Analyzing the impact of different household variables on household size in different regions of the Philippines

Yue Xing

Qinyuan Ye

Jingyi Zhu

Chengyang Dong

Contents

Introduction

Explanatory Data Analysis

Model Selection

Assumption Check

Conclusion

Further Work

Introduction

We used GLMs to construct the relationship between house size and other variables.

To select which variables are important. We used correlation analysis, AIC, and cross-validation for model selection, and residual analysis for model validation.

Research Question:

Which household related variables influence the number of people living in a household?

Explanatory Data Analysis

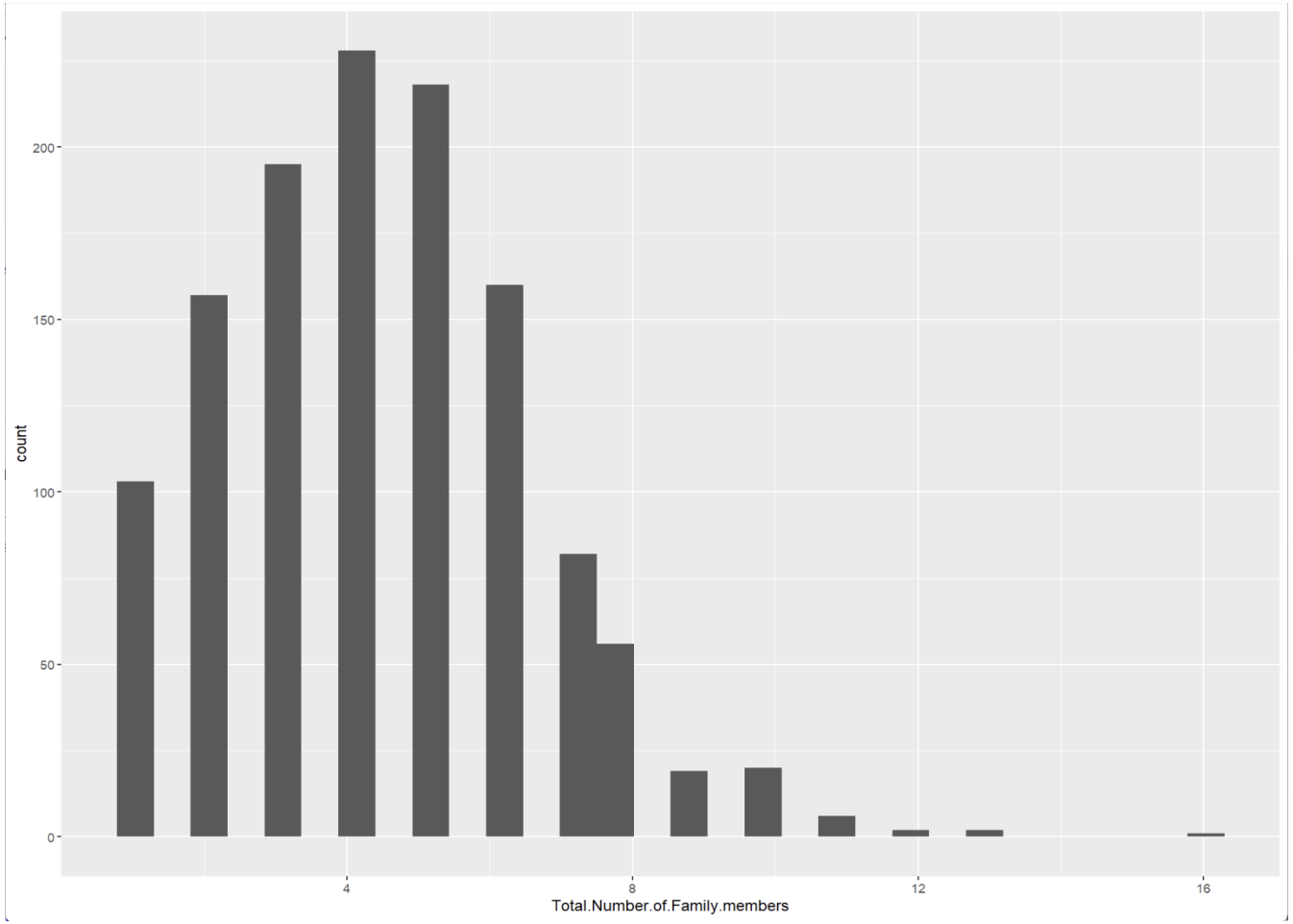


Figure 1: Histogram on the distribution of the total number of family members

Explanatory Data Analysis

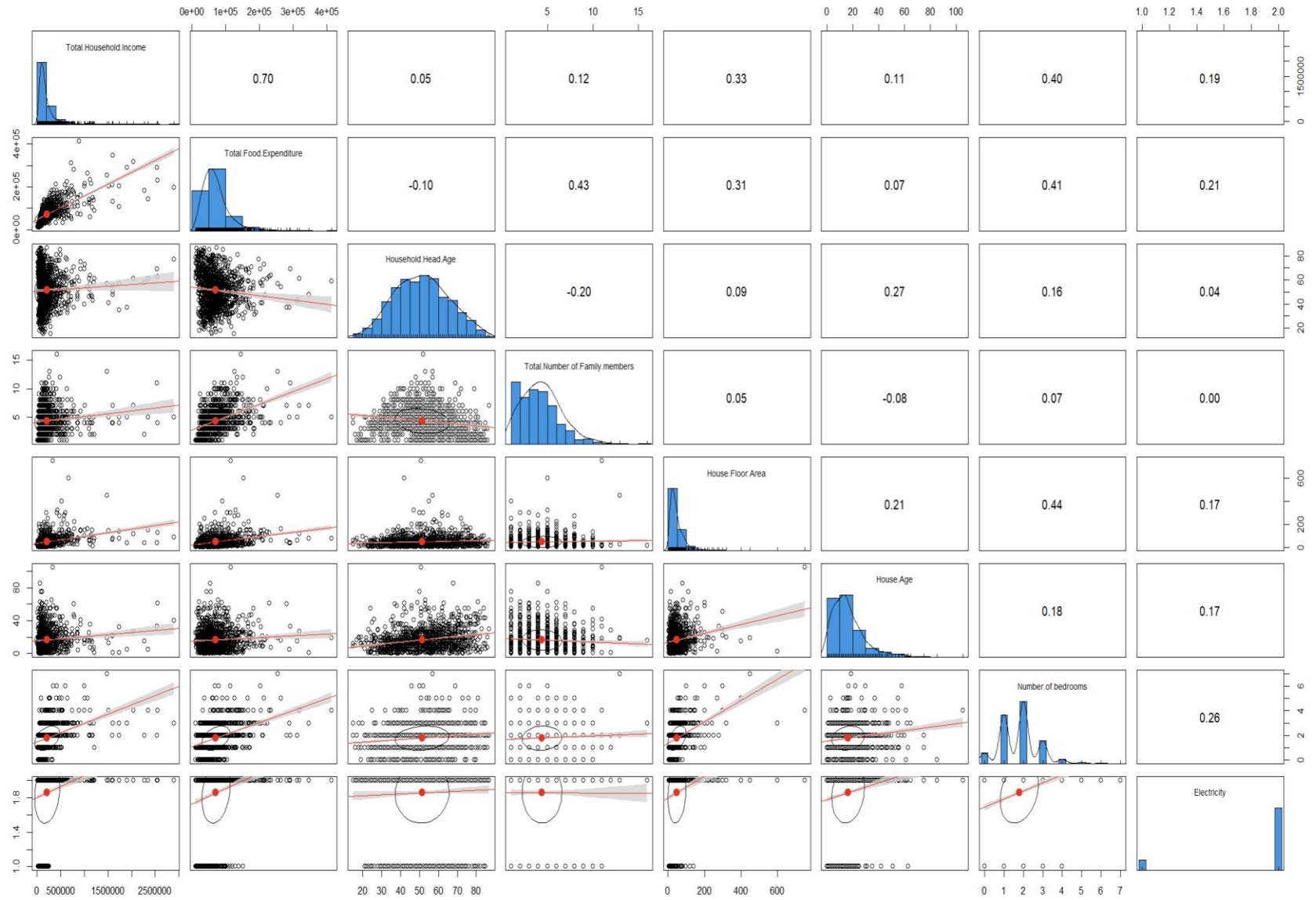


Figure 2: Visualization of multivariate relationships

Explanatory Data Analysis

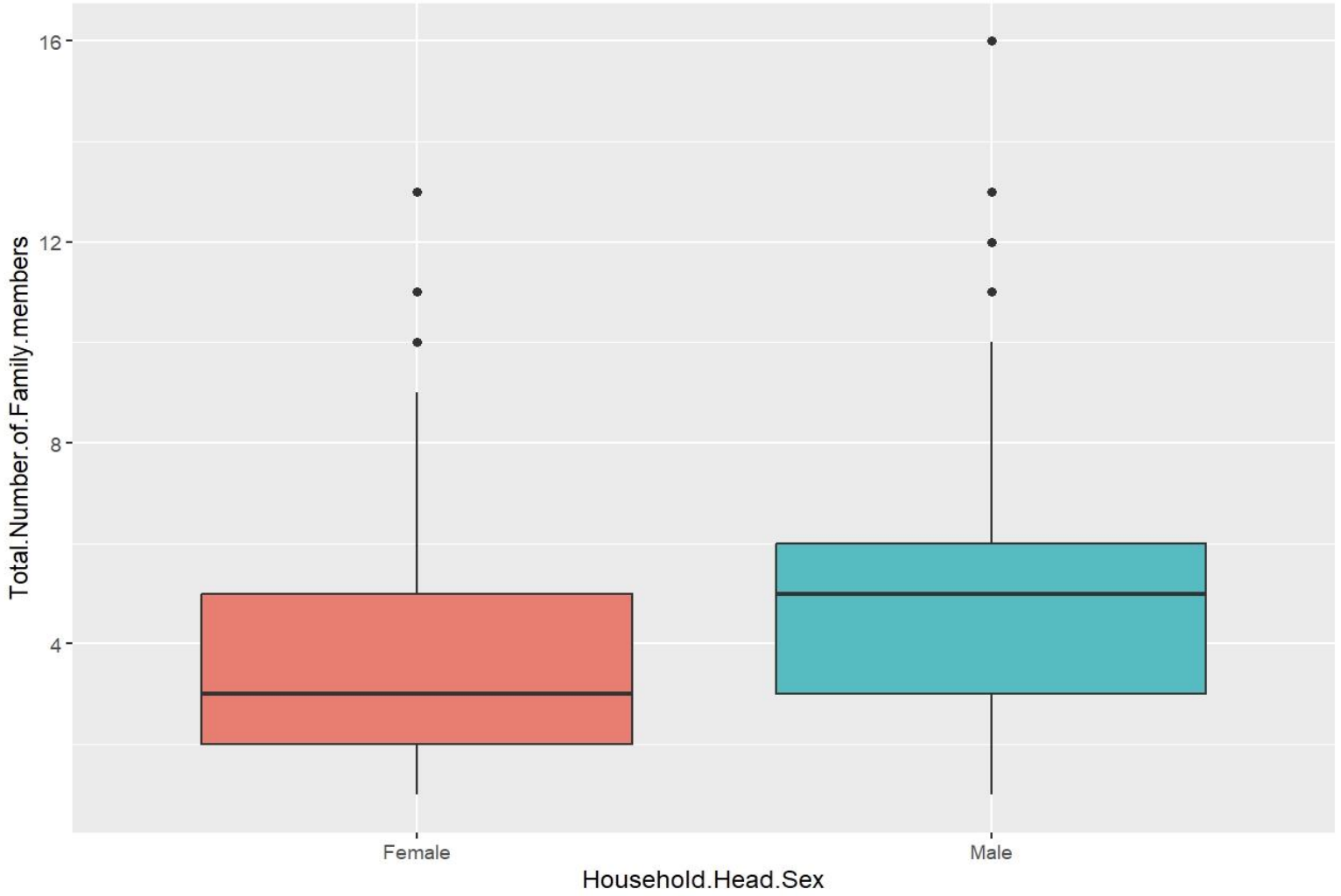


Figure 3: Boxplot for Electricity and total number of family member by household head sex

