DAS Project 2

Group 03

```
library(MASS)
library(tidyverse)
library(dplyr)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(gt)
library(GGally)
library(gridExtra)
library(caret)
```

1 Introduction

With the data from the FIES (Family Income and Expenditure Survey) recorded in the Philippines, the analysis aims at finding which household related variables influence the number of people living in a household?

```
house_data <- read.csv("dataset03.csv")
```

Table 1: house data overview Part I

Total.Household.Income	Region	Total.Food.Expenditure
89359	X - Northern Mindanao	54537
108400	X - Northern Mindanao	56611
51982	X - Northern Mindanao	30827
76623	X - Northern Mindanao	43639
135232	X - Northern Mindanao	59614

Table 2: house data overview Part II

Household.Head.Sex	Household.Head.Age	Type.of.Household	Total.Number.of.Family.members
Female	34	Single Family	10
Male	55	Single Family	8
Male	26	Single Family	5
Male	53	Single Family	6
Male	55	Single Family	6
Male	38	Single Family	5

Table 3: house data overview Part III

House.Floor.Area	House.Age	Number.of.bedrooms	Electricity
64	11	1	0
60	13	3	1
48	13	1	0
42	5	2	0
56	5	2	1
56	8	1	1

Among the data we have such variables:

1.1 Outcome variable:

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

1.2 Explanatory variables:

`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

`House.Floor.Area`: Floor area of the house (in !)

`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)

2 Data Processing

We first have a numerical summary and data visualization to get an overview of the distribution of each variable, and detect if there may exist relationships between variables.

2.1 Data Numerical Summary

Table 4: Numerical Summary for Total Number of Family Members

mean	median	std_dev
4.68	4.00	2.30

`Total.Number.of.Family.members` holds a mean value as 4.68 and a median value as 4.00, which means there are more than 4 peoples in each house on average, and the median value of the number of family members is 4. The median value is lower than mean value, indicating the distribution may be right-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 2.30, indicating the distribution is not dispersed.

Table 5: Numerical Summary for Total Household Income

mean	median	std_dev
214,057.78	131,806.00	232,931.78

`Total.Household.Income` holds a mean value as 214057.78 and a median value as 131806.00, which means the total household income in each family is 214057.78 on average and the median value of the household income in each family is 131806.00. The median value is lower than mean value, indicating the distribution may be right-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 232931.78, indicating the distribution may be dispersed.

Table 6: Numerical Summary for Total Food expenditure

mean	median	std_dev
64,112.59	54,594.00	39,497.08

`Total.Food.Expenditure` holds a mean value as 64112.59 and a median value as 54594.00, which means each family will spend 64112.59 on food on average and the median value of the food expenditure of each family is 54594.00. The median value is lower than mean value, indicating the distribution may be right-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 39497.08, indicating the distribution may be dispersed.

Table 7: Numerical Summary for Household Head Age

mean	median	std_dev
51.52	51.00	13.81

`Household.Head.Age` holds a mean value as 51.52 and a median value as 51.00, which means the householder's age is bigger than 51 on average and the median value of the householder's age is 51. The median value is slightly lower than mean value, indicating the distribution may be slightly right-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 13.81, indicating the distribution may be dispersed.

Table 8: Numerical Summary for House Floor Area

mean	median	std_dev
59.81	50.00	47.00

`House.Floor.Area` holds a mean value as 59.81 and a median value as 50.00, which means the house floor area is 59.81 on average and the median value of floor area in each family is 50.00. The median value is lower than mean value, indicating the distribution may be right-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 47.00, indicating the distribution may be dispersed.

Table 9: Numerical Summary for House Age

mean	median	std_dev
19.50	16.00	13.15

`House.Age` holds a mean value as 19.50 and a median value as 16.00, which means the mean value of each house is 19.50 on average while the median value of the houses is 16. The median value is lower than mean value, indicating the distribution may be right-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 13.15, indicating the distribution may be dispersed.

Table 10: Numerical Summary for Number of Bedrooms

mean	median	std_dev
1.94	2.00	1.08

`Number.of.bedrooms` holds a mean value as 1.94 and a median value as 2.00, which means number of bedrooms in each house is less than 2 on average and the median value of bedrooms in each house is 2.00. The median value is slightly higher than mean value, indicating the distribution may be slightly left-skewed and affecting the mean value to deviate it from the median value. The standard deviation value is 1.08, indicating the distribution may not be dispersed.

2.2 Missing Values

Brfore the formal modeling, we first check if there exists missing values in the data.

Table 11: missing value Part I

Total.Household.Income	Region	Total.Food.Expenditure
0	0	0

Table 12: missing value Part II

Household.Head.Sex	Household.Head.Age	Type.of.Household	Total.Number.of.Family.members
0	0	0	0

Table 13: missing value Part III

House.Floor.Area	House.Age	Number.of.bedrooms	Electricity
0	0	0	0

The result show that there is no missing values in the data, we can directly move to the next step to visualize and process the data, and construct the data.

3 Data Visualization

Before the formal modeling phase, we draw pairs graph and box-plots to get an overview of the relationships between the variables. Then we draw a histogram of the outcome variable to check the distribution of thr outcome variables.

We draw a pairs graph Figure 1 for the outcome variable and all non-categorical variables to check if there may exists relationship between the variables. The correlation values and sactter-plots in the pairs graph indicates that there may exists relationship between the `Total.Number.of.Family.members` and the `Total.Food.Expenditure`, and there may also exists weak relationship between the `Total.Number.of.Family.members` and `Total.Household.Income`, `Household.Head.Age`, `Number.of.bedrooms`.

Then we draw box-plots Figure 2 for all the category variables to see if the outcome variable performs differently through the different category. Through the box-plots we can tell that `Total.Number.of.Family.members` seems performs differently in different sex of `Household.Head.Sex` and in different type of `Type.of.Household`. There seems no difference between `Total.Number.of.Family.members` in whether the house have electricity.

Finally we draw a histogram Figure 3 for the outcome variable `Total.Number.of.Family.members` to get an overview of the distribution of the outcome variable. The distribution shows an obvious poisson distribution.

3.1 Data Wrangling

As there exists three different categories in the explanatory variable `Type.of.Household`: `Single Family`, `Extended Family`, `Two or More Nonrelated Persons/Members`, we split the variable in to two columns of dummy variables to test the influence of different category.

```
house <- house_data %>%
  mutate(Type.of.House.Single =
           ifelse(Type.of.Household == "Single Family",
                  1, 0)) %>%
  mutate(Type.of.House.Extend =
           ifelse(Type.of.Household == "Extended Family",
                  1, 0)) %>%
  dplyr::select(-Region, -Type.of.Household)

house.x <- house %>%
  dplyr::select(-Total.Number.of.Family.members)

house.y <- house$Total.Number.of.Family.members
```

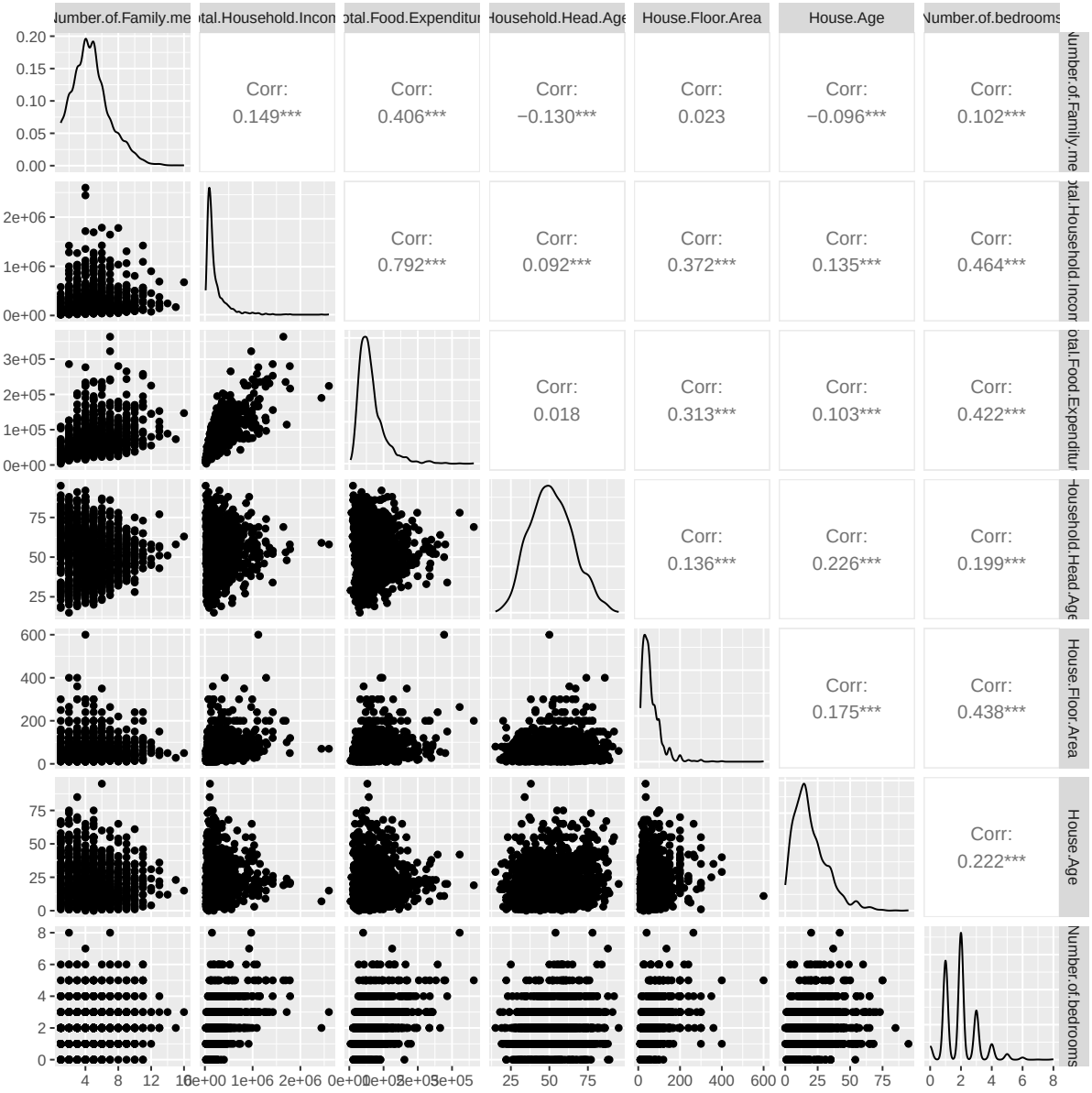

Figure 1: pairs graph of the house data

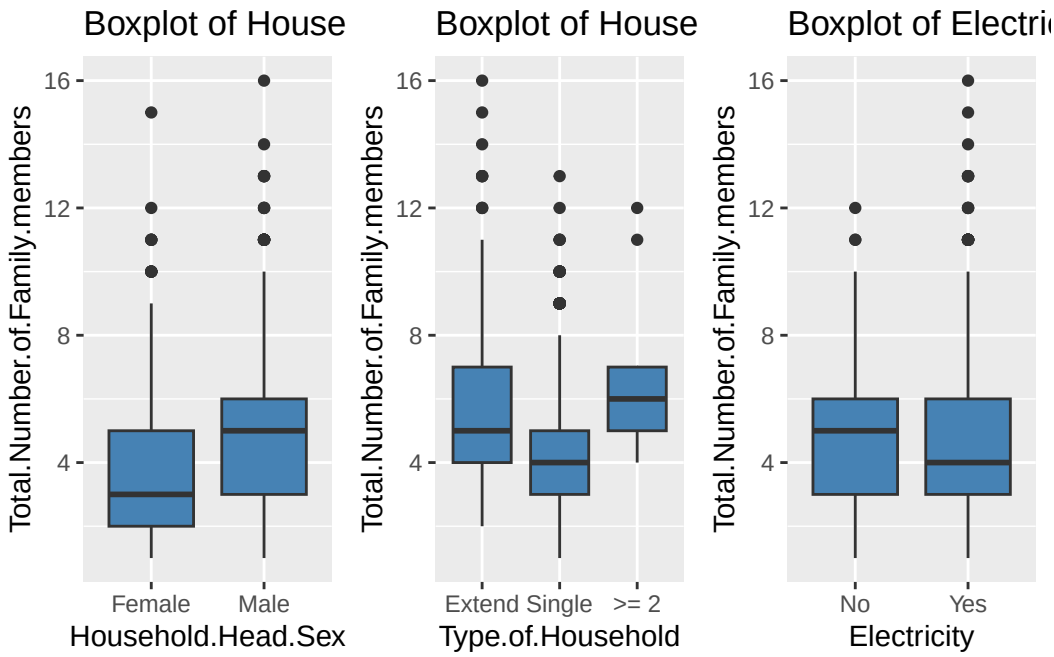

Figure 2: boxplot of Categories variables

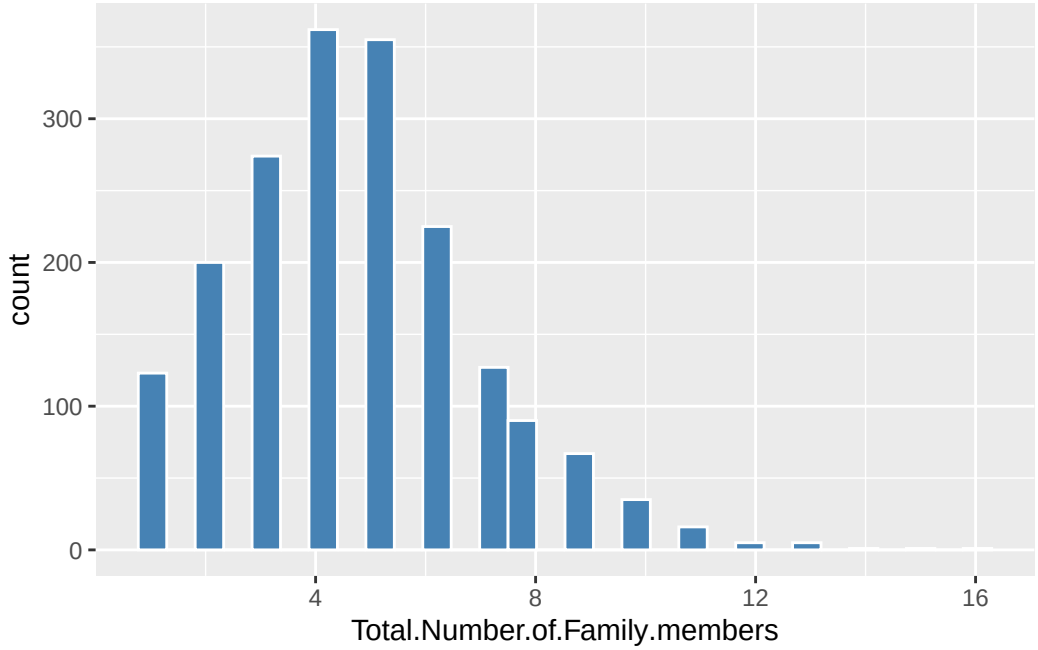

Figure 3: histogram of total number of family members

4 Generalized Linear Model

Preparing the dataset for analysis with a generalized linear model (GLM) by transforming categorical variables into factors. In R, factors are used to represent categorical data and are essential for fitting a GLM because they enable the model to treat these variables appropriately.

```
house$Household.Head.Sex <- as.factor(house$Household.Head.Sex)
house$Type.of.House.Extend <- as.factor(house$Type.of.House.Extend)
house$Type.of.House.Single <- as.factor(house$Type.of.House.Single)
```

4.1 Gaussian Model

4.1.1 Model Construction

```
model_gaussian_1 <- glm(Total.Number.of.Family.members ~
                          Total.Household.Income +
                          Total.Food.Expenditure +
                          Household.Head.Sex +
                          Household.Head.Age +
                          House.Floor.Area +
                          House.Age +
                          Number.of.bedrooms +
                          Electricity +
                          Type.of.House.Single +
                          Type.of.House.Extend,
                        data = house,
                        family = gaussian)

model_gaussian_1 %>%
  summary()
```

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

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.078e+00  6.603e-01   9.205  < 2e-16 ***
Total.Household.Income -3.492e-06  3.140e-07 -11.123  < 2e-16 ***
Total.Food.Expenditure  3.686e-05  1.838e-06  20.047  < 2e-16 ***
Household.Head.SexMale  7.205e-01  1.079e-01   6.678 3.18e-11 ***
Household.Head.Age     -2.121e-02  3.366e-03  -6.300 3.70e-10 ***
House.Floor.Area       -1.079e-03  1.034e-03  -1.043 0.297170
House.Age              -1.375e-02  3.418e-03  -4.021 6.02e-05 ***
Number.of.bedrooms      4.530e-02  4.874e-02   0.929 0.352790
Electricity1           -4.679e-01  1.307e-01  -3.580 0.000353 ***
Type.of.House.Single1  -2.303e+00  6.179e-01  -3.727 0.000200 ***
Type.of.House.Extend1  -8.069e-01  6.206e-01  -1.300 0.193707
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3.390283)

    Null deviance: 9962.5  on 1886  degrees of freedom
Residual deviance: 6360.2  on 1876  degrees of freedom
AIC: 7671.9

Number of Fisher Scoring iterations: 2
```

In the model output, each coefficient indicates the expected change in the number of family members for each unit change in the predictor variables:

- **Total.Household.Income**: More income correlates with fewer family members.
- **Total.Food.Expenditure**: More spending on food is associated with more family members.
- **Household.Head.SexMale**: Male-headed households tend to have more members.
- **Household.Head.Age** and **House.Age**: Older household heads and older houses are linked to fewer family members.
- **Number.of.bedrooms**: The number of bedrooms doesn't significantly affect family size.
- **Electricity**: Having electricity is related to fewer family members.
- **Type.of.House.Single**: Single-type households tend to have fewer members, while **Type.of.House.Extend** does not show a significant effect. This output is from a generalized linear model (GLM) with a Gaussian family, indicating that the response variable (`Total.Number.of.Family.members`) is assumed to be normally distributed. The

model estimates the relationship between the number of family members and several explanatory variables.

- **Significance**: The three asterisks next to the predictors' p-values denote a high level of statistical significance ($p < 0.001$).
- **Model Fit**: The model has a null deviance of 9962.5 and a residual deviance of 6360.2. The lower residual deviance compared to the null deviance indicates that the model explains a significant portion of the variability in the response variable.
- **AIC**: The AIC of the model is 7671. The AIC is a measure of the relative quality of statistical models for a given set of data. Lower AIC values generally indicate a model is more parsimonious.

4.1.2 Model selection

```
model_gaussian_1.step_model <-
  stepAIC(glm(Total.Number.of.Family.members ~ .,
              data = house, family = gaussian),
          direction = "both")
```

```
Start:  AIC=7671.91
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms + Electricity + Type.of.House.Single +
    Type.of.House.Extend

                         Df Deviance    AIC
- Number.of.bedrooms      1   6363.1 7670.8
- House.Floor.Area        1   6363.9 7671.0
- Type.of.House.Extend    1   6365.9 7671.6
<none>                        6360.2 7671.9
- Electricity             1   6403.6 7682.7
- Type.of.House.Single    1   6407.3 7683.8
- House.Age               1   6415.0 7686.1
- Household.Head.Age      1   6494.7 7709.4
- Household.Head.Sex      1   6511.4 7714.2
- Total.Household.Income  1   6779.6 7790.4
- Total.Food.Expenditure  1   7722.7 8036.2

Step:  AIC=7670.77
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
```

```
    Household.Head.Sex + Household.Head.Age + House.Floor.Area +
    House.Age + Electricity + Type.of.House.Single + Type.of.House.Extend

                         Df Deviance    AIC
- House.Floor.Area        1   6365.3 7669.4
- Type.of.House.Extend    1   6368.7 7670.4
<none>                        6363.1 7670.8
+ Number.of.bedrooms      1   6360.2 7671.9
- Electricity             1   6404.1 7680.9
- Type.of.House.Single    1   6410.1 7682.6
- House.Age               1   6415.9 7684.4
- Household.Head.Age      1   6494.9 7707.4
- Household.Head.Sex      1   6516.4 7713.7
- Total.Household.Income  1   6782.2 7789.1
- Total.Food.Expenditure  1   7743.2 8039.2

Step:  AIC=7669.43
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Age + Electricity +
    Type.of.House.Single + Type.of.House.Extend

                         Df Deviance    AIC
- Type.of.House.Extend    1   6370.6 7669.0
<none>                        6365.3 7669.4
+ House.Floor.Area        1   6363.1 7670.8
+ Number.of.bedrooms      1   6363.9 7671.0
- Electricity             1   6407.6 7679.9
- Type.of.House.Single    1   6411.5 7681.1
- House.Age               1   6421.1 7683.9
- Household.Head.Age      1   6501.3 7707.3
- Household.Head.Sex      1   6518.7 7712.4
- Total.Household.Income  1   6810.9 7795.1
- Total.Food.Expenditure  1   7743.2 8037.2

Step:  AIC=7668.99
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Age + Electricity +
    Type.of.House.Single

                         Df Deviance    AIC
<none>                        6370.6 7669.0
+ Type.of.House.Extend    1   6365.3 7669.4
+ House.Floor.Area        1   6368.7 7670.4
```

```
+ Number.of.bedrooms      1   6369.1 7670.6
- Electricity             1   6412.2 7679.3
- House.Age               1   6427.9 7683.9
- Household.Head.Age      1   6508.5 7707.4
- Household.Head.Sex      1   6522.7 7711.5
- Total.Household.Income  1   6815.9 7794.5
- Type.of.House.Single    1   7180.5 7892.8
- Total.Food.Expenditure  1   7748.5 8036.5
```

```
# Print the summary of the final model
summary(model_gaussian_1.step_model)
```

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

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.298e+00  2.684e-01  19.740  < 2e-16 ***
Total.Household.Income -3.490e-06  3.045e-07 -11.460  < 2e-16 ***
Total.Food.Expenditure  3.692e-05  1.831e-06  20.160  < 2e-16 ***
Household.Head.SexMale  7.219e-01  1.078e-01   6.699 2.76e-11 ***
Household.Head.Age     -2.122e-02  3.328e-03  -6.377 2.26e-10 ***
House.Age              -1.389e-02  3.378e-03  -4.111 4.10e-05 ***
Electricity1           -4.518e-01  1.289e-01  -3.504  0.00047 ***
Type.of.House.Single1  -1.519e+00  9.826e-02 -15.456  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3.390405)

    Null deviance: 9962.5  on 1886  degrees of freedom
Residual deviance: 6370.6  on 1879  degrees of freedom
AIC: 7669

Number of Fisher Scoring iterations: 2
```

Model Selection Process: In the series of steps provided, we can see: The model starts with a relatively high AIC of 7671.91. Variables like Number.of.bedrooms, House.Floor.Area,

and Type.of.House.Extend are evaluated for removal because removing them might lead to a lower AIC. In successive steps, different variables are considered for addition or removal. For example, House.Floor.Area is added back in and then considered for removal again. Throughout the process, variables that do not contribute significantly to the model based on AIC are removed. For instance, Type.of.House.Extend is consistently considered for removal.The aim is to simplify the model without sacrificing the explanatory power. This is achieved by ending up with variables that contribute meaningfully to the model..

final model Analysis: The summary shows the results from a GLM where the response variable `Total.Number.of.Family.members` is being modeled as a function of various predictors. The family is set to `gaussian`, which suggests that the response variable is being treated as continuous with a normal distribution.All predictors are statistically significant, given the p-values are all less than 0.05.`Total.Household.Income` ,`House.Age` and `Electricity1`have negative coefficients, indicating that as these variables increase, the expected number of family members decreases.`Total.Food.Expenditure` and `Household.Head.SexMale`, have positive coefficients, suggesting that increases in these predictors are associated with an increase in the number of family members.`Type.of.House.Single1` has a large negative coefficient, which means that being a single-type household is associated with a significant decrease in the number of family members compared to the baseline household type (which could be 'extended' or another type not shown here).

4.2 Poisson Model

4.2.1 Model Construction

```
model_poisson_1 <- glm(Total.Number.of.Family.members ~
                         Total.Household.Income +
                         Total.Food.Expenditure +
                         Household.Head.Sex +
                         Household.Head.Age +
                         House.Floor.Area +
                         House.Age +
                         Number.of.bedrooms +
                         Electricity +
                         Type.of.House.Single +
                         Type.of.House.Extend,
                       data = house,
                       family = poisson)

model_poisson_1 %>%
  summary()
```

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

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.837e+00  1.403e-01  13.098  < 2e-16 ***
Total.Household.Income -6.932e-07  8.117e-08  -8.539  < 2e-16 ***
Total.Food.Expenditure  6.760e-06  4.147e-07  16.300  < 2e-16 ***
Household.Head.SexMale  1.773e-01  2.900e-02   6.114 9.71e-10 ***
Household.Head.Age     -5.455e-03  8.734e-04  -6.245 4.23e-10 ***
House.Floor.Area       -2.240e-04  2.569e-04  -0.872 0.383225
House.Age              -3.298e-03  8.958e-04  -3.682 0.000232 ***
Number.of.bedrooms      1.123e-02  1.225e-02   0.917 0.359385
Electricity1           -9.849e-02  3.275e-02  -3.007 0.002638 **
Type.of.House.Single1  -4.350e-01  1.282e-01  -3.392 0.000693 ***
Type.of.House.Extend1  -1.163e-01  1.288e-01  -0.903 0.366485
---
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: 1398.4  on 1876  degrees of freedom
AIC: 7630.1

Number of Fisher Scoring iterations: 4
```

This is a Poisson generalized linear model predicting the number of family members. Here's a concise analysis:

- The negative coefficient for **Total.Household.Income** suggests fewer family members in higher-income households.
- **Total.Food.Expenditure** has a positive coefficient, indicating that families with higher food spending tend to be larger.
- Male-headed households (**Household.Head.SexMale**) are associated with a higher number of family members.
- An increase in **Household.Head.Age** and **House.Age** correlates with fewer family members.

- **House.Floor.Area** is not statistically significant in this model, implying it may not be a good predictor for the number of family members.
- Having electricity (**Electricity1**) is negatively associated with family size.
- **Type.of.House.Single** households have significantly fewer family members, while **Type.of.House.Extend** is not significantly associated with family size.

The model has a substantial AIC and the deviance information suggests the model fits the data well compared to the null model. The significance codes indicate that most variables are highly significant predictors, except for **House.Floor.Area** and **Number.of.bedrooms**, which do not show a clear association with the number of family members in this model. ### Model selection

```
model_poisson_1.step_model <-
  stepAIC(glm(Total.Number.of.Family.members ~ .,
              data = house, family = poisson),
          direction = "both")
```

```
Start:  AIC=7630.12
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms + Electricity + Type.of.House.Single +
    Type.of.House.Extend

                         Df Deviance    AIC
- House.Floor.Area        1   1399.1 7628.9
- Type.of.House.Extend    1   1399.1 7628.9
- Number.of.bedrooms      1   1399.2 7629.0
<none>                        1398.3 7630.1
- Electricity             1   1407.2 7637.0
- Type.of.House.Single    1   1408.4 7638.2
- House.Age               1   1412.1 7641.9
- Household.Head.Sex      1   1437.0 7666.8
- Household.Head.Age      1   1437.5 7667.2
- Total.Household.Income  1   1478.1 7707.8
- Total.Food.Expenditure  1   1655.0 7884.8

Step:  AIC=7628.89
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Age + Number.of.bedrooms +
    Electricity + Type.of.House.Single + Type.of.House.Extend

                         Df Deviance    AIC
- Number.of.bedrooms      1   1399.6 7627.3
```

```
- Type.of.House.Extend    1   1399.8 7627.6
<none>                        1399.1 7628.9
+ House.Floor.Area        1   1398.3 7630.1
- Electricity             1   1408.0 7635.8
- Type.of.House.Single    1   1409.0 7636.8
- House.Age               1   1413.3 7641.0
- Household.Head.Sex      1   1438.0 7665.8
- Household.Head.Age      1   1438.8 7666.6
- Total.Household.Income  1   1481.7 7709.5
- Total.Food.Expenditure  1   1655.2 7883.0

Step:  AIC=7627.34
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Age + Electricity +
    Type.of.House.Single + Type.of.House.Extend

                         Df Deviance    AIC
- Type.of.House.Extend    1   1400.3 7626.0
<none>                        1399.6 7627.3
+ Number.of.bedrooms      1   1399.1 7628.9
+ House.Floor.Area        1   1399.2 7629.0
- Electricity             1   1408.0 7633.8
- Type.of.House.Single    1   1409.4 7635.2
- House.Age               1   1413.3 7639.1
- Household.Head.Age      1   1438.9 7664.6
- Household.Head.Sex      1   1438.9 7664.7
- Total.Household.Income  1   1483.6 7709.3
- Total.Food.Expenditure  1   1659.8 7885.6

Step:  AIC=7626.04
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + House.Age + Electricity +
    Type.of.House.Single

                         Df Deviance    AIC
<none>                        1400.3 7626.0
+ Type.of.House.Extend    1   1399.6 7627.3
+ Number.of.bedrooms      1   1399.8 7627.6
+ House.Floor.Area        1   1399.9 7627.7
- Electricity             1   1408.6 7632.4
- House.Age               1   1414.3 7638.1
- Household.Head.Sex      1   1439.3 7663.0
- Household.Head.Age      1   1440.1 7663.9
```

```
- Total.Household.Income  1   1484.2 7708.0
- Type.of.House.Single    1   1583.8 7807.6
- Total.Food.Expenditure  1   1660.4 7884.2
```

