Group_01_Analysis.Rmd

Group_01

3/14/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(moderndive)
library(skimr)
library(readr)
library(Stat2Data)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(broom)
library(sjPlot)
library(DescTools)
library(knitr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
```

Introduction

Dataset comes from the FIES (Family Income and Expenditure Survey) recorded in the Philippines. The survey, which is undertaken every three years, is aimed at providing data on family income and expenditure. In this study, we are going to identify the most influential variables on the number of people living in a household using a Generalised Linear Model. Below you can see an overview of the data and variables

```
## Rows: 1,717
## Columns: 10
## $ Income           <int> 480332, 198235, 82785, 107589, 189322, 152883, 198621~
## $ Expenditure      <int> 117848, 67766, 61609, 78189, 94625, 73326, 104644, 95~
## $ Gender           <fct> Female, Male, Male, Male, Male, Male, Male, Male, Fem~
## $ Age              <int> 49, 40, 39, 52, 65, 46, 45, 33, 17, 53, 49, 35, 38, 5~
## $ Type             <fct> Extended Family, Single Family, Single Family, Single~
## $ Number_of_Family <int> 4, 3, 6, 3, 4, 4, 5, 5, 2, 6, 4, 7, 7, 3, 2, 4, 5, 8,~
## $ Area             <int> 80, 42, 35, 30, 54, 40, 35, 35, 35, 70, 40, 35, 35, 5~
## $ HouseAge         <int> 75, 15, 12, 15, 16, 7, 18, 48, 8, 12, 9, 17, 5, 43, 7~
## $ Bedrooms         <int> 3, 2, 1, 1, 3, 2, 1, 2, 1, 3, 2, 3, 1, 3, 1, 1, 1, 1,~
## $ Electricity      <int> 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,~
```

$FoodExpenditure$ is the annual expenditure by the household on food (in Philippine peso)
$Gender$ is the head of the households sex
Age is the head of the households age (in years)
$Type$ is the relationship between the group of people living in the house
$Number_of_Family$ is the number of people living in the house
$Area$ is the floor area of the house (in m^2)
$HouseAge$ is the age of the building (in years)
$bedrooms$ is the number of bedrooms in the house
$Electricity$ indicates that if the house have electricity? (1=Yes, 0=No)

Exploratory Data Analysis

The following tables and graphs are produced to provide statistical summaries and graphs to see the distribution variables and their relationship and identify any possible outliers.

Table 1: Summary Statistics

Variable	Min	Mean	SD	Median	Max
Income	11988	269524.8	275079.4	188050	6042860
Expenditure	6781	80249.4	41241.7	73459	327724
Age	17	52.2	14.5	52	99
Number_of_Family	1	4.7	2.3	4	15
Area	5	90.9	99.3	54	900
HouseAge	0	22.9	15.3	20	100
Bedrooms	0	2.3	1.4	2	9
Electricity	0	0.9	0.3	1	1

Referring to 1, the difference between mean and median indicates the presence of outliers in the data. The following pair plots fuhrer support the presence of outliers.

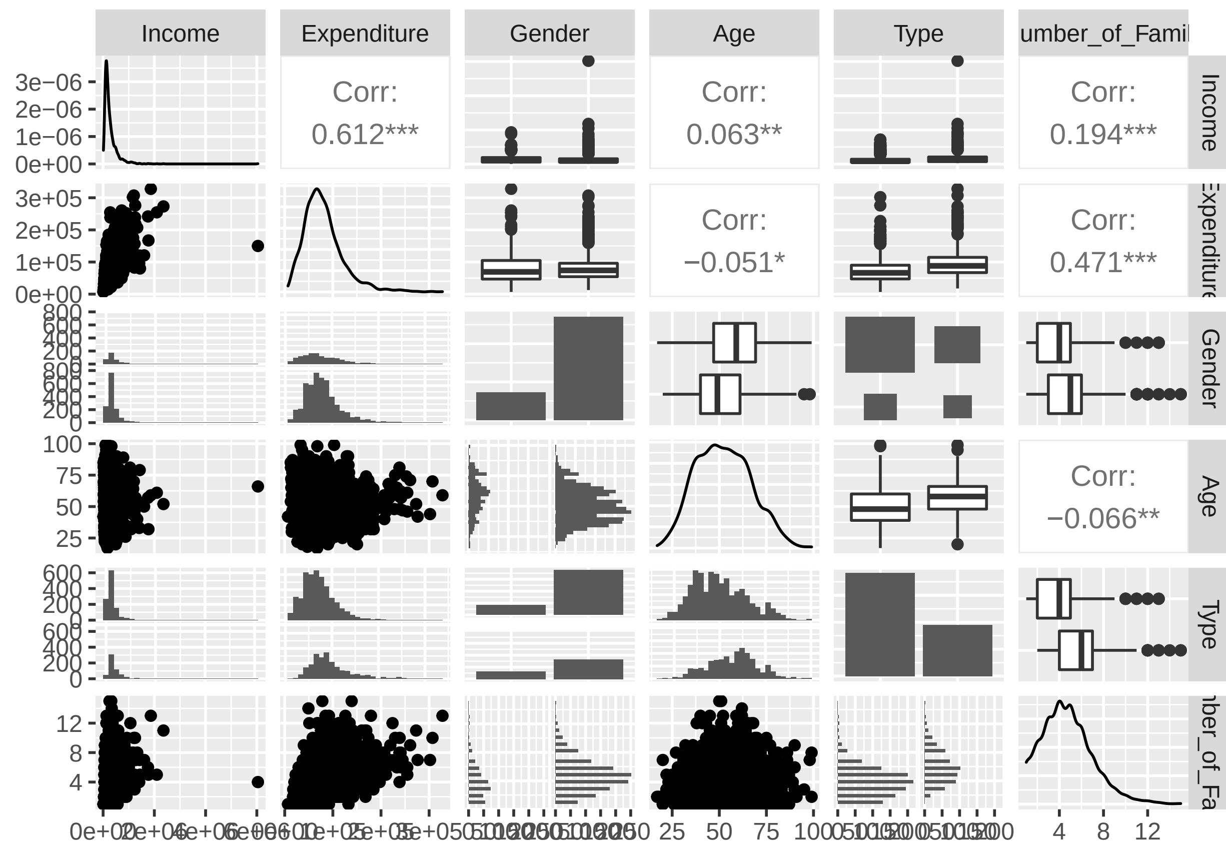


Figure 1: The pair plot (first five variables) shows the relationship between each of the two variables

Figure 3 presents the distribution of outcome variables with regards to each of the continuous variables. The boxplots can help us identify the outliers

Once the outliers are spotted and removed, the skewness in data is decreased. The Figure 4 displays the box plots after removing outliers and log transforming Income, Expenditure and Area.

