Modeling Family Size: Literacy and Marriage Age in Rural Portugal

A Generalized Linear Model Approach

Shaotong (Max) Li

November 17, 2024

1 Introduction

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.

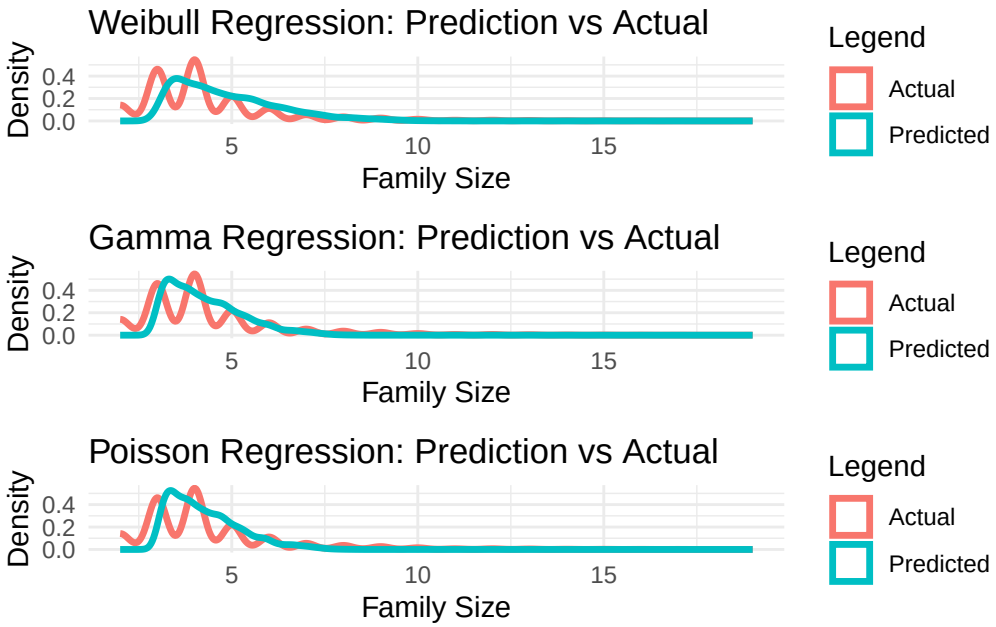
literacy (Binary): A factor variable indicating whether an individual is literate (yes) or not (no). This variable is included as a second independent variable, as literacy is hypothesized to influence family planning and size.

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.

Model Comparison				
Model	AIC	BIC	Log_Likelihood	RMSE
Poisson	19292.89	19312.53	-9643.446	1.612
Gamma	17168.52	17194.71	-8580.262	1.614
Weibull	17996.81	18022.99	-8994.403	1.769

Note:
AIC, BIC, Log-Likelihood, and RMSE for different regression models.

2.2 Generalized Linear Models



2.3 Compare

3 Result

3.1 Generalized Gamma Linear Model

$$\text{Gamma GLM: } \mathbb{E}[Y \mid X_1, X_2] = \exp(\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2) \tag{1}$$

where:

- Y is the response variable, following a Gamma distribution.
- X_1 and X_2 are predictor variables.

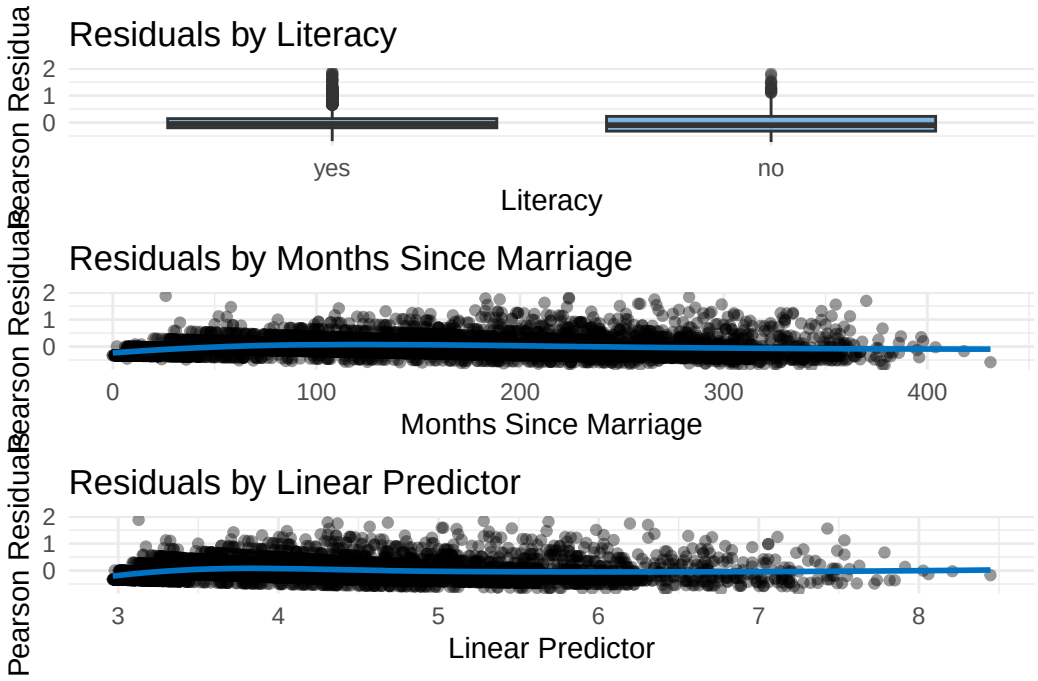
- β_1 and β_2 are the coefficients associated with X_1 and X_2.
- β_0 is the intercept term.

$$\text{Link Function: } g(\mathbb{E}[Y]) = \log(\mathbb{E}[Y]) \tag{2}$$

where:

- The log link function ensures $\mathbb{E}[Y] > 0$.
- The linear predictor is given by $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.
- The expectation of Y is modeled as $\mathbb{E}[Y] = \exp(\eta)$.

3.2 Model Summary and Overdispersion



Gamma Model Summary				
Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.08727	0.0089854	121.0	< 2.2e-16
literacyno	0.19404	0.0154244	12.6	< 2.2e-16
monthsSinceM	0.00204	0.0000515	39.6	< 2.2e-16
Dispersion Parameter	0.112			
Null Deviance	776.149			
Residual Deviance	532.031			
Null DF	5147.000			
Residual DF	5145.000			

Note:

Coefficient estimates, standard errors, t-values, significance levels, and dispersion metrics for the Gamma model.