```r
# Print the summary of the final model
summary(model_poisson_1.step_model)
```

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

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.727e+00  6.701e-02  25.770  < 2e-16 ***
Total.Household.Income -6.891e-07  7.854e-08  -8.773  < 2e-16 ***
Total.Food.Expenditure  6.773e-06  4.131e-07  16.396  < 2e-16 ***
Household.Head.SexMale  1.778e-01  2.896e-02   6.140 8.25e-10 ***
Household.Head.Age     -5.439e-03  8.625e-04  -6.306 2.87e-10 ***
House.Age              -3.293e-03  8.857e-04  -3.718 0.000201 ***
Electricity1           -9.439e-02  3.235e-02  -2.918 0.003527 **
Type.of.House.Single1  -3.234e-01  2.360e-02 -13.701  < 2e-16 ***
---
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: 1400.3  on 1879  degrees of freedom
AIC: 7626

Number of Fisher Scoring iterations: 4
```

Model Selection Process: Starting Model: The model begins with an AIC of 7630.12, with all the candidate variables included. Stepwise Adjustments: The algorithm evaluates the addition or subtraction of variables to find a model with a lower AIC. Variables considered for removal or addition include House.Floor.Area, Type.of.House.Extend, and Number.of.bedrooms. Throughout these steps, variables are assessed one at a time to determine their impact on the model's AIC. Model Refinement: The AIC fluctuates slightly as the stepwise algorithm adds and removes variables. This is a typical behavior in stepwise procedures as the algorithm

explores the variable space. Final Steps: The algorithm reaches a point where the addition or subtraction of variables does not meaningfully lower the AIC, indicating a convergence towards an optimal set of variables. In the final steps, we observe that the variable Type.of.House.Single remains in the model, suggesting its significant contribution to explaining the variability in the number of family members. Conclusion: The lowest AIC achieved in the steps shown is 7626.04. This suggests that the final model includes Total.Household.Income, Total.Food.Expenditure, Household.Head.Sex, Household.Head.Age, House.Age, Electricity, and Type.of.House.Single as predictors. The model does not include House.Floor.Area or Type.of.House.Extend, as their inclusion does not improve the AIC. **final model Analysis**: The Poisson regression model provided aims to predict the `Total.Number.of.Family.members` based on various household factors. Here's a concise analysis:

- **Intercept**: The expected log count of family members is about 1.727 when all other variables are zero.
- **Total.Household.Income**: The negative coefficient suggests that with each unit increase in income, the log count of family members decreases, implying fewer family members in wealthier households.
- **Total.Food.Expenditure**: The positive coefficient indicates that higher food expenditure is associated with a greater number of family members.
- **Household.Head.SexMale**: The presence of a male household head is positively associated with the number of family members.
- **Household.Head.Age**: Older household heads are associated with fewer family members.
- **House.Age**: Similarly, greater house age is associated with fewer family members.
- **Electricity**: The negative coefficient suggests that households with electricity have fewer family members, though this relationship is less strong (as indicated by the double asterisk denoting a p-value between 0.001 and 0.01).
- **Type.of.House.Single**: Single-type households are associated with significantly fewer family members.

All variables included are statistically significant, as indicated by p-values less than 0.05, except for **House.Floor.Area** which was not included in this final model.

The model's AIC is 7626, which is a measure of the relative quality of the model; a lower AIC indicates a better fit to the data while penalizing for the number of explanatory variables used.

The null deviance and residual deviance show a considerable reduction when going from a model with only an intercept (null model) to the current model, indicating that the predictors provide substantial information in explaining the variability of the number of family members.

4.3 Final Model Selection

Considering the distribution of outcome variable `Total.Number.of.Family.members` and the sample size, we finally decide to use the poisson model.

