 variables:
 $ Total.Number.of.Family.members: int  10 8 5 6 6 5 5 5 8 2 ...
 $ Total.Household.Income        : int  89359 108400 51982 76623 135232 73522 60369 60146 103
 $ Total.Food.Expenditure        : int  54537 56611 30827 43639 59614 47563 48962 47482 5566
 $ Household.Head.Sex            : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 2 1 2 2 2 .
 $ Household.Head.Age            : int  34 55 26 53 55 38 50 32 49 60 ...
 $ Type.of.Household             : Factor w/ 3 levels "Extended Family",..: 2 2 2 2 2 2 2 2 1
 $ House.Floor.Area              : int  64 60 48 42 56 56 48 42 48 20 ...
 $ House.Age                     : int  11 13 13 5 5 8 5 13 3 19 ...
 $ Number.of.bedrooms            : int  1 3 1 2 2 1 1 1 2 1 ...
 $ Electricity                   : Factor w/ 2 levels "0","1": 1 2 1 1 2 2 1 1 1 2 ...
```

```r
summary(FIES) # Get summary statistics of all variables
```

```
 Total.Number.of.Family.members Total.Household.Income Total.Food.Expenditure
 Min.   : 1.000                 Min.   :  16238        Min.   :  3704
 1st Qu.: 3.000                 1st Qu.:  85545        1st Qu.: 38311
 Median : 4.000                 Median : 131806        Median : 54594
 Mean   : 4.677                 Mean   : 214058        Mean   : 64113
 3rd Qu.: 6.000                 3rd Qu.: 249176        3rd Qu.: 77068
 Max.   :16.000                 Max.   :2598050        Max.   :363572
 Household.Head.Sex Household.Head.Age
 Female: 400        Min.   :15.00
 Male  :1487        1st Qu.:41.00
                    Median :51.00
                    Mean   :51.52
                    3rd Qu.:61.00
                    Max.   :95.00
                              Type.of.Household House.Floor.Area   House.Age
 Extended Family                       : 567    Min.   : 10.00   Min.   : 0.0
 Single Family                         :1311    1st Qu.: 30.00   1st Qu.:10.0
 Two or More Nonrelated Persons/Members:   9    Median : 50.00   Median :16.0
                                                Mean   : 59.81   Mean   :19.5
                                                3rd Qu.: 80.00   3rd Qu.:26.0
                                                Max.   :600.00   Max.   :95.0
 Number.of.bedrooms Electricity
 Min.   :0.000      0: 257
 1st Qu.:1.000      1:1630
 Median :2.000
 Mean   :1.945
 3rd Qu.:2.000
 Max.   :8.000
```

```
dim(FIES) # Check dataset dimensions (number of rows and columns)
```

```
[1] 1887   10
```

Numerical summaries and Data visualization

Now we can take a look at the numerical summaries and data visualization of **dependent variable** shown in the following tables and plots:

Table 1: Summary statistics for 'Total.Number.of.Family.members'

Mean	Median	Std. Dev	Minimum	Maximum	Interquartile Range	Sample Size
4.68	4.00	2.30	1.00	16.00	3.00	1,887.00

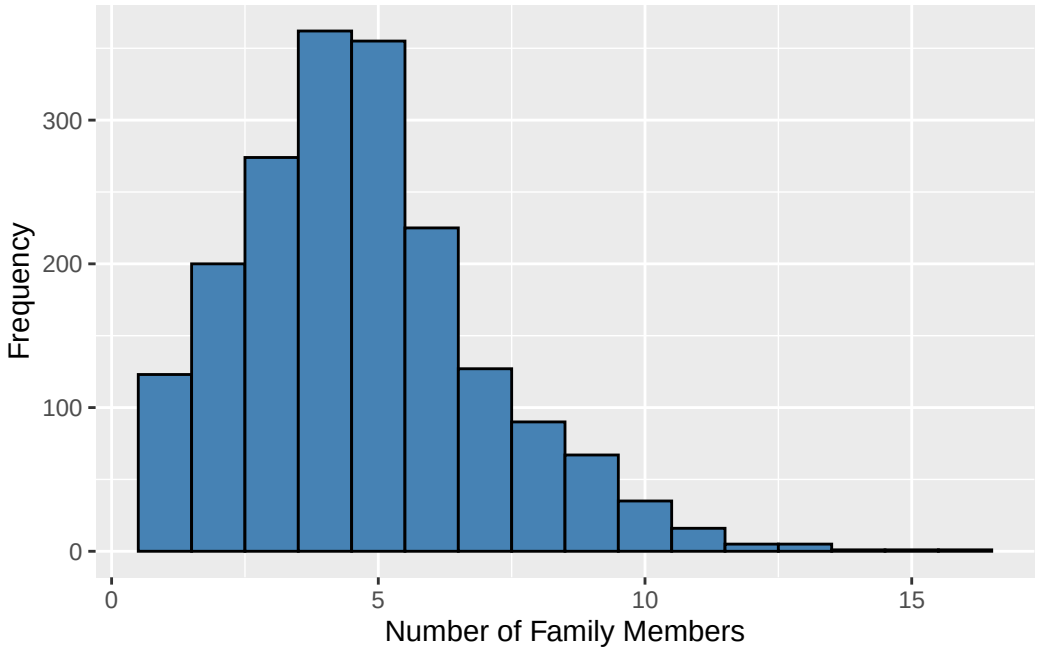

Figure 1: Histogram of 'Total.Number.of.Family.members'

From Table 1, we can see our dataset includes 1887 samples, which is a sufficiently large sample size to ensure reliability. The mean value (4.68) and median (4.00) are very close, suggesting a roughly symmetric distribution. However, the median is slightly lower than the mean hints

at a mild right skew in the data. The standard deviation (2.30) indicates moderate variability around the mean.

Figure 1 shows a strong right-skewed distribution of frequency data. The highest bar (around 300) is concentrated on the left side of the x-axis, indicating that most values are in the lower ranges. The frequency sharply decreased toward the right, with the far-right bars approaching 0, shows that rare occurrences in higher-value intervals.

Then check the variance of the dependent variable and compare with the mean:

