Household Insights: Income, Expenditure & Housing Analysis

GROUP：2

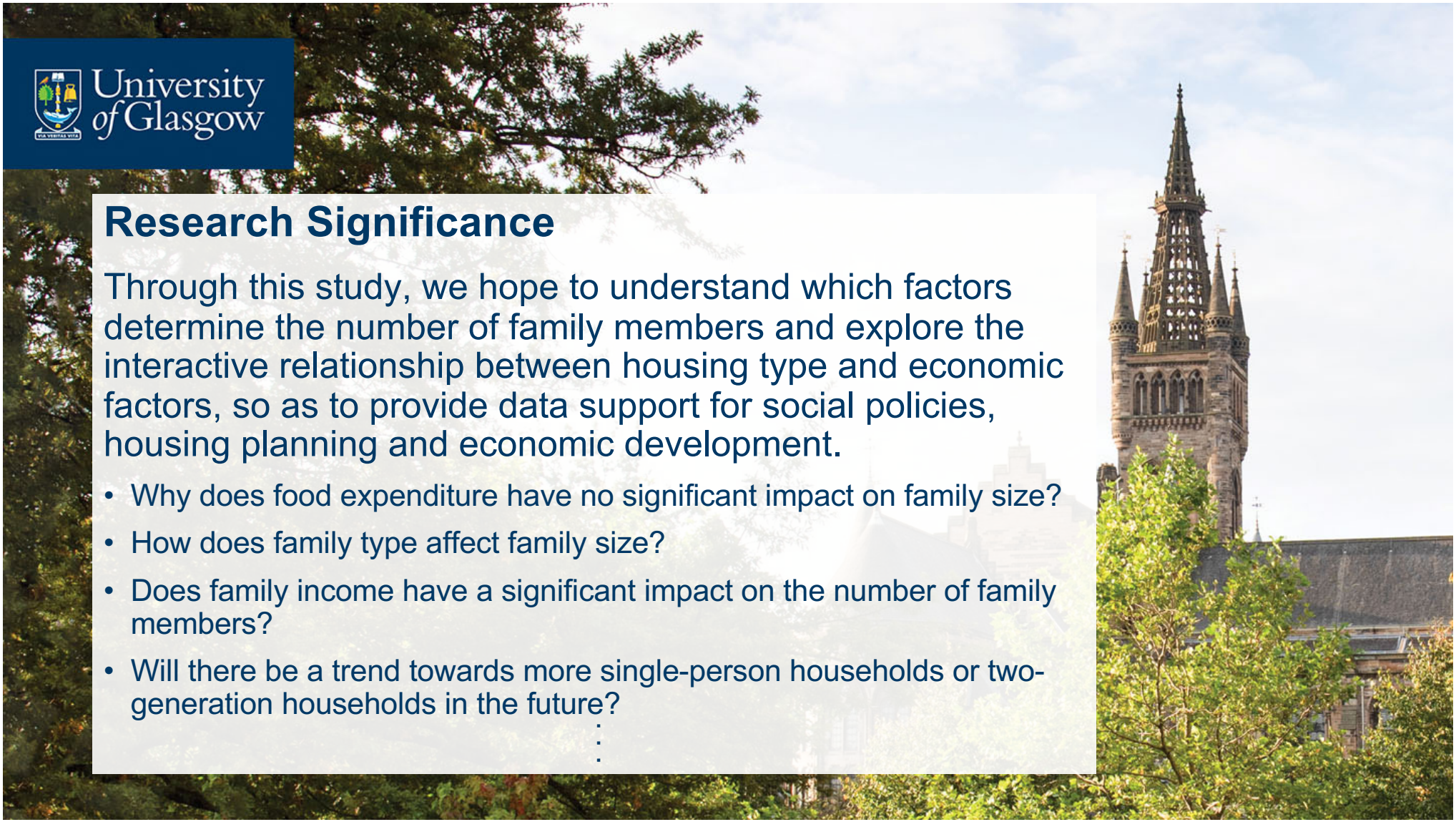
· Jiale Wang · Jinbo LIN · Xinyao Fu · Jie XIU · Aravindan Thombrakud Saju

School of Mathematics and Statistics
University of Glasgow
Correspondence: 2980789W@student.gla.ac.uk

Friday,28 March 2025

Introduction and Research Background

- This study aims to explore the relationship between household income, consumption expenditure and housing conditions, and analyze how different economic factors affect family size and living patterns. Through data analysis, we studied the impact of food expenditure, age of household head and family type on **the number of family members**, and used **Poisson Regression , Negative Binomial Regression** and **Gamma Regression** to model and reveal the connection between key variables.
- The economic status of a family is closely related to housing conditions. Income level determines consumption capacity, while housing conditions affect the living patterns of family members. In socioeconomic development, different types of families present different characteristics in consumption behavior, housing choice, and member structure. For example, **single families** and **two or more non-relatives families** have fewer members, while **extended families** have more members. In addition, whether **food expenditure** is affected by family size and whether **the age of the head of the household** affects the number of family members are still issues worthy of further study.



Research Significance

Through this study, we hope to understand which factors determine the number of family members and explore the interactive relationship between housing type and economic factors, so as to provide data support for social policies, housing planning and economic development.

- Why does food expenditure have no significant impact on family size?
- How does family type affect family size?
- Does family income have a significant impact on the number of family members?
- Will there be a trend towards more single-person households or two-generation households in the future?

⋮

Data Exploration

The total number of family members follows a **right-skewed** distribution, with most families having 4 to 6 members. Influenced by food expenditure, household head age, and family type, it was modeled using Poisson Regression to identify key factors.

- The dependent variable (number of family members) is suitable for **Poisson regression** modeling. If there is **overdispersion**, **negative binomial regression** can be used.

