Group_02_Analysis

Group 02

Data

In this project, we used a dataset to study the impact of other variables on Total.Number.of.Family.members. Next, we will use a series of methods to identify a few suitable variables to begin research and perform GLM fitting and evaluation. These variables include Total.Food.Expenditure, Household.Head.Sex, and Type.of.Household.

Data Exploration

In order to identify a few suitable research variables for easier study, we will use functions like cor, ANOVA, and others to select two appropriate numerical variables and one categorical variable.

We have decided to select the two most correlated numerical variables and the categorical variable with the smallest p-value as our research subjects.

The three selected variables are: **Total.Food.Expenditure, Household.Head.Age, and Type.of.Household.**

Variable Distribution Visualization

Next, we will perform visualization to better understand how to process the data in the following sections and which models to use for fitting. Through visualization, we can gain insights into the relationships between variables, detect potential outliers, and decide on the most suitable modeling approach.

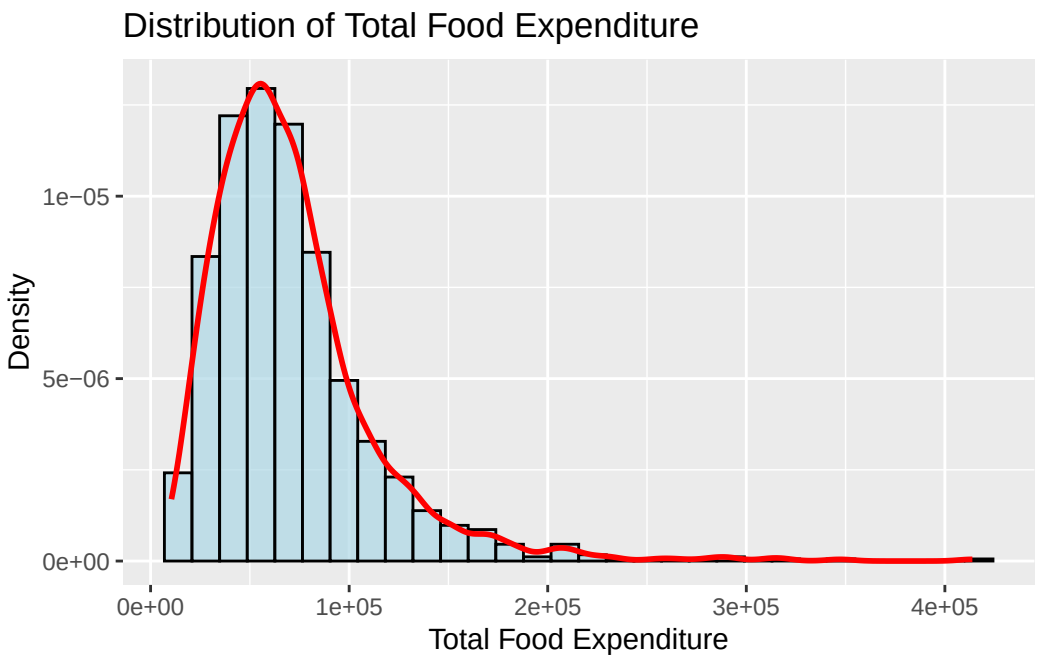


Figure 1: Histogram of Total.Food.Expenditure

The variable distribution of Total.Food.Expenditure is highly right-skewed (long right tail), meaning that most households have low food expenditures, but there are some extremely high values.

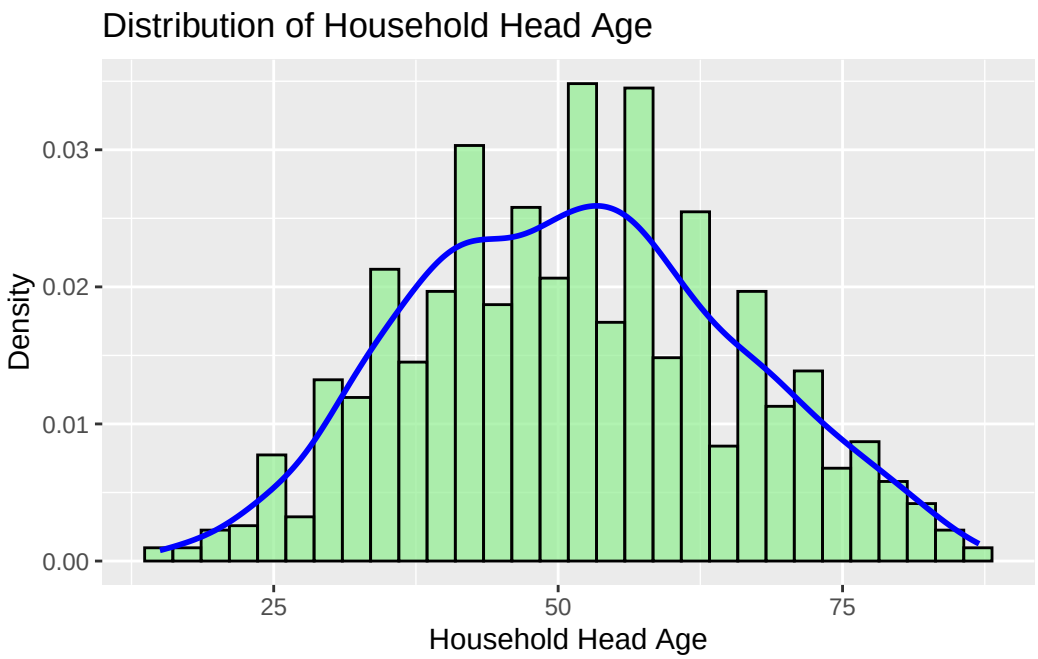


Figure 2: Histogram of Household.Head.Age

The distribution of Household.Head.Age is approximately normal.The data is well-distributed and can be used directly.

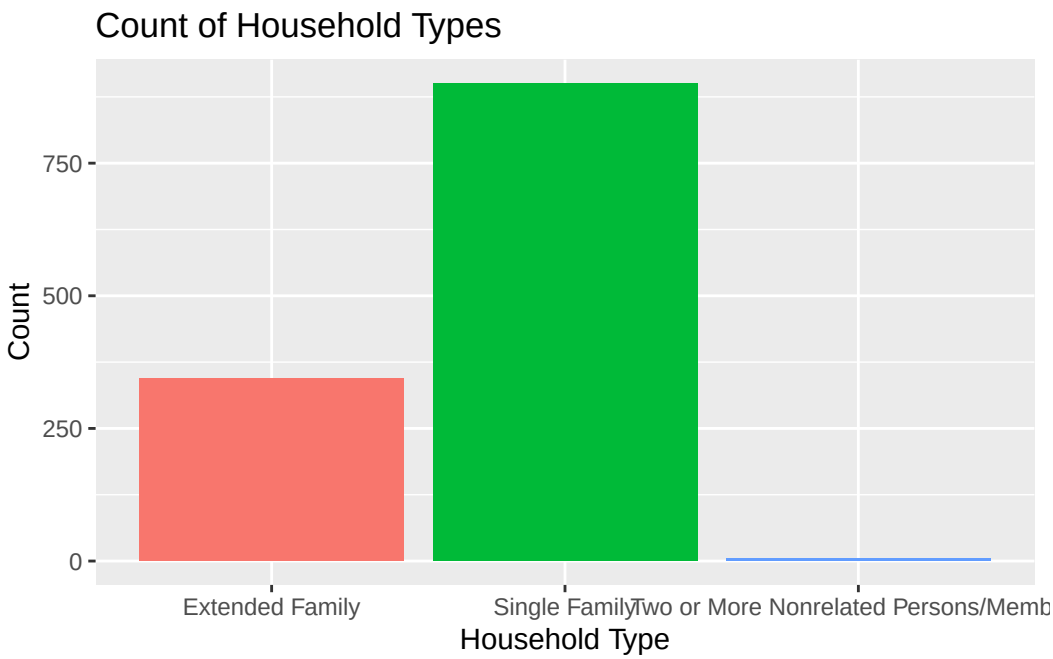


Figure 3: Bar chart of Type.of.Household

We can easily observe that the "Two or More Nonrelated Persons/Members" category has very few samples, which may affect model stability in GLM fitting.

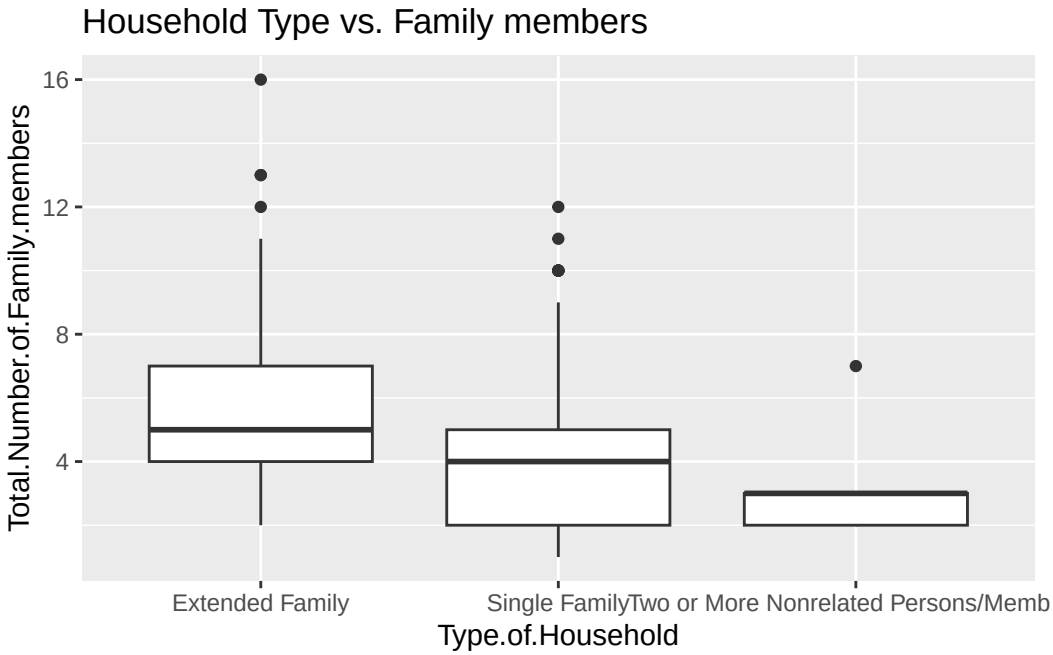


Figure 4: Box plot of Type.of.Household and Total.Number.of.Family.members

From the above figure, we can see that Extended Family usually has the highest number of family members, Single Family has fewer, and households with nonrelated persons have the least. However, there are outliers present. In the subsequent data processing, we will consider using the IQR method to remove these outliers.

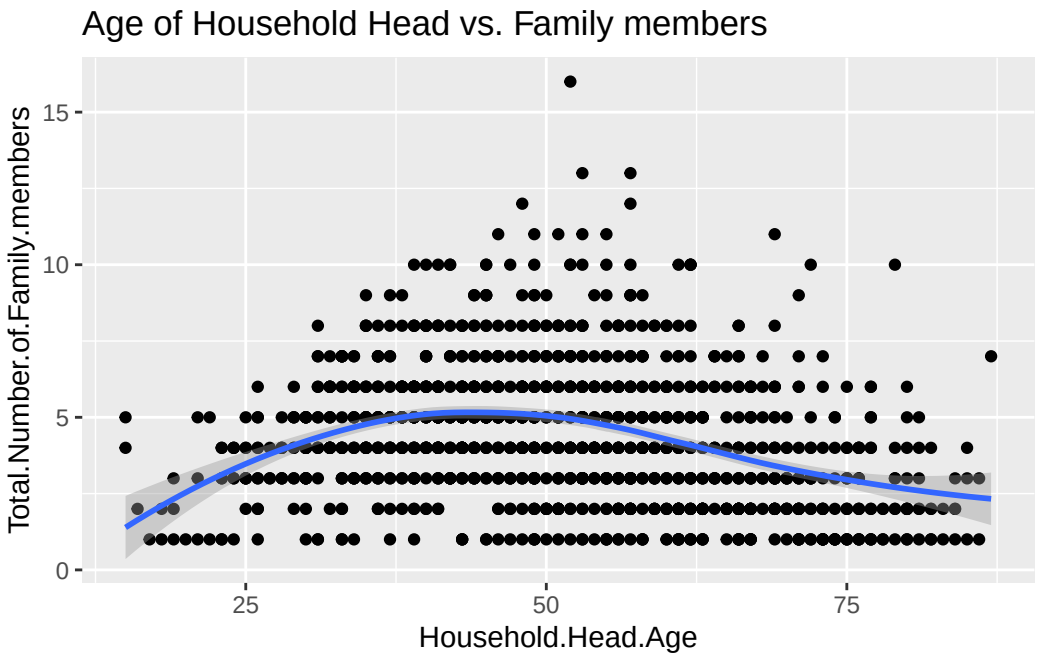


Figure 5: Histogram of Household.Head.Age vs. Total.Number.of.Family.members

The LOESS curve exhibits a nonlinear trend (first rising and then falling), whereas Poisson regression assumes a linear relationship. Therefore, we will consider polynomial regression in the subsequent model construction. Additionally, there are some outliers that we may consider removing.

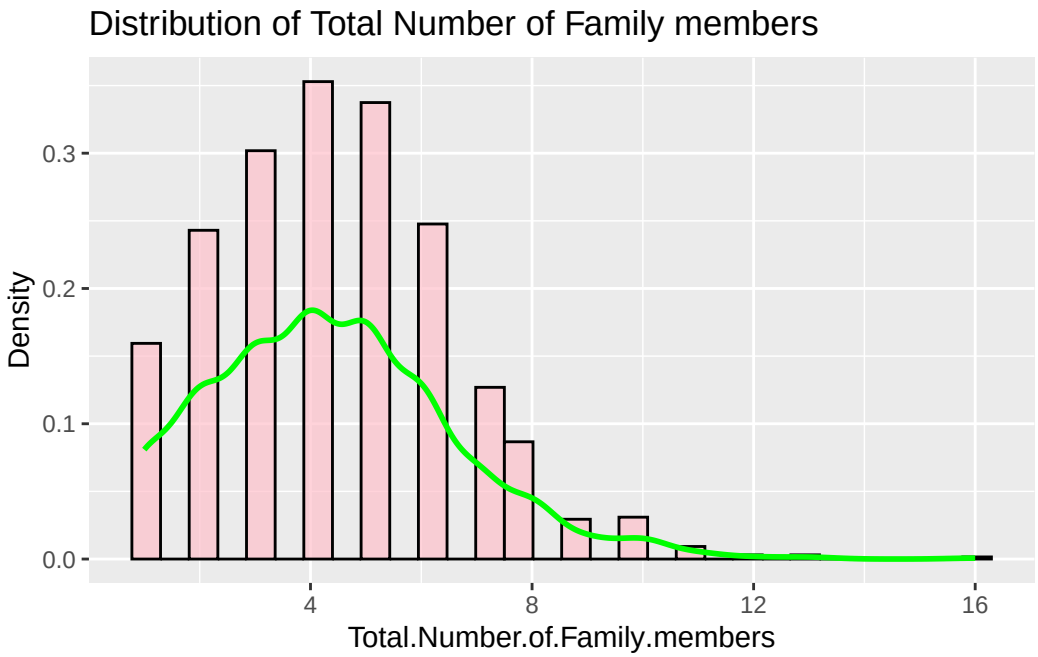


Figure 6: Histogram of Total.Number.of.Family.members

We found that the data is right-skewed and contains a few high-member outliers. We will consider whether to remove these outliers.

Data Preprocessing

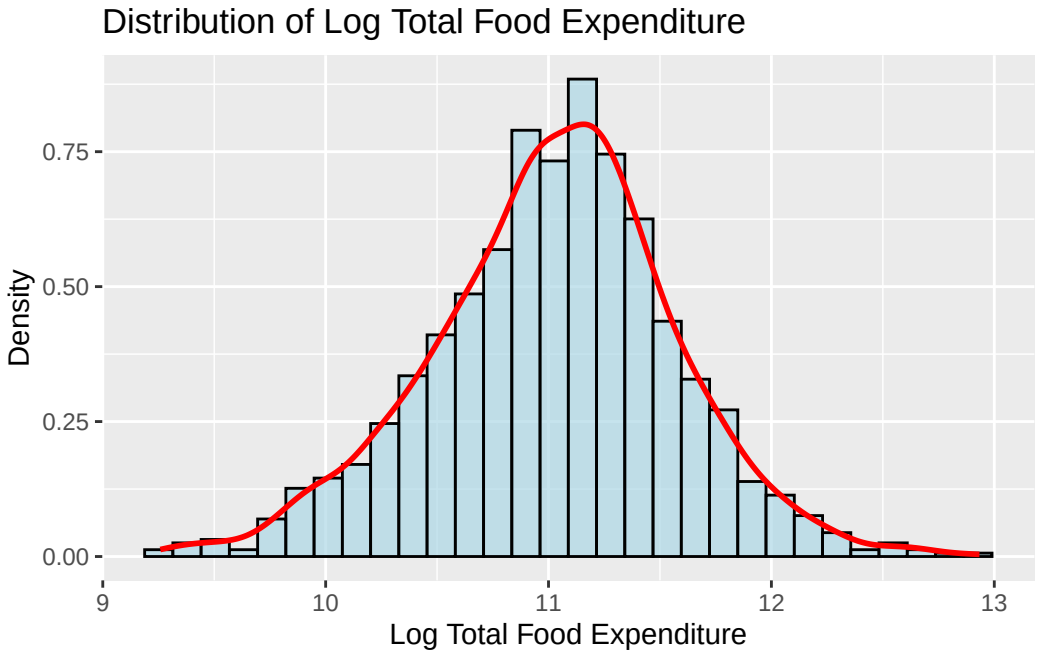


Figure 7: Histogram of log_Total.Food.Expenditure

It is evident that the data follows a normal distribution more closely after applying the logarithm transformation, achieving our goal. Next, we will consider using this transformed data for model fitting.

Modeling with GLM Poisson, Negative binomial and Gamma regression

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

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                        1.922e+00  6.701e-02
```

```
Total.Food.Expenditure                                    3.310e-06  2.775e-07
Household.Head.Age                                       -8.874e-03  1.009e-03
Type.of.HouseholdSingle Family                           -3.475e-01  3.002e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -6.119e-01  2.441e-01
                                                         z value Pr(>|z|)
(Intercept)                                               28.676   <2e-16 ***
Total.Food.Expenditure                                    11.927   <2e-16 ***
Household.Head.Age                                        -8.799   <2e-16 ***
Type.of.HouseholdSingle Family                           -11.578   <2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members   -2.507   0.0122 *
---
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:  988.69  on 1244  degrees of freedom
AIC: 5027.5

Number of Fisher Scoring iterations: 4
```

Total.Food.Expenditure has a p-value $<$ 2e-16, indicating that it is statistically significant. Household.Head.Age has a p-value $<$ 2e-16, indicating that it is statistically significant. Type.of.Household has a p-value $=$ 0.0122, indicating that it is statistically significant, however, the effect size is relatively small. Next, we will consider Negative Binomial Regression.

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Food.Expenditure +
    Household.Head.Age + Type.of.Household, data = Data, init.theta = 70960.5884,
    link = log)

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              1.922e+00  6.701e-02
Total.Food.Expenditure                                   3.310e-06  2.775e-07
Household.Head.Age                                      -8.874e-03  1.009e-03
Type.of.HouseholdSingle Family                          -3.475e-01  3.002e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -6.119e-01  2.441e-01
                                                        z value Pr(>|z|)
(Intercept)                                              28.674   <2e-16 ***
```

```
Total.Food.Expenditure                                    11.927   <2e-16 ***
Household.Head.Age                                        -8.798   <2e-16 ***
Type.of.HouseholdSingle Family                           -11.578   <2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members   -2.507   0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(70960.59) family taken to be 1)

    Null deviance: 1373.55  on 1248  degrees of freedom
Residual deviance:  988.62  on 1244  degrees of freedom
AIC: 5029.6

Number of Fisher Scoring iterations: 1

              Theta:  70961
          Std. Err.:  332492
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -5017.557


   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   4.395   6.000  16.000



Call:
glm(formula = Total.Number.of.Family.members ~ Total.Food.Expenditure +
    Type.of.Household + Household.Head.Age, family = Gamma(link = "log"),
    data = Data)

Coefficients:
                                                           Estimate Std. Error
(Intercept)                                                1.957e+00  6.483e-02
Total.Food.Expenditure                                     4.689e-06  3.083e-07
Type.of.HouseholdSingle Family                            -3.706e-01  2.919e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -6.449e-01  1.972e-01
Household.Head.Age                                        -1.129e-02  9.005e-04
                                                          t value Pr(>|t|)
(Intercept)                                                30.183   <2e-16 ***
Total.Food.Expenditure                                     15.211   <2e-16 ***
Type.of.HouseholdSingle Family                            -12.697   <2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members   -3.271   0.0011 **
```

```
Household.Head.Age                                      -12.542   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.191361)

    Null deviance: 362.57  on 1248  degrees of freedom
Residual deviance: 258.20  on 1244  degrees of freedom
AIC: 4966.1

Number of Fisher Scoring iterations: 6
```

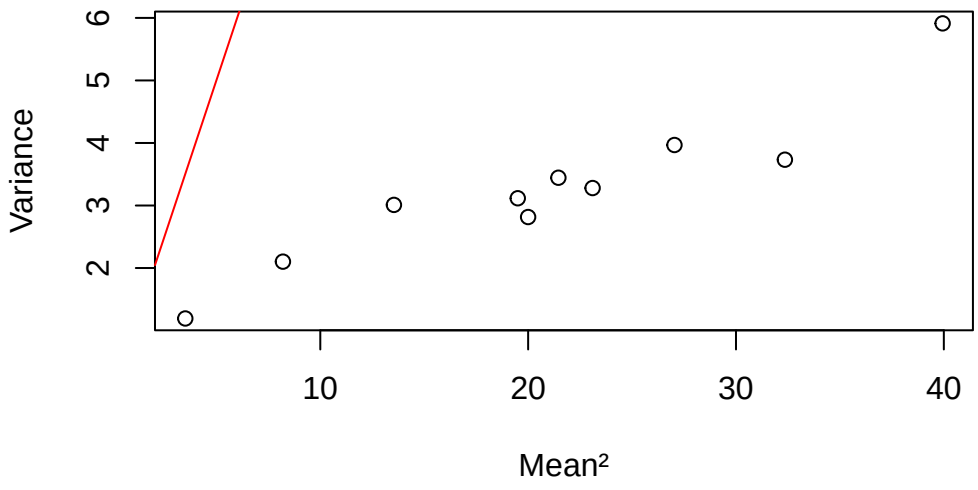
After obtaining three different models, we consider adding the previously log-transformed data.

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

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                            -2.295000   0.313509
log_Total.Food.Expenditure                              0.392760   0.026379
Household.Head.Age                                     -0.007308   0.001033
Type.of.HouseholdSingle Family                         -0.305287   0.030273
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.512654   0.243627
                                                       z value Pr(>|z|)
(Intercept)                                             -7.320 2.47e-13 ***
log_Total.Food.Expenditure                              14.889  < 2e-16 ***
Household.Head.Age                                      -7.073 1.52e-12 ***
Type.of.HouseholdSingle Family                         -10.084  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -2.104   0.0354 *
---
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:  894.31  on 1244  degrees of freedom
AIC: 4933.2
```

```
Number of Fisher Scoring iterations: 4
```

It can be observed that the p-values improve to varying degrees. Therefore, we mainly focus on the first three models.

#Check the Gamma model



GLM Regression Analysis Results

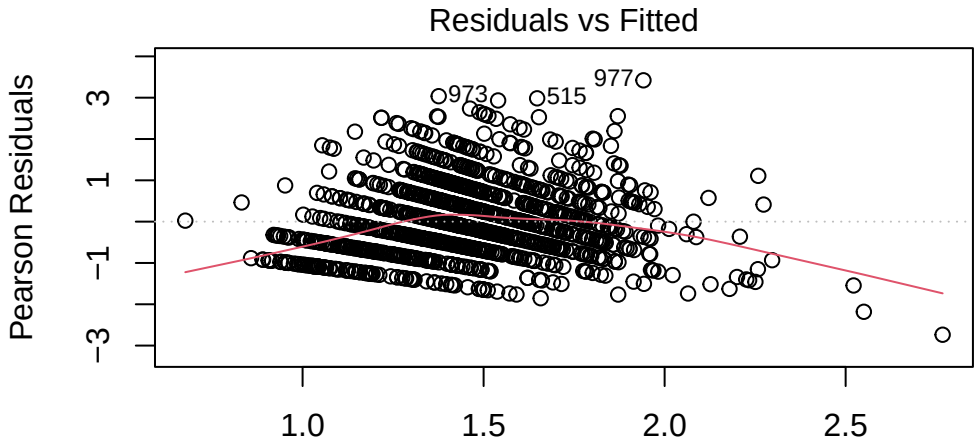
```
              df      AIC
poisson_model  5 4933.170
nb_model       6 5029.557
gamma_model    6 4966.077

Likelihood ratio test

Model 1: Total.Number.of.Family.members ~ log_Total.Food.Expenditure +
    Household.Head.Age + Type.of.Household
Model 2: Total.Number.of.Family.members ~ Total.Food.Expenditure + Household.Head.Age +
    Type.of.Household
  #Df  LogLik Df  Chisq Pr(>Chisq)
```

```
1   5 -2461.6
2   6 -2508.8  1 94.388  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                                                (Intercept)
                                                  6.8316969
                                       Total.Food.Expenditure
                                                  1.0000033
                                           Household.Head.Age
                                                  0.9911651
                                 Type.of.HouseholdSingle Family
                                                  0.7064262
Type.of.HouseholdTwo or More Nonrelated Persons/Members
                                                  0.5423123
```

```
Overdispersion = 0.718901
```

```
                                GVIF Df GVIF^(1/(2*Df))
log_Total.Food.Expenditure 1.084252  1        1.041274
Household.Head.Age         1.095155  1        1.046497
Type.of.Household          1.147790  2        1.035060
```



By comparing the AIC values of the three models,it can be seen that model Gamma has the lowest AIC value.Although the minimum AIC is the reference index, the matching between the data and the hypothesis of the model is the fundamental basis. Therefore, we tested the relationship between the mean value and the variance and found that the trend was not linearly positive and did not meet the core hypothesis of the gamma distribution, so we could not use gamma to fit.

So we back to Poisson and negative binomial models. For the AIC value of the poisson model is lower and after compared with the two models p valve(0.9075) is not lower than 0.05 so we should not change the poisson model into negative binomial model.

For the poisson model its Overdispersion = 0.7947643 close to 1 that shows the data not Significant overdispersion.The IRR part shows IRR (Total.Food.Expenditure)=1.0000033 effect can be ignored,IRR(Household.Head.Age)=0.991 weakly negative effect,IRR(Type.of.Household_Single Family)= 0.706,IRR(Two or More Nonrelated Persons)= 0.542 the type of family is the key factor affecting the number of family members, and the reduction effect of multiple non-relatives is the most significant.

The GVIF values of all variables are close to 1, and the adjusted values are all < 1.04, indicating that there is no multicollinearity problem and the model coefficients are reliable.