```
mean = 4.677266
```

```
var = 5.28232
```

```
var_ratio = 1.129361
```

var/mean=5.28/4.68=1.13 1 indicates that there is no overdispersion problem, so Poisson regression model may be appropriate.

Then we separate independent variables into categorical variables and numerical variables for analysis.

1. **Categorical Variables**

Table 2: Summary statistics on 'Total.Number.of.Family.members' by 'Household.Head.Sex'

Household.Head.Sex	Mean	Median	Std. Dev	Minimum	Maximum	Interquartile Range	Samp
Female	3.83	3.00	2.33	1.00	15.00	3.00	
Male	4.91	5.00	2.23	1.00	16.00	3.00	1,

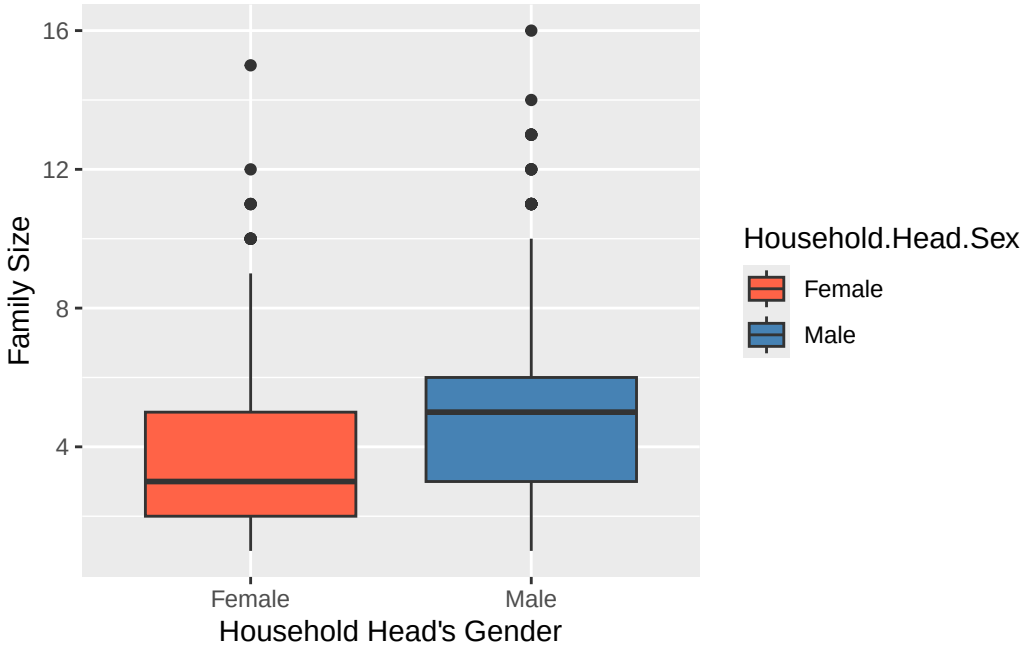

Figure 2: Boxplot of 'Total.Number.of.Family.members' by 'Household.Head.Sex'

Table 2 compares family size statistics between female-headed and male-headed households. Female-headed households exhibit smaller family size, with a mean of 3.83 and a median of 3.00, while male-headed households show significantly larger families (mean = 4.91, median = 5.00). Despite similar variability in both groups (standard deviations of nearly 2.3), the male-headed households display a broader range. Notably, the dataset is heavily skewed toward male-headed households (1487 samples vs. 400 female samples), which could influence the result. Both groups share identical interquartile ranges (IQR = 3.00), suggesting comparable central clustering of data.

Figure 2 shows that male-headed households exhibit a higher median family size of 5 members, compared to female-headed households with a median of 4 members. Both groups shows moderate variability in their distribution, but male-headed families display a wider range, with extreme outliers reaching up to 16 members which higher than the maximum of 12 members observed in female-headed households. This visualizes the tendency for male-led families to hold a larger household size.

Table 3: Summary statistics on 'Total.Number.of.Family.members' by 'Type.of.Household'

Type.of.Household	Mean	Median	Std. Dev	Minimum	Maximum	Inter
Single Family	4.14	4.00	2.03	1.00	13.00	
Extended Family	5.88	5.00	2.40	2.00	16.00	
Two or More Nonrelated Persons/Members	6.89	6.00	2.80	4.00	12.00	

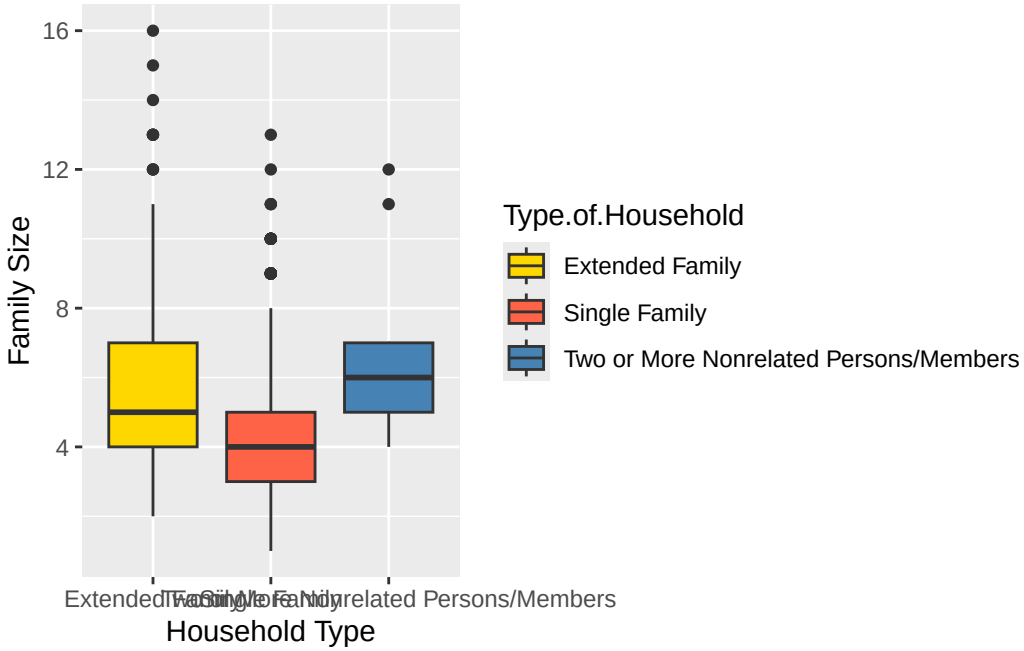

Figure 3: Boxplot of 'Total.Number.of.Family.members' by 'Type.of.Household'

Table 3 summarizes family size statistics across three household types. Single-family households have the smallest average family size (mean = 4.14, median = 4.00) with a large sample size (1311). Extended families show significantly larger family sizes (mean=5.88, median = 5.00) and a broader spread (max = 16). Households with two or more unrelated members report the highest average (mean = 6.89, median - 6.00), but with a extremely small sample size (9) weakens reliability.

