Group_2_Analysis

Group 2

2022/3/20

```
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(MASS)
library(kableExtra)
library(olsrr)
#library(qcc)
```

```
#import data
data<-read.csv("dataset2.csv")

#processing discrete data
data[, 4] <- as.factor(data[, 4])
data[, 6] <- as.factor(data[, 6])
data[, 11] <- as.factor(data[, 11])

data = data[, -2]
```

Introduction

The Family Income and Expenditure Survey (FIES) is a survey of every households in a country which is taken every three years. This gives information on the levels of living and disparities in income of each family and spending patterns.

In this project, we use the pre-downloaded FIES data of a single region of Philippines. It is Mimaropa, former designated as Region IV-B and formally known as the southwestern Tagalog region. There are 1249 recorded households. Each of them contains 11 following variables:

· `Total.Household.Income` is the Annual household income (in Philippine peso)
· `Region` is the region of the Philippines which a household is in
· `Total.Food.Expenditure` is the annual expenditure by the household on food (in Philippine peso)
· `Household.Head.Sex` is the head of the households sex
· `Household.Head.Age` is the head of the households age (in years)
· `Type.of.Household` is the relationship between the group of people living in the house
· `Total.Number.of.Family.members` is the number of people living in the house
· `House.Floor.Area` is the floor area of the house (in square meter)
· `House.Age` is the age of the building (in years)
· `Number.of.bedrooms` is the number of bedrooms in the house
· `Electricity` is the electricity status of the house (1=Yes, 0=No)

where "head of the household" is the person who is in charge of that house.

The Generalised Linear Model (GLM) method will be used as an analysing tool. We are interested in the number of people living in a household (`Total.Number.of.Family.members`). The other variables having influences will be investigated.

Exploratory Data Analysis

Modelling and Results

Because the dependent variable of the data of this fitting model is the counting variable (the total number of families), and the independent variable is the continuity or category variable. In addition, the variable data are measured every three years, and the length of the whole observation concentration is unchanged. This study decided to use Poisson regression to fit the model. Poisson regression mainly has two assumptions. Firstly, the human time risk of different objects with the same characteristics and at the same time is homogeneous. Secondly, when the sample size is larger and larger, the mean of frequency tends to variance.

Preliminary fitting model

fitting model

```
model<-glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household.Head.S
summary(model)
```

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

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.SexMale                               2.418e-01  3.739e-02
Household.Head.Age                                  -5.818e-03  1.080e-03
Type.of.HouseholdSingle Family                      -3.732e-01  3.047e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -5.036e-01  2.447e-01
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
Electricity1                                        -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.SexMale                                6.467 1.00e-10 ***
Household.Head.Age                                   -5.386 7.21e-08 ***
Type.of.HouseholdSingle Family                      -12.250  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -2.058   0.0396 *
House.Floor.Area                                     -0.299   0.7653
```

```
House.Age                                                -2.082   0.0374 *
Number.of.bedrooms                                       -1.409   0.1589
Electricity1                                             -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
```

The stepwise method was used to complete the screening of independent variables

```
step(model)
```

```
Start:  AIC=4931.87
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + Type.of.Household +
    House.Floor.Area + House.Age + Number.of.bedrooms + Electricity

                        Df Deviance    AIC
- House.Floor.Area       1   881.10 4930.0
- Electricity            1   882.67 4931.5
- Number.of.bedrooms     1   883.00 4931.9
<none>                       881.01 4931.9
- House.Age              1   885.41 4934.3
- Household.Head.Age     1   910.02 4958.9
- Total.Household.Income 1   916.63 4965.5
- Household.Head.Sex     1   924.80 4973.7
- Type.of.Household      2  1028.11 5075.0
- Total.Food.Expenditure 1  1033.71 5082.6

Step:  AIC=4929.96
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
    Household.Head.Sex + Household.Head.Age + Type.of.Household +
    House.Age + Number.of.bedrooms + Electricity

                        Df Deviance    AIC
- Electricity            1   882.78 4929.6
<none>                       881.10 4930.0
- Number.of.bedrooms     1   883.59 4930.4
- House.Age              1   885.80 4932.7
- Household.Head.Age     1   910.07 4956.9
- Total.Household.Income 1   917.76 4964.6
- Household.Head.Sex     1   924.93 4971.8
- Type.of.Household      2  1028.11 5073.0
- Total.Food.Expenditure 1  1033.71 5080.6

Step:  AIC=4929.64
Total.Number.of.Family.members ~ Total.Household.Income + Total.Food.Expenditure +
```

```
    Household.Head.Sex + Household.Head.Age + Type.of.Household +
    House.Age + Number.of.bedrooms

                        Df Deviance    AIC
<none>                       882.78 4929.6
- Number.of.bedrooms      1   886.06 4930.9
- House.Age               1   888.38 4933.2
- Household.Head.Age      1   911.64 4956.5
- Total.Household.Income  1   919.96 4964.8
- Household.Head.Sex      1   927.56 4972.4
- Type.of.Household       2  1030.52 5073.4
- Total.Food.Expenditure  1  1033.99 5078.8

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

Coefficients:
                                          (Intercept)
                                            1.636e+00
                                Total.Household.Income
                                           -4.333e-07
                                Total.Food.Expenditure
                                            5.211e-06
                                Household.Head.SexMale
                                            2.441e-01
                                    Household.Head.Age
                                           -5.808e-03
                        Type.of.HouseholdSingle Family
                                           -3.739e-01
Type.of.HouseholdTwo or More Nonrelated Persons/Members
                                           -5.039e-01
                                             House.Age
                                           -2.707e-03
                                    Number.of.bedrooms
                                           -2.859e-02

Degrees of Freedom: 1248 Total (i.e. Null);  1240 Residual
Null Deviance:      1374
Residual Deviance: 882.8    AIC: 4930
```

Use a better model

```
model.better<-glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure +Household
family = "poisson")
```

Look for outliers in the model

```
library(car)
outlierTest(model.better)
```

```
     rstudent unadjusted p-value Bonferroni p
944 -5.065151         4.0808e-07   0.00050969
```

Remove the row of outliers

```
data<-data[-944,]
model.better<-glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure +Household
family = "poisson")
outlierTest(model.better)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
709 -2.89874          0.0037467           NA
```

Without outliers, the best model is obtained

```
summary(model.better)
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7839  -0.6516  -0.1001   0.4892   2.7201

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                         1.565e+00  7.977e-02
Total.Household.Income                             -5.150e-07  7.839e-08
Total.Food.Expenditure                              6.114e-06  4.521e-07
Household.Head.SexMale                              2.415e-01  3.733e-02
Household.Head.Age                                 -5.273e-03  1.089e-03
Type.of.HouseholdSingle Family                     -3.694e-01  3.038e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -5.151e-01  2.449e-01
House.Age                                          -2.896e-03  1.152e-03
Number.of.bedrooms                                 -2.997e-02  1.575e-02
                                                   z value Pr(>|z|)
(Intercept)                                         19.622  < 2e-16 ***
Total.Household.Income                              -6.570 5.03e-11 ***
Total.Food.Expenditure                              13.524  < 2e-16 ***
Household.Head.SexMale                               6.470 9.82e-11 ***
Household.Head.Age                                  -4.842 1.29e-06 ***
Type.of.HouseholdSingle Family                     -12.159  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -2.104   0.0354 *
House.Age                                           -2.515   0.0119 *
Number.of.bedrooms                                  -1.902   0.0571 .
---
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 1247  degrees of freedom
Residual deviance:  857.2  on 1239  degrees of freedom
AIC: 4900.6

Number of Fisher Scoring iterations: 4
```

Test the goodness of fit of Poisson model

```
library(epiDisplay)
poisgof(model.better)
```

```
$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 857.1986

$df
[1] 1239

$p.value
[1] 1
```

The p value is 1, which indicates that the goodness of fit of the model is good.

Coefficient and interpretation of model

