Modeling Family Size: Literacy and Marriage Age in Rural Portugal

A Generalized Linear Model Approach

Shaotong (Max) Li

November 17, 2024

1 Introduction

Understanding the factors influencing family size is of paramount importance in demographic research, as family size has significant implications for economic development, resource allocation, and population growth. In rural areas, where socioeconomic conditions often shape family dynamics, investigating the determinants of family size is especially relevant. Portugal in the late 20th century, characterized by pronounced economic disparities and relatively low levels of education, provides an intriguing context for studying these dynamics. Rural Portugal, in particular, exhibited distinct family patterns shaped by varying levels of literacy and cultural norms surrounding marriage, making it an ideal setting to explore these relationships.

Education is widely recognized as a key determinant of family size, operating through both direct and indirect pathways. Studies have demonstrated that higher levels of education are strongly correlated with smaller family sizes, as education expands access to information about family planning, increases participation in the labor force, and delays the timing of marriage. The impact of education on family size is nuanced: while the direct effects are more evident at lower levels of education, indirect effects, such as those mediated through labor force participation and age at marriage, become increasingly significant at higher education levels, particularly for women. Furthermore, these effects are often asymmetrical between genders, with the wife's education playing a more dominant role in shaping family outcomes than the husband's education.

The complex interplay between literacy, age at marriage, and family size highlights the need for targeted statistical analysis to disentangle these relationships. Using data from a fertility survey conducted in rural Portugal in 1979, this study aims to model the effects of literacy and marriage age on family size through the application of generalized linear models (GLMs). This approach not only facilitates the estimation of direct effects but also enables the identification of additional variation in birth rates after accounting for known explanatory variables. The

findings from this study will contribute to the broader understanding of demographic behavior in rural settings and provide insights into the socioeconomic determinants of family size that remain relevant for contemporary policy planning.

2 Methods

2.1 Clean Data Process

In this study, we began by carefully selecting the variables most relevant to our research question: understanding how literacy and marriage age affect family size in rural Portugal. From the dataset, we identified three key variables:

children (Numerical): This variable represents the number of children in a family and serves as the basis for calculating the dependent variable, family_size.

ageMarried (Categorical): This variable captures the marriage age of individuals, categorized into meaningful intervals: 0to15, 15to18, 18to20, 20to22, 22to25, 25to30, and 30toInf. It reflects the social and demographic variation in marriage age and is included as an independent variable in the model.

literacy (Binary): A factor variable indicating whether an individual is literate (yes) or not (no). This variable is included as a second independent variable, as literacy is hypothesized to influence family planning and size.

To address our research objective, the children variable was transformed to create a new variable, family_size, defined as the total number of children in a family plus two. This transformation assumes a baseline family size of two individuals (e.g., parents) and ensures consistency in defining the dependent variable.

2.2 Generalized Linear Models

2.2.1 Normal Distribution Model

$$\text{Model 1 (Normal Regression): family\_size} = \beta_1 \cdot \text{ageMarried} + \beta_2 \cdot \text{literacy} + \beta_3 \cdot \text{monthsSinceM} + \beta_0 \tag{1}$$

where:

- ageMarried is a categorical variable representing different marriage age groups.
- literacy $\in \{\text{yes}, \text{no}\}$, indicating whether an individual is literate.
- monthsSinceM $\in \mathbb{R}^+$ represents the number of months since marriage (a continuous variable).

- β_1, β_2, and β_3 are the coefficients for the independent variables.
- β_0 is the intercept term.
- This model assumes family_size follows a normal distribution.

2.2.2 Binomial Distribution Model

Model 2 (Binomial Regression): $\log\left(\frac{P(\text{large\_family}=1)}{1-P(\text{large\_family}=1)}\right) = \beta_1 \cdot \text{ageMarried} + \beta_2 \cdot \text{literacy} + \beta_3 \cdot \text{monthsSince}$ (2)

where:

- large_family $\in \{0, 1\}$, where 1 indicates a family size greater than 5.
- ageMarried is a categorical variable representing different marriage age groups.
- literacy $\in \{\text{yes}, \text{no}\}$, indicating whether an individual is literate.
- monthsSinceM $\in \mathbb{R}^+$ represents the number of months since marriage (a continuous variable).
- β_1, β_2, and β_3 are the coefficients for the independent variables.
- β_0 is the intercept term.
- The dependent variable follows a binomial distribution.
- The logit function transforms probabilities into a linear function of the predictors.

2.2.3 Poisson Distribution Model

Model 3 (Poisson Regression): $\log(\mu) = \beta_1 \cdot \text{ageMarried} + \beta_2 \cdot \text{literacy} + \beta_3 \cdot \text{monthsSinceM} + \beta_0$ (3)

where:

- $\mu = E(\text{family\_size})$ represents the expected family size.
- ageMarried is a categorical variable representing different marriage age groups.
- literacy $\in \{\text{yes}, \text{no}\}$, indicating whether an individual is literate.
- monthsSinceM $\in \mathbb{R}^+$ represents the number of months since marriage (a continuous variable).
- β_1, β_2, and β_3 are the coefficients for the independent variables.
- β_0 is the intercept term.