```
## TableGrob (3 x 2) "arrange": 6 grobs
##   z     cells    name           grob
```

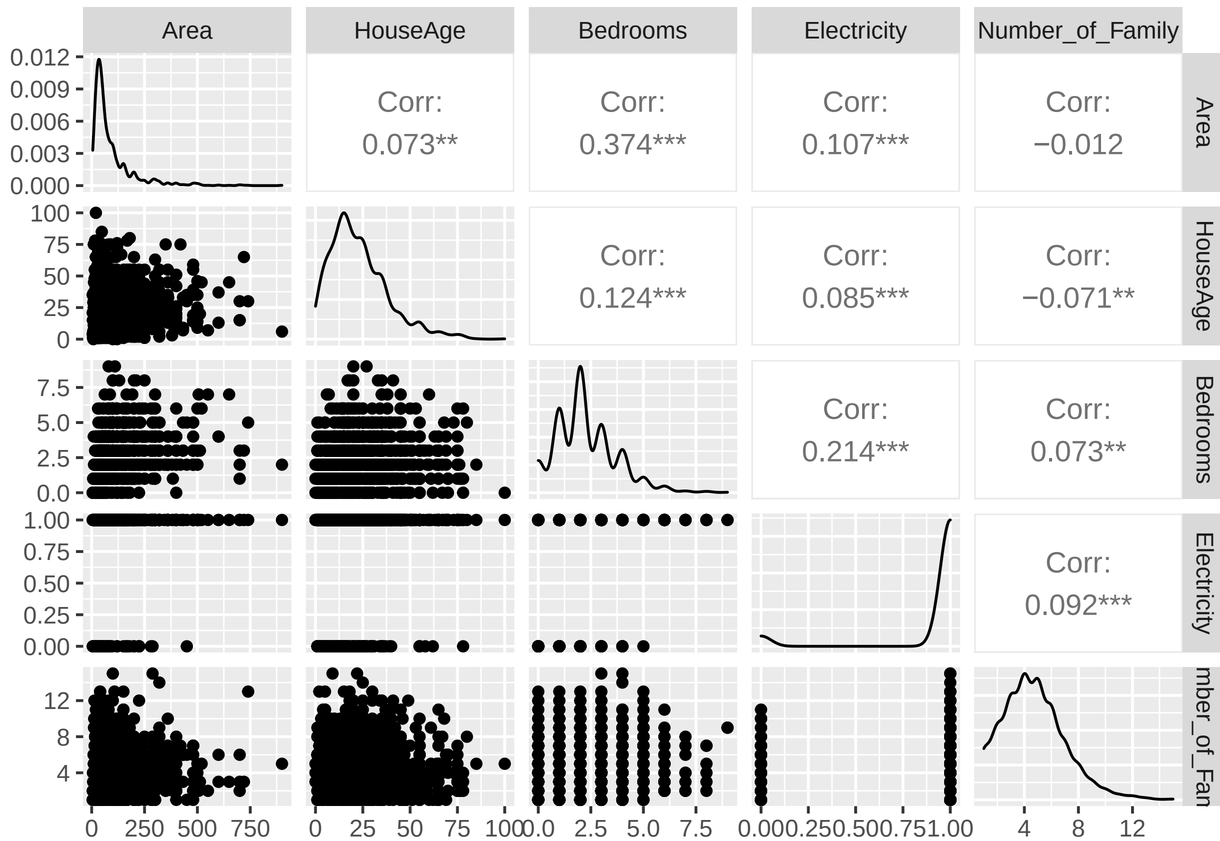


Figure 2: The pair plot (second five variables) shows the relationship between each of the two variables

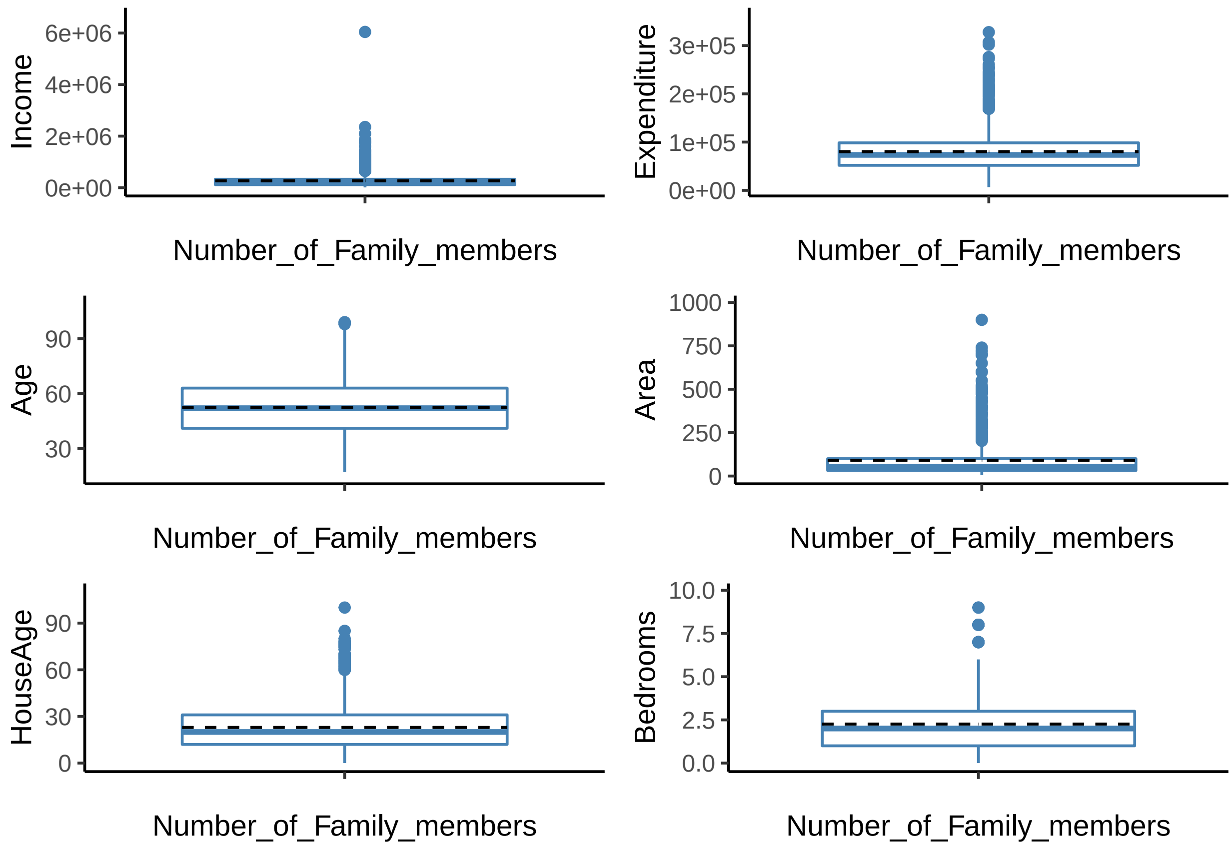


Figure 3: The boxplot of the outcome variables vs each of the continuous explanatory variables

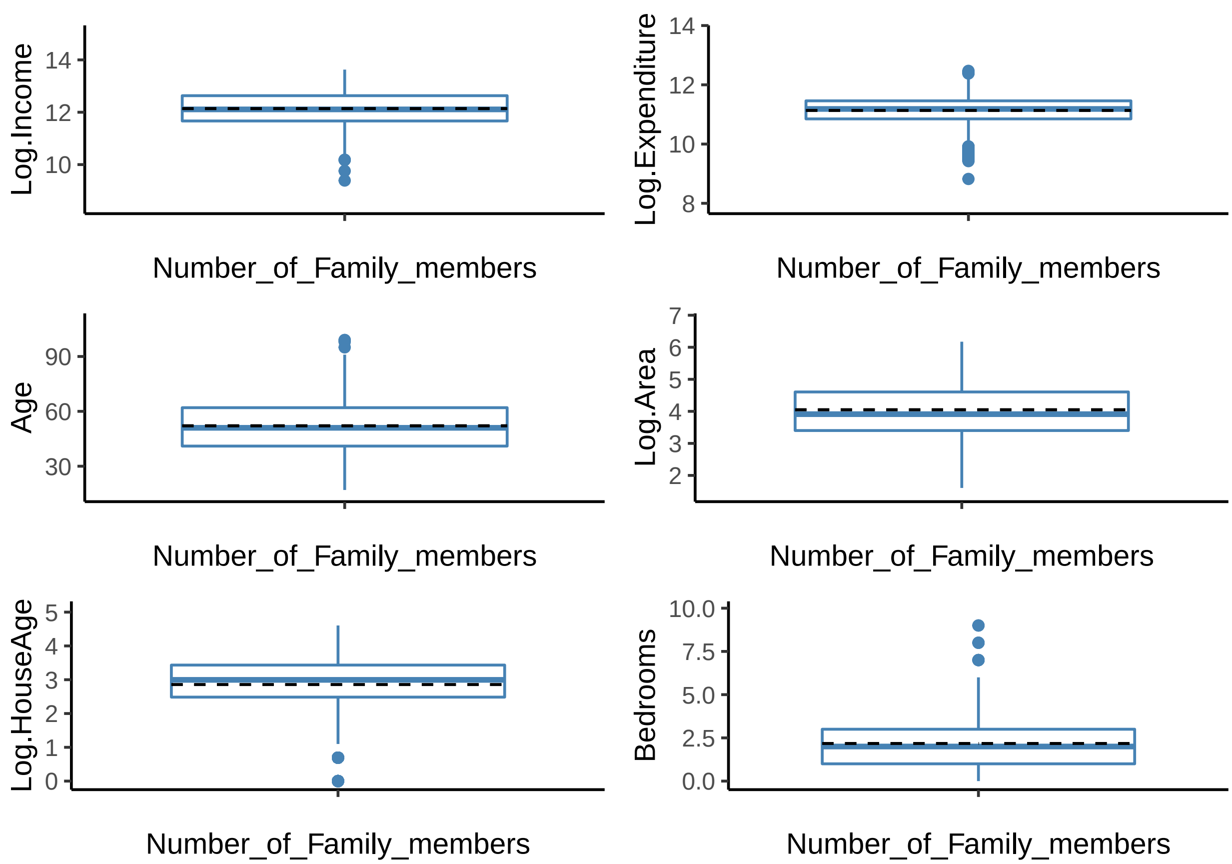


Figure 4: The boxplot of the outcome variables vs each of the continuous explanatory variables after removing outliers and applying log transformation

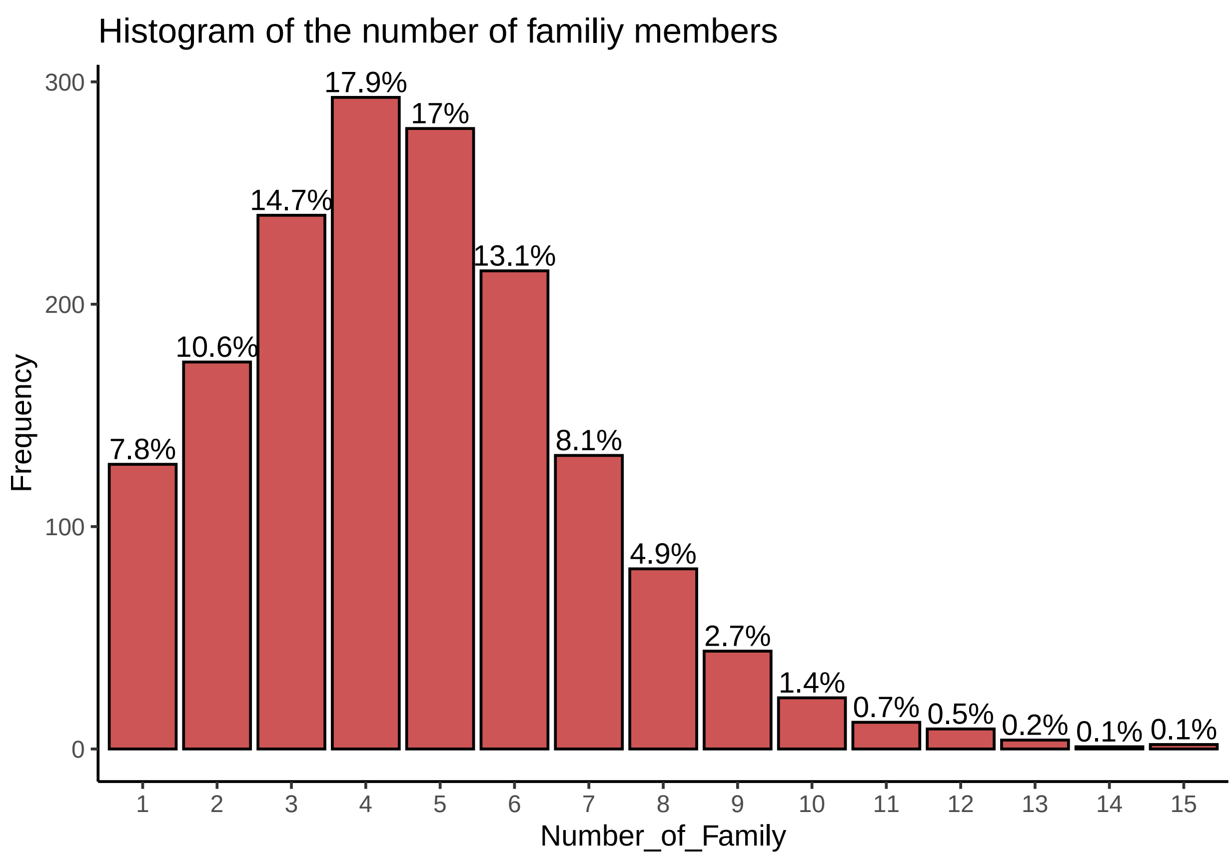
```
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (2-2,1-1) arrange gtable[layout]