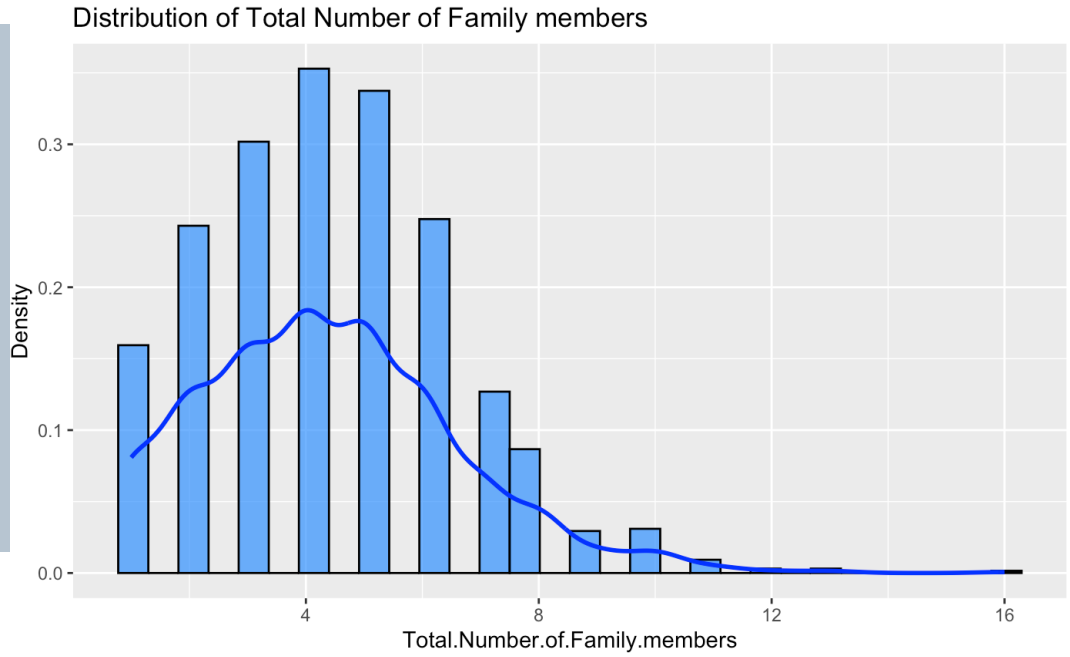


Figure1

- The research results can be used to optimize housing policies, formulate food subsidy policies, and analyze the impact of family size on consumption patterns.

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.

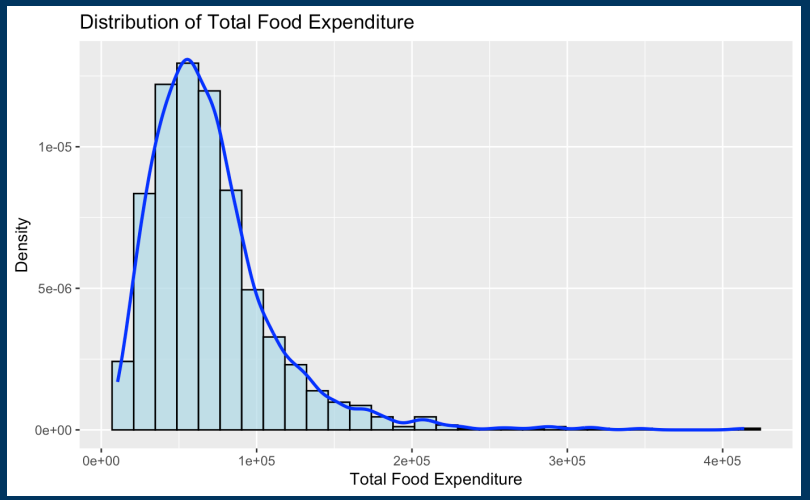


Figure2

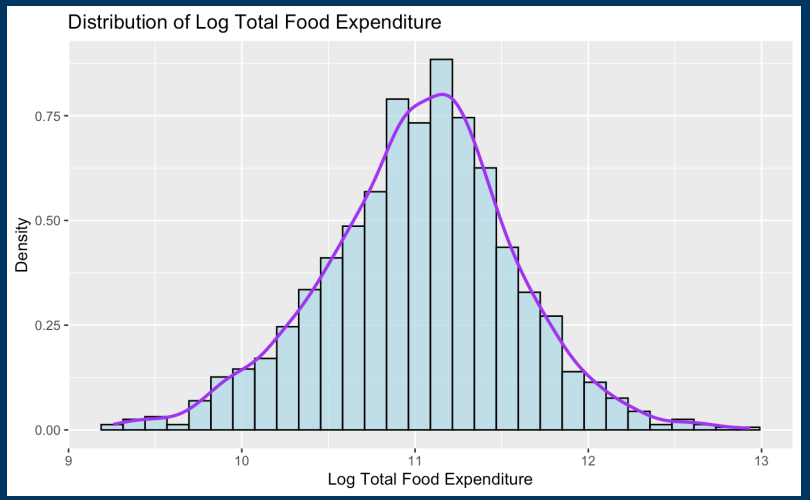


Figure3

- The distribution of **Total Food Expenditure** is **highly right-skewed**, indicating that most households have low food expenditures while a few have extremely high values. However, after applying a **logarithm transformation**, the data becomes more **normally distributed**, reducing skewness and improving suitability for statistical modeling and regression analysis.

Data Exploration

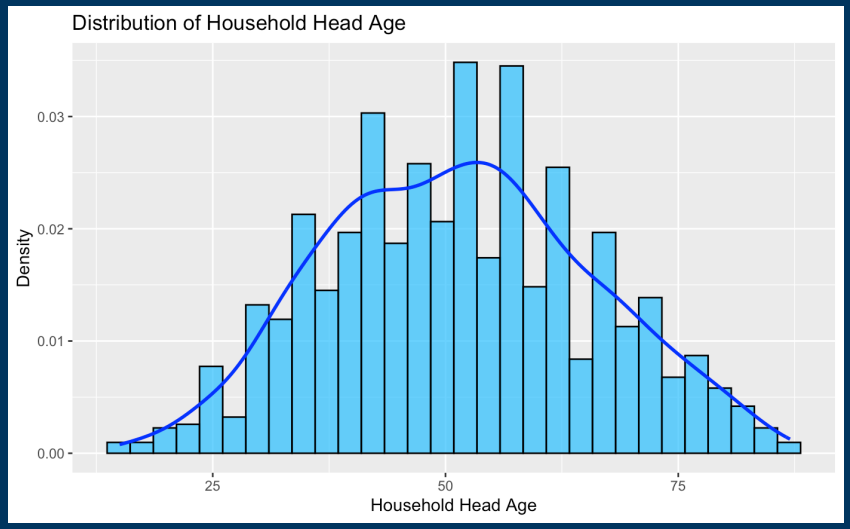


Figure4

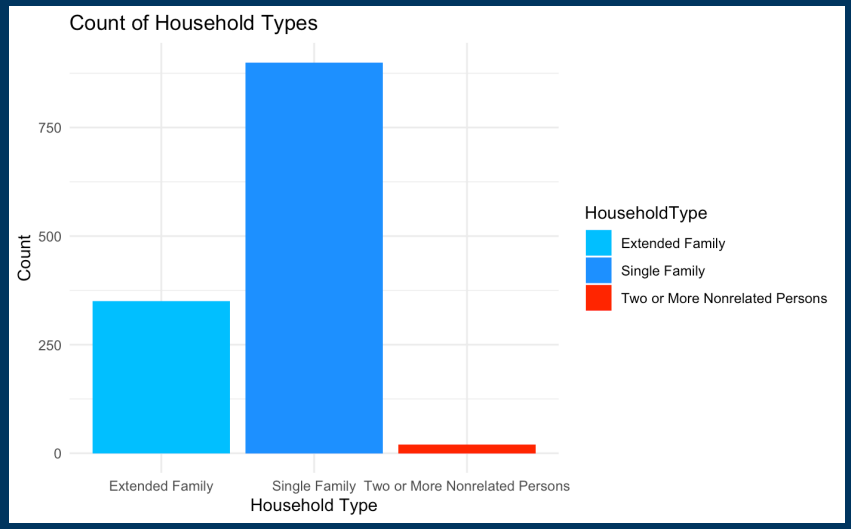


Figure5

The distribution of household head age is approximately normal, slightly right-skewed, with most household heads concentrated between 30 and 70 years old.

Single households are the most common, extended families have grown to some extent but one type is missing, while two or more non-relatives families are rare, likely due to economic, cultural, or housing policy factors, and their small sample size may impact GLM model stability.

Data Exploration

- Household size varies significantly across household types, with **extended families** having the largest number of members due to multi-generational living, **single families** showing moderate size but diverse structures, and **non-relative households** being the smallest, as they are typically shared by friends or tenants; however, the presence of outliers suggests that the IQR method may be needed for data refinement.

- 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 consider removing.

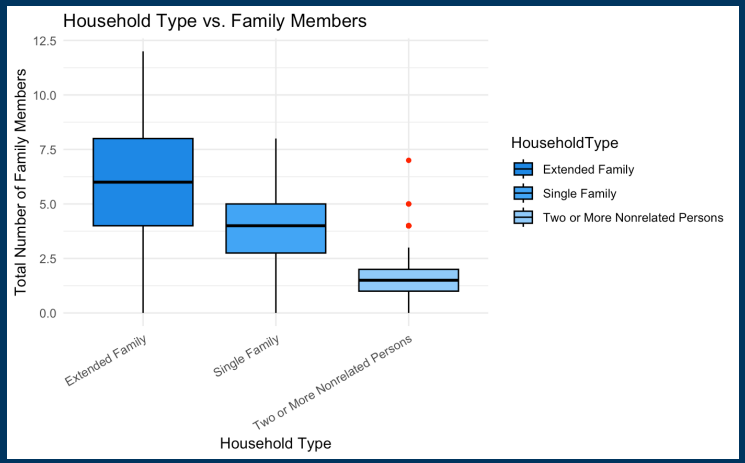


Figure6

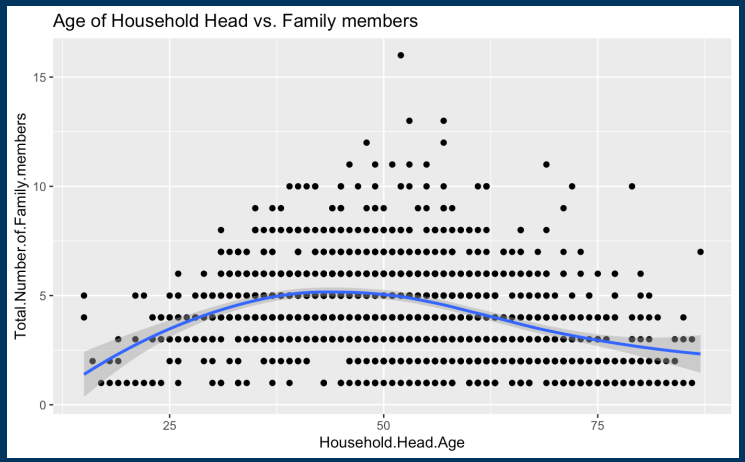


Figure7

Data Exploration Summary

The type of family determines the number of family members, with single-person families accounting for the highest proportion.

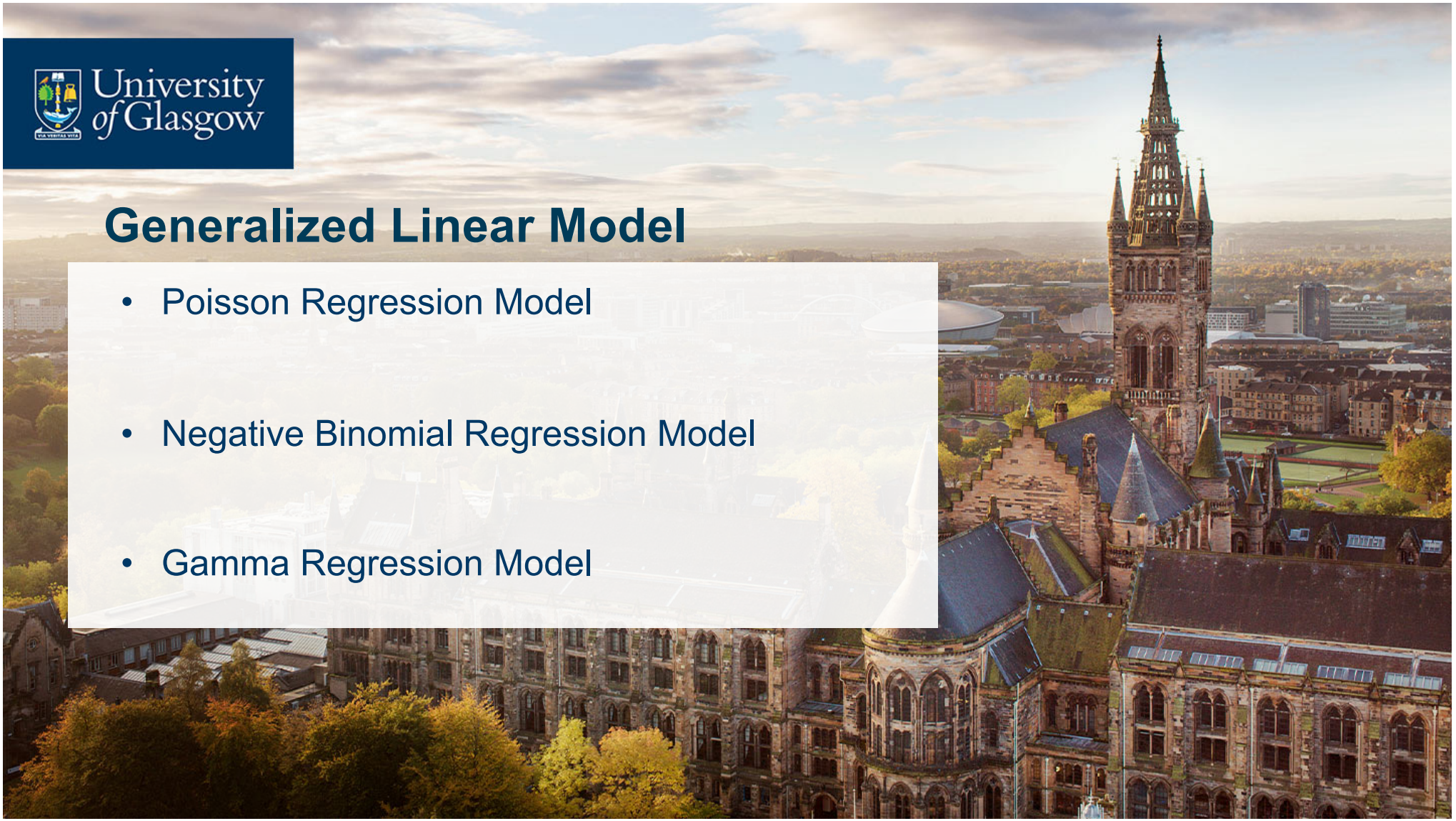
- Extended families have the largest number of members, often including multiple generations or additional relatives.
- Single families account for the highest proportion, but the number of members varies widely, reflecting the diversity of family structures.
- Two or More Nonrelated Persons are the smallest and rarest, which may be related to housing policies and economic conditions.

There is a nonlinear relationship between the age of the household head and the size of the household.

- The number of family members gradually increases when the age of the household head is relatively low, and reaches a peak after reaching middle age (about 50 years old).
- After the age of 50, the number of family members decreases, which may be due to the fact that children live independently after adulthood, family splitting or other social factors.

Housing type has a significant impact on family structure and number of members.

- ANOVA statistical analysis shows that there are significant differences in the number of family members in different family types.
- Single-person households are the largest, followed by extended families, and non-relative co-living households are the least, indicating that housing structure is affected by economic, social culture and policies.
- Housing policies, rental markets and social trends may be key factors affecting changes in family structure.



Generalized Linear Model

- Poisson Regression Model
- Negative Binomial Regression Model
- Gamma Regression Model

Poisson Regression Model

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 \times \text{Total.Food.Expenditure} + \beta_2 \times \text{Household.Head.Age} + \beta_3 \times \text{Type.of.Household} \quad \cdots \quad (1)$$

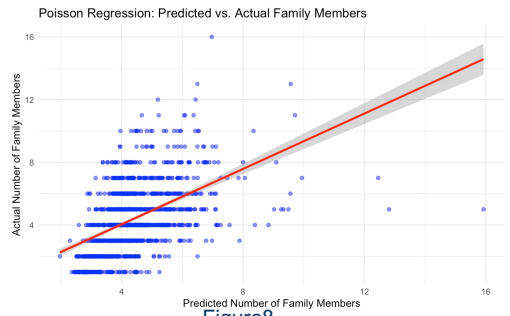


Figure8

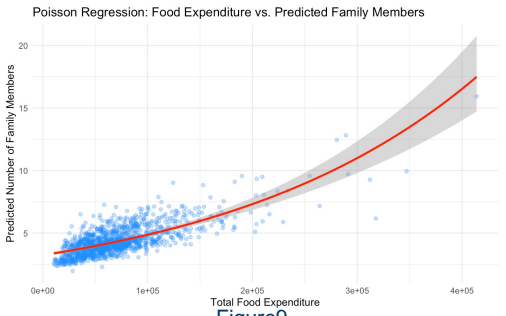


Figure9

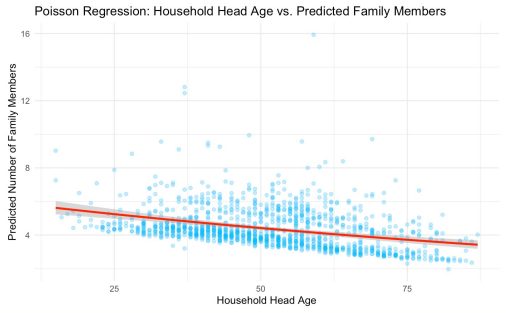


Figure10

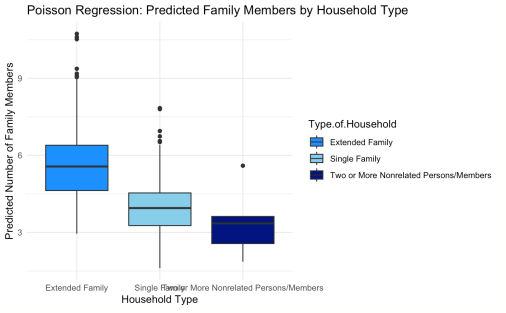


Figure11

- Conclusion: Food expenditure is an important **positive** variable affecting the number of family members, while the age of the household head has a **negative** impact. The Poisson regression model predicts that extended families generally have more members than single families, with a wider distribution and more outliers.

Poisson Regression Model

Table1:Possion Regression Coefficients

AIC: 5009.7

Variable	Estimate	Std.Error	Z-Value	Pr(> | Z |)
Intercept	1.578e+00	5.627e-02	28.041	<2e-16 ***
Total.Food.Expenditure	3.292e-06	2.809e-07	11.720	<2e-16 ***
Household.Head.Age	-8.930e-03	1.011e-03	-8.829	<2e-16 ***
Type.of.Household Extended Family	3.484e-01	3.004e-02	11.595	<2e-16 ***

- The Poisson regression model found that（P-Value：<2e-16）food expenditure and family type (extended family) have a significant positive impact on the total number of family members, while the age of the head of household has a significant negative impact, indicating that as the age of the head of household increases, the number of family members may decrease, while higher food expenditure and extended family structure often correspond to more family members. The residual deviation (987.81) and AIC (5009.7) of the model show that the fitting effect is good, but the overdispersion problem needs to be further tested, so the next step will consider Negative Binomial Regression for comparative analysis.

Negative Binomial Regression Model

$$\log(\mu_i) = \beta_0 + \beta_1 \text{Total.Food.Expenditure}_i + \beta_2 \text{Household.Head.Age}_i + \beta_3 \text{Type.of.Household}_i \quad \cdots (2)$$

$$Y_i \sim \text{Negative Binomial}(\mu_i, \theta) \quad \cdots (3)$$

$$P(Y_i = k) = \frac{\Gamma(k+\theta)}{k!\Gamma(\theta)} \left(\frac{\mu_i}{\mu_i + \theta}\right)^k \left(\frac{\theta}{\mu_i + \theta}\right)^\theta \quad \cdots (4)$$

Explanation:

- The **first equation** represents the log-link function used in Negative Binomial regression.
- The **second equation** states that the dependent variable Y_i follows a Negative Binomial distribution with mean μ_i and dispersion parameter θ.
- The **third equation** is the probability mass function (PMF) of the Negative Binomial distribution, where:
  - Γ(·) is the gamma function.
  - μ_i is the mean number of family members predicted by the model.
  - θ is the dispersion parameter.

Negative Binomial Regression Model

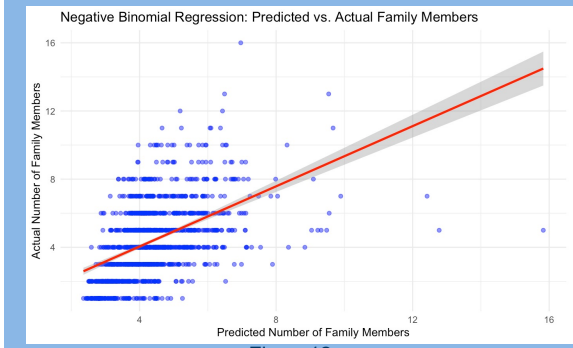


Figure12

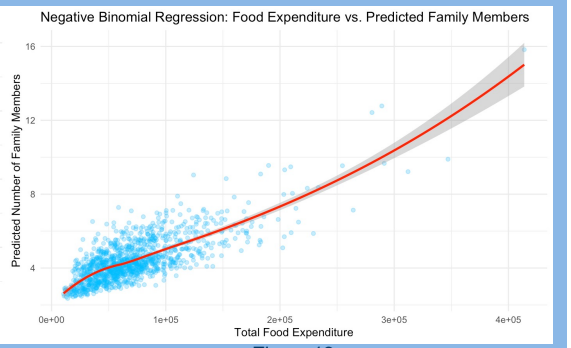


Figure13

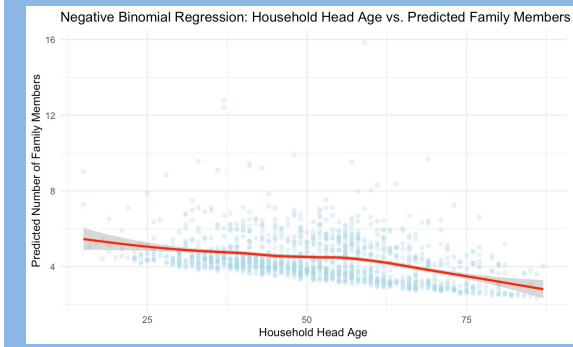


Figure14

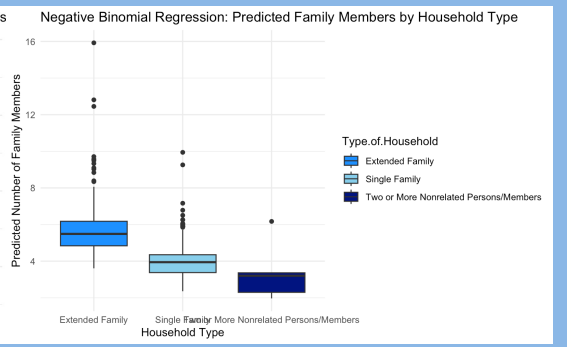


Figure15

- The Negative Binomial Regression model shows a strong correlation between predicted and actual family members, with most points aligning along the red regression line.
- There is a positive relationship between total food expenditure and predicted family size, indicating that households with higher food expenditures tend to have more family members.
- The model suggests a slight negative correlation, where households with older.
- The Negative Binomial Regression model predicts that extended families generally have more members than single-family households, with a higher median and greater variability in household size.

Negative Binomial Regression Model

Table2:Negative binomial Regression Coefficients

AIC:5029.6

Variable	Estimate	Std.Error	Z-Value	Pr(> | Z |)
Intercept	1.922e+00	6.701e-02	28.674	<2e-16 ***
Total.Food.Expenditure	3.310e-06	2.775e-07	11.927	<2e-16 ***
Household.Head.Age	-8.874e-03	1.009e-03	-8.798	<2e-16 ***
Type.of.HouseholdSingle Family	-3.475e-01	3.002e-02	-11.578	<2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-6.119e-01	2.441e-01	-2.507	**0.0122** *

- The negative binomial regression confirms that total food expenditure positively affects （P-Value：<2e-16 ）family size, while household head age has a negative impact .The effect for nonrelated households is weaker (P-Value=0.0122).
- Single-family and nonrelated households tend to have fewer members than extended families, but the model shows minimal improvement over the Poisson regression due to low dispersion.

Gamma Regression Model

$$\log(E(\text{T.N.of.F.m})) = \beta_0 + \beta_1 \times \text{T.F.E} + \beta_2 \times \text{Single F} + \beta_3 \times \text{T or MNPersons/Members} + \beta_4 \times \text{Household.H.Age} \quad \ldots (5)$$

$$E(\text{T.N.of.F.m}) = \exp(\beta_0 + \beta_1 \times \text{T.F.E} + \beta_2 \times \text{Single F} + \beta_3 \times \text{Tor MNPersons/Members} + \beta_4 \times \text{Household.H.Age}) \quad \ldots (6)$$

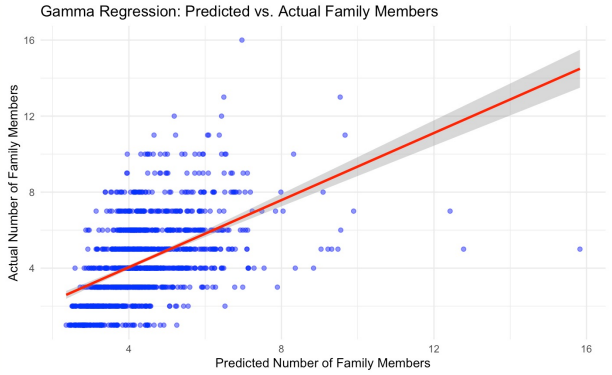


Figure16

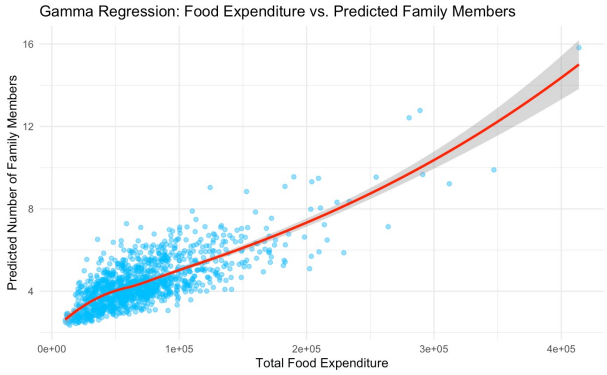


Figure17

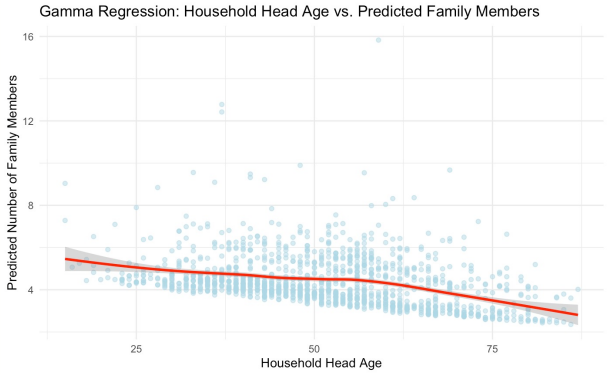


Figure18

- These graphs illustrate the Gamma regression results, showing a strong correlation between predicted and actual family members, a positive relationship between food expenditure and family size, and a negative association between household head age and family size.

