# Analyzing Influencing Factors on Family Size in Portugal*

Jiadong Wang, Yanzun Jiang

2025-02-03

## Contents



*Code and data supporting this paper is available at: https://github.com/Stary54264/factors_affect_family_size_in_portugal

# 1 Introduction

Understanding the factors that influence family size is crucial for shaping social and economic policies. In many societies, family size is often linked to cultural and socioeconomic factors such as literacy levels and the age at which individuals marry. Portugal, despite being a European country, had a GDP per capita in 1980 comparable to that of Mexico, making it an interesting case study for fertility patterns. Prior research suggests that rural families tend to have more children than urban families, and birth rates may vary based on educational attainment and marital timing. Building on these findings, our study aims to answer this question: **"How do literacy and age of a marriage affect family size?"**.

Previous studies have shown that higher literacy levels among women are associated with reduced fertility rates due to increased awareness of family planning and career aspirations (Kassim and Ndumbaro 2022). Similarly, research on marriage timing suggests that early marriage is linked to larger family sizes due to prolonged reproductive periods (Kassim and Ndumbaro 2022; Abdallah, Mohammed, and Mohamed 2023). Comparative studies in China and India further indicate that shifting social norms and economic conditions play a crucial role in fertility decisions (He and Xie 2023). These findings highlight the importance of investigating how these variables interact within the Portuguese context.

To analyze this relationship, we will use Generalized Linear Models (GLMs) as they are well-suited for count data like family size. Specifically, Poisson regression or negative binomial regression will be considered based on the presence of overdispersion. Exploratory data analysis will be conducted to summarize key variables, followed by model selection techniques to identify the best-fitting statistical model. R (R Core Team 2023) will be used to conduct this, and packages including `tidyverse` (Wickham et al. 2019), `here` (Müller 2020), `knitr` (Xie 2014), `kableExtra` (Zhu 2024), `patchwork` (Pedersen 2024), `MASS` (Venables and Ripley 2002), and `glmmTMB` (Brooks et al. 2017) will also be used. The final model will be interpreted and draw meaningful conclusions about the relationship between literacy, age at marriage, and family size in Portugal.

# 2 Methods

To investigate the relationship between literacy, age at marriage, and family size, we will employ Generalized Linear Models (GLMs), specifically Poisson or negative binomial regression models. Childbirth follows the Poisson process, which assumes that events occur independently and at a constant rate over time, and this makes Poisson distribution an appropriate choice for the data. However, if the data exhibit overdispersion (the variance is significantly greater than the mean), a negative binomial regression model may be considered, as it introduces a dispersion parameter to allow for greater variability.

The primary predictor variables of interest are literacy and age at marriage, as they might influence family size. Interaction effects will also be considered to assess whether the relationship between literacy and family size depends on marital age. These variables are relevant since prior research suggests that rural families tend to have more children, and economic conditions may influence both literacy levels and marriage timing.

The significance of predictors will be determined using hypothesis tests and t-tests on coefficients, with p-values and confidence intervals providing statistical evidence. By conducting an ANOVA test, both explained and unexplained variations would be determined.

# 3 Results

The dataset includes key variables such as Number of Children in a Family (the response variable), Age at Marriage, and Literacy Status (the predictors). Statistical summaries (Table 1) provide insights into the distribution and central tendencies of these variables, such as the mean number of children (2.3) and the frequency of age-at-marriage intervals. Visualizations (Figure 1) further illustrate relationships, such as how family size varies with literacy status or age at marriage. These analyses guide model selection and interpretation, helping to identify patterns and trends in the data.

The number of children in a family, follows a right-skewed distribution, with most families having between 0 and 5 children. The center of this distribution appears to be around 2 children, while the spread extends

Table 1: Descriptive statistics of the number of children in a family, the distribution of literacy status, and age at marriage among the study population. The mean number of children is 2.3, indicating that, on average, families in the study population have slightly more than two children. Literacy status shows a majority of respondents reporting "yes", while age at marriage is distributed across various intervals, with the highest frequency in the 22 to 25 age group.

| Mean | Variance | Median | Minimum | Maximum | IQR |
|---|---|---|---|---|---|
| 2.3 | 3.5 | 2 | 0 | 17 | 2 |

| Literacy Status | Count |
|---|---|
| yes | 4567 |
| no | 581 |

| Age at Marriage | Count |
|---|---|
| 0to15 | 52 |
| 15to18 | 452 |
| 18to20 | 910 |
| 20to22 | 1126 |
| 22to25 | 1468 |
| 25to30 | 923 |
| 30toInf | 217 |

up to approximately 15 children, though values beyond 10 are rare. The long right tail suggests that while large families exist, they are uncommon in the sample population. Literacy status is categorical and shows a strong imbalance, with the majority of individuals being literate. The frequency of literate individuals is significantly higher than illiterate individuals, indicating a high overall literacy rate in the sample. Age at marriage follows a roughly normal distribution, peaking between the ages of 22 and 25. The center of this distribution is likely around 22 to 25 years, with a spread covering a range from below 15 to above 30 years. The shape suggests that while some individuals marry very early or late, most tend to marry between 18 and 30 years. The relationship between literacy and age at marriage could be explored further to determine if literacy impacts marital age trends. A bivariate analysis could help assess whether literacy influences the likelihood of early or delayed marriage in the sample population.

The Poisson GLM was developed to investigate the relationship between literacy and age at marriage while accounting for the natural log of the duration of marriage through an offset term. The model includes literacy status and categorized age at marriage as predictor variables, with the reference group consisting of literate individuals who married between the ages of 22 and 25. The results (Table 2) indicate that illiteracy is a significant predictor ($p = 0.000$), suggesting that being illiterate is strongly positively associated with changes in the family size compared to literate individuals. Among the marriage age groups, individuals who married between 15-18 years ($p = 0.025$) and 18-20 years ($p = 0.026$) show significant differences from the reference category, implying that marrying at these ages has a notable impact. Conversely, marriage before 15, between 20-22, between 25-30, and after 30 do not show significant differences from the reference group. These findings highlight the role of literacy and specific marriage age ranges in influencing the outcome, though additional modeling techniques or interactions could be explored to refine the analysis further.

We can see from the summary table that the variance of the number of children is significantly greater than the mean, suggesting the presence of overdispersion. This makes the Poisson GLM unsuitable, as it assumes equal mean and variance. A new model, the Negative Binomial GLM, accommodates overdispersion, was therefore employed. From Table 3, we can find that the intercept term is significant, with a value of -1.783 and a 95% confidence interval ranging from -1.822 to -1.744. Literacy shows a positive effect on family size, with an estimate of 0.148 and a confidence interval from 0.095 to 0.200, indicating that illiterate individuals tend to have more children. The age at marriage categories also show varying effects: marrying between

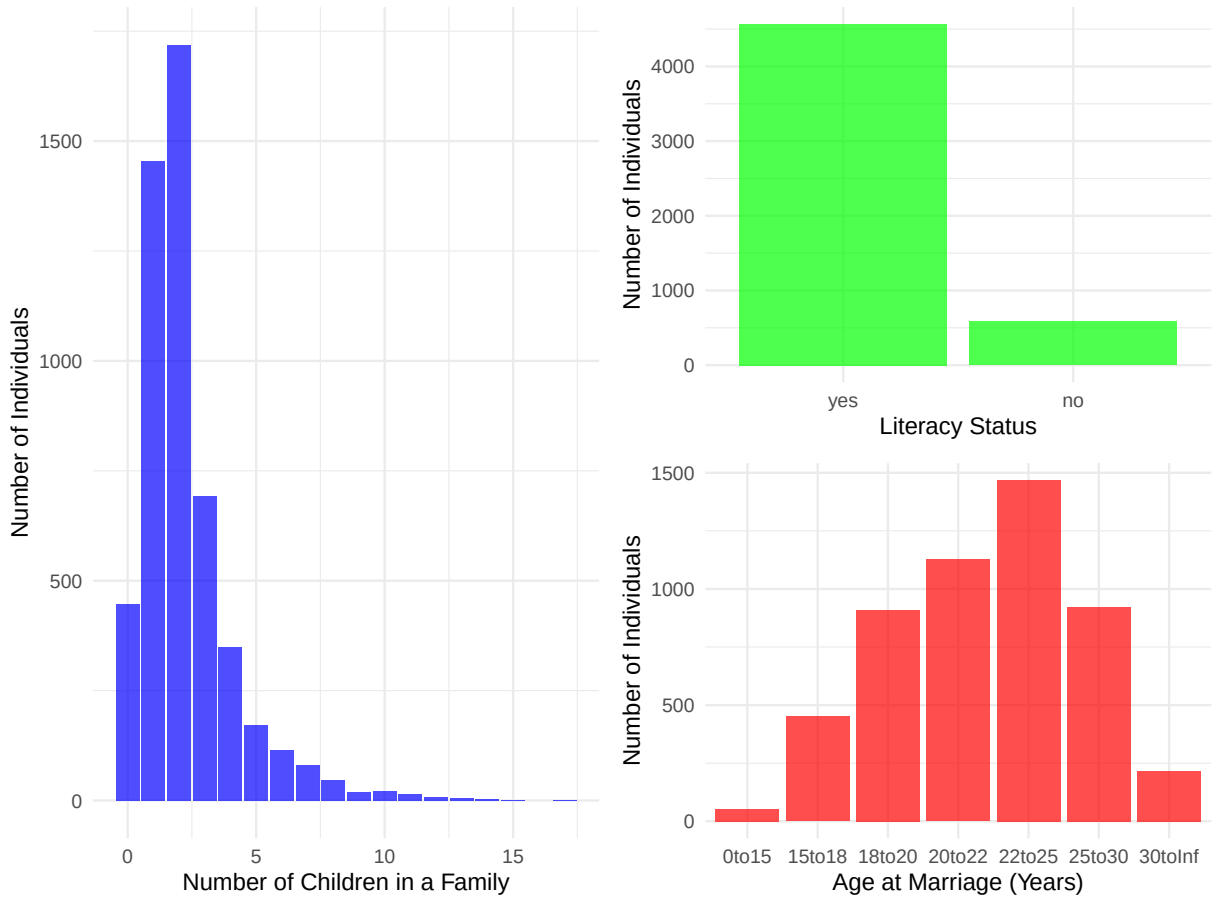


Figure 1: Overview of key variables: number of children in a family, literacy status, and age at marriage. The left panel (blue) displays the distribution of the number of children per family, showing a right-skewed pattern where most families have fewer children, with the highest frequency at 1 to 3 children. The top-right panel (green) represents literacy status, categorized as 'yes' (literate) or 'no' (illiterate). The majority of individuals are literate, with a much smaller proportion classified as illiterate. The bottom-right panel (red) illustrates the distribution of age at marriage, indicating that the most common age range for marriage is 22 to 25 years, followed by 20 to 22 and 25 to 30 years. Fewer individuals marry at very young (0 to 15, 15 to 18) or older ages (30 and above). These distributions provide critical context for subsequent Generalized Linear Model (GLM) analysis, as they highlight demographic patterns that may influence relationships between these variables.

Table 2: Logistic regression results showing the relationship between literacy and age at marriage with the response variable with significance level 0.05. Literacy is a significant predictor (p = 0.000), indicating a strong effect. Marriage between ages 15-18 and 18-20 also shows significance (p = 0.025 and 0.026, respectively), suggesting an impact on the outcome. Other age groups do not exhibit significant effects.