## 4 4 (2-2,2-2) arrange gtable[layout]
## 5 5 (3-3,1-1) arrange gtable[layout]
## 6 6 (3-3,2-2) arrange gtable[layout]
```

The table 2 shows the summary statistics after removing outliers and transforming the variables of Income, Expenditure and Aare using log transformation. The difference between medians and means are now narrower.

Table 2: Summary Statistics after outliers removed

Variable	Min	Mean	SD	Median	Max
Log.Income	9.4	12.1	0.7	12.1	13.6
Log.Expenditure	8.8	11.1	0.5	11.2	12.5
Age	17.0	52.1	14.6	51.0	99.0
Number_of_Family	1.0	4.6	2.3	4.0	15.0
Log.Area	1.6	4.0	0.9	3.9	6.2
Log.HouseAge	0.0	2.9	0.8	3.0	4.6
Bedrooms	0.0	2.2	1.4	2.0	9.0
Electricity	0.0	0.9	0.3	1.0	1.0

The Figure **??** shows the distribution of household sizes:



The histogram (Figure **??**) of the family size resembles a Poisson distribution.

Formal Data Analysis

In this section we model the data to identify the most influential factors on the number of family members using Poisson distribution since the the outcome variabl is a count data. Then the goodness of fit comparison is made to select the best model based on AIC and Deviance. However, for the a Poisson to completly hold the variance and mean shall be equal. The variance and the mean are 5.4 and 4.6. We can say that they are roughly equal.