Figure 3 shows that extended families exhibit the highest median family size (5) with a broader range which is up to 16 members, indicating potential outliers. Single families shows a lower median (4) and tighter clustering of data. It reflects a more consistent household size. Households with two or more unrelated members have a median of 6 members. However it only have a sample size of 9 which weakens the reliability of this category.

Table 4: Summary statistics on 'Total.Number.of.Family.members' by 'Electricity'

Electricity	Mean	Median	Std. Dev	Minimum	Maximum	Interquartile Range	Sample Size
0	4.70	5.00	2.47	1.00	12.00	3.00	257.00
1	4.67	4.00	2.27	1.00	16.00	3.00	1,630.00

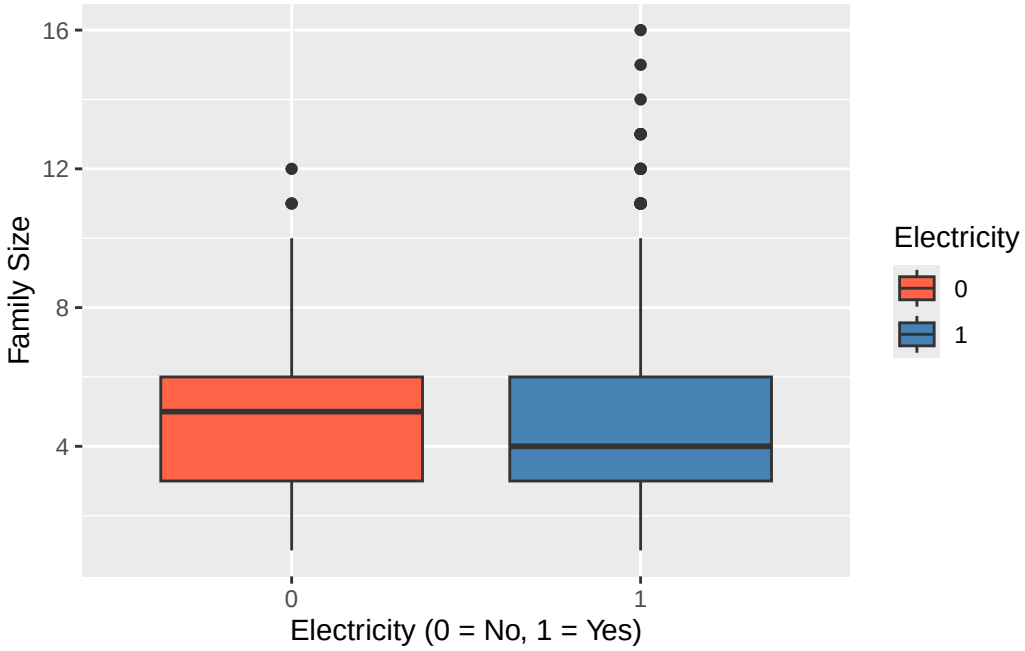

Figure 4: Boxplot of 'Total.Number.of.Family.members' by 'Electricity'

From Table 4, we can see households with electricity (1,630) are about six times more than those without (257). The mean values are nearly identical (4.70 vs. 4.67), but the median is slightly higher for households without electricity (5.00 vs. 4.00). The standard deviation is also slightly larger in the non-electric group (2.47 vs. 2.27), indicating a bit more variability. The maximum value is higher in the electricity group (16.00 vs. 12.00), suggesting a wider range. Both groups have the same interquartile range (3.00), meaning their middle 50% distributions are similar. These differences can be visualized more clearly with boxplots.

Figure 4 shows that households without electricity tend to have a slightly larger median family size compared to those with electricity. However, the distributions of family sizes in both groups are quite similar. The spread of family sizes in both groups is also comparable, as indicated by the interquartile range. Additionally, both groups exhibit several outliers, representing families with unusually large sizes, as shown by the points beyond the "whiskers" of the boxplots.

2. **Numerical Variables**

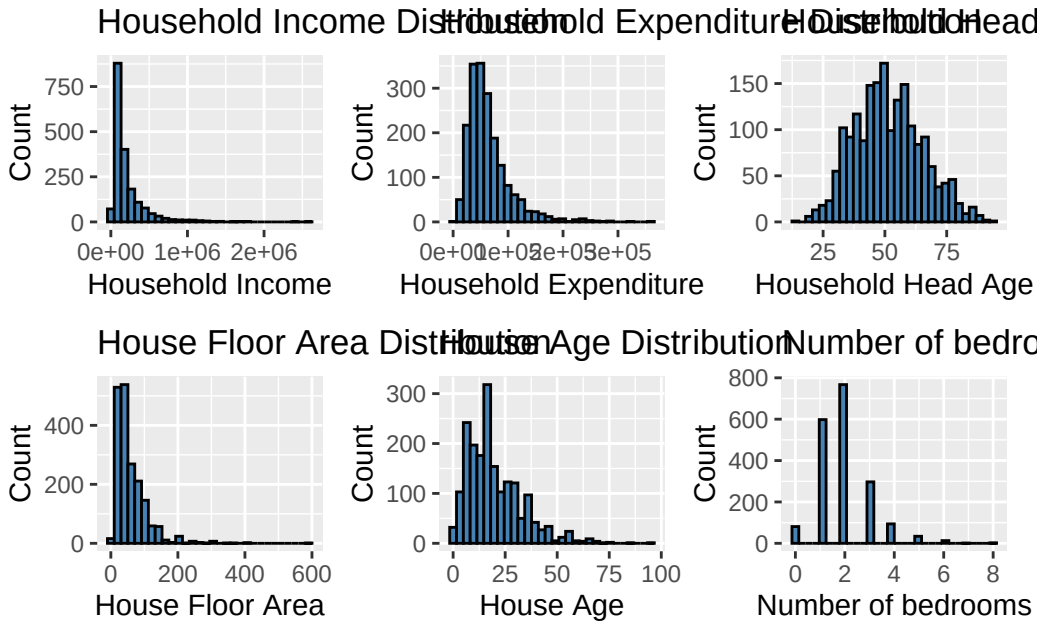

Figure 5: Histograms for Numerical Variables

From Figure 5, we can see the distributions of 'Total.Household.Income', 'Total.Food.Expenditure', and 'House.Floor.Area' are right-skewed, so applying a log transformation would be beneficial when fitting a Poisson model.

"Household.Head.Age" follows an approximately normal distribution, with most household heads falling within the 30-60 age range. This suggests that middle-aged individuals are the primary decision-makers in households.

The distribution of 'House.Floor.Area' is strongly right-skewed, with most houses having relatively small areas, while a few have significantly larger ones. The scarcity of large houses may be due to their higher costs.

The distribution of 'House.Age' exhibits a slight bimodal pattern, indicating the presence of two types of houses: newly houses and older houses.

The 'Number.of.bedrooms' is a discrete variable, with most houses having 2 or 3 bedrooms, while houses with 4 or more bedrooms are less common. So we can covert the 'Number.of.bedrooms' into categorical variable:

```r
# Convert 'Number.of.bedrooms' to categorical variable
FIES$Bedroom.Category <- cut(FIES$Number.of.bedrooms,
         c(-1, 1, 3, Inf),
         labels = c("Small", "Medium", "Large"),right = TRUE)
# Convert categorical variable into factor
FIES$Bedroom.Category <- factor(FIES$Bedroom.Category)
```

Then we can use the heat map to check the correlation between numerical variables and dependent variable:

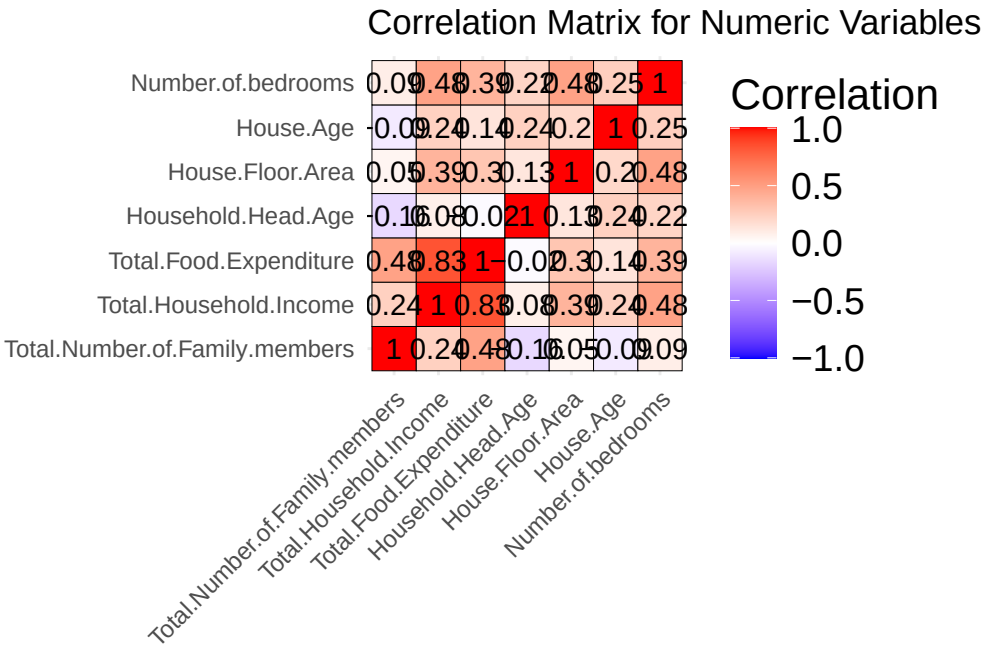

Figure 6: Heat map for numerical variables by 'Total.Number.of.Family.members'

Form Figure 6, we can see 'Total.Household.Income' and 'Total.Food.Expenditure' are strongly correlated, reinforcing the expected relationship that higher-income households tend to spend more. 'Total.Number.of.Family.members' influences 'Total.Food.expenditure', but its impact on 'Total.Household.Income' is weaker. 'Household.head.age' has a minor effect on most other variables, with a slight negative correlation with family size. 'House.Floor.Area' shows some correlation with income, but not significantly with 'Household.head.age'.

Formal Data Analysis

Since the dependent variable 'Total.Number.of.Family.members' is a typical count variable with a mean and variance that are approximately equal, Poisson regression was chosen for modeling. Based on the results of EDA, some variables were log-transformed to improve linear relationships and reduce heteroscedasticity. All selected variables were then included in the model, and stepwise regression using drop1(poisson_model, test = "F") was performed to assess variable significance, gradually eliminating insignificant variables to ensure the final model's robustness and explanatory power.