4.3.1 Model Visualization

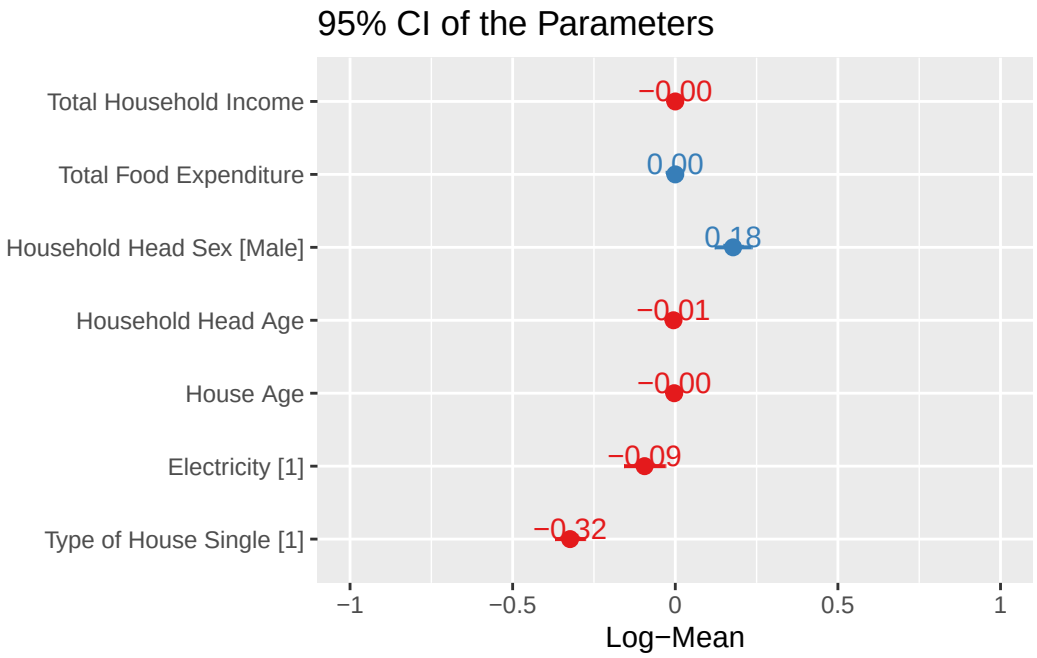

Figure 4: results-plots

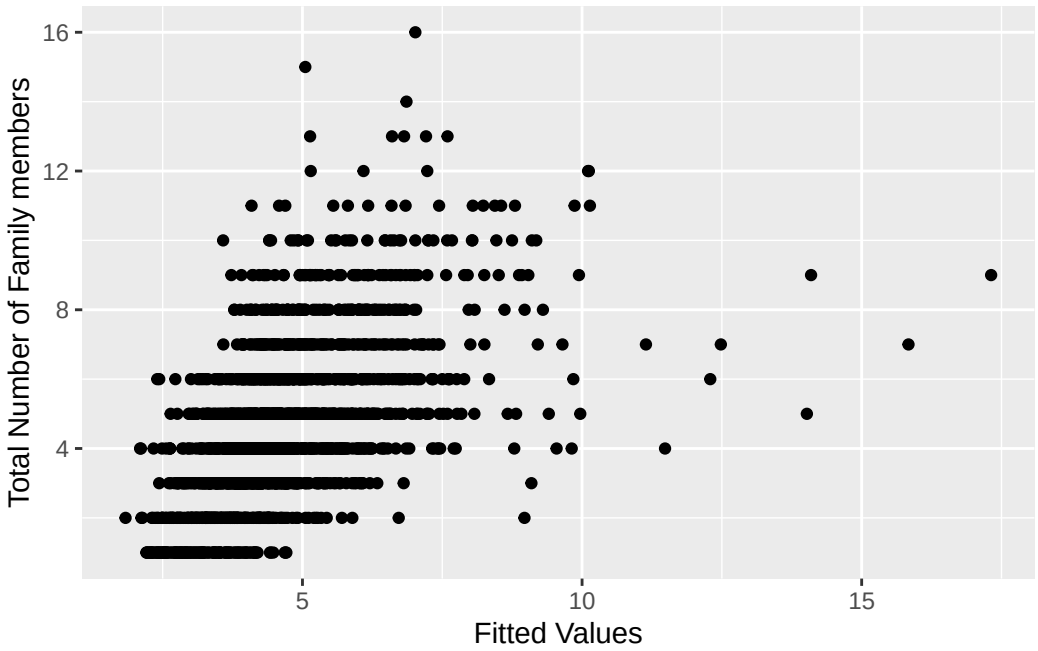

Figure 5: results-plots

The plot shows the 95% confidence intervals (CI) of the log-mean coefficients from a Poisson regression model. Each horizontal line represents the range within which we can be 95% confident the true value of the coefficient lies. The points are the estimated coefficients, and their position along the x-axis indicates the magnitude and direction of the relationship between each predictor and the log-count of family members.

- **Total Household Income** and **Electricity** [**1**] have negative coefficients, suggesting that as income increases and when electricity is available, the log-count of family members decreases.
- **Total Food Expenditure** and **Household Head Sex** [**Male**] have positive coefficients, indicating an increase in the log-count of family members with higher food expenditure and in male-headed households.
- **Household Head Age** and **House Age** also have negative coefficients, indicating fewer family members in households with older heads and older houses.
- **Type of House Single** [**1**] has the largest negative effect, suggesting that single-type houses are associated with a much lower log-count of family members.

The fact that the confidence intervals for some coefficients (such as Total Household Income and Household Head Age) do not cross the zero line indicates that these effects are statistically significant at the 95% confidence level. This plot is useful for quickly assessing the significance and strength of predictors in your model.

The plot is a scatter plot of observed versus fitted values from a Poisson regression model. In this type of plot:

- The x-axis represents the fitted values (predicted counts of family members) from your model.
- The y-axis represents the observed counts of family members from your dataset.

The plot is typically used to assess the model's goodness of fit. Ideally, if the model fits well, we would expect to see the points forming a diagonal line from the bottom left to the top right.

From the plot, it looks like for lower predicted values, the model fits quite well as the points are closely packed and increase linearly. However, as the fitted values increase, the variance also increases, and the points start to scatter more widely. This pattern could suggest potential overdispersion in the data, which is when the observed variance is higher than the variance predicted by the model.

4.4 Assumption Check

4.4.1 Residual plots code

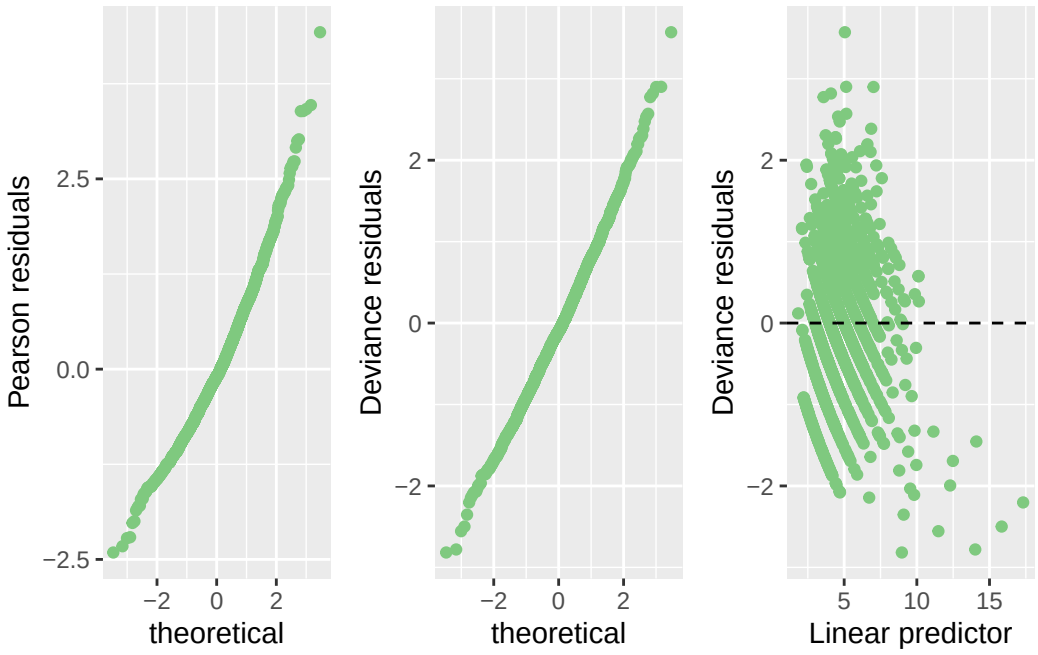

Figure 6: residual-plots

These are plots for a Poisson regression model, specifically checking the residuals of the model:

1. **QQ Plots of Pearson and Deviance Residuals**: The first two plots are quantile-quantile (QQ) plots for Pearson and deviance residuals, respectively. They compare the residuals to a theoretical normal distribution. In a well-fitted model, points should approximately follow the 45-degree line.

