Project 2

group 2

1 Exploratory Data Analysis

1.1 Dataset Description:

- **Total.Household.Income** – Annual household income (in Philippine peso)
- **Region** – The region of the Philippines which you have data for
- **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
- **Total.Number.of.Family.members** – Number of people living in the house
- **House.Floor.Area** – Floor area of the house (in m2)
- **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)

1.2 Summary Analysis

Table 1: Summary statistics

Variable	n	Mean	SD	Min	Q1	Median	Q3	Max
Total Household Income	1249	216685.12	263207.20	18784	92101	140483	230402	2891788
Total Household Expenditure	1249	70760.29	41638.03	10488	43751	62590	86708	413844
Household Head Age	1249	51.37	14.24	15	41	51	61	87
Number of Family Members	1249	4.39	2.19	1	3	4	6	16
Floor Area	1249	48.95	49.43	5	20	36	60	750
House Age	1249	16.49	12.51	0	8	14	22	105
Number of Bedrooms	1249	1.78	0.98	0	1	2	2	7

According to the descriptive statistic, there is not any missing value in this database and the number of observations is 1249. There are large standard deviations observed in total household income and total household expenditure, which means the range between rich and poor in this survey is quite considerable. The average capacity of a family is more than 4 people, which ranges from 1 to 16. Most houses have a floor area lies in 5 to 60 square meters, while some houses are drastically big and can be measured up to 750 square meters. Three quarters of the houses are under 22 years old. However, the oldest house could be dated back to 105 years ago. The number of bedrooms ranges from 0 to 7, which has an average of 1.78.

1.3 Boxplots

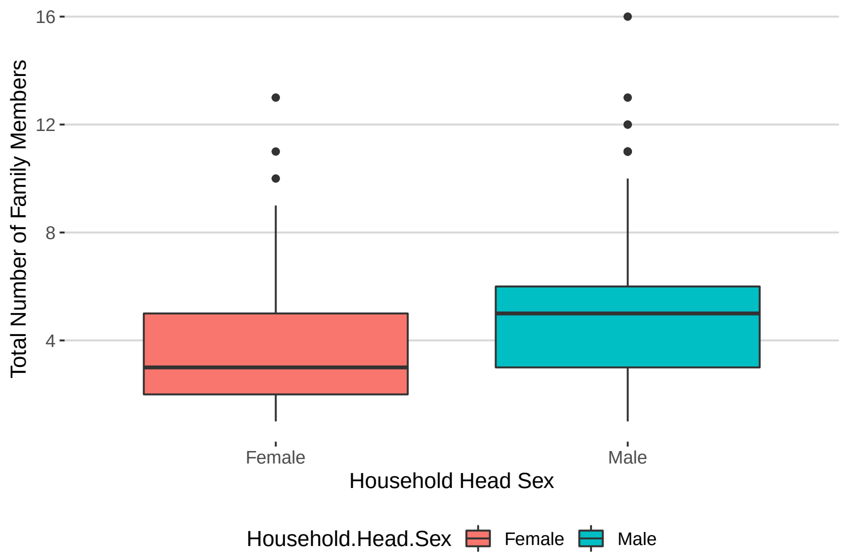

Figure 1: Number of Family Members by Family Head Sex.

The households that have a male head tend to have a larger family, i.e. the median and interquartile range of male head household are larger compared to female head household.