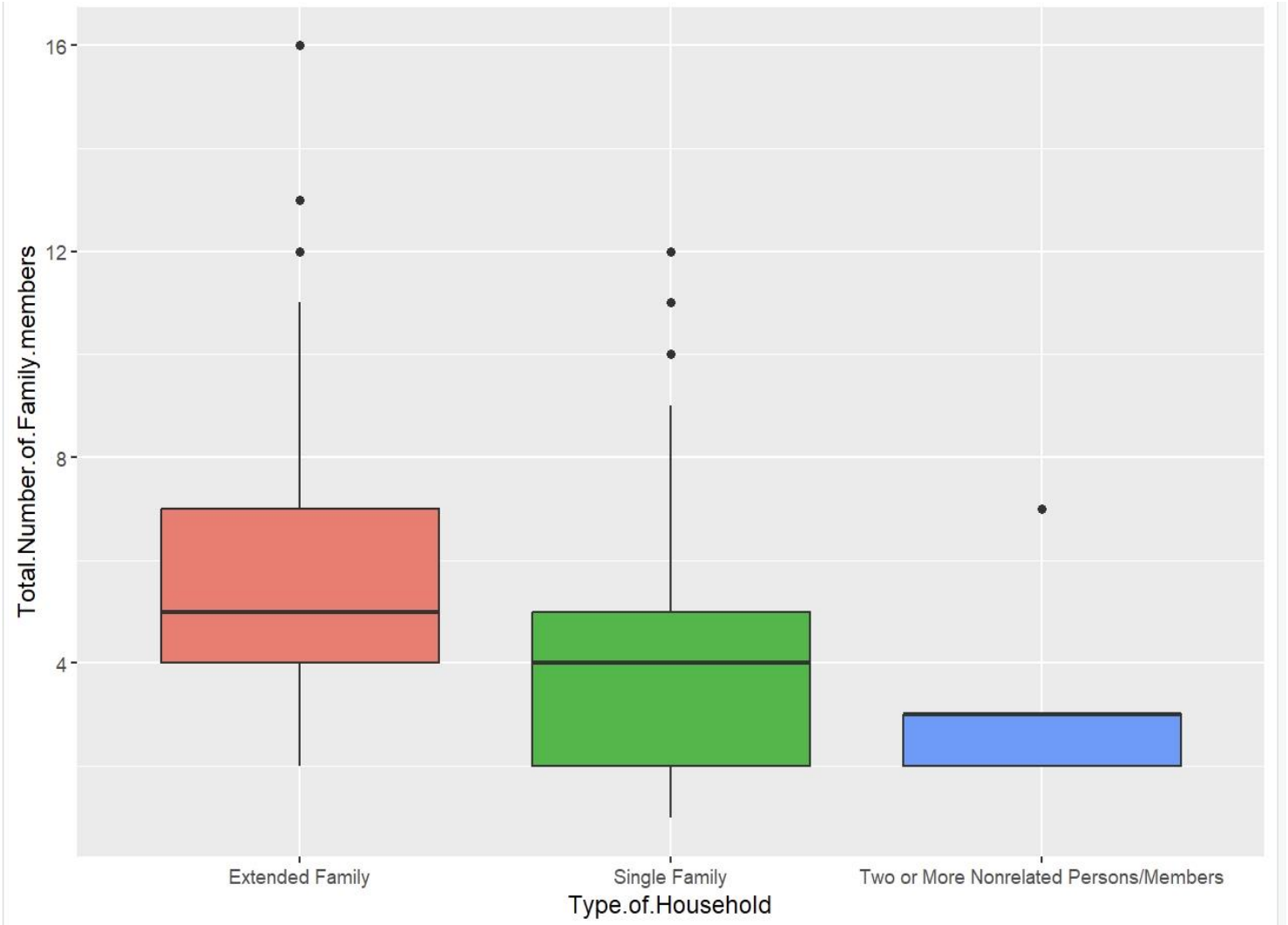


Figure 4: Boxplot of Total Number of Family Members by Type of Household

Explanatory Data Analysis

Explanatory Data Analysis

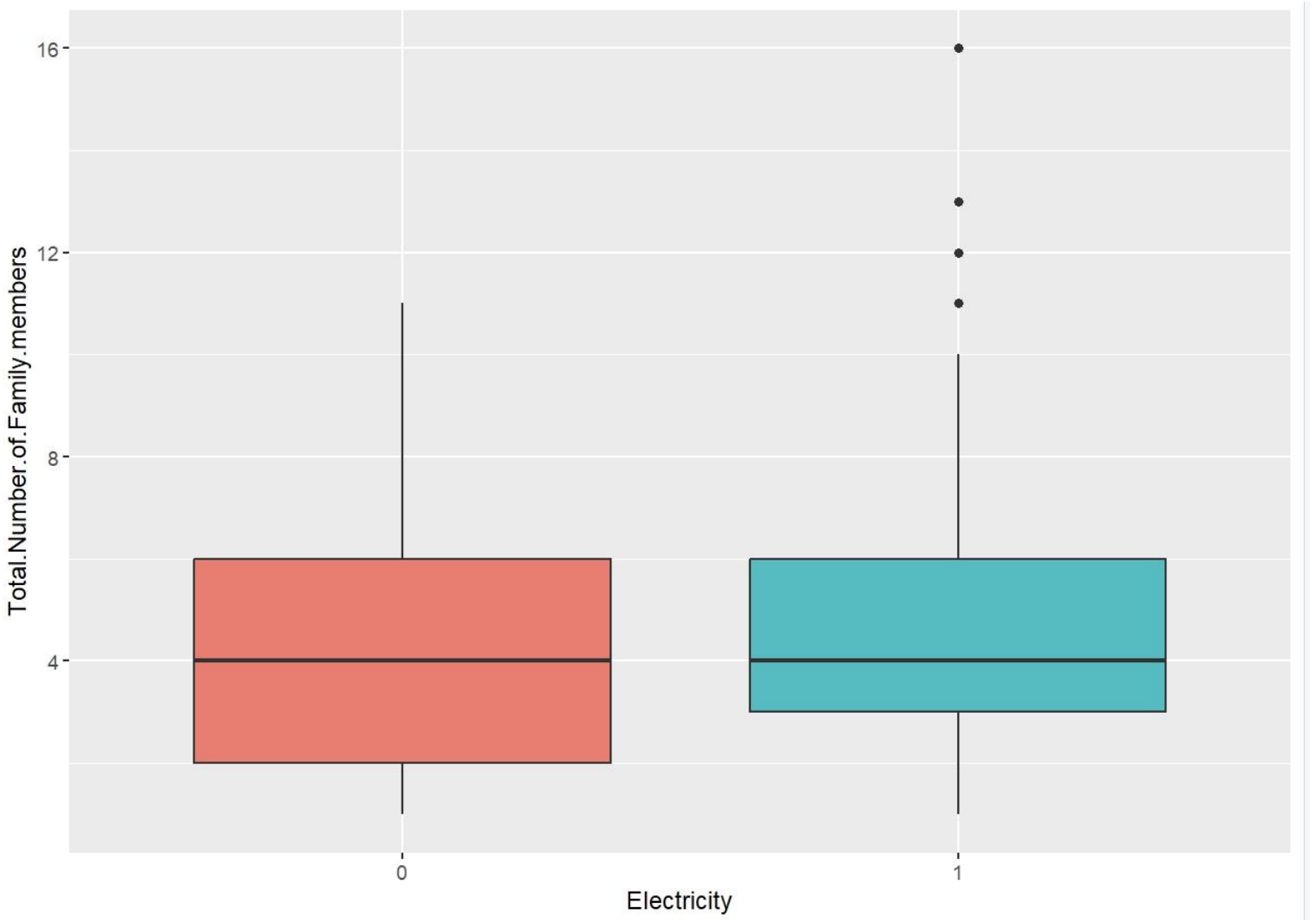


Figure 5: Boxplot of Total Number of Family Members by Electricity

Explanatory Data Analysis

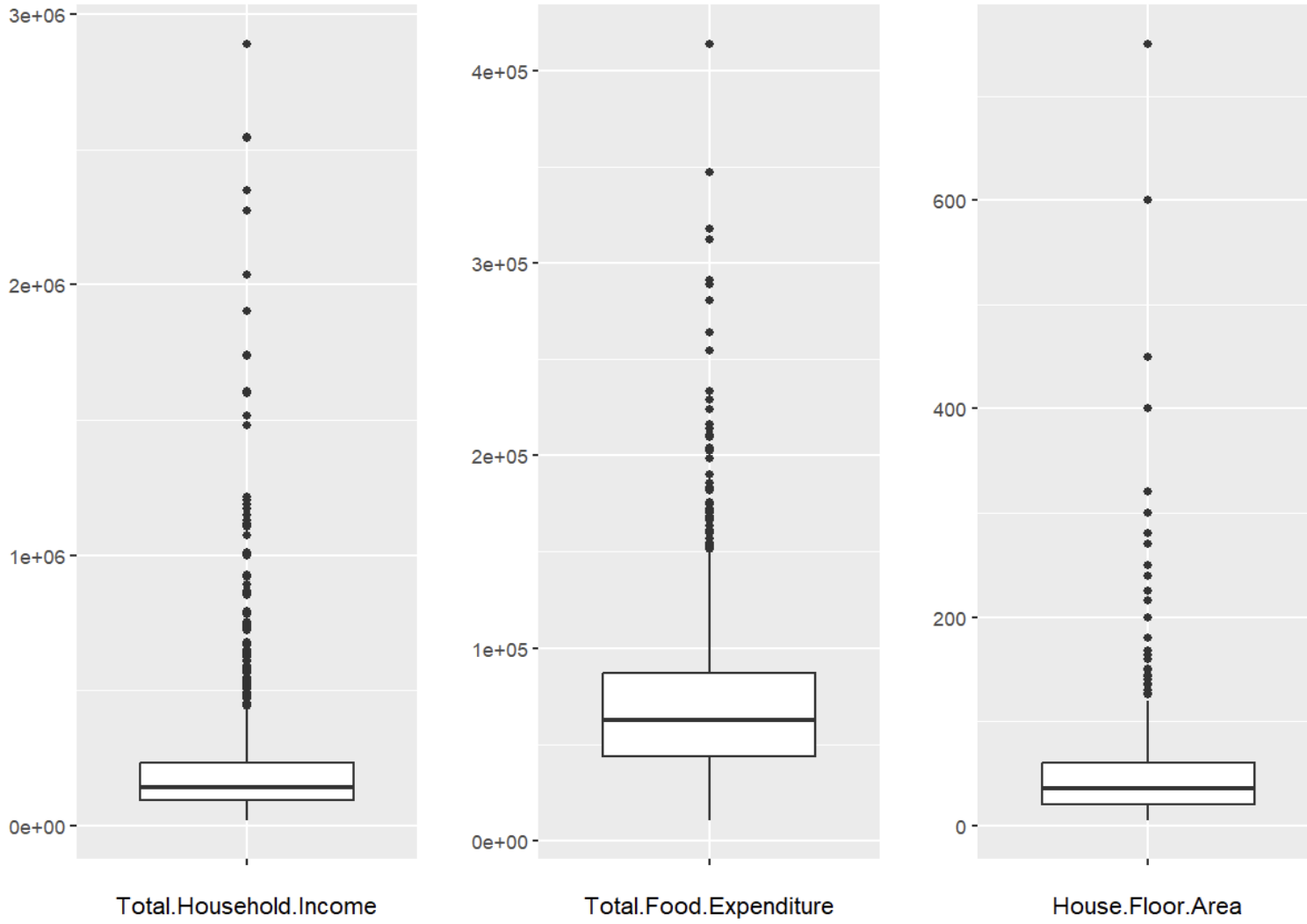


Figure 6: Boxplot of Total Number of Family Members by other continuous variables

Model Selection

Model1		Model2	
AIC 4931.87	P value	AIC 4929.96	
Continuous Variables		**Continuous Variables**	
Household.income	<0.05	Household.income	<0.05
Food Expenditure	<0.05	Food Expenditure	<0.05
Household.head.age	<0.05	Household.head.age	<0.05
Bedroom numbers	0.16	Bedroom numbers	0.16
House.age	<0.05	House.age	<0.05
Floor area	0.77		
Categorical variables		**Categorical variables**	
Household type_single	<0.05	Household type_single	<0.05
Household type_Two or more	<0.05	Household type_Two or more	<0.05
Electricity	0.20	Electricity	0.20
Household_sex	<0.05	Household_sex	<0.05

Table 1: Summary of Model 1 and Model 2

Model Selection

Model3		Model4	
AIC 4929.64	P value	AIC 4930.92	
Continuous Variables		**Continuous Variables**	
Household.income	<0.05	Household.income	<0.05
Food Expenditure	<0.05	Food Expenditure	<0.05
Household.head.age	<0.05	Household.head.age	<0.05
Bedroom numbers	0.16		
House.age	<0.05	House.age	<0.05
Categorical variables		**Categorical variables**	
Household type_single	<0.05	Household type_single	<0.05
Household type_Two or more	<0.05	Household type_Two or more	<0.05
Household_sex	<0.05	Household_sex	<0.05

Table 2: Summary of Model 3 and Model 4

Model Selection

• AIC value

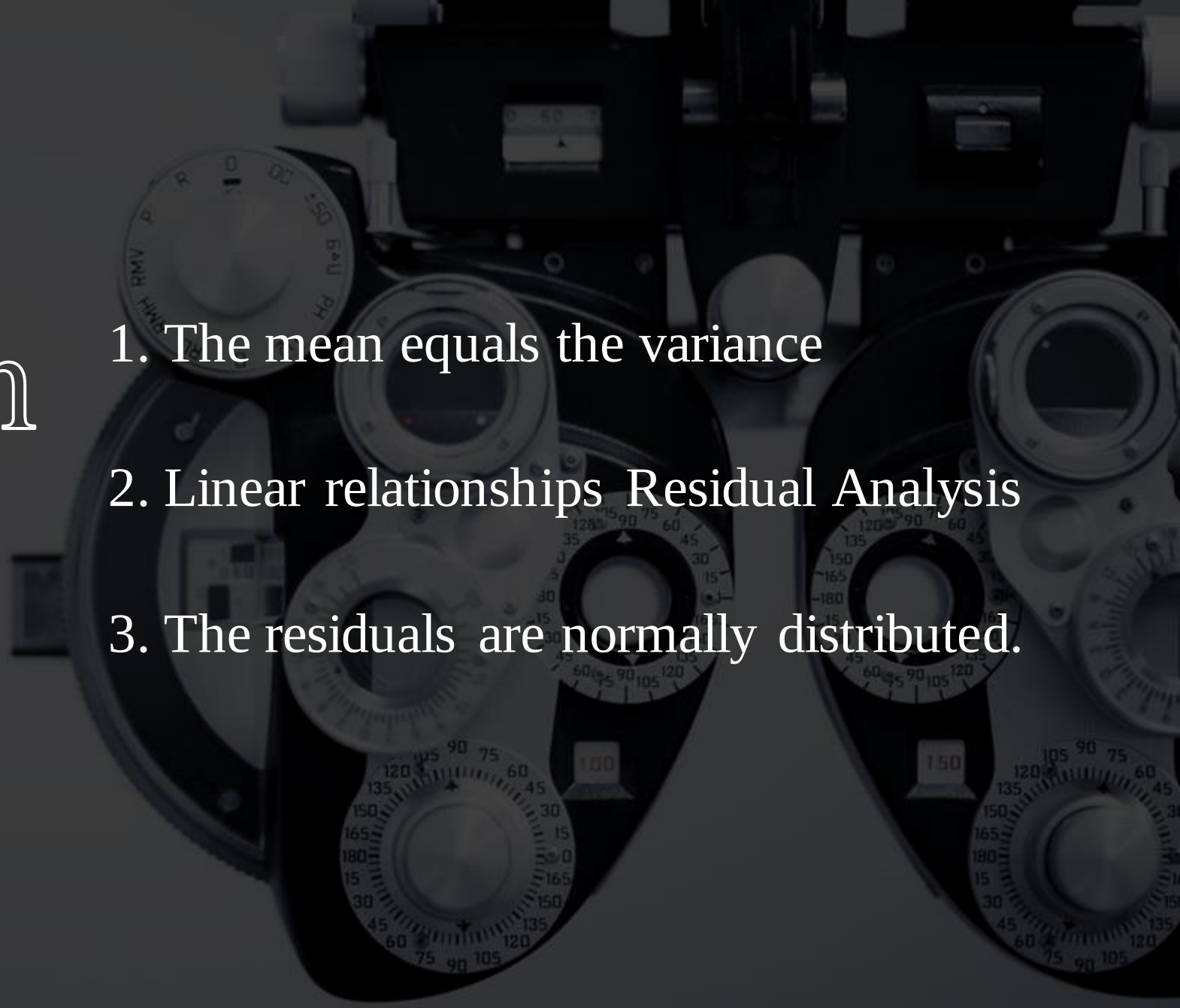
The AIC value of Model 3 is smaller.

• Cross validation

Variable: Number.of.bedrooms

residual_deviance_with_bed	882.7785
residual_deviance_without_bed	886.0647

Assumption check



1. The mean equals the variance

2. Linear relationships Residual Analysis

3. The residuals are normally distributed.

Assumption Check

1. The assumption that the mean equals the variance

Overdispersion:

print(dispersion_index)

[1] 0.7079346

The dispersion index is below 1.

2. Assumption of linear relationships Residual Analysis

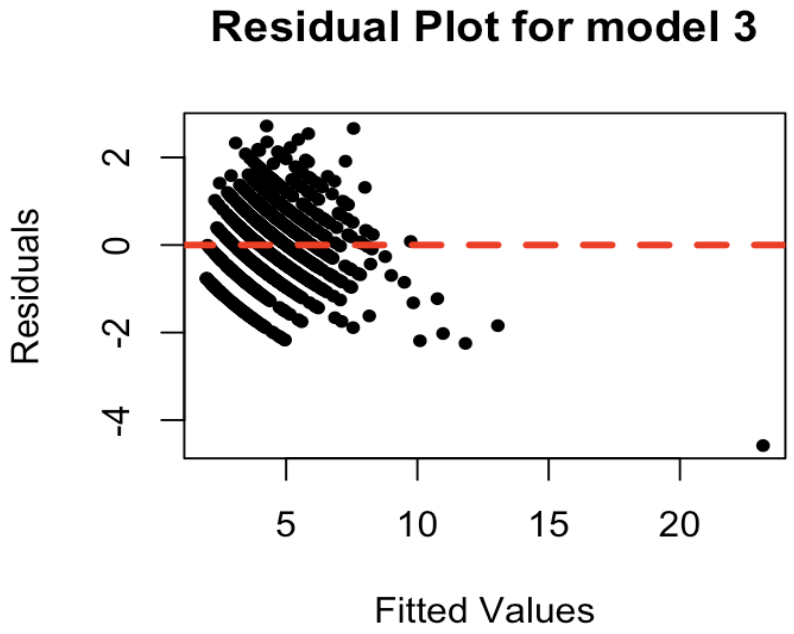


Figure 7: Scatter plot for Residuals VS Fitted

The scatter is mostly above and below the red line, indicating that the model fits these data points well.