    - In both QQ plots, the residuals show a clear deviation from the theoretical normal distribution line, especially at the tails. This suggests that the residuals are not normally distributed in total and indicates that the model may not be adequately capturing all aspects of the data's structure—common in count data models like Poisson when overdispersion is present.

2. **Residuals vs. Linear Predictor**: The third plot shows the deviance residuals against the fitted values (linear predictor).

    - Ideally, you would want to see a random scatter of points without any discernible pattern. However, the plot shows a fanning out effect (increasing spread of residuals with the fitted values), which is a sign of overdispersion or nonlinearity in the data.

5 Conclusion

In this analysis, General Linear model is applied to detect the relationship between the total number of family members in each house and other variables and to find which related variables influence the number of people living in a household.

Numerical summary and data visualization are used to overview the data distribution and relationship, then the GLM is applied in both Poisson and Gaussian model, considering about the distribution of outcome variables, the Poisson model is finally chosen. The explanatory variables `Total.Household.Income`, `Total.Food.Expenditure`, `Household.Head.Sex`, `Household.Head.Age`, `House.Age`, `Electrivity`, and `Type.of.House` in `Single` is retained in the final model. The plot of 95% CI of the Parameters is drawn. The assumption check is conducted and the assumptions are proved to be valid.

6 Further Work

6.1 New Model Attempt

Since when checking the residuals of the final model, it was found that there was a certain linear relationship between the residuals. So we tried a new model: negative binomial and a new way of checking the model: cross validation.

```r
# Fitting a negative binomial model
model_negbin <- glm.nb(Total.Number.of.Family.members ~
                         Total.Household.Income +
                         Total.Food.Expenditure +
                         Household.Head.Sex +
                         Household.Head.Age +
                         House.Age +
                         Electricity +
                         Type.of.House.Single, data = house)

# Summarizing the model
summary(model_negbin)
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    House.Age + Electricity + Type.of.House.Single, data = house,
    init.theta = 86087.92182, link = log)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.727e+00  6.701e-02  25.769  < 2e-16 ***
Total.Household.Income -6.891e-07  7.855e-08  -8.773  < 2e-16 ***
Total.Food.Expenditure  6.773e-06  4.131e-07  16.395  < 2e-16 ***
Household.Head.SexMale  1.778e-01  2.896e-02   6.140 8.26e-10 ***
Household.Head.Age     -5.439e-03  8.625e-04  -6.306 2.87e-10 ***
House.Age              -3.293e-03  8.857e-04  -3.718 0.000201 ***
Electricity1           -9.439e-02  3.235e-02  -2.918 0.003528 **
Type.of.House.Single1  -3.234e-01  2.361e-02 -13.700  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(86087.92) family taken to be 1)

    Null deviance: 2129.9  on 1886  degrees of freedom
Residual deviance: 1400.2  on 1879  degrees of freedom
AIC: 7628.1

Number of Fisher Scoring iterations: 1

              Theta:  86088
```

```
          Std. Err.:  311520
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -7610.059
```

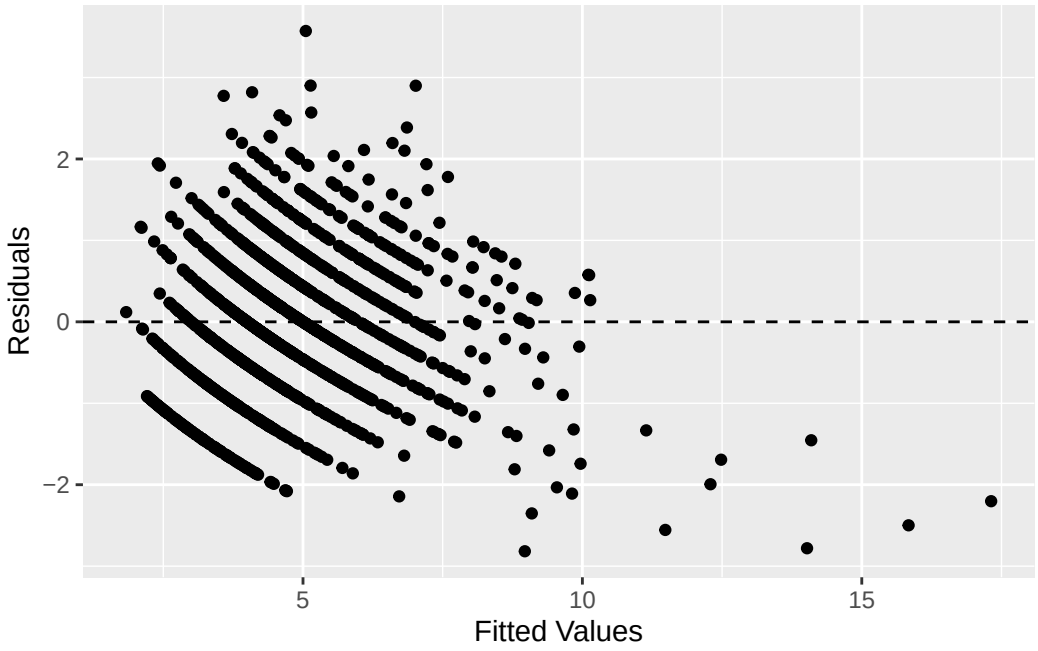

Figure 7: final-plot

```r
set.seed(123) # for reproducibility

# Define training control
train_control <- trainControl(method = "cv", number = 10)

# Fit the model using caret's train function
model_negbin_cv <- train(Total.Number.of.Family.members ~ .,
                         data = house,
                         method = "glm.nb",
                         trControl = train_control)

# Print the results
print(model_negbin_cv)
```

```
Negative Binomial Generalized Linear Model
```

```
1887 samples
  10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1697, 1699, 1698, 1699, 1698, 1698, ...
Resampling results across tuning parameters:

  link      RMSE      Rsquared   MAE
  identity  1.833275  0.3753705  1.400174
  log       1.919441  0.3148094  1.466200
  sqrt      1.883312  0.3387501  1.438976

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was link = identity.
```