

Analyzing Influencing Factors on Family Size in Portugal*

Jiadong Wang, Yanzun Jiang

2025-02-03

Contents

1	Introduction	2
2	Methods	2
3	Results	2
4	Conclusion	3
	References	5

*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), `xMASS` (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, including mean, variance, median, minimum, maximum, and IQR are shown. The distribution of literacy status and age at marriage among the study population are also presented. 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

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.

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

4 Conclusion

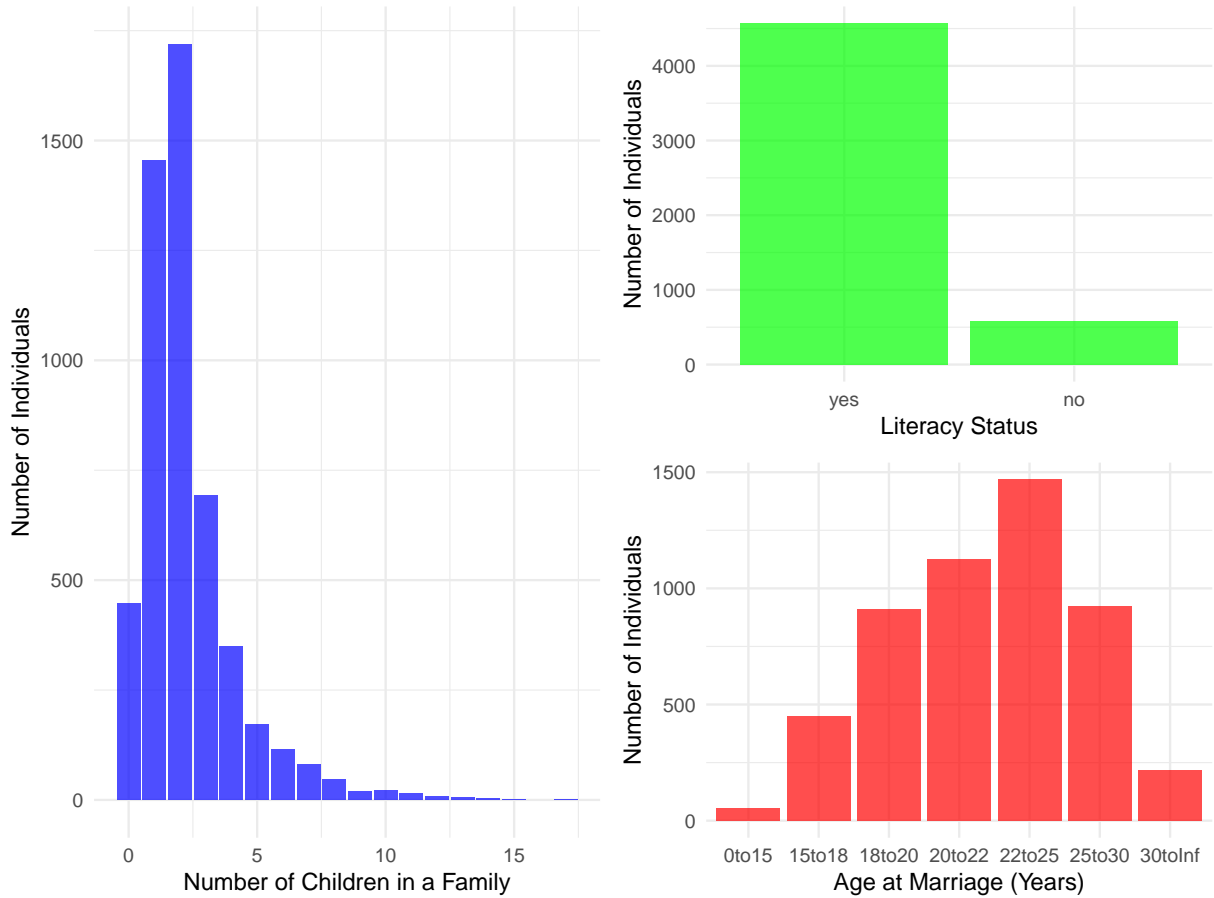


Figure 1: Overview of key demographic variables: number of children in a family, literacy status, and age at marriage. The left panel (blue) displays the distribution of the number of children per family, showing a right-skewed pattern where most families have fewer children, with the highest frequency 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.

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 Golemund, 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>.