- This model assumes family_size follows a Poisson distribution.
- The log-link function ensures positive expected values.

2.2.4 Gamma Distribution Model

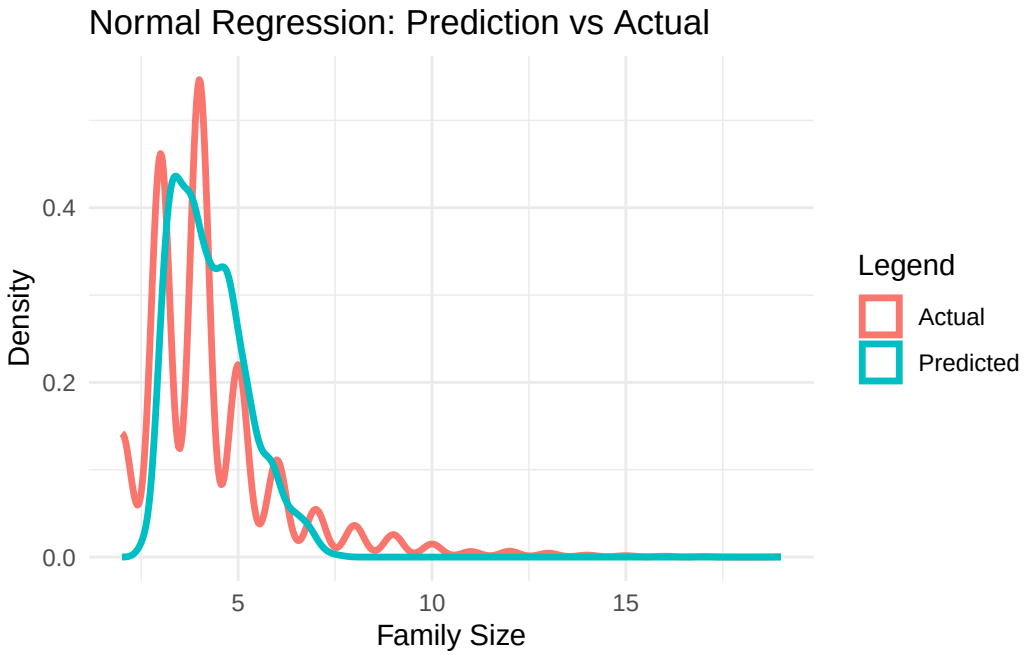
Model 4 (Gamma Regression): $\log(\mu) = \beta_1 \cdot \text{ageMarried} + \beta_2 \cdot \text{literacy} + \beta_3 \cdot \text{monthsSinceM} + \beta_0$ (4)

where:

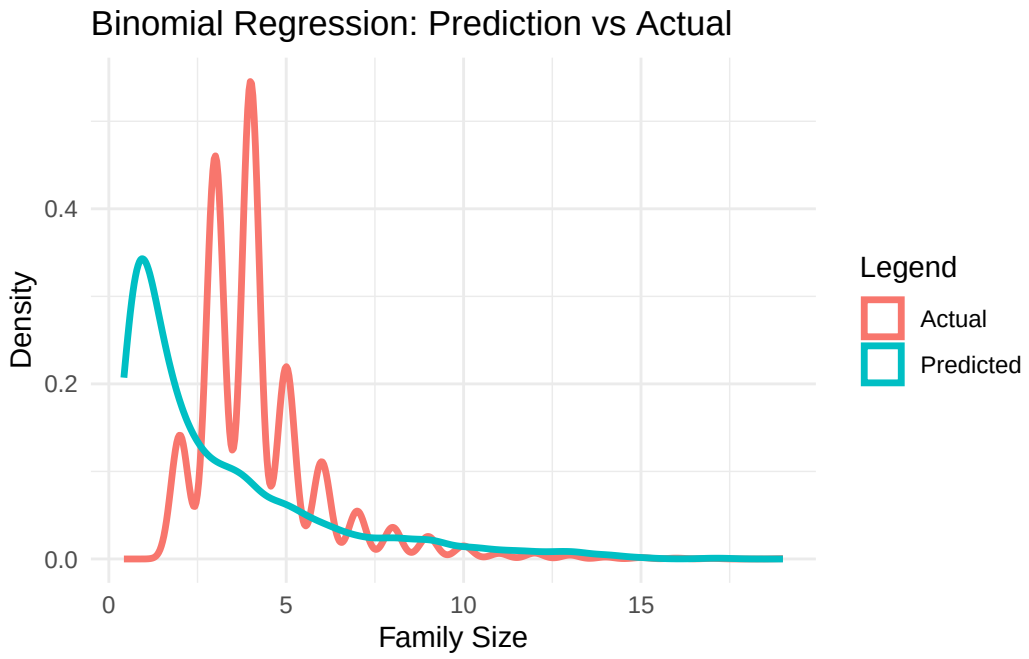
- $\mu = E(\text{family\_size})$ represents the expected family size.
- ageMarried is a categorical variable representing different marriage age groups.
- literacy $\in \{\text{yes}, \text{no}\}$, indicating whether an individual is literate.
- monthsSinceM $\in \mathbb{R}^+$ represents the number of months since marriage (a continuous variable).
- β_1, β_2, and β_3 are the coefficients for the independent variables.
- β_0 is the intercept term.
- This model assumes family_size follows a Gamma distribution.
- The log-link function ensures positive expected values.