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -1.802 | 0.018 | -99.076 | 0.000 |
| literacyno | 0.159 | 0.024 | 6.770 | 0.000 |
| ageMarried0to15 | 0.050 | 0.080 | 0.624 | 0.533 |
| ageMarried15to18 | 0.076 | 0.034 | 2.236 | 0.025 |
| ageMarried18to20 | 0.062 | 0.028 | 2.228 | 0.026 |
| ageMarried20to22 | 0.029 | 0.026 | 1.105 | 0.269 |
| ageMarried25to30 | 0.013 | 0.029 | 0.468 | 0.640 |
| ageMarried30toInf | 0.022 | 0.059 | 0.369 | 0.712 |

Table 3: Logistic regression results showing the relationship between literacy and age at marriage with the response variable with significance level 0.05. Literacy is a significant predictor (p = 0.000), indicating a strong effect. Marriage between ages 15-18 and 18-20 also shows significance (p = 0.025 and 0.026, respectively), suggesting an impact on the outcome. Other age groups do not exhibit significant effects.

| | Estimate | 2.5 % | 97.5 % |
|---|---|---|---|
| (Intercept) | -1.783 | -1.822 | -1.744 |
| literacyno | 0.148 | 0.095 | 0.200 |
| ageMarried0to15 | 0.068 | -0.109 | 0.244 |
| ageMarried15to18 | 0.084 | 0.009 | 0.158 |
| ageMarried18to20 | 0.069 | 0.009 | 0.129 |
| ageMarried20to22 | 0.035 | -0.022 | 0.092 |
| ageMarried25to30 | 0.011 | -0.051 | 0.072 |
| sd | 0.265 | 0.235 | 0.298 |

15 to 18 years and 18 to 20 years have positive and significant estimates of 0.084 and 0.069, respectively, suggesting that marrying at these ages is associated with larger family sizes. Confidence intervals of other age groups contains 0, which suggests that the coefficient is not significant. The standard deviation of the random effects is 0.265, with a confidence interval from 0.235 to 0.298, indicating variability in family size not explained by the predictors. Overall, the model suggests that both literacy status and age at marriage are significant predictors of family size, with the Negative Binomial GLM appropriately accounting for the observed overdispersion.

# 4 Conclusion

In conclusion, the analysis comparing Poisson and Negative Binomial GLMs for predicting family size based on literacy status and age at marriage, with years since marriage as an offset, revealed significant insights. The Poisson model, which assumes equal mean and variance, was found to be unsuitable due to overdispersion in the data, as indicated by the variance exceeding the mean. The Negative Binomial model, which accounts for overdispersion, provided a better fit and more reliable estimates. The results from the Negative Binomial GLM (Table 3) showed that literacy status significantly affects family size, with individuals lacking literacy having a positive and significant effect (estimate = 0.148, 95% CI: 0.095 to 0.200). This suggests that lower literacy levels are associated with larger family sizes. Additionally, age at marriage categories, particularly marrying between 15 to 18 years and 18 to 20 years, also showed positive and significant effects, indicating that earlier marriage is associated with increased family size.

The findings align with existing literature that highlights the influence of socio-economic factors, such as

literacy, cultural norms, and age at marriage, on fertility rates. The positive association between lower literacy levels and larger family sizes is consistent with studies suggesting that education often leads to delayed marriage and smaller family sizes due to increased awareness and access to family planning resources (Kassim and Ndumbaro 2022). Similarly, the effect of age at marriage on family size supports the notion that earlier marriage extends the reproductive period, thereby increasing the likelihood of having more children. These results underscore the importance of considering socio-economic and cultural factors in demographic studies and policy-making aimed at addressing population growth and family planning. The insights from this study can be used to inform targeted interventions aimed at improving literacy and delaying marriage to manage family size and promote sustainable development.

# References


Abdallah, Ahmed Saied Rahama, Mohammed Omar Musa Mohammed, and Adel Ali Ahmed Mohamed. 2023. "Early Marriage and Its Association with the Socioeconomic and Sociocultural Factors of Women in Sudan: A Predictive Model." *The Open Public Health Journal* 16 (1).

Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal* 9 (2): 378–400. https://doi.org/10.32614/RJ-2017-066.

He, Yuchen, and Yu Xie. 2023. "The Shift in Ideal Family Size: Examining the Impact of Marriage and Parenthood." *Chinese Journal of Sociology* 9 (4): 553–73.

Kassim, Mohamed, and Faraja Ndumbaro. 2022. "Factors Affecting Family Planning Literacy Among Women of Childbearing Age in the Rural Lake Zone, Tanzania." *BMC Public Health* 22 (1): 646.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.