Gamma Regression Model

Table3:Gamma Regression Coefficients

AIC: 4966.1

Variable	Estimate	Std.Error	t-Value	Pr(> | t |)
Intercept	1.957e+00	6.483e-02	30.183	<2e-16 ***
Total.Food.Expenditure	3.310e-06	3.083e-07	15.211	<2e-16 ***
Type.of.HouseholdSingle Family	-3.706e-01	2.919e-02	-12.697	<2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-6.449e-01	1.972e-01	-3.271	0.0011 **
Household.Head.Age	-1.129e-02	9.005e-04	-12.542	<2e-16 ***

- The Gamma regression model shows that higher food expenditure ($p < 2e\text{-}16$) is associated with larger family size, while single-family ($p < 2e\text{-}16$) and nonrelated person households ($p = 0.0011$) tend to have fewer members.
- Household head age ($p < 2e\text{-}16$) is negatively correlated with family size. The model fits well with a residual deviance of 258.20.

Add the Previously Log-transformed Data

$$\log(\mathbb{E}[\text{TN of FM}]) = \beta_0 + \beta_1 \log(\text{Total Food Expenditure}) + \beta_2 \text{Household Head Age} + \beta_3 \text{Type of Household} \quad \cdots \quad (7)$$

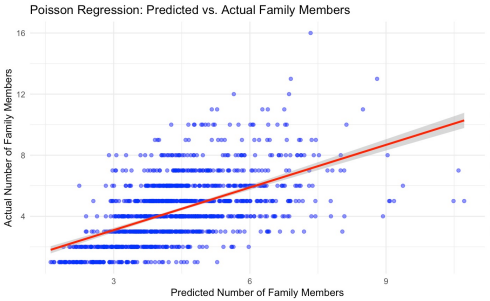


Figure19

Figure20

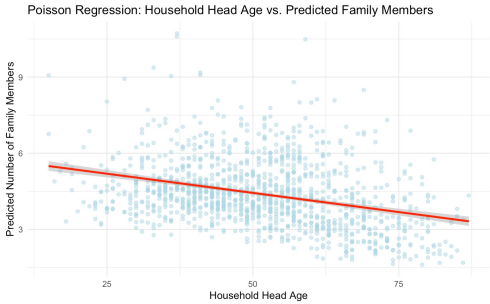


Figure21

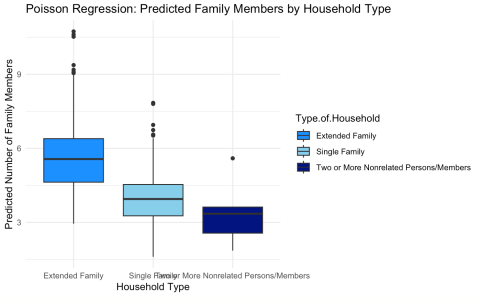


Figure22

- Taking the logarithm of food expenditure improves the model by stabilizing variance, reducing skewness, and making relationships more linear, which enhances interpretability and model performance. Additionally, the log transformation mitigates the influence of extreme values, leading to more reliable coefficient estimates.

Add the Previously Log-transformed Data

Table4:Add the Previously Log-transformed Data Coefficients

AIC: 4915

Variable	Estimate	Std.Error	Z-Value	Pr(> | Z |)
Intercept	-2.576372	0.305562	-8.432	<2e-16 ***
Total.Food.Expenditure	0.390907	0.026506	14.748	<2e-16 ***
Household.Head.Age	-0.007380	0.001036	-7.122	1.06e-12 ***
Type.of.HouseholdSingle Family	0.306279	0.030290	10.112	<2e-16 ***

- **Log Transformation Improves Model Fit** – Using the logarithm of total food expenditure results in a lower residual deviance (893.15) compared to the original model, indicating better model performance.
- **Significant Predictors** – Log-transformed food expenditure, household head age, and household type (extended family) are all highly significant ($p < 0.001$), suggesting strong relationships with the number of family members.
- **Positive and Negative Effects** – Food expenditure and extended family households positively impact family size, while household head age has a negative effect, meaning older household heads tend to have fewer family members.

Model Selection and Test

$$\text{Model 1:} \quad \text{TN of FM} = \beta_0 + \beta_1 \log(\text{TFE}) + \beta_2 \text{Household Head Age} + \beta_3 \text{Type of Household} + \varepsilon \quad \ldots (8)$$

$$\text{Model 2:} \quad \text{TN of FM} = \alpha_0 + \alpha_1 \text{TFE} + \alpha_2 \text{Household Head Age} + \alpha_3 \text{Type of Household} + \varepsilon \quad \ldots (9)$$

i. Likelihood Ratio Test (LRT)

- **Log-Likelihood Values:** Model 1: -2461.6 Model 2: -2508.8
- **Chi-Square Statistic:** 94.388
- **Degrees of Freedom (Df):** 1
- **P-value:** < 2.2e-16 (Highly significant)

ii. Since the p-value is very small **(< 0.001)**, Model 1 (with log-transformed Total Food Expenditure) is **significantly better** than Model 2. This suggests that **log-transformation** improves model fit, making it preferable.

