The Influence of Women's Literacy and Marriage Age on Fertility Rates: A Statistical Analysis Using Generalized Linear Models*

Sakura Hu and Zhanyi Wang

Introduction

The family size in a given region, particularly the number of children in each family, often reflects the local birth rate. A persistently low birth rate can lead to significant shifts in both societal implications and overall population size (Fauser et al. 2024). As many countries face a sharp decline in population, studying what factors affect family size has become a major concern for many policymakers and researchers. Previous studies have shown that some factors related to family situations might influence family size. The research of Saurabh, Sarkar, and Pandey (2013) illustrates that there is an inverse relationship between female literacy rates and crude birth rates in India (Saurabh, Sarkar, and Pandey 2013), which means that in households where women are literate, the number of children tends to be lower. In another study examining birth rate rates in India, Talwar indicates that later marriage delays the age of childbirth for women and leads to fewer children being born(Talwar 1967). However, Song's research on Korean birth rates suggests a different view from Talwar's research, showing that the fertility rate of women in their 30s is significantly higher than that of women in their 20s (Song et al. 2018). These findings highlight the complex interplay between literacy, marriage age, and childbirth rate, which vary across different cultural and socioeconomic contexts. This study, building upon the findings of the three preceding papers, will focus on Portugal and employ Generalized linear regression analysis to model the existing data from Demographic and Health Surveys (DHS) Program (2024). The primary objective is to examine the relationship between women's age at marriage, literacy rates, and fertility rates to derive meaningful insights. By investigating these factors, the study aims to provide valuable conclusions that can serve as a reference for future demographic research.

Method

This study examines how women's literacy and age at marriage influence fertility rates by modeling the number of children per family. The response variable is the number of children in a family, which is a count variable. Since count data often follows a Poisson distribution, a Poisson Generalized Linear Model is first applied.

Primary Predictor Variables of Interest:

- **ageMarried (Age at marriage)**: Previous research suggests that delayed marriage may influence fertility rates. The reference category is ages 25-30, as this aligns with findings from Song et al. (2018), which indicate that women in their 30s have the highest fertility rates.
- **literacy (Literacy status)**: Prior studies suggest that higher female literacy rates are associated with lower birth rates.

Since fertility rates naturally depend on how long a woman has been married, an offset term is included in the model to account for exposure time without treating it as a predictor variable. The offset is defined

*Code and data are available at: https://github.com/xycww/sta305.

as log(years since marriage), where months since marriage is converted to years and log-transformed. This ensures that the model estimates fertility rates rather than just child counts.

After fitting the Poisson model, the study assesses model appropriateness by comparing the variance and mean of the response variable. If the variance is significantly greater than the mean, this indicates overdispersion, suggesting that the Poisson assumption may not be appropriate. In such cases, a Negative Binomial model is applied, as it introduces an additional dispersion parameter that better accounts for variability in the data.

To evaluate the significance of predictors and the variability in the response variable, hypothesis tests are conducted for the coefficients of both the Poisson and Negative Binomial models. Predictor significance is assessed using z-values and p-values, with a significance threshold of 0.05. Variables with p-values below 0.05 are considered to have a statistically significant impact on fertility rates. Additionally, 1/sqrt(theta) is examined in the Negative Binomial model to quantify the level of overdispersion. A comparison of confidence intervals between the two models is also made to assess how adjusting for overdispersion affects the reliability of coefficient estimates.

This modeling approach provides a structured way to determine whether female literacy and age at marriage significantly impact family size, while also ensuring that the chosen statistical model appropriately accounts for data characteristics such as overdispersion.

Result

Figure 1 presents the statistical summary of the response variable "Children." The number of children per family ranges from 0 to 17, indicating that some families have no children, while others have as many as 17. The distribution is concentrated around 2 to 3 children and right-skewed, indicating that fewer families having very high numbers of children. The mean number of children is 2.26, the median is 2, and the standard deviation is 1.86.

Figure 2 provides the statistical summary of the independent variables `Age at marriage` and `Literacy Status`. Among the 5,148 samples, the majority of individuals marry between ages 20 to 25. Specifically, 1,126 individuals married between 20 to 22, and 1,468 individuals married between 22 to 25. In contrast, marriage before age 15 and after 30 is uncommon, with only 52 individuals marrying between 0 to 15 and 217 marrying after 30. The distribution of `Age at marriage` follows an approximately bell-shaped pattern, slightly left-skewed, with the peak occurring in the 22 to 25 category. This suggests that most individuals marry in early adulthood, but fewer marry very young or much later.

For `Literacy Status`, the majority of individuals are literate. Out of the 5,148 samples, 4,567 individuals reported being literate, while only 581 reported otherwise. The distribution is highly skewed, with literacy being the dominant category.

Figure1: Statistical Summary and Distribution of the Response Variable 'Children'. This figure provides an overview of the distribution of the number of children per family. The table on the left summarizes key statistics, including the mean, median, standard deviation, minimum, and maximum number of children. The histogram on the right visually represents the distribution, showing that the data is right−skewed, with most families having 2 to 3 children.