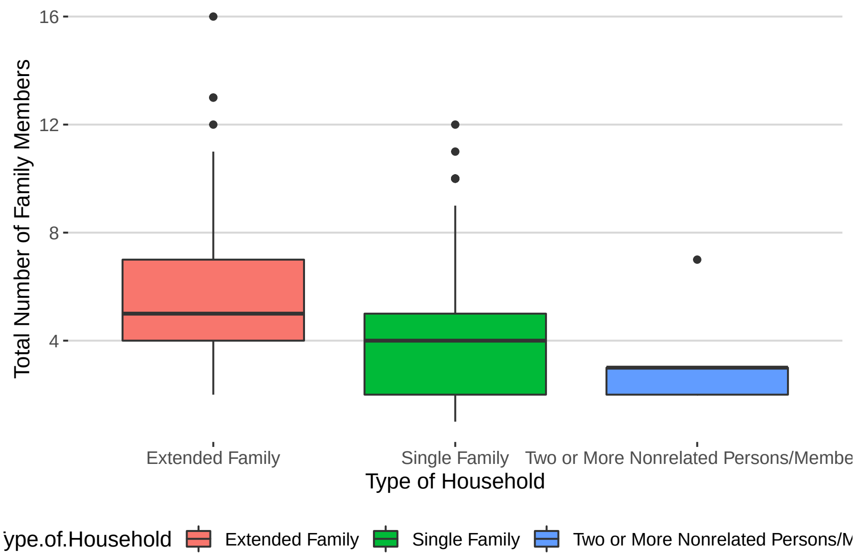

Figure 2: Number of Family Members by Type of Household.

The extended families accommodate more family members compared to other two types, which have a larger median as well. Single families are slightly smaller but still larger than families made with two or more non-related persons.

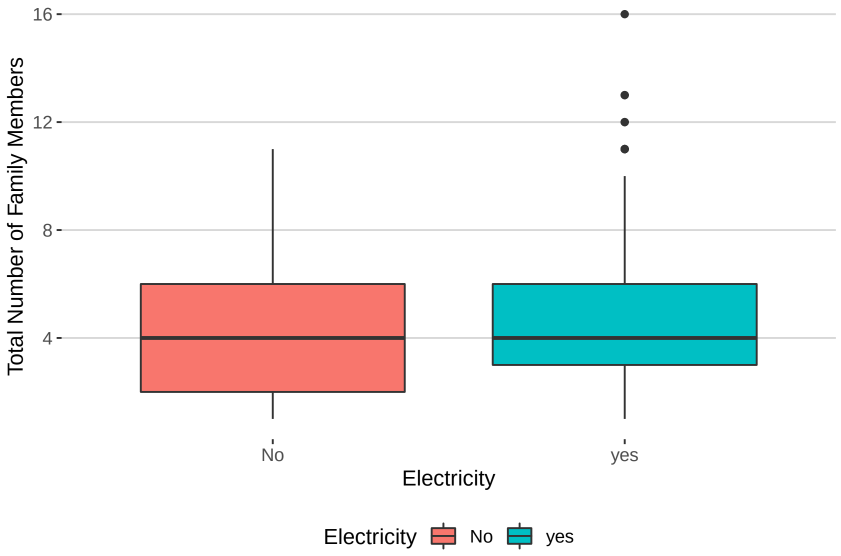

Figure 3: Number of Family Members by Electricity.

The influence of electricity to family capacity is not clear since the median values are same between households with and without electricity. However, the interquatile range of no electricity families is larger which means more minor families live without electricity.

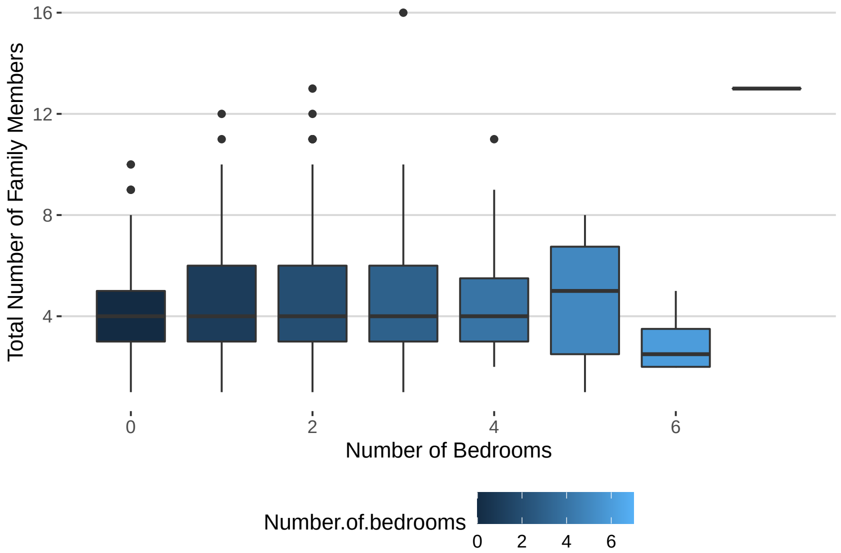

Figure 4: Number of Family Members by Number of Bedrooms.