```
Call:
glm(formula = Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex +
    Head_Age + Household_Type + log(Floor_Area) + log(House_Age +
    0.1) + Bedrooms + Electricity, family = poisson(link = "log"),
    data = FIES)

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                       -2.7764518  0.2533804
log(Income)                                       -0.2750530  0.0282186
log(Food_Exp)                                      0.7366394  0.0386852
Head_SexMale                                       0.1461315  0.0290117
Head_Age                                          -0.0042208  0.0008871
Household_TypeSingle Family                       -0.2853906  0.0239477
Household_TypeTwo or More Nonrelated Persons/Members  0.0947877  0.1288759
log(Floor_Area)                                    0.0066325  0.0182464
log(House_Age + 0.1)                              -0.0450949  0.0140560
Bedrooms                                          -0.0016833  0.0123272
Electricity1                                      -0.1097757  0.0336198
                                                  z value Pr(>|z|)
(Intercept)                                       -10.958  < 2e-16 ***
log(Income)                                        -9.747  < 2e-16 ***
log(Food_Exp)                                      19.042  < 2e-16 ***
Head_SexMale                                        5.037 4.73e-07 ***
Head_Age                                           -4.758 1.96e-06 ***
Household_TypeSingle Family                       -11.917  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members   0.735  0.46204
log(Floor_Area)                                     0.363  0.71623
log(House_Age + 0.1)                               -3.208  0.00134 **
Bedrooms                                           -0.137  0.89138
Electricity1                                       -3.265  0.00109 **
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2130.1  on 1886  degrees of freedom
Residual deviance: 1218.5  on 1876  degrees of freedom
AIC: 7450.3