Assumption Check

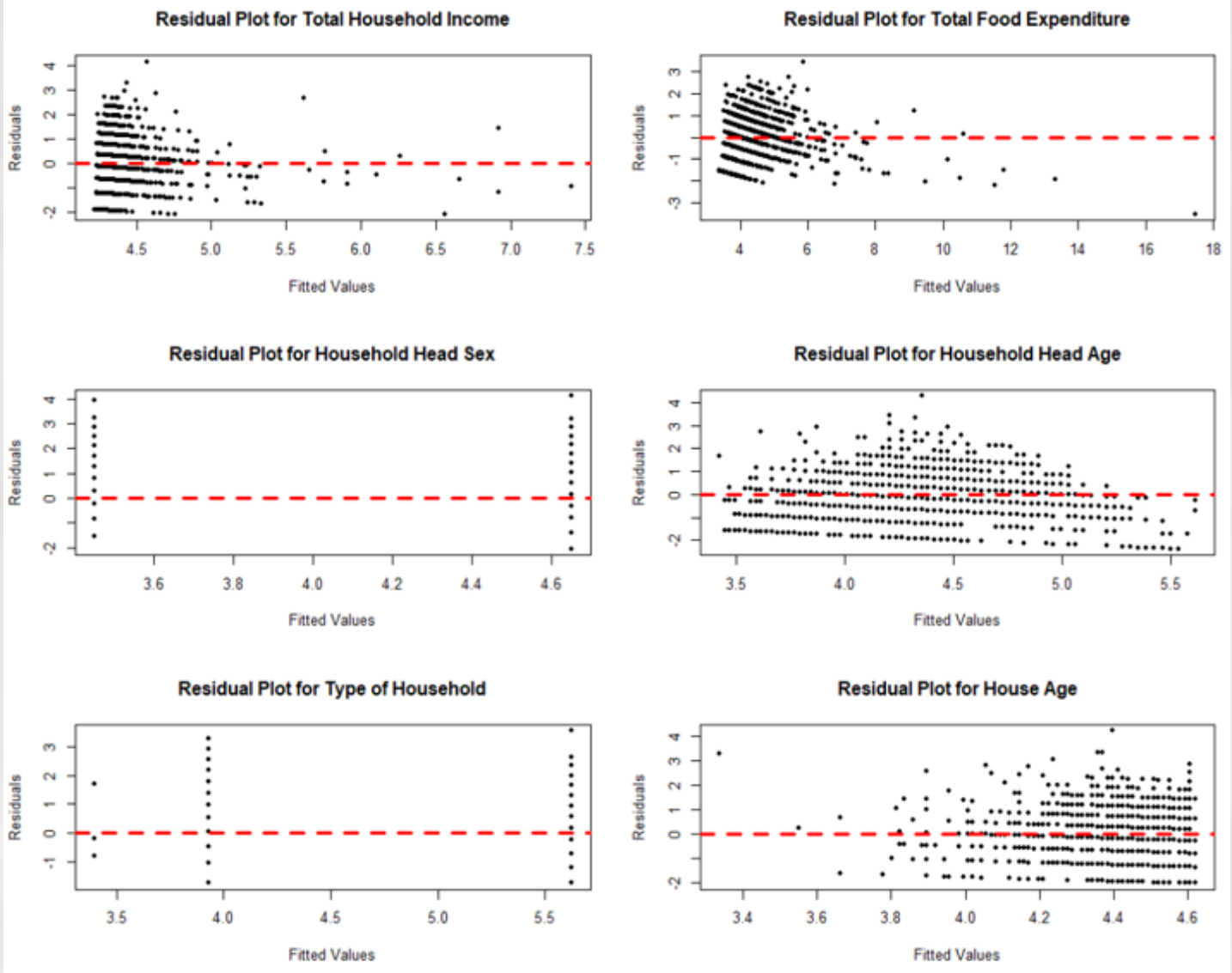


Figure 8: Residuals Plot for different variables.

For Total household income and Total food expenditure, the lower the income or food expenditure, the better the predictions of the model.

For household head age and household age, there may be heteroskedasticity.

Assumption Check

3. The Q-Q Plot

The residuals are normally distributed.

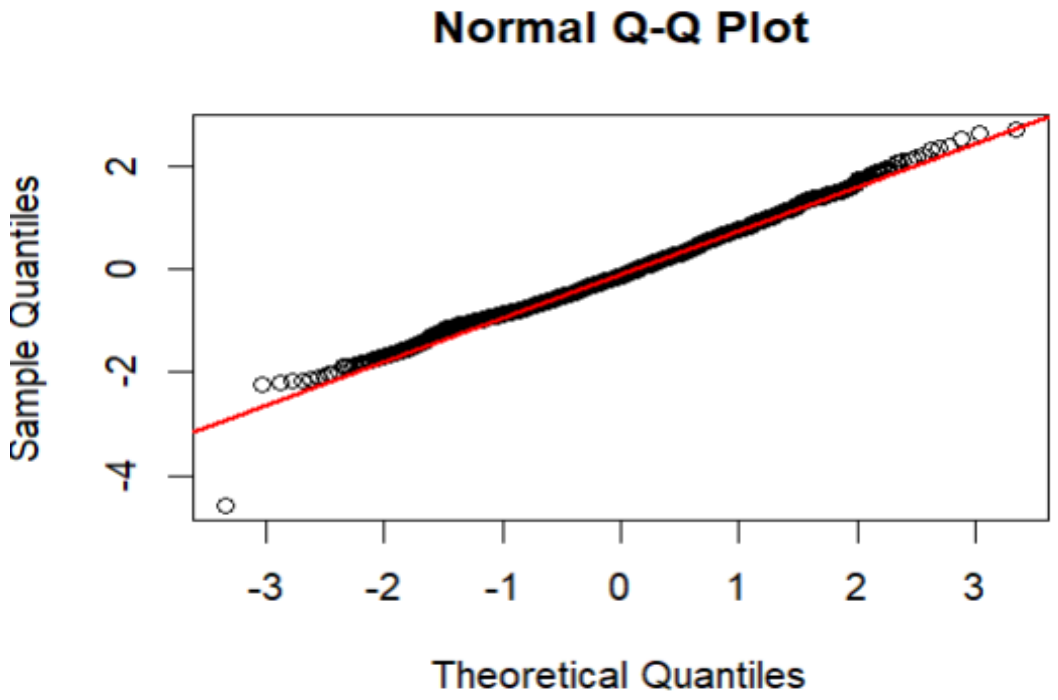


Figure 9: Q-Q Plot of Residuals

The mean equals the variance, consistent with the assumptions of Poisson regression.

Conclusion

The overall fit of the model is satisfactory for the research problem. However, there are some noteworthy problems. The correlations between the variables are low and the removal of these variables has a negligible effect on the model fit. Therefore, further research is needed to gain deeper insights.

Further Work

1. The addition of extra variables can uncover potential relationships between different variables, broadening the dataset for feature selection and enhancing the performance of the model.

2. Collect data from different time periods and incorporate time variables. By dividing the dataset into different time periods, such as three years, and applying time-series methods can improve the accuracy of predictions.