The median family members in household with five bedrooms shows the largest, which is 5, while which stay 4 in households have under five bedrooms. Households with five bedrooms also have the largest interquartile range, shows a big difference in family capacities. Nevertheless, households with six bedrooms shows the lowest median and range.

1.4 ggplots

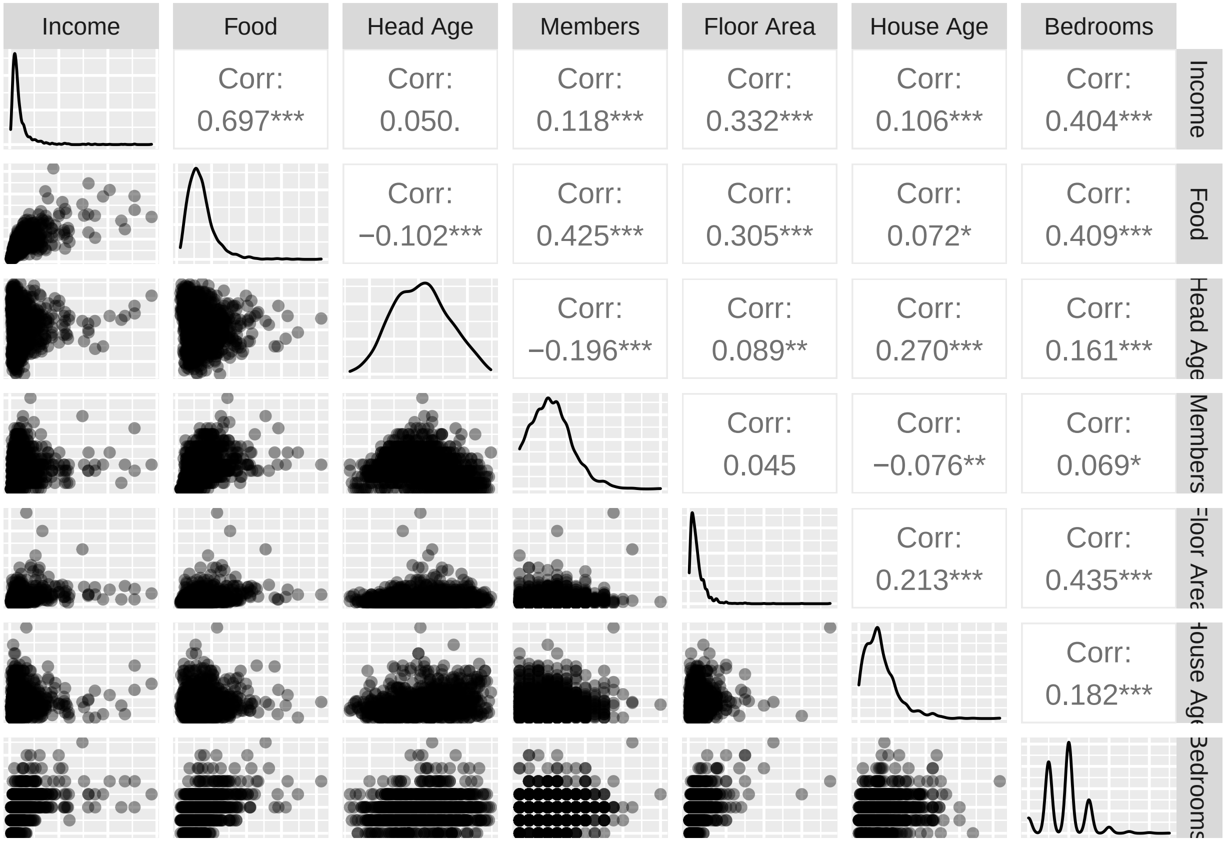

Figure 5: Correlation Plot

The correlation between household food expenditure and household income is large, which indicates the multicollinearity may need take into consideration. (bedroom&income/food/floor area?) According to the correlation in this graph, the number of family members has a moderate positive relationship with household food expenditure, while other relationships are not obvious and need further research.