Number of Fisher Scoring iterations: 4

Single term deletions

Model:
Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex + Head_Age +
    Household_Type + log(Floor_Area) + log(House_Age + 0.1) +
    Bedrooms + Electricity
                     Df Deviance    AIC  F value    Pr(>F)
<none>                    1218.5 7450.3
log(Income)           1   1315.2 7545.0 148.8405 < 2.2e-16 ***
log(Food_Exp)         1   1588.8 7818.6 570.1446 < 2.2e-16 ***
Head_Sex              1   1244.6 7474.4  40.1738 2.903e-10 ***
Head_Age              1   1241.2 7471.0  34.9168 4.077e-09 ***
Household_Type        2   1362.3 7590.1 110.6954 < 2.2e-16 ***
log(Floor_Area)       1   1218.7 7448.4   0.2034    0.6520
log(House_Age + 0.1)  1   1228.7 7458.5  15.6560 7.880e-05 ***
Bedrooms              1   1218.5 7448.3   0.0287    0.8654
Electricity           1   1229.0 7458.7  16.1106 6.210e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For Floor_Area, since it is a continuous variable, a log transformation was applied during the EDA process to improve linearity and reduce heteroscedasticity. However, it remained insignificant in the Poisson regression model even after transformation, so it is considered for removal. As for Bedrooms, it is a discrete integer variable with a large number of zero values. Given the potential categorical effect, converting it into a categorical variable may be more appropriate to better capture its impact on family size.

```r
FIES$Bedroom.Category <- cut(FIES$Bedrooms,
                             breaks = c(-1, 1, 3, Inf),
                             labels = c("Small", "Medium", "Large"),
                             right = TRUE)
```

```r
FIES$Bedroom.Category <- factor(FIES$Bedroom.Category)
```

```
Call:
lm(formula = Family_Size ~ Bedroom.Category, data = FIES)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0208 -1.8102 -0.3962  1.1898 11.1898

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.39617    0.08785  50.041  < 2e-16 ***
Bedroom.CategoryMedium 0.41398    0.11244   3.682 0.000238 ***
Bedroom.CategoryLarge  0.62466    0.21002   2.974 0.002974 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.289 on 1884 degrees of freedom
Multiple R-squared:  0.008977,  Adjusted R-squared:  0.007925
F-statistic: 8.533 on 2 and 1884 DF,  p-value: 0.0002045
```

```
Call:
glm(formula = Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex +
    Head_Age + Household_Type + log(House_Age + 1) + Bedroom.Category +
    Electricity, family = poisson(link = "log"), data = FIES)

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                       -2.7447427  0.2518361
log(Income)                                       -0.2719520  0.0276354
log(Food_Exp)                                      0.7340369  0.0386511
Head_SexMale                                       0.1436677  0.0290257
Head_Age                                          -0.0042813  0.0008873
Household_TypeSingle Family                       -0.2843104  0.0238796
Household_TypeTwo or More Nonrelated Persons/Members  0.1026948  0.1287961
log(House_Age + 1)                                -0.0519835  0.0159625
Bedroom.CategoryMedium                             0.0295704  0.0254386
Bedroom.CategoryLarge                             -0.0171166  0.0471953
Electricity1                                      -0.1169705  0.0339937
```

```
                                                    z value Pr(>|z|)
(Intercept)                                        -10.899  < 2e-16 ***
log(Income)                                         -9.841  < 2e-16 ***
log(Food_Exp)                                       18.991  < 2e-16 ***
Head_SexMale                                         4.950 7.43e-07 ***
Head_Age                                            -4.825 1.40e-06 ***
Household_TypeSingle Family                        -11.906  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members  0.797  0.42525
log(House_Age + 1)                                  -3.257  0.00113 **
Bedroom.CategoryMedium                               1.162  0.24506
Bedroom.CategoryLarge                               -0.363  0.71685
Electricity1                                        -3.441  0.00058 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2130.1  on 1886  degrees of freedom
Residual deviance: 1216.0  on 1876  degrees of freedom
AIC: 7447.8

Number of Fisher Scoring iterations: 4

Single term deletions

Model:
Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex + Head_Age +
    Household_Type + log(House_Age + 1) + Bedroom.Category +
    Electricity
                  Df Deviance    AIC  F value    Pr(>F)
<none>                 1216.0 7447.8
log(Income)        1   1314.7 7544.4 152.1694 < 2.2e-16 ***
log(Food_Exp)      1   1584.2 7814.0 567.9627 < 2.2e-16 ***
Head_Sex           1   1241.2 7471.0  38.8521 5.632e-10 ***
Head_Age           1   1239.4 7469.1  35.9972 2.366e-09 ***
Household_Type     2   1359.7 7587.5 110.8329 < 2.2e-16 ***
log(House_Age + 1) 1   1226.6 7456.3  16.2519 5.767e-05 ***
Bedroom.Category   2   1218.5 7446.2   1.8675    0.1548
Electricity        1   1227.7 7457.4  17.9213 2.414e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After converting Bedrooms into a categorical variable and refitting the model, it remained

insignificant, so it was ultimately removed.

```
Call:
glm(formula = Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex +
    Head_Age + Household_Type + log(House_Age + 1) + Electricity,
    family = poisson(link = "log"), data = FIES)

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                       -2.7453484  0.2411600
log(Income)                                       -0.2731779  0.0268843
log(Food_Exp)                                      0.7355788  0.0386450
Head_SexMale                                       0.1454000  0.0289923
Head_Age                                          -0.0041994  0.0008776
Household_TypeSingle Family                       -0.2847395  0.0238574
Household_TypeTwo or More Nonrelated Persons/Members  0.0961818  0.1286671
log(House_Age + 1)                                -0.0512741  0.0158517
Electricity1                                      -0.1092392  0.0335128
                                                  z value Pr(>|z|)
(Intercept)                                       -11.384  < 2e-16 ***
log(Income)                                       -10.161  < 2e-16 ***
log(Food_Exp)                                      19.034  < 2e-16 ***
Head_SexMale                                        5.015 5.30e-07 ***
Head_Age                                           -4.785 1.71e-06 ***
Household_TypeSingle Family                       -11.935  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members   0.748  0.45475
log(House_Age + 1)                                 -3.235  0.00122 **
Electricity1                                       -3.260  0.00112 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2130.1  on 1886  degrees of freedom
Residual deviance: 1218.5  on 1878  degrees of freedom
AIC: 7446.2

Number of Fisher Scoring iterations: 4
```

Although the classification of Household_Type is not significant statistically, it is retained based on the theoretical justification from the original data classification. This ensures that its impact is still considered during model interpretation.

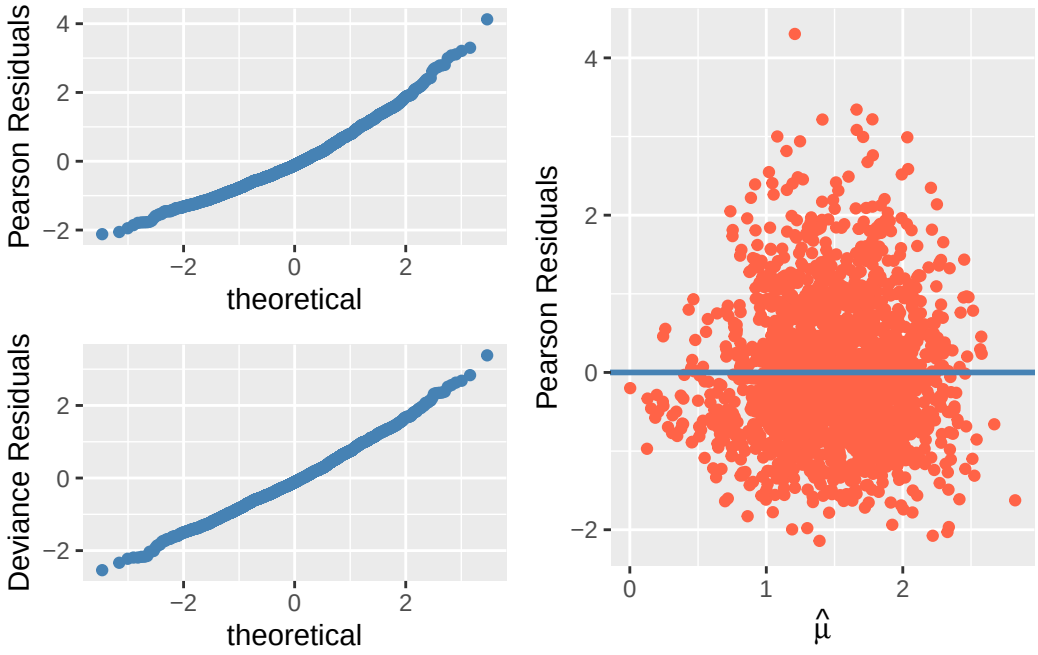

Figure 7: The Pearson and deviance residuals against the linear predictor

By plotting the QQ plots of Pearson and Deviance, we found that the data conform to the normality assumption. To further assess whether there is significant overdispersion in the Deviance distribution, we compared the performance of the original Poisson model and the Quasi-Poisson model in handling overdispersion. By comparing the standardized Pearson residuals of both models, we found that the original Poisson model performed better, indicating no significant overdispersion. Therefore, no adjustment was made to the standard errors of the parameters, as overdispersion was not evident.

```
Call:
glm(formula = Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex +
    Head_Age + Household_Type + log(House_Age + 1) + Electricity,
    family = poisson(link = "log"), data = FIES)

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                       -2.7453484  0.1956354
log(Income)                                       -0.2731779  0.0218092
log(Food_Exp)                                      0.7355788  0.0313499
Head_SexMale                                       0.1454000  0.0235193
Head_Age                                          -0.0041994  0.0007119
```

```
Household_TypeSingle Family                             -0.2847395  0.0193537
Household_TypeTwo or More Nonrelated Persons/Members     0.0961818  0.1043781
log(House_Age + 1)                                      -0.0512741  0.0128593
Electricity1                                            -0.1092392  0.0271865
                                                        z value Pr(>|z|)
(Intercept)                                             -14.033  < 2e-16 ***
log(Income)                                             -12.526  < 2e-16 ***
log(Food_Exp)                                            23.464  < 2e-16 ***
Head_SexMale                                              6.182 6.32e-10 ***
Head_Age                                                 -5.899 3.66e-09 ***
Household_TypeSingle Family                             -14.712  < 2e-16 ***
Household_TypeTwo or More Nonrelated Persons/Members      0.921    0.357
log(House_Age + 1)                                       -3.987 6.68e-05 ***
Electricity1                                             -4.018 5.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 0.6580882)

