Investigating the Effect of Literacy and Marriage Age on Family Size*

Ariza Hossain, Tanmay Shinde

February 09, 2024

Contents

1 Introduction **1**
1.1 Literature Review and Importance of Study . . . 1
1.2 Research Objective and Statistical Overview . . . 2

2 Methods **2**

3 Results **3**
3.1 Statistical Summaries of Predictor Variables . . . 3
3.2 Modeling Process . . . 4
3.3 Model Results . . . 7

4 Conclusion **8**
4.1 Interpretation of Model Results . . . 8
4.2 Comparison of Model Effects with Existing Literature . . . 8
4.3 Closing Remarks . . . 8

References **9**

1 Introduction

1.1 Literature Review and Importance of Study

Understanding the relationship between literacy, age at marriage, and family size is crucial for addressing key societal challenges. Studies show that smaller family sizes are associated with better child health outcomes due to increased parental investment per child, leading to improved nutrition, education, and healthcare access. People who grew up in relatively small families tend to have higher educational attainment, earn more income, and accumulate greater household wealth (Parr 2012). From an economic perspective, a population with a larger proportion of working-age adults—facilitated by smaller family sizes—can drive economic growth by enhancing workforce productivity and reducing dependency ratios (Bloom, Canning, and Sevilla 2003). Additionally, promoting female education and delaying marriage are powerful tools for advancing gender equality, empowering women to pursue education and career opportunities, and improving their decision-making autonomy within households (Bates, Maselko, and Schuler 2007). These factors highlight the relevance of this study in informing policies aimed at fostering sustainable development, improved public health, and social equity.

Previous research has established strong correlations between literacy, marriage age, and fertility outcomes. One study found that literate women were significantly more likely to use contraception (54% vs. 43%, $p = 0.008$), undergo sterilization (94% vs. 44%, $p < 0.001$), and have fewer children. Additionally, they demonstrated better family planning knowledge and a reduced preference for male children (51% vs. 73%, p

*Code and data are available at: EffectOfLiteracyAndMarriageAgeOnFamilySize

= 0.029) (Suliankatchi et al. 2012). Another study found that rural women expressed a higher preference for large families, with 32.1% desiring six or more children. Education played a crucial role in fertility preferences, as 72% of women and 83.6% of men who favored smaller families had at least a secondary education (Kahansim, Hadejia, and Sambo 2013). A third study highlighted the impact of early marriage, showing that women who married before the age of 20 had, on average, one more child than those who married between 25 and 29. Early marriage was linked to higher fertility due to prolonged childbearing periods and lower contraceptive efficiency. Regression analysis indicated that family size was influenced by fecundity, contraceptive use, and social selection factors, reinforcing the association between early marriage and larger families (Busfield 1972).

1.2 Research Objective and Statistical Overview

Building on these findings, our study aims to investigate the extent to which literacy and age at marriage influence family size in Portugal, a developing economy and is that varies between urban and rural areas. Specifically, we seek to determine whether higher literacy rates and delayed marriage lead to smaller family sizes and to what degree these factors interact. By synthesizing insights from prior research which were also mostly based in developing economies, we aim to contribute to a more nuanced understanding of how educational and marital timing interventions can shape reproductive behavior and demographic trends in Portugal. To achieve this, we will employ a generalized linear model (GLM) to analyze the impact of literacy and marriage age on family size. This statistical approach will allow us to quantify the effects of these variables while controlling for other relevant factors such as socioeconomic status, geographic location, and gender composition of children. By examining the coefficients derived from the model, we can assess the relative influence of literacy and marital timing on fertility outcomes, providing evidence-based recommendations for policymakers and public health initiatives.

2 Methods

Our study will use the number of children per woman as the main response variable. The key predictor variables will be the age of marriage, literacy rate, region, and number of sons. These predictors are included based on our literature review which suggests that literacy and delayed marriage reduce fertility, and regional differences can capture cultural and economic factors influencing family size while the desire to have more sons may increase the family size. This study utilizes data from the Portuguese Fertility Survey (1979-80) (Conim 1986). We use the statistical programming language R (R Core Team 2023) to perform our analysis using various helpful packages like, readr (Wickham and Hester 2023), tidyverse (Wickham et al. 2023b), and jtools (Long 2022). Further, libraries like ggplot2 (Wickham et al. 2023a), knitr (Xie 2023), and kable (Zhu 2023) were used to analyze the data and create visualizations and tables.