2 Formal Analysis

Let's have a look at the GLM Poisson model of Response against all Explanatory variables.

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms + Type.of.Household + Household.Head.Sex +
    Electricity, family = poisson, data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6392  -0.6578  -0.1209   0.5018   2.7098
```

```
Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              1.671e+00  8.230e-02
Total.Household.Income                                  -4.266e-07  7.596e-08
Total.Food.Expenditure                                   5.239e-06  4.066e-07
Household.Head.Age                                      -5.818e-03  1.080e-03
House.Floor.Area                                        -9.056e-05  3.033e-04
House.Age                                               -2.451e-03  1.177e-03
Number.of.bedrooms                                      -2.366e-02  1.680e-02
Type.of.HouseholdSingle Family                          -3.732e-01  3.047e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -5.036e-01  2.447e-01
Household.Head.SexMale                                   2.418e-01  3.739e-02
Electricityyes                                          -5.232e-02  4.048e-02
                                                        z value Pr(>|z|)
(Intercept)                                              20.299  < 2e-16 ***
Total.Household.Income                                   -5.616 1.96e-08 ***
Total.Food.Expenditure                                   12.886  < 2e-16 ***
Household.Head.Age                                       -5.386 7.21e-08 ***
House.Floor.Area                                         -0.299   0.7653
House.Age                                                -2.082   0.0374 *
Number.of.bedrooms                                       -1.409   0.1589
Type.of.HouseholdSingle Family                          -12.250  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -2.058   0.0396 *
Household.Head.SexMale                                    6.467 1.00e-10 ***
Electricityyes                                           -1.293   0.1961
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1373.63  on 1248  degrees of freedom
Residual deviance:  881.01  on 1238  degrees of freedom
AIC: 4931.9

Number of Fisher Scoring iterations: 4
```

Here it can been that few explanatory variables have insignificant p value. And confident intervals also contains zero so Let’s work on these variables. Let’s look at individual models, response against explanatory variables.

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Household.Head.Age + Number.of.bedrooms + Household.Head.Sex,
    family = poisson, data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4190  -0.7879  -0.1182   0.5704   3.9549

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.491e+00  6.638e-02  22.460  < 2e-16 ***
Total.Household.Income  1.737e-07  5.019e-08   3.461 0.000538 ***
Household.Head.Age     -6.054e-03  1.006e-03  -6.021 1.74e-09 ***
```

```
Number.of.bedrooms      3.035e-02  1.508e-02   2.013 0.044131 *
Household.Head.SexMale  2.499e-01  3.695e-02   6.764 1.34e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1373.6  on 1248  degrees of freedom
Residual deviance: 1247.3  on 1244  degrees of freedom
AIC: 5286.1

Number of Fisher Scoring iterations: 4
```

From this analysis we can see that the variables given defines the model better so we will go with these variables and let's have a look at model fit.

