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.

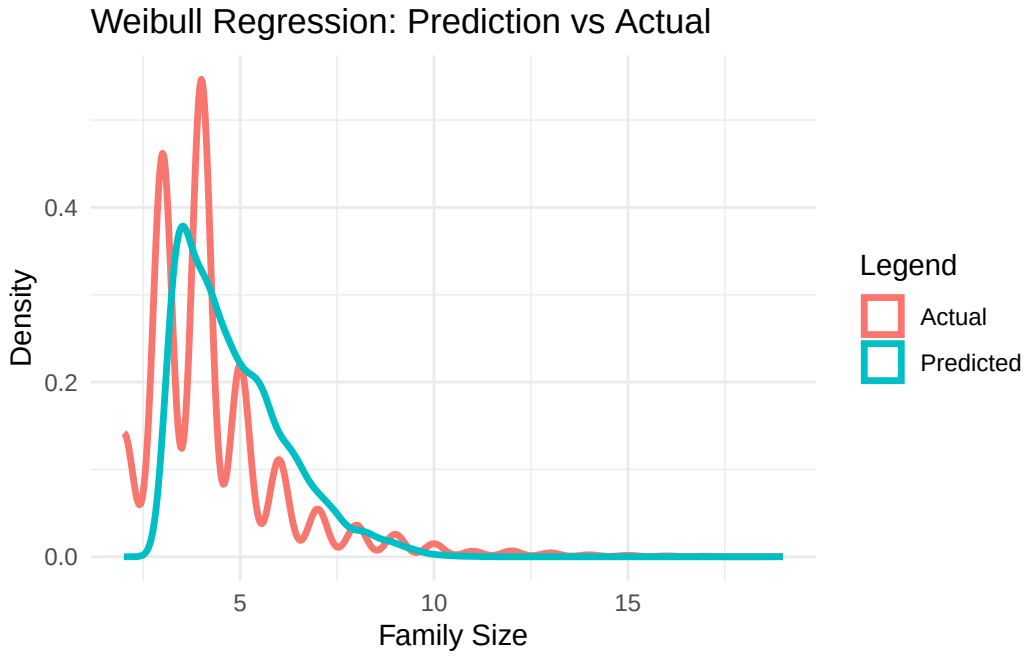
literacy (Binary): A factor variable indicating whether an individual is literate (yes) or not (no). This variable is included as a second independent variable, as literacy is hypothesized to influence family planning and size.

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.

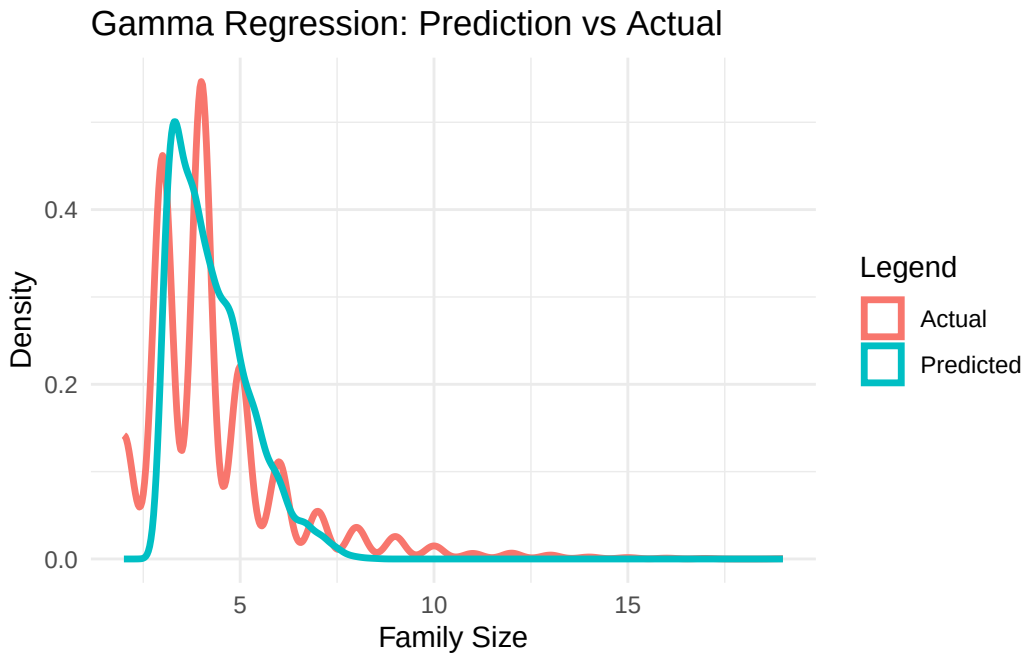
2.2 Generalized Linear Models

3 Result

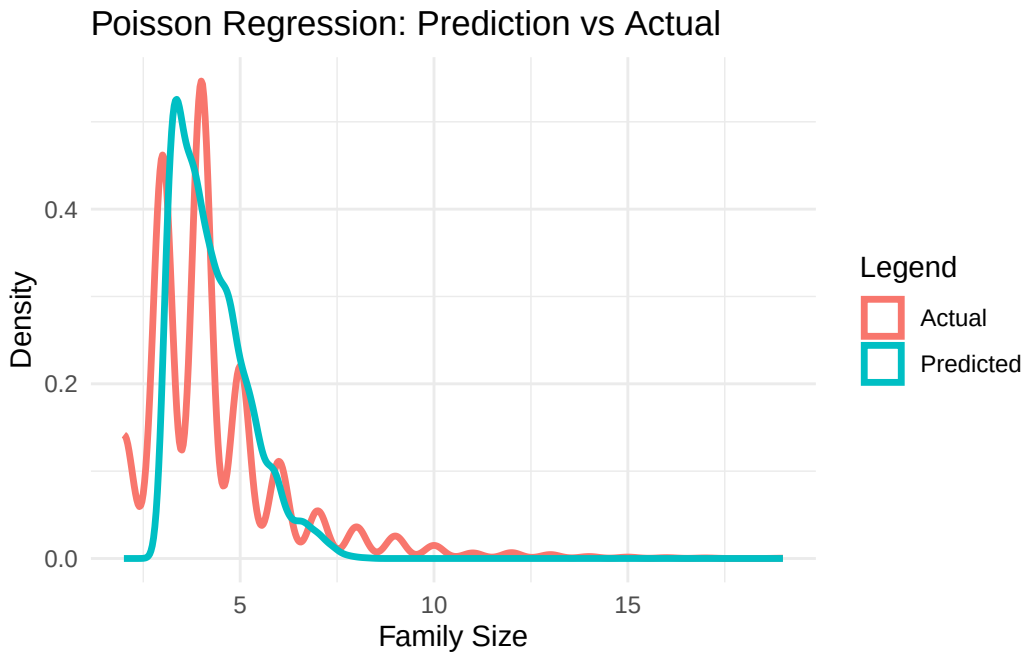
3.1 Weibull Distribution Model



3.2 Gamma Distribution Model



3.3 Poisson Distribution Model



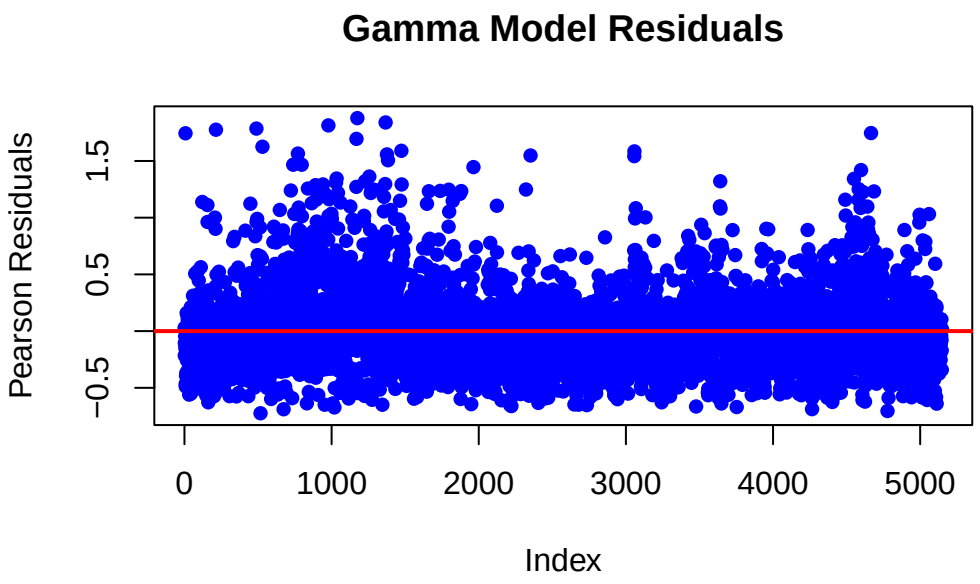
3.4 Compare

```
    Model      AIC      BIC Log_Likelihood     RMSE
1 Poisson 19292.89 19312.53      -9643.446 1.611659
2   Gamma 17168.52 17194.71      -8580.262 1.613784
3 Weibull 17996.81 18022.99      -8994.403 1.768889
```

3.5 Overdispersion

```
    Mean_Family_Size Variance_Family_Size
            4.260490             3.463704
```

```
Gamma_Dispersion
       0.1122859
```



3.6 Interpret

```
Call:
glm(formula = family_size ~ literacy + monthsSinceM, family = Gamma(link = "log"),
    data = portugal)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  1.087e+00  8.985e-03  121.00   <2e-16 ***
literacyno   1.940e-01  1.542e-02   12.58   <2e-16 ***
monthsSinceM 2.039e-03  5.154e-05   39.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1122859)

    Null deviance: 776.15  on 5147  degrees of freedom
Residual deviance: 532.03  on 5145  degrees of freedom
AIC: 17169

Number of Fisher Scoring iterations: 4
```