To model the relationship between these variables, we will initially fit a Poisson generalized linear model (GLM). A Poisson model is appropriate because our response variable consists of count data, which are non-negative and right-skewed. We assume that the number of children born to each woman follows a homogeneous Poisson process, meaning that each birth is independent and uniformly distributed over time. This assumption justifies using the Poisson distribution, which is commonly applied to model count data in similar demographic studies.

We will further conduct exploratory data analysis and assess the relationships between predictors and the need for potential interaction variables. We will also check for overdispersion by examining the means and variances of each group of predictors to ensure they satisfy the assumption of the Poisson model. Furthermore, we will analyze scatter plots to determine if our data is comparable across different predictor variables. This will help us decide whether an offset term should be included in our model to account for any differences.

To refine our model, we will fit multiple models with different combinations of predictors, offsets and interaction terms based on our exploratory findings. To determine the best model, we will use the partial F-test to compare nested models and assess how much additional variation is explained by each new predictor or interaction term. The partial F-test will help us determine whether including an interaction term significantly improves model fit or if a simpler model should be preferred. By systematically comparing models, we aim to select the one that provides the best explanatory power while maintaining statistical significance.

Finally, we will compare model fits using statistical criteria, selecting the model that best accounts for overdispersion, interaction effects, and any additional structural patterns in the data. We will look at the confidence intervals and their overlap with zero along with the z-values for each predictor to confirm if our predictors are statistically significant. By systematically evaluating these considerations, we aim to produce a robust and interpretable model for understanding the influence of literacy and marriage age on family size.

3 Results

3.1 Statistical Summaries of Predictor Variables

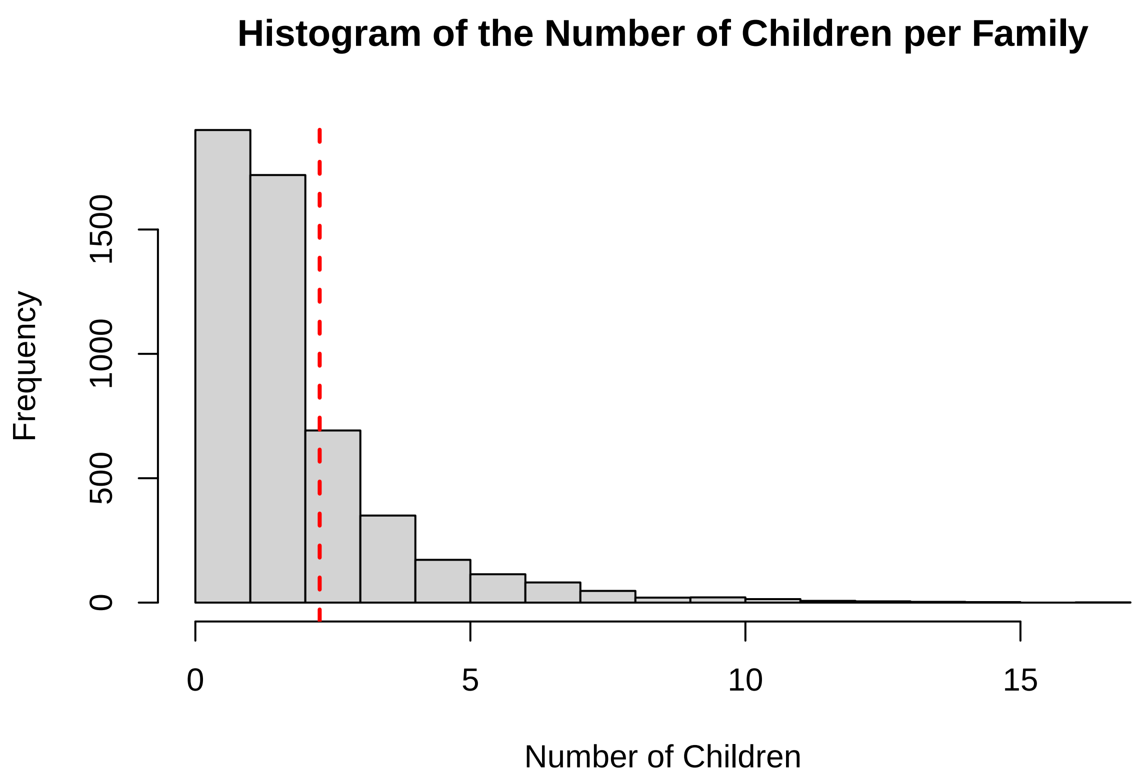


Figure 1: Histogram showing the distribution of the outcome variable, number of children per family. The data exhibit variability in family size, with most families clustering around lower values. This visualization provides insight into the overall spread and central tendency of family sizes in the sample.

Figure 1 shows a right-skewed distribution of the number of children per family, with most families having 0 or 1 children, and fewer families having larger numbers. The dashed line shows the mean number of children per family in Portugal in the 1980s which is about 2.26. The spread is wide with a variance of 3.46 and we can see that some families having more than 10 children, though such cases are rare, indicating potential outliers. The long right tail suggests that while most families have a small number of children, a few have significantly more. This distribution is relevant in our study as it is the response variable that will be modeled as a function of literacy and the age a woman married. It helps us understand the trend surrounding the number of children people had in Portugal in the 1980s.

Figure 2 shows the univariate summaries of the predictor variables. For the number of sons per family, the data is right-skewed, with the mean around 1 son per family, indicating that most families have fewer sons, but a few have significantly more. The age at marriage distribution is more uniform, with the majority of observations (mode) falling within the 22-25 age range, and the mean marriage age is between 20-22, reflecting a tendency toward earlier marriages. The region variable is right-skewed as well, with most observations (mode) coming from regions with populations of fewer than 10,000, suggesting a concentration in smaller areas. Finally, the literacy data is highly imbalanced, with the majority of observations indicating "yes" (over 4000), and a smaller proportion of "no" responses (less than 1000), pointing to a skewed distribution toward higher literacy in the population.

3.2 Modeling Process

To investigate the factors influencing the number of children per family, we first fit a Poisson regression model using literacy, age at marriage, and region as predictor variables:

$$\mathbf{num\_children} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{literacy} + \beta_\mathbf{2}\mathbf{age\_married} + \beta_\mathbf{3}\mathbf{region}$$

The initial model allows us to assess the main effects of these predictors on the response variable (number of children). The exponentiated coefficients provide the rate ratios, which indicate how each predictor affects the expected count of children.

We then introduce the variable sons as a potential confounder, creating an updated model:

$$\mathbf{num\_children} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{literacy} + \beta_\mathbf{2}\mathbf{age\_married} + \beta_\mathbf{3}\mathbf{region} + \beta_\mathbf{4}\mathbf{sons}$$

To evaluate whether adding sons significantly improves model fit, we conduct a partial F test between the model with the confounder variable and our initial model. The result of the ANOVA test suggests that the larger model is very significant with a large chi-square statistic and a highly significant p-value ($p < 2.2e\text{-}16$), suggesting that including sons significantly reduces residual deviance, indicating a better model fit.

Next, we use boxplots to visually explore relationships between predictor variables and the number of children. By examining trends across groups, we can assess whether certain predictors interact with each other. If the spread of the number of children differs across literacy levels within different regions, for example, it suggests the need for an interaction term in the model.

From Figure 3, we can see that there is a difference in the mean number of children per family across the same age at marriage or region population category when the literacy status is varied. This suggests that we need to consider interaction terms between ageMarried and literacy as well as literacy and region. Thus, next we consider two models with the above mentioned interaction terms:

1. Interaction between literacy and age at marriage:

$$\mathbf{num\_children} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{literacy} + \beta_\mathbf{2}\mathbf{age\_married} + \beta_\mathbf{3}\mathbf{region} + \beta_\mathbf{4}\mathbf{sons} + \beta_\mathbf{5}\mathbf{literacy} * \mathbf{age\_married}$$

2. Interaction between literacy and region:

$$\mathbf{num\_children} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{literacy} + \beta_\mathbf{2}\mathbf{age\_married} + \beta_\mathbf{3}\mathbf{region} + \beta_\mathbf{4}\mathbf{sons} + \beta_\mathbf{5}\mathbf{literacy} * \mathbf{region}$$

The partial F tests for both models show no significant improvement ($p = 0.9084$ for the model with interaction between literacy and ageMarried) and ($p = 0.5229$ for the model with interaction between literacy and region), suggesting that the interaction terms are not meaningful and do not lead to a significant chunk of the variation in the dataset being explained by the models with the interaction terms. This suggests that literacy effects do not vary meaningfully by region or age at marriage.

To account for the varying durations of marriage, we introduce an offset term to the model to see if it helps us explain more variation in the data. To do this, we first convert the monthsSinceMarried variable in the

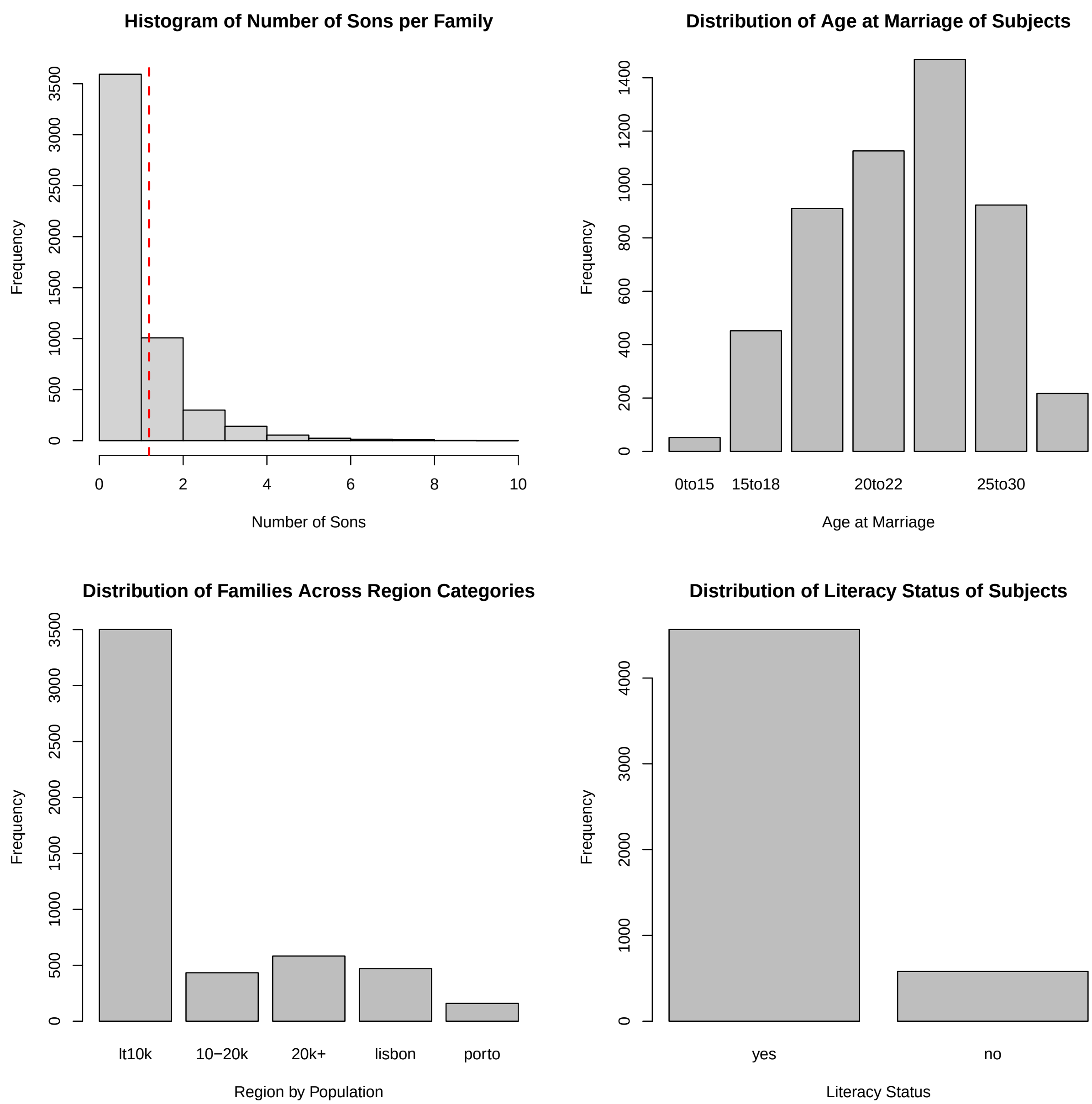
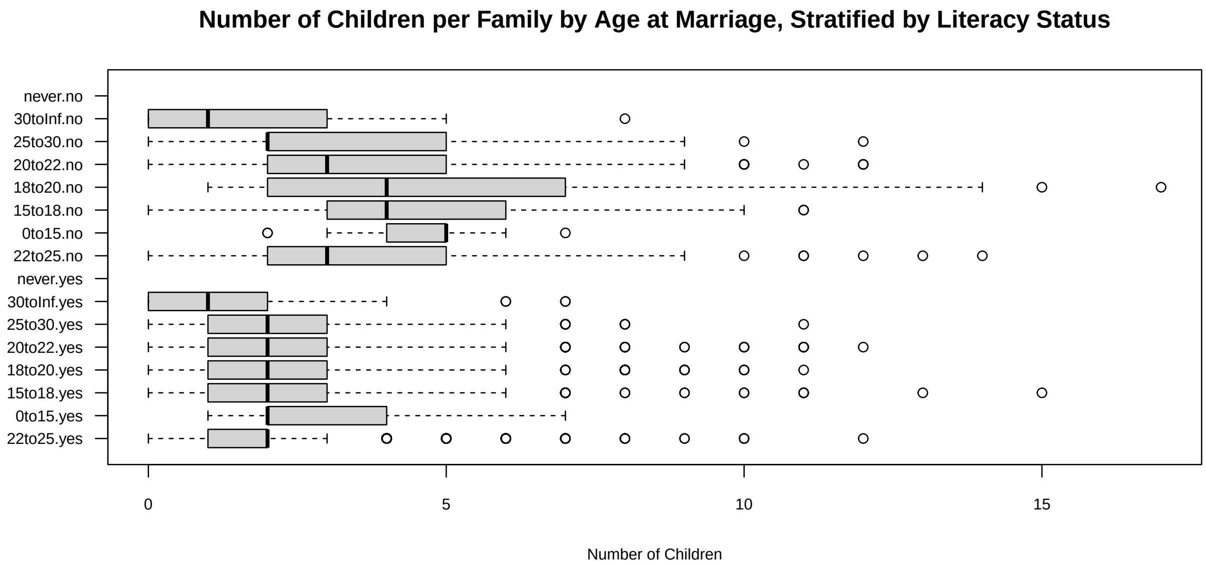


Figure 2: Histogram of the number of sons per family, with the mean indicated by a red dashed line. Bar plots display the distribution of age at marriage, regional family distribution by population size, and literacy status of subjects, providing insights into key demographic characteristics of the surveyed population.



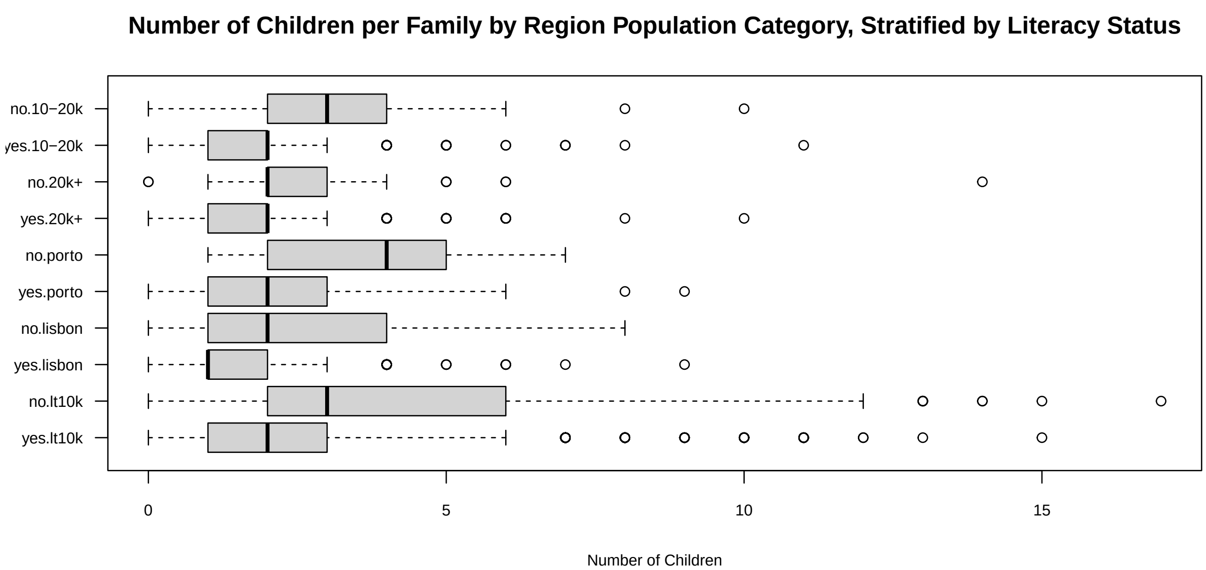


Figure 3: Boxplot of Number of Children per Family by Age at Marriage and Region Population Category, Stratified by Literacy Status. Visualizing differences in the mean number of children per family across age at marriage and region population categories when stratified by literacy status. The observed variation suggests the need to include interaction terms between age at marriage and literacy as well as literacy and region in the analysis.

dataset to yearsSinceMarried and then take the log of that. We use logYrsMarried i.e. the log of number of years for whcih the subject has been married. Thus giving us the new model:

$$\mathbf{num\_children} = \mathbf{offset(logYrsMarried)} + \beta_0 + \beta_1\mathbf{literacy} + \beta_2\mathbf{age\_married} + \beta_3\mathbf{region} + \beta_4\mathbf{sons}$$

However, we notice that the inclusion of the offset terms leads to all our coefficients becoming non-significant. Thus, we choose to avoid the offset term and go ahead with the model with literacy, ageMarried, region, and number of sons as predictors.

From the model summary for this model, we see that most of the coefficients for the ageMarried category are not statistically significant. Thus, we group them together into new categories: Below20, 20to30, Above30. The original categories contained smaller sample sizes, particularly in certain age groups, which might have led to unstable or insignificant results. By consolidating these age groups into broader, more meaningful ranges, we ensure that each category has a larger sample size, leading to more reliable coefficients.

Next, we check for overdispersion in the model by comparing the means and variances across groups of predictor variables to check and see if they are roughly equal and satisfy the assumption of the Poisson model. Figure 4 shows us that the mean and variances for each group are roughly equal and thus there is no need to consider overdispersion in our model. The dispersion ratio also comes out to 0.829 suggesting that the ratio of residual deviance to residual degrees of freedom is close to 1 and a Poisson model is the appropriate choice.

AgeMarried	Literacy	Mean	Variance
20to30	yes	2.02	2.21
Above30	yes	1.42	1.79
Below20	yes	2.25	3.34
20to30	no	3.75	6.90
Above30	no	1.68	3.41
Below20	no	4.64	9.74

Thus, our modeling process demonstrates that literacy, age at marriage, region, and number of sons significantly impact the number of children. The inclusion of sons as a confounder greatly improves model fit, while interaction terms do not add substantial explanatory power. The dispersion analysis confirms that our Poisson regression model is well-specified for this dataset.

3.3 Model Results

Figure 5 shows the model summary for our final model:

$$\mathbf{num\_children} = \beta_0 + \beta_1\mathbf{literacy} + \beta_2\mathbf{age\_married} + \beta_3\mathbf{region} + \beta_4\mathbf{sons}$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.336	0.016	20.662	0.000
literacyno	0.209	0.025	8.188	0.000
ageMarriedAbove30	-0.269	0.057	-4.702	0.000
ageMarriedBelow20	0.051	0.020	2.494	0.013
regionlisbon	-0.170	0.037	-4.595	0.000
regionporto	0.039	0.053	0.721	0.471
region20k+	-0.136	0.033	-4.146	0.000
region10-20k	-0.049	0.036	-1.363	0.173
sons	0.312	0.005	58.243	0.000

4 Conclusion

4.1 Interpretation of Model Results

Our findings confirm that marriage age is a significant predictor of family size. Holding all other factors constant, women who married before 20 are expected to have 5.2% more children on average compared to those who married between 20 to 30 years old. Conversely, women who married between 25 to 30 years old are expected to have 23.6% fewer children on average than those in the 20-30 age group. These results suggest that earlier marriage leads to higher fertility, likely due to prolonged childbearing periods.