Poisson model

```
model.poisson = glm(data = data, Number_of_Family ~ Log.Income + Log.Expenditure +
                       Gender + Age + Type + Log.Area + Log.HouseAge + Bedrooms + Electricity,
                     family = poisson(link = "log"))
initial.poisson.AIC = model.poisson$aic
summary(model.poisson)
```

```
##
## Call:
## glm(formula = Number_of_Family ~ Log.Income + Log.Expenditure +
##     Gender + Age + Type + Log.Area + Log.HouseAge + Bedrooms +
##     Electricity, family = poisson(link = "log"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4622  -0.6427  -0.1092   0.4471   3.9148
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.9645515  0.3234877 -12.256  < 2e-16 ***
## Log.Income          -0.1040666  0.0278387  -3.738 0.000185 ***
## Log.Expenditure      0.6030269  0.0364237  16.556  < 2e-16 ***
## GenderMale           0.1940268  0.0308884   6.282 3.35e-10 ***
## Age                 -0.0012899  0.0009081  -1.420 0.155472
## TypeExtended Family  0.3007800  0.0257492  11.681  < 2e-16 ***
## Log.Area            -0.0173178  0.0165715  -1.045 0.296007
## Log.HouseAge        -0.0184174  0.0139917  -1.316 0.188071
## Bedrooms            -0.0197678  0.0106700  -1.853 0.063931 .
## Electricity         -0.0290303  0.0484405  -0.599 0.548974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1919.9  on 1636  degrees of freedom
## Residual deviance: 1131.5  on 1627  degrees of freedom
## AIC: 6511.5
##
## Number of Fisher Scoring iterations: 4
```

The full model data in terms of AIC, BIC and can be seen Table 3

Table 3: Full model data

Model	AIC	BIC	pseudo_R2
Poisson	6511.5	6565.5	0.108

We now try to see if we can improve model AIC and decrease the deviance using step function. In this method we start from the full model and every time drop one variable and calculate the AIC. This procedure is continued until no further reduction in AIC is observed.

```
## Start:  AIC=6511.53
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Age +
##     Type + Log.Area + Log.HouseAge + Bedrooms + Electricity
##
##                   Df Deviance    AIC
## - Electricity      1   1131.9 6509.9
## - Log.Area         1   1132.6 6510.6
## - Log.HouseAge     1   1133.3 6511.3
## <none>                 1131.5 6511.5
## - Age              1   1133.6 6511.5
## - Bedrooms         1   1135.0 6513.0
## - Log.Income       1   1145.6 6523.6
## - Gender           1   1172.5 6550.5
## - Type             1   1266.4 6644.4
## - Log.Expenditure  1   1413.2 6791.2
##
## Step:  AIC=6509.89
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Age +
##     Type + Log.Area + Log.HouseAge + Bedrooms
##
##                   Df Deviance    AIC
## - Log.Area         1   1133.0 6509.0
## - Log.HouseAge     1   1133.8 6509.8
## - Age              1   1133.8 6509.8
## <none>                 1131.9 6509.9
## - Bedrooms         1   1135.5 6511.5
## - Log.Income       1   1146.5 6522.4
## - Gender           1   1172.8 6548.8
## - Type             1   1266.4 6642.4
## - Log.Expenditure  1   1413.4 6789.4
##
## Step:  AIC=6509.02
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Age +
##     Type + Log.HouseAge + Bedrooms
##
##                   Df Deviance    AIC
## - Age              1   1135.0 6509.0
## - Log.HouseAge     1   1135.0 6509.0
## <none>                 1133.0 6509.0
## - Bedrooms         1   1139.2 6513.2
## - Log.Income       1   1150.2 6524.1
## - Gender           1   1173.8 6547.7
## - Type             1   1266.6 6640.6
## - Log.Expenditure  1   1422.3 6796.2
```

```
##
## Step:  AIC=6508.99
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Type +
##     Log.HouseAge + Bedrooms
##
##                   Df Deviance    AIC
## <none>                 1135.0 6509.0
## - Log.HouseAge     1   1137.5 6509.5
## - Bedrooms         1   1142.1 6514.1
## - Log.Income       1   1153.1 6525.1
## - Gender           1   1179.6 6551.6
## - Type             1   1270.5 6642.5
## - Log.Expenditure  1   1447.0 6819.0
```