2.1 Residual Plots

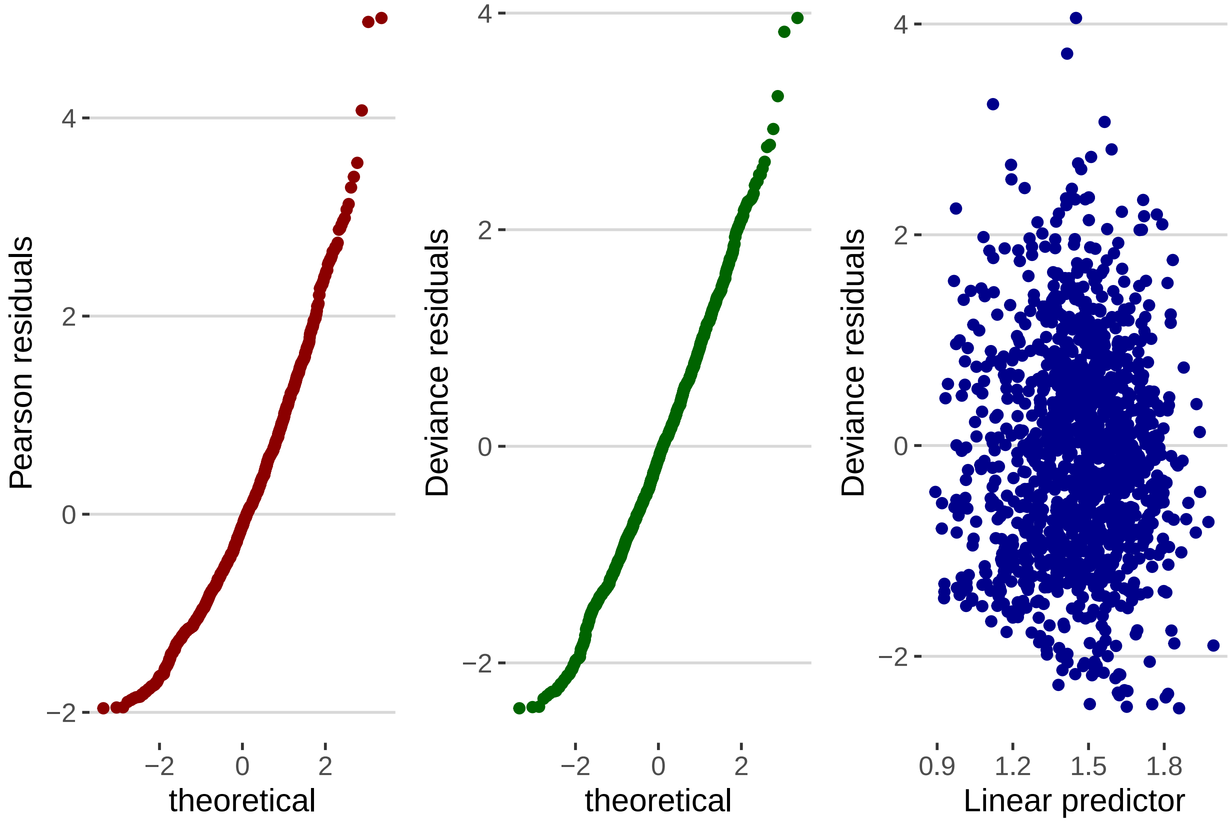

Figure 6: Residual Plots

Here in the grid, Probability plots are showing no deviation from the line. The third plot shows that there is no obvious pattren in our residuals.

2.2 Mean and Variance of Response

Let's have a look at the mean and the variance see the over dispersion.

```
[1] 1247.271
```

Table 2: Mean and Variance of Response variable

	Response
Var	4.784
Mean	4.395

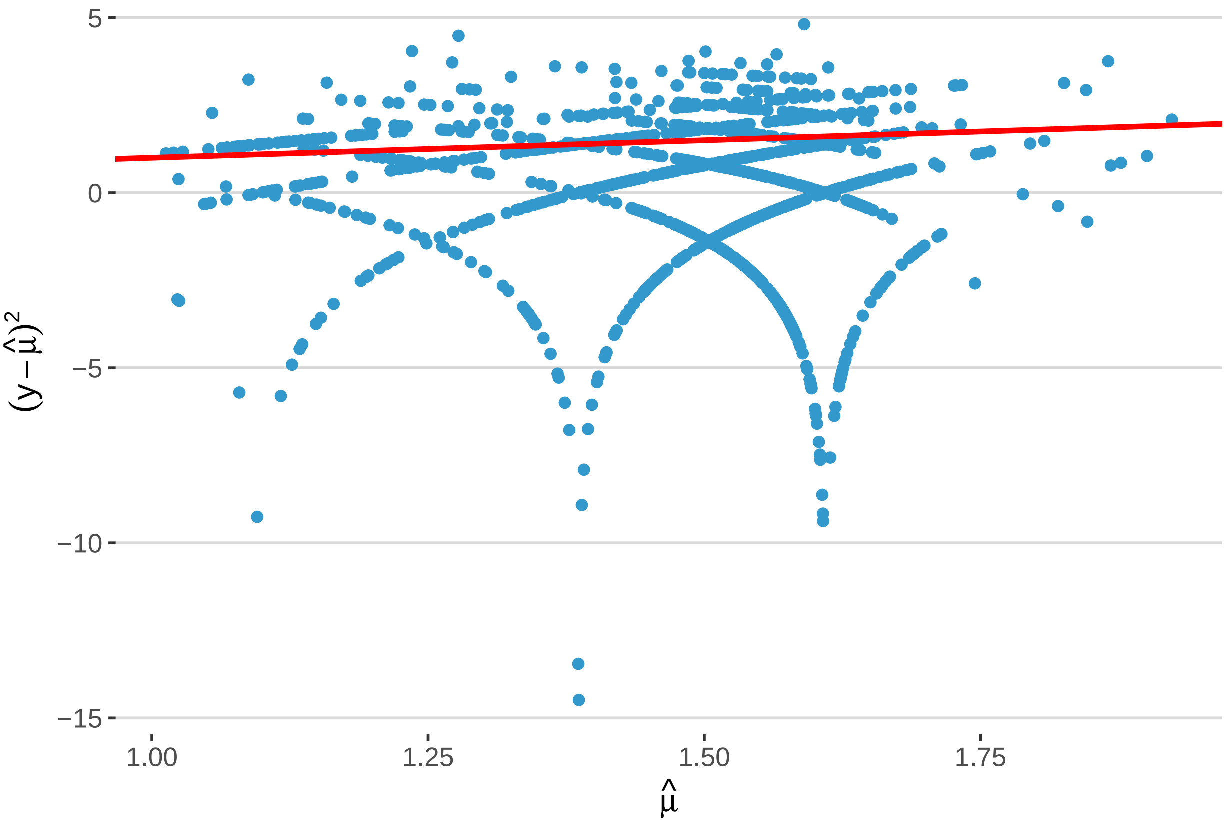

Figure 7: Mean vs Variance

The residual deviance of model is 1247.3 and degree of freedom is 1244. Mean and Variance of response variable have a little difference of 0.4 which can indicate to overdispersion. Furthermore the plot between fitted values and σ^2 we can see that there is no problem of dispersion as the residuals have almost equal spread agaist the line. We will proceed to with chi square test, and dispersion test from *AER* library to see how good is our model.

2.3 Goodness of fit

Table 3: Chi Square Test: Null Model vs Complete Model

	values
D0-D1	126.360
X^2 (10)	9.488

Table 4: Chi Square Test: Residual Deviance

	values
resid D	1247.271
X^2 (1238)	1327.166

Table 5: Dispersion

	values
dispersion	1.013
Disp Parameter	1.016

From the comparison of null model and full model in table 3 we can see that the $\chi^2(4)$ value is less than the difference between null deviance and residual deviance. which can indicate lack of fit. Secondly from the table 4 we can see that residual deviance is greater than $\chi^2(1244)$, which can also lead to lack of fit. lastly, we can see that dispersion test and dispersion parameter value is same So we can try Quasipoisson using **dispersion parameter**.

2.4 Dispersion

2.4.1 Dispersion Parameeter

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

As because of dispersion parameter Wald statistic is not valid, so to check the significance of coefficients we go for "**F**" test.

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Household.Head.Age + Number.of.bedrooms + Household.Head.Sex,
    family = poisson, data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4190  -0.7879  -0.1182   0.5704   3.9549

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.491e+00  6.690e-02  22.286  < 2e-16 ***
```

```
Total.Household.Income  1.737e-07  5.058e-08   3.434 0.000594 ***
Household.Head.Age     -6.054e-03  1.013e-03  -5.974 2.32e-09 ***
Number.of.bedrooms      3.035e-02  1.520e-02   1.997 0.045803 *
Household.Head.SexMale  2.499e-01  3.724e-02   6.712 1.92e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for poisson family taken to be 1.015724)

    Null deviance: 1373.6  on 1248  degrees of freedom
Residual deviance: 1247.3  on 1244  degrees of freedom
AIC: 5286.1