```
exp(coef(model.better))
```

```
                                        (Intercept)
                                          4.7835932
                                Total.Household.Income
                                          0.9999995
                                Total.Food.Expenditure
                                          1.0000061
                                Household.Head.SexMale
                                          1.2731562
                                    Household.Head.Age
                                          0.9947412
                        Type.of.HouseholdSingle Family
                                          0.6911276
Type.of.HouseholdTwo or More Nonrelated Persons/Members
                                          0.5974600
                                          House.Age
                                          0.9971080
                                   Number.of.bedrooms
                                          0.9704752
```

Because the sample data of Total.Household.Income and Total.Food.Expenditure is too large, their coefficient is too close to 1. It is impossible to know whether the interval contains 1, that is, whether the variable is significant. Therefore, we try to use the logarithm of these two variables to repeat the above steps to fit the model.

Change variable fitting model

```
model<-glm(Total.Number.of.Family.members~log(Total.Household.Income)+log(Total.Food.Expenditure)+Housel
summary(model)
```

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.19376  -0.58252  -0.09119   0.41915   2.65231

Coefficients:
                                                    Estimate Std. Error
(Intercept)                                        -3.156e+00  3.498e-01
log(Total.Household.Income)                        -2.523e-01  3.382e-02
log(Total.Food.Expenditure)                         7.270e-01  4.658e-02
Household.Head.SexMale                              1.921e-01  3.756e-02
Household.Head.Age                                 -3.411e-03  1.114e-03
Type.of.HouseholdSingle Family                     -3.251e-01  3.074e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -4.589e-01  2.440e-01
House.Floor.Area                                   -8.945e-05  3.042e-04
House.Age                                          -2.296e-03  1.174e-03
Number.of.bedrooms                                 -2.993e-02  1.704e-02
Electricity1                                       -7.605e-02  4.138e-02
                                                   z value Pr(>|z|)
(Intercept)                                         -9.024  < 2e-16 ***
log(Total.Household.Income)                         -7.461 8.59e-14 ***
log(Total.Food.Expenditure)                         15.607  < 2e-16 ***
Household.Head.SexMale                               5.114 3.16e-07 ***
Household.Head.Age                                  -3.062   0.0022 **
Type.of.HouseholdSingle Family                     -10.576  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -1.880   0.0601 .
House.Floor.Area                                    -0.294   0.7687
House.Age                                           -1.957   0.0504 .
Number.of.bedrooms                                  -1.756   0.0790 .
Electricity1                                        -1.838   0.0661 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1373.55  on 1247  degrees of freedom
Residual deviance:  749.66  on 1237  degrees of freedom
```

```
AIC: 4797

Number of Fisher Scoring iterations: 4
```

The stepwise method was used to complete the screening of independent variables.

```
step(model)
```

```
Start:  AIC=4797.04
Total.Number.of.Family.members ~ log(Total.Household.Income) +
    log(Total.Food.Expenditure) + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + House.Floor.Area + House.Age + Number.of.bedrooms +
    Electricity

                             Df Deviance    AIC
- House.Floor.Area            1   749.75 4795.1
<none>                            749.66 4797.0
- Number.of.bedrooms          1   752.75 4798.1
- Electricity                 1   753.00 4798.4
- House.Age                   1   753.54 4798.9
- Household.Head.Age          1   759.03 4804.4
- Household.Head.Sex          1   776.78 4822.2
- log(Total.Household.Income) 1   807.08 4852.5
- Type.of.Household           2   860.16 4903.5
- log(Total.Food.Expenditure) 1  1000.58 5046.0

Step:  AIC=4795.13
Total.Number.of.Family.members ~ log(Total.Household.Income) +
    log(Total.Food.Expenditure) + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + House.Age + Number.of.bedrooms + Electricity

                             Df Deviance    AIC
<none>                            749.75 4795.1
- Electricity                 1   753.08 4796.5
- Number.of.bedrooms          1   753.45 4796.8
- House.Age                   1   753.90 4797.3
- Household.Head.Age          1   759.11 4802.5
- Household.Head.Sex          1   776.89 4820.3
- log(Total.Household.Income) 1   809.51 4852.9
- Type.of.Household           2   860.16 4901.5
- log(Total.Food.Expenditure) 1  1001.24 5044.6

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

Coefficients:
                                       (Intercept)
                                         -3.144768
                          log(Total.Household.Income)
                                         -0.253972
                          log(Total.Food.Expenditure)
```

```
                                                0.727582
                                    Household.Head.SexMale
                                                0.192173
                                       Household.Head.Age
                                                -0.003409
                             Type.of.HouseholdSingle Family
                                                -0.324931
Type.of.HouseholdTwo or More Nonrelated Persons/Members
                                                -0.460081
                                                House.Age
                                                -0.002346
                                      Number.of.bedrooms
                                                -0.031361
                                            Electricity1
                                                -0.076011

Degrees of Freedom: 1247 Total (i.e. Null);  1238 Residual
Null Deviance:      1374
Residual Deviance: 749.8    AIC: 4795
```

Use a better model.

```
model.best<-glm(Total.Number.of.Family.members~log(Total.Household.Income)+log(Total.Food.Expenditure) +
family = "poisson")
outlierTest(model.best)
```

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
977 2.681894          0.0073207           NA
```

```
summary(model.best)
```

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.19539  -0.58469  -0.09141   0.42019   2.66999

Coefficients:
                                                        Estimate Std. Error
(Intercept)                                            -3.144768   0.347530
log(Total.Household.Income)                            -0.253972   0.033361
log(Total.Food.Expenditure)                             0.727582   0.046545
Household.Head.SexMale                                  0.192173   0.037564
Household.Head.Age                                     -0.003409   0.001114
Type.of.HouseholdSingle Family                         -0.324931   0.030739
Type.of.HouseholdTwo or More Nonrelated Persons/Members -0.460081   0.244011
House.Age                                              -0.002346   0.001161
```

```
Number.of.bedrooms                                        -0.031361   0.016328
Electricity1                                              -0.076011   0.041378
                                                          z value Pr(>|z|)
(Intercept)                                                -9.049  < 2e-16 ***
log(Total.Household.Income)                                -7.613 2.68e-14 ***
log(Total.Food.Expenditure)                                15.632  < 2e-16 ***
Household.Head.SexMale                                      5.116 3.12e-07 ***
Household.Head.Age                                         -3.061  0.00221 **
Type.of.HouseholdSingle Family                            -10.571  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members    -1.885  0.05936 .
House.Age                                                  -2.021  0.04325 *
Number.of.bedrooms                                         -1.921  0.05478 .
Electricity1                                               -1.837  0.06621 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1373.55  on 1247  degrees of freedom
Residual deviance:  749.75  on 1238  degrees of freedom
AIC: 4795.1

Number of Fisher Scoring iterations: 4
```

Without outliers, the best model is obtained.

Test the goodness of fit of Poisson model

```
library(epiDisplay)
poisgof(model.best)
```

```
$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 749.7502

$df
[1] 1238

$p.value
[1] 1
```

The p value is 1, which indicates that the goodness of fit of the model is good.