3 Result

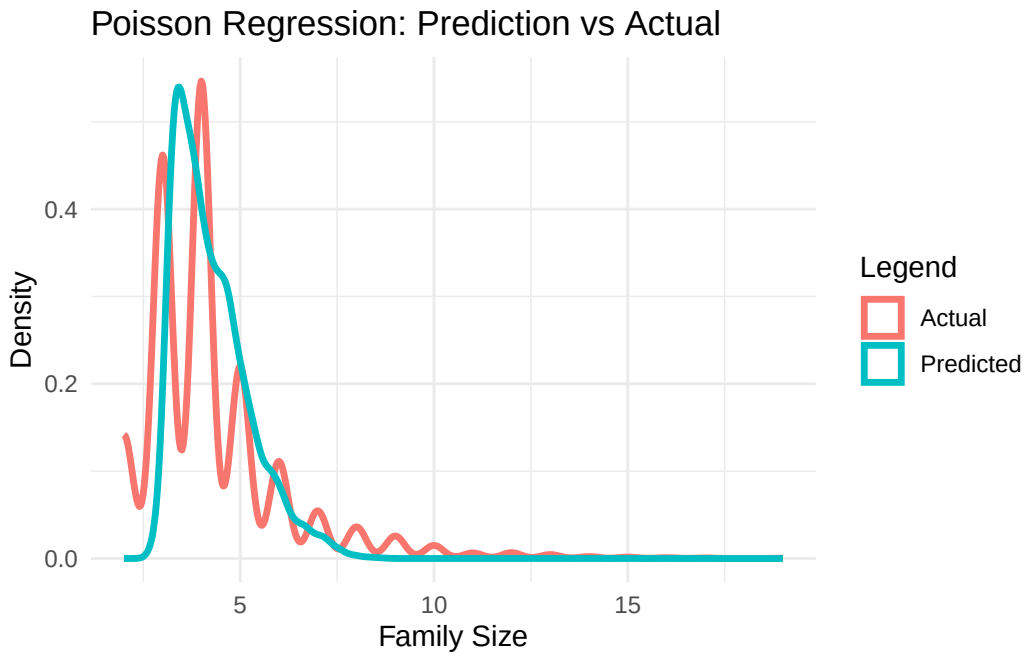
3.1 Normal Distribution Model



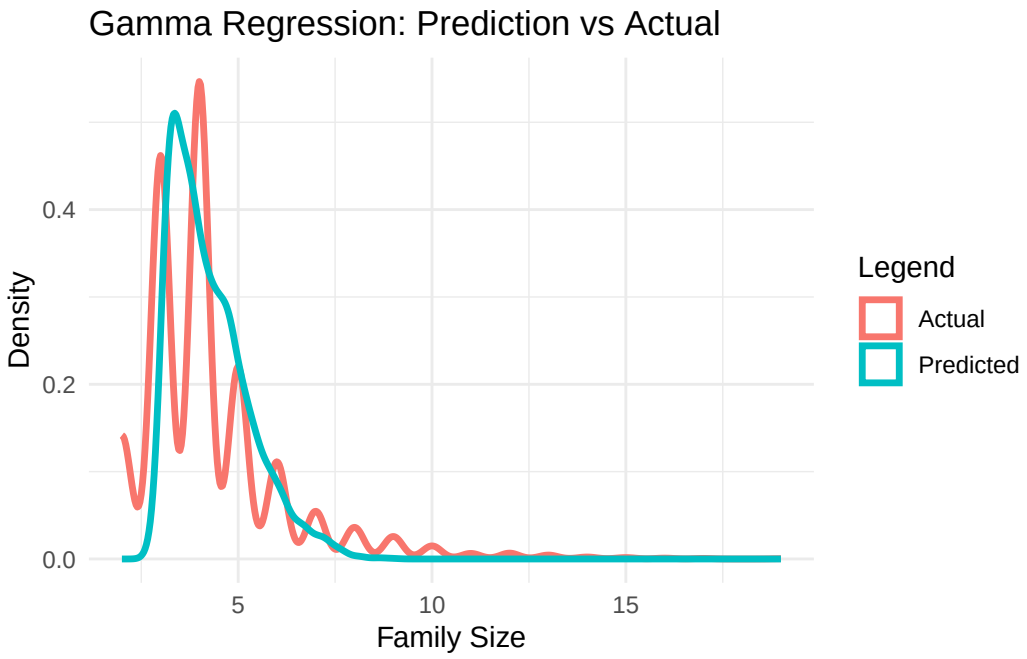
3.2 Binomial Distribution Model



3.3 Poisson Distribution Model



3.4 Gamma Distribution Model



4 Conclusion