```
##
## Call:  glm(formula = Number_of_Family ~ Log.Income + Log.Expenditure +
##     Gender + Type + Log.HouseAge + Bedrooms, family = poisson(link = "log"),
##     data = data)
##
## Coefficients:
##         (Intercept)           Log.Income      Log.Expenditure
##            -4.12194             -0.11468              0.61611
##          GenderMale  TypeExtended Family         Log.HouseAge
##             0.19999              0.28829             -0.02198
##            Bedrooms
##            -0.02603
##
## Degrees of Freedom: 1636 Total (i.e. Null);  1630 Residual
## Null Deviance:       1920
## Residual Deviance: 1135  AIC: 6509
```

```
##
## Call:
## glm(formula = Number_of_Family ~ Log.Income + Log.Expenditure +
##     Gender + Type + Bedrooms + Log.HouseAge, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4907  -0.6526  -0.1064   0.4607   3.8772
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.121939   0.307149 -13.420  < 2e-16 ***
## Log.Income          -0.114685   0.027026  -4.243 2.20e-05 ***
## Log.Expenditure      0.616108   0.035435  17.387  < 2e-16 ***
## GenderMale           0.199987   0.030547   6.547 5.87e-11 ***
## TypeExtended Family  0.288294   0.024596  11.721  < 2e-16 ***
## Bedrooms            -0.026034   0.009801  -2.656   0.0079 **
## Log.HouseAge        -0.021982   0.013793  -1.594   0.1110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 1919.9  on 1636  degrees of freedom
## Residual deviance: 1135.0  on 1630  degrees of freedom
## AIC: 6509
##
## Number of Fisher Scoring iterations: 4
```

By removing Electricity and Area and Age, the AIC is reduced from 6511.5 to 6509. We now check for the goodness of fit by comparing it against the null model. The 95 percent $\chi^2(5)$ equals 11.1. Taking the difference in deviances (likelihood ratio test) results in the value of 782.4 which is significant when compared to 11.1. Therefore, there is no deviance of lack of fit with our model after removing the variables. The Log.HouseAge is not also significantly different from zero and by removing it, the AIC remains almost constant(6509.5). Given above, The formula below is proposed to model expected count:

$$log(Number\_of\_Family) = \beta_0 + \beta_1 \cdot log(Income) + \beta_2 \cdot log(FoodExpenditure) + \beta_3 \cdot Gender + \beta_4 \cdot Bedrooms + \beta_5 \cdot Type$$

Table 4 presents AIC, BIC and Pseudo-R2 for the final model:

Table 4: Final model data

Model	AIC	BIC	pseudo_R2
Poisson	6509.5	6541.9	0.107

Parameter estimates

Table 5 displays the parameter estimates. The main explanatory variables are significantly different from zero

Table 5: parameter estimates of the Poisson regression model

	Estimate	Std. Error	Test Statistic	P.value
(Intercept)	-4.203	0.303	-13.890	0.000
Log.Income	-0.116	0.027	-4.277	0.000
Log.Expenditure	0.619	0.035	17.503	0.000
GenderMale	0.204	0.030	6.686	0.000
TypeExtended Family	0.285	0.025	11.633	0.000
Bedrooms	-0.028	0.010	-2.934	0.003

The rate ratio are obtained by exponentiating the coefficients(Table 6). Figure **??** exhibits the rate ration for the different explantory variables

Table 6: Rate ratios 95 percent confidence interval

	Estimate	Lower CI	Upper CI
(Intercept)	0.01	0.01	0.03
Log.Income	0.89	0.84	0.94
Log.Expenditure	1.86	1.73	1.99
GenderMale	1.23	1.16	1.30
TypeExtended Family	1.33	1.27	1.40

	Estimate	Lower CI	Upper CI
Bedrooms	0.97	0.95	0.99

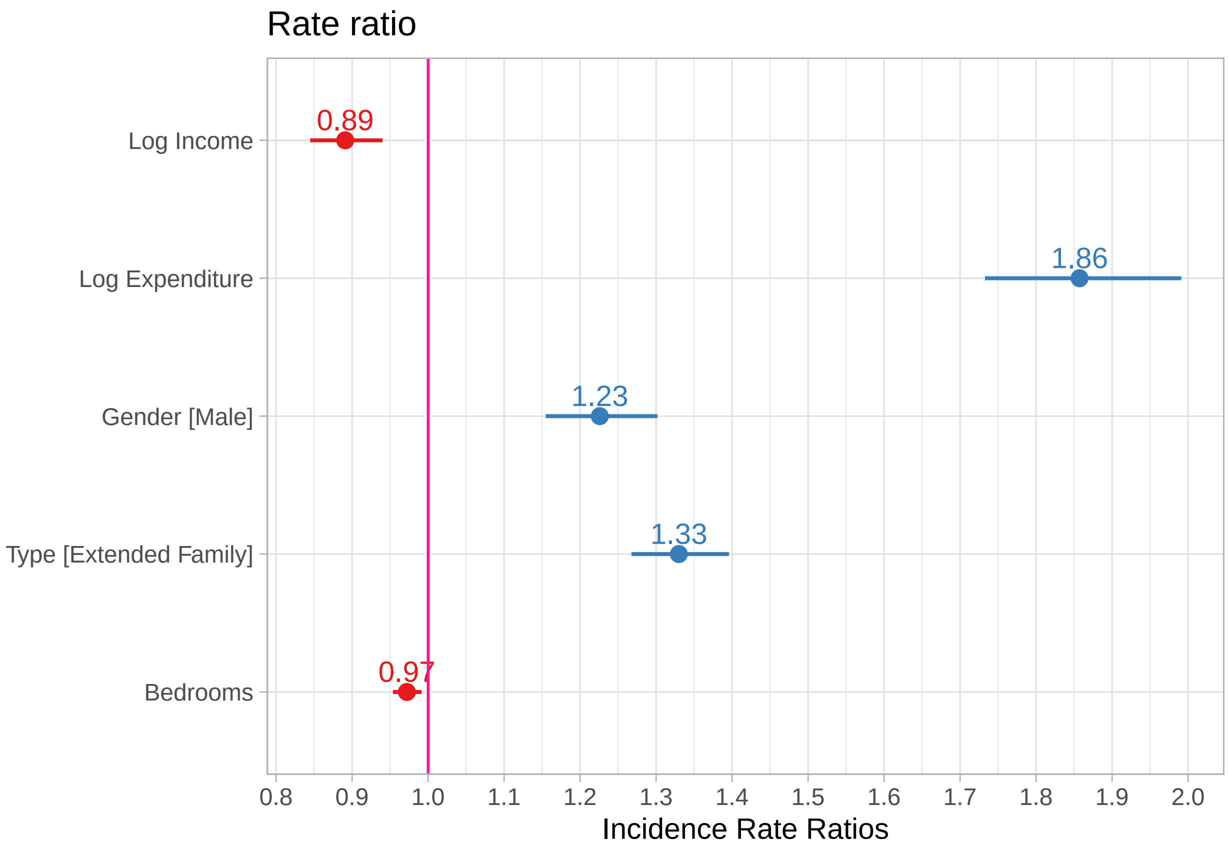


Figure 5 shows the relationship between each of the variables and the outcome variable

Figure 6 illustrates the predictions of the model. The difference between real values and predictions can be seen in Figure 7