    Null deviance: 2130.1  on 1886  degrees of freedom
Residual deviance: 1218.5  on 1878  degrees of freedom
AIC: 7446.2

Number of Fisher Scoring iterations: 4
```

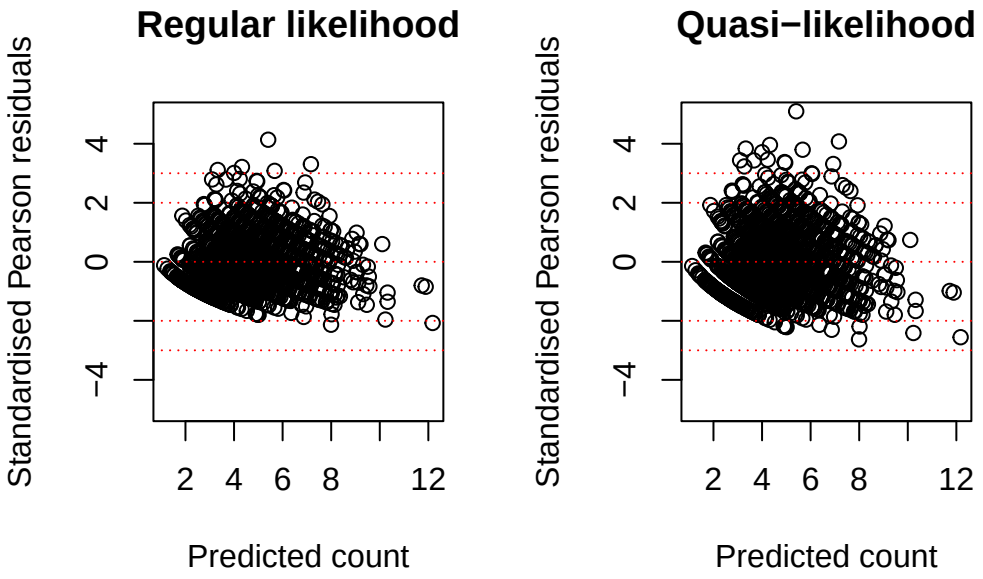

Figure 8: Regular likelihood and Quasi-likelihood

Based on the model comparison and LRT test results, and considering the high correlation observed in the EDA heatmap, an interaction term between Income and Food_Exp was introduced. After comparing the original model with the model including the interaction term, it was found that the Deviance decreased by 47.118 after adding the interaction term, indicating that the inclusion of the interaction term significantly improved the model's fit. Therefore, the conclusion can be made that introducing the interaction term effectively enhanced the model's explanatory power and fit.

```
Analysis of Deviance Table

Model 1: Family_Size ~ log(Income) + log(Food_Exp) + Head_Sex + Head_Age +
    Household_Type + log(House_Age + 1) + Electricity
Model 2: Family_Size ~ log(Income) * log(Food_Exp) + Head_Sex + Head_Age +
    Household_Type + log(House_Age + 1) + Electricity
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1878     1218.5
2      1877     1171.3  1   47.118 6.685e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then we can check the Relative Risk(RR):

```
                                        (Intercept)
                                       3.507819e-10
                                        log(Income)
                                       3.790341e+00
                                      log(Food_Exp)
                                       1.124965e+01
                                       Head_SexMale
                                       1.147725e+00
                                           Head_Age
                                       9.962935e-01
                          Household_TypeSingle Family
                                       7.537601e-01
Household_TypeTwo or More Nonrelated Persons/Members
                                       1.085301e+00
                                 log(House_Age + 1)
                                       9.461484e-01
                                       Electricity1
                                       8.608042e-01
                       log(Income):log(Food_Exp)
                                       8.679292e-01
```

Based on the results of Relative Risk(RR), the following key conclusions can be summarized:

- Income's Impact on Family Size: When income increases by 1%, the average family size increases by approximately 3.79 times. This indicates a positive correlation between income and family size, suggesting that higher income may be associated with larger families.
- Food Expenditure's Impact on Family Size: When food expenditure increases by 1%, the average family size increases by approximately 11.25 times. This suggests that higher food spending is closely related to an increase in family size.
- House Age's Impact on Family Size: When house age increases by 1%, the average family size decreases by approximately 5.39%. This implies that older houses may be associated with smaller family sizes, possibly due to living conditions or other factors.
- Interaction Between Income and Food Expenditure: When both income and food expenditure increase by 1%, the average family size decreases by approximately 13.21%. This suggests that while both variables have a positive individual effect, their interaction shows a negative association, possibly indicating that higher income combined with higher food expenditure may not lead to a larger family size.
- Head Sex's Impact on Family Size: Families with male heads tend to have approximately 15% more members than those with female heads. This indicates that families with male heads may be larger than those with female heads.

- Head Age's Impact on Family Size: For each additional year in the head's age, the average family size decreases by approximately 0.4%. This suggests that older heads of households may have smaller families.
- Household Type's Impact on Family Size: If the household type is a single-parent family, the average family size is about 25% smaller compared to other household types. Single-parent families generally have smaller family sizes.

Conclusion

In conclusion, these results show that family size is closely related to various factors, particularly income, food expenditure, head sex, and household type. Changes in family size are influenced not only by individual variables but also by the interactions between them.