Model Selection and Test

Table5:Negative Binomial Model Coefficients (Incidence Rate Ratios - IRR)

Predictor	Estimate	IRR (exp(β))
Intercept	6.8317	exp(6.8317)=922.83
Total.Food.Expenditure	1.0000033	exp(1.0000033)=2.72
Household.Head.Age	0.9912	exp(0.9912)=2.69
Type.of.Household (Single Family)	0.7064	exp(0.7064)=2.03
Type.of.Household (Two or More Nonrelated Persons/Members)	0.5423	exp(0.5423)=1.72

- **Total.Food.Expenditure** has a greater impact, with each unit increase in the number of family members **increasing by 2.72 times** .
- The impact of **Household.Head.Age** is similar, with the number of family members **increasing by 2.69 times for every 1 year older** .
- **The family type** has a significant impact, with the number of family members in single-parent families and families with non-relative members **increasing by 2.03 and 1.72 times**, respectively, compared with the baseline group.

Model Selection and Test

- **Overdispersion Check**

Overdispersion = 0.718901. Since the overdispersion value is **less than 1**, there is no significant overdispersion, meaning the negative binomial model is a good choice.

- **Multicollinearity Check**

Table6:Multicollinearity Check (Generalized Variance Inflation Factor - GVIF)

Predictor	GVIF	DF	Adjusted GVIF
log_Total.Food.Expenditure	1.084	1	1.041
Household.Head.Age	1.095	1	1.046
Type.of.Household	1.148	2	1.035

i. **VIF < 5** for all variables, meaning **no serious multicollinearity**.
ii. Adjusted GVIF confirms that predictors are stable.

Model Selection and Test

Conclusion

This plot is used to check for overdispersion and shows that the variance increases as the square of the mean increases, indicating that the data may violate the Poisson assumption and is suitable for Negative Binomial regression.

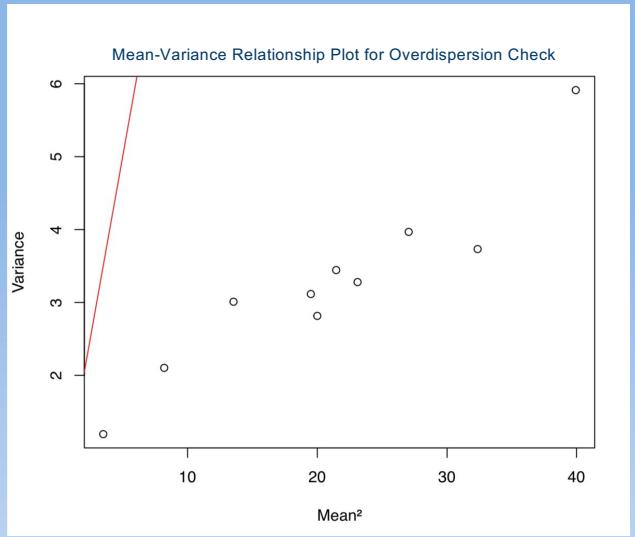


Figure23

- 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.

Model Selection and Test

Conclusion

- 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.

Residuals vs Fitted Plot for Negative Binomial Model

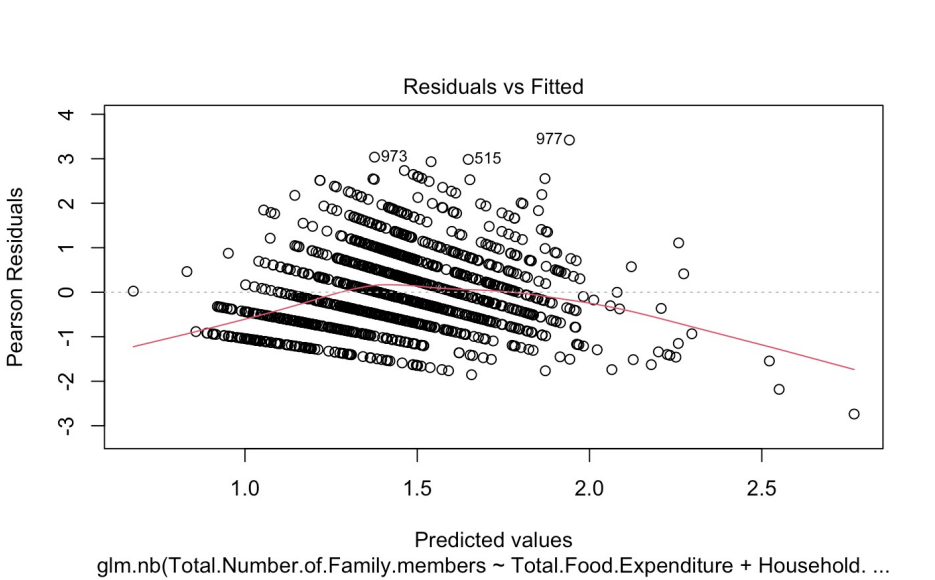


Figure24

- Evaluating the fit of the Negative Binomial regression model.
- Examining the relationship of Pearson residuals to predicted values.
- Determine whether the model has systematic bias, heteroskedasticity, or underfit.

- The Negative Binomial model performs well in the low prediction value part, but there is a large fluctuation in the high prediction value part. There is a certain degree of heteroscedasticity, and we will choose a suitable method for further optimization in the future.

Future Research Directions

- **Improve model fitting methods**

Use zero-inflated negative binomial model (ZINB) or generalized additive model (GAM) to optimize the fit of high prediction value areas and reduce the systematic bias of residuals.

- **Explore additional explanatory variables**

Introduce additional variables that may affect the number of family members (such as occupation type, geographical location, etc.) to improve the explanatory power of the model.

- **Deal with data heteroskedasticity**

Use log transformation or standardized variables to reduce residual fluctuations in high prediction value areas and improve model stability.

- **Test model robustness and external validation**

Test the model on different data sets and use cross-validation or external data to evaluate the model generalization ability to ensure its applicability in different samples.

⋮


University of Glasgow
Closing slide – If you want to know more about our research project, you can email our corresponding author , thank you!
INCOME, EXPENDITURE & HOUSING
Group_02@UofGlasgow