Number of Fisher Scoring iterations: 4
```

Table 6: F Test

	Df	Deviance	AIC	F value	Pr(>F)
<none>	NA	1247.271	5286.128	NA	NA
Total.Household.Income	1	1258.455	5295.312	11.155	0.001
Household.Head.Age	1	1283.605	5320.462	36.239	0.000
Number.of.bedrooms	1	1251.307	5288.164	4.025	0.045
Household.Head.Sex	1	1295.408	5332.265	48.011	0.000

Table 7: Confidence Interveals

	2.5 %	97.5 %
(Intercept)	1.360	1.621
Total.Household.Income	0.000	0.000
Household.Head.Age	-0.008	-0.004
Number.of.bedrooms	0.001	0.060
Household.Head.SexMale	0.178	0.323

According to the confidence intervals and F statistic, we can see that there is no more insignificant variables. To deal with a little bit of dispersion in model we go for Quasipoisson of negative binomial model and observe results

2.4.2 Quasipoisson

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Household.Head.Age + Number.of.bedrooms + Household.Head.Sex,
    family = quasipoisson, data = dataset2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4190  -0.7879  -0.1182   0.5704   3.9549

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

```
(Intercept)             1.491e+00  6.690e-02  22.285  < 2e-16 ***
Total.Household.Income  1.737e-07  5.058e-08   3.434 0.000614 ***
Household.Head.Age     -6.054e-03  1.013e-03  -5.974 3.02e-09 ***
Number.of.bedrooms      3.035e-02  1.520e-02   1.997 0.046024 *
Household.Head.SexMale  2.499e-01  3.724e-02   6.712 2.92e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.015753)

    Null deviance: 1373.6  on 1248  degrees of freedom
Residual deviance: 1247.3  on 1244  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

2.4.3 Negative Binomial

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Household.Head.Age + Number.of.bedrooms + Household.Head.Sex,
    data = dataset2, init.theta = 9993.84637, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4185  -0.7877  -0.1182   0.5702   3.9533

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.491e+00  6.639e-02  22.455  < 2e-16 ***
Total.Household.Income  1.737e-07  5.020e-08   3.460 0.000539 ***
Household.Head.Age     -6.054e-03  1.006e-03  -6.020 1.75e-09 ***
Number.of.bedrooms      3.035e-02  1.508e-02   2.012 0.044190 *
Household.Head.SexMale  2.499e-01  3.696e-02   6.763 1.35e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for Negative Binomial(9993.846) family taken to be 1)

    Null deviance: 1373.0  on 1248  degrees of freedom
Residual deviance: 1246.7  on 1244  degrees of freedom
AIC: 5288.1

Number of Fisher Scoring iterations: 1

              Theta:  9994
          Std. Err.:  153469
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -5276.129
```

2.5 Model Comparison

We can compare the Poisson and Negative Binomial models by looking at their deviance and AIC scores. Below in the table we can see that AIC can't be calculated for Quasipoisson model and the deviance much smaller for the negative binomial model so the Negative Binomial model is preferred over the Poisson model

Table 8: Model Comparison

	Deviance	AIC
QuasiPoisson	1247.271	NA
Negative Binomial	1246.723	5288.129

2.6 Final Results

Here in the table below, we can see that the estimates and their confidence intervals which explains the significance of model parameters.

Table 9: Results

	Estimate	2.5 %	97.5 %
(Intercept)	4.441	3.897	5.056
Total.Household.Income	1.000	1.000	1.000
Household.Head.Age	0.994	0.992	0.996
Number.of.bedrooms	1.031	1.001	1.062
Household.Head.SexMale	1.284	1.195	1.381

2.6.1 Model Equation

$$\widehat{Members}_i = 4.441 + 1.00 \cdot \text{income}_i + 0.994 \cdot \text{age}_i + 1.031 \cdot \text{bedrooms}_i + 1.284 \cdot I_{sex}\ (i)$$

$$I_{sex}(i) = \begin{cases} 1 & \text{if } ith \text{ observation is male,} \\ 0 & \text{Otherwise.} \end{cases}$$

$members_i$ – Response variable; Number of people living in the house
$income_1$ – Annual household income (in Philippine peso)
age_i – Head of the households age (in years)
$bedrooms$ – Number of bedrooms in the house
sex_i – Head of the households sex