```
## Warning: Removed 2 rows containing missing values (geom_col).
```

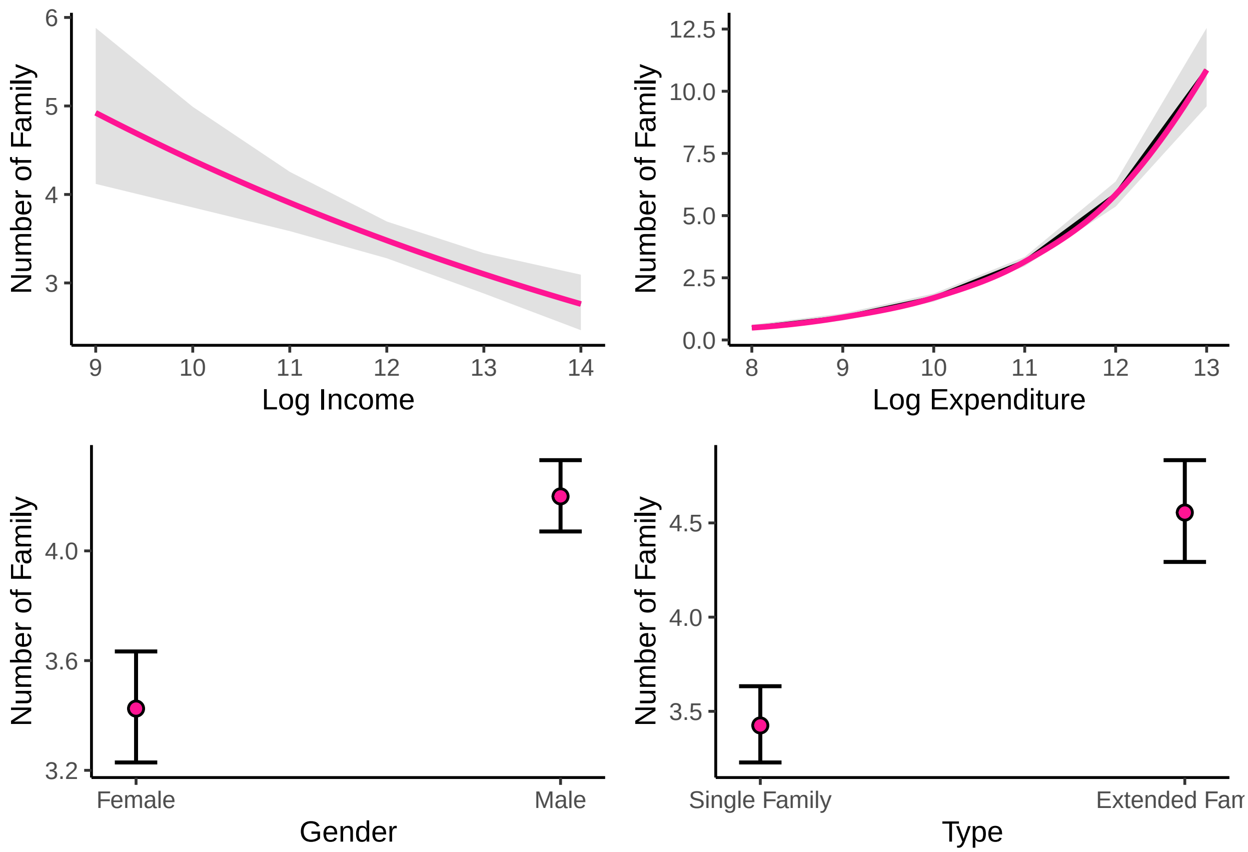


Figure 5: The relationship between explanatory variables and the number of family members