Coefficient and interpretation of model

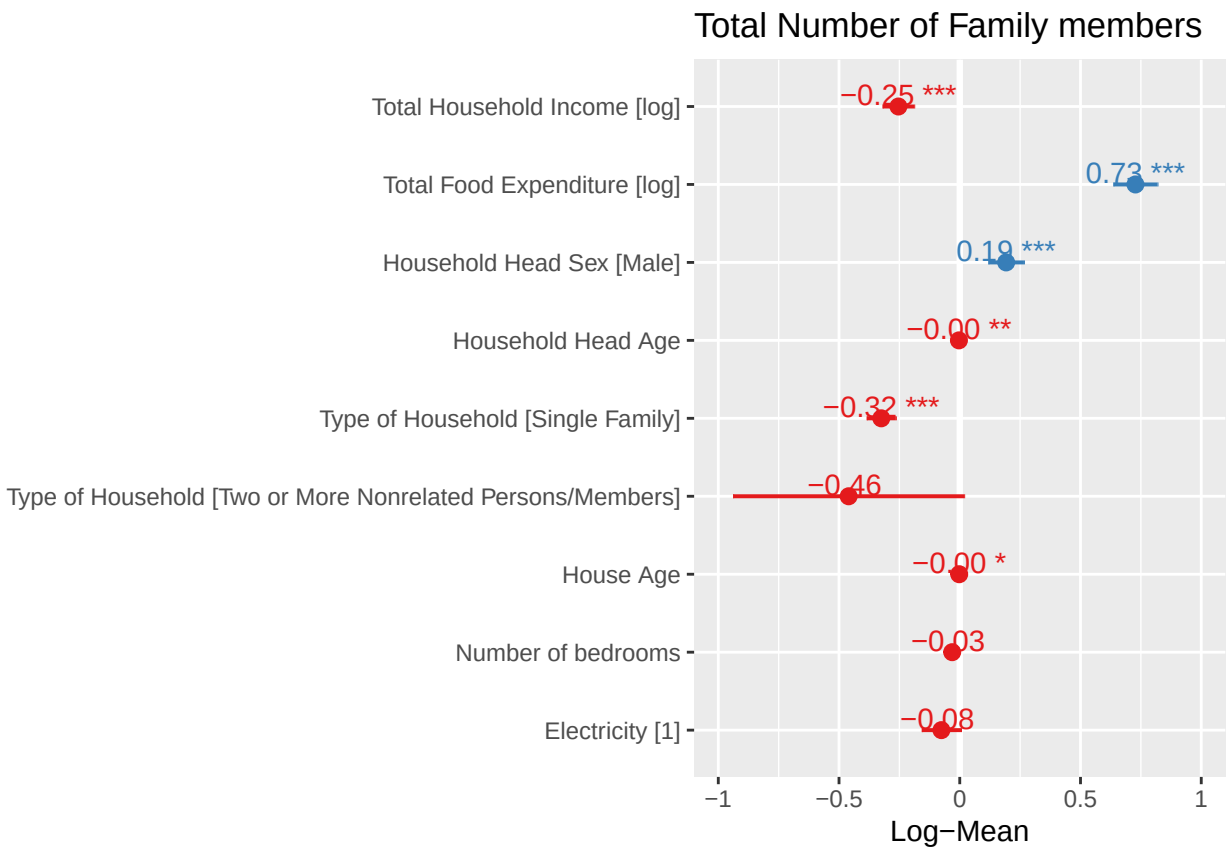
```
exp(coef(model.best))
```

```
                                          (Intercept)
                                            0.04307693
                              log(Total.Household.Income)
                                            0.77571384
                             log(Total.Food.Expenditure)
                                            2.07006853
                                  Household.Head.SexMale
                                            1.21188068
                                     Household.Head.Age
                                            0.99659634
                           Type.of.HouseholdSingle Family
                                            0.72257687
Type.of.HouseholdTwo or More Nonrelated Persons/Members
                                            0.63123227
                                           House.Age
                                            0.99765704
                                    Number.of.bedrooms
                                            0.96912536
                                          Electricity1
                                            0.92680617
```

In the MIMAROPA region, all variables except Number.of.bedrooms and Electricity show significance. While keeping other variables unchanged, the number of people living in the house will be multiplied by 0.7757 for every 1 unit increase in the logarithm of the family's annual income (Philippine Peso).the number of people living in the house will be multiplied by 2.0701 for every 1 unit increase in the logarithm of Annual expenditure by the household on food (inmPhilippine peso).If the gender of head of the houses sex is male, the number of people living in the house will be multiplied by 1.2119, indicating that the owner is male, which has a positive impact on the increase of the number of people living in the room. The number of people living in the house will be multiplied by 0.9966 for each additional year of head of the houses age. In the relationship between the group of people living in the house,two or more nonrelated persons / members have no significant effect on the number of residents and single family will have a negative impact on the increase of the number of people living in the room. The number of people living in the house will be multiplied by 0.9977 for each year of age of the building.

Poisson regression predicting Total.Number.of.Family.members

```
plot_model(model.best, transform = NULL, show.p = T, show.values = T)
```



It can be seen from the above figure that Total.Household.Income,Total.Food.Expenditure,Household.Head.Sex ,Household.Head.Age,Single Family in Type.of.Household and House.Age all have a significant impact on Total.Number.of.Family.That is, the increase of annual household food expenditure (Philippine Peso), the gender of the head of household is male, which has a positive impact on the number of people living in this house, and the increase of annual household income (in Philippine Peso), the age of the head of household (in), single families and construction age (in) have a negative impact on the number of people living in this house

Conclusions and Future Work

After selecting models, we have found that the influential variables of the number of people living in a household are : a, b, c, d

We may select more regions of Philippines to compare these variables, or select year as one of the explanatory variables since this data is collected every three years.