Our final model further indicates that illiteracy is associated with an increased expected number of children (~23% more) compared to literate women, reinforcing the importance of education in shaping fertility choices. Regional differences also play a key role, with fewer children observed in Lisbon and other cities with populations over 10,000 compared to rural areas and small towns. Lastly, having more sons is strongly associated with larger total family sizes, suggesting a continued preference for male children.

Variable	Estimate	Exp(Estimate)
(Intercept)	0.336	1.399
literacyno	0.209	1.232
ageMarriedAbove30	-0.269	0.764
ageMarriedBelow20	0.051	1.052
regionlisbon	-0.170	0.844
regionporto	0.039	1.040
region20k+	-0.136	0.873
region10-20k	-0.049	0.952
sons	0.312	1.366

4.2 Comparison of Model Effects with Existing Literature

These findings align with existing literature. Prior studies have established that literate women are more likely to use contraception, have smaller families, and exhibit less preference for male children (Suliankatchi et al., 2012). Additionally, research has shown that rural women tend to prefer larger families, with education playing a critical role in shaping fertility preferences (Kahansim et al., 2013). Our results support these trends, highlighting that early marriage, lower education levels, and rural residence contribute to higher fertility rates, consistent with past research (Busfield, 1972).

4.3 Closing Remarks

From a policy perspective, these findings emphasize the need for targeted interventions in rural areas and among women with lower education levels. Policymakers can leverage these insights to promote female education, discourage early marriage, and enhance family planning programs, ultimately contributing to reduced fertility rates and improved maternal and child health outcomes.

References

Bates, Lisa M., Joanna Maselko, and Sidney Ruth Schuler. 2007. "Women's Education and the Timing of Marriage and Childbearing in the Next Generation: Evidence from Rural Bangladesh." *Studies in Family Planning* 38 (2): 101–12.

Bloom, David E., David Canning, and Jaypee Sevilla. 2003. *The Demographic Dividend: A New Perspective on the Economic Consequences of Population Change*. RAND Corporation.

Busfield, Joan. 1972. "Age at Marriage and Family Size: Social Causation and Social Selection Hypotheses." *Journal of Biosocial Science* 4 (1): 117–34. https://doi.org/10.1017/S0021932000008385.

Conim, C. 1986. "Evaluation of the Portugal Fertility Survey 1979-80." 81. WFS Scientific Reports. World Fertility Survey.

Kahansim, M. L., I. S. Hadejia, and M. N. Sambo. 2013. "A Comparative Study of Factors Influencing Decisions on Desired Family Size Among Married Men and Women in Bokkos, a Rural Local Government Area in Plateau State." *African Journal of Reproductive Health* 17 (1): 149–57.

Long, Jacob. 2022. *Jtools: Analysis and Presentation of Social Scientific Data*. https://CRAN.R-project.org/package=jtools.

Parr, N. 2012. "Do Children from Small Families Do Better?" *Journal of Population Research* 23: 1–25. https://doi.org/10.1007/BF03031865.

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

Suliankatchi, R., A. Kankaria, R. Roy, R. Upadhyay, P. Chinnakali, V. Chellaiyan, and S. Babu. 2012. "Effect of Literacy on Family Planning Practices Among Married Women in Rural South India." *International Journal of Medicine and Public Health* 2 (4). https://doi.org/10.5530/ijmedph.2.4.5.

Wickham, Hadley et al. 2023a. *Ggplot2: Elegant Graphics for Data Analysis*. https://CRAN.R-project.org/package=ggplot2.

——— et al. 2023b. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, and Jim Hester. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.