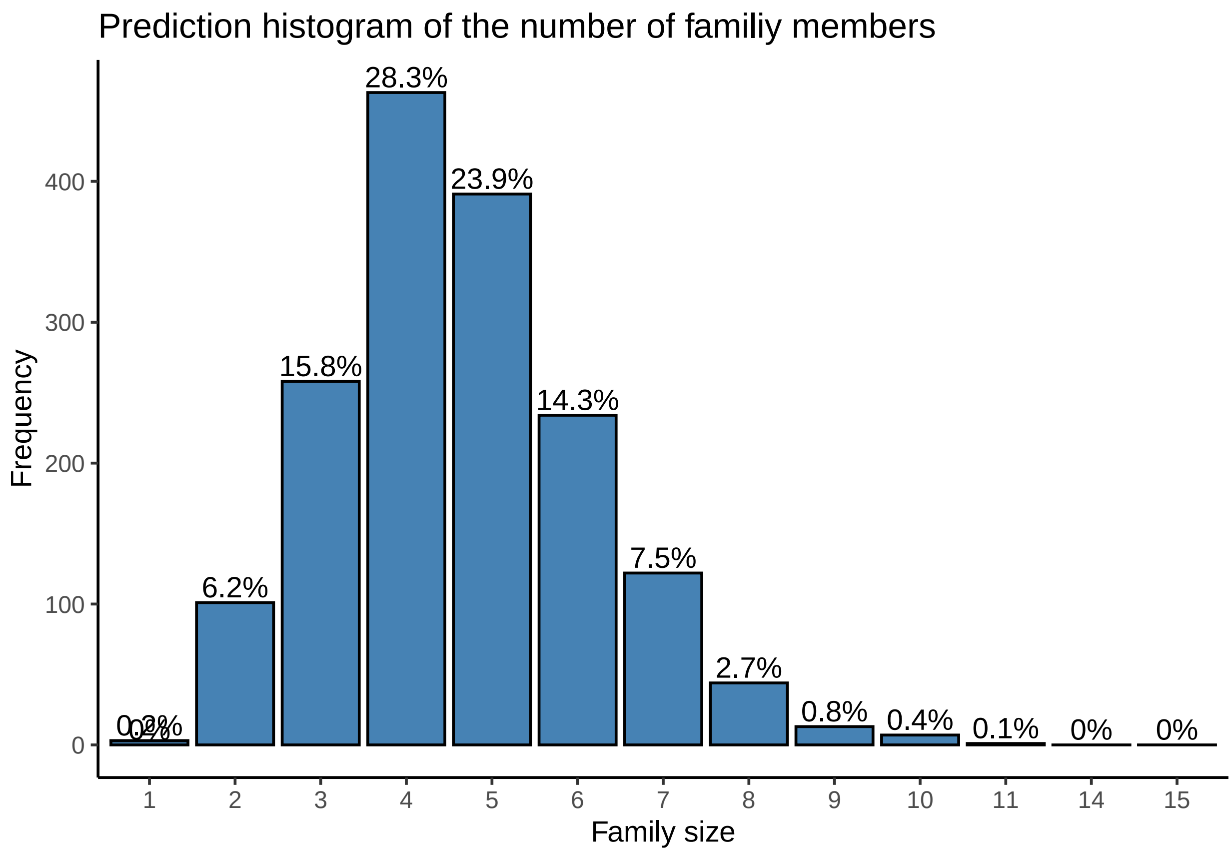


Figure 6: Histogram of predictions of the number of family members

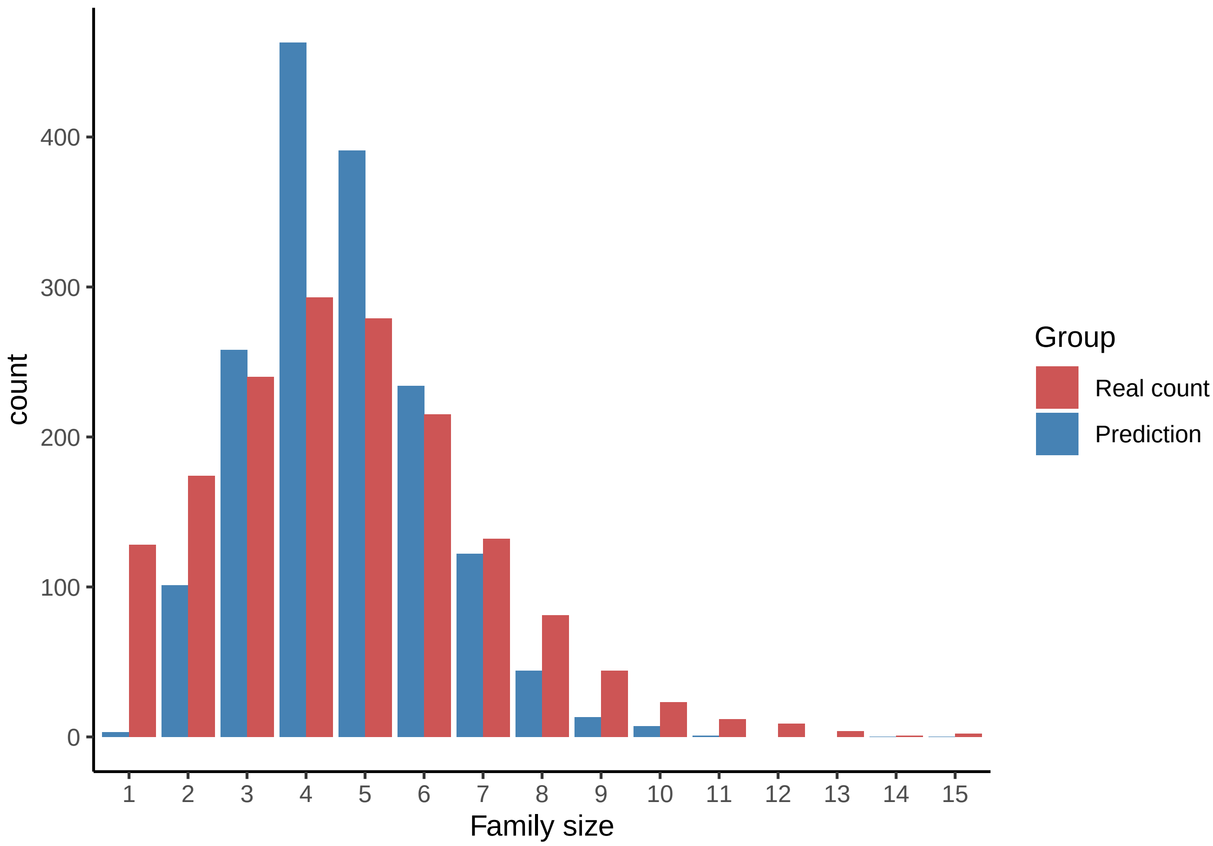


Figure 7: Comparison between real data and predictions