	Mean	Median	SD	Min	Max
1	2.26	2	1.86	0	17

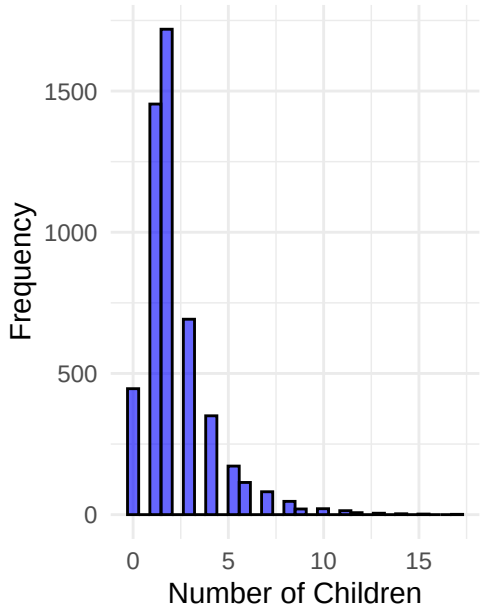
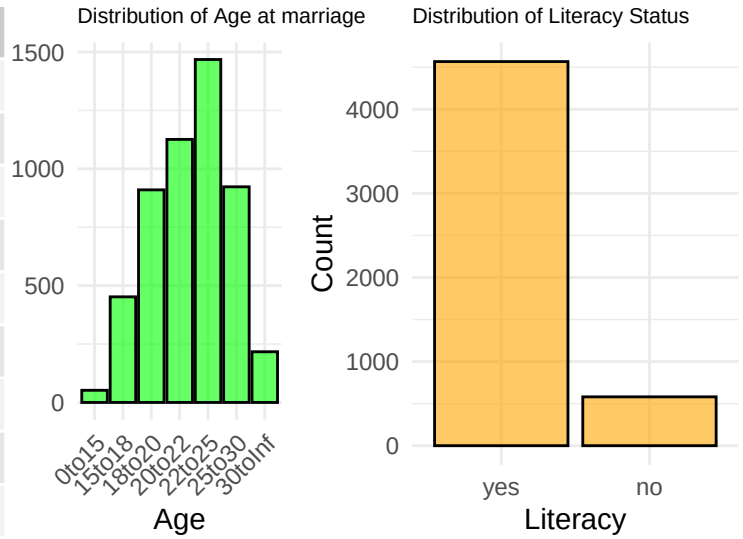


Figure 2: Statistical Summary and Distribution of Predictor Variables 'Age at marriage' and 'Literacy status'.The table on the left summarizes the number of individuals in each category, showing that the majority of individuals marry between ages 20 to 25, with 1,126 marrying between 20 to 22 and 1,468 marrying between 22 to 25. Marriages before age 15 and after 30 are less common. The bar plot for 'Age Married' illustrates a slightly left−skewed distribution, peaking in the 22 to 25 category. The bar plot for 'Literacy Status' illustrates the proportion of literate and illiterate individuals, showing that 4,567 out of 5,148 individuals in the dataset are literate.

Variable	Category	Count
Age Married	0to15	52
Age Married	15to18	452
Age Married	18to20	910
Age Married	20to22	1126
Age Married	22to25	1468
Age Married	25to30	923
Age Married	30toInf	217
Literacy	yes	4567
Literacy	no	581



To address the research question, the data was first modeled using a Poisson regression with an offset for the log-transformed years married. The offset is included to account for the duration of marriage, which may influence the number of children over time. The model is specified as follows:

$$Y_i \sim \text{Poisson}(\pi_i)$$

$$\log(\pi_i) = \beta_1 \cdot \text{literacy}_i + \beta_2 \cdot \text{ageMarried}_i + \log(O_i)$$

Where:

- π_i represents the expected number of children per individual,
- β_1 and β_2 are the coefficients for literacy and age at marriage, respectively,
- O_i is the offset term, defined as:

$$O_i = \frac{\max(1, \text{monthsSinceM}_i)}{12}$$

Table 1 presents the summary results of the Poisson regression model. The predictor variable `literacyno` (indicating illiteracy) has a statistically significant effect on the number of children. The estimated coefficient is 0.159 with a standard error of 0.024, yielding a z-value of 6.770 and a p-value < 0.001. Since this p-value is below the significance level of 0.05, there is strong evidence against the null hypothesis, suggesting that being illiterate (compared to being literate) is associated with a higher log count of children. The 95% confidence interval ranges from 1.120 to 1.228, indicating that, all else being equal, illiterate women are expected to have approximately 1.120 to 1.228 times more children than literate women. In contrast, the coefficients for age at marriage categories do not show statistically significant effects, as all levels in this category have

p-values greater than 0.05. Furthermore, their 95% confidence intervals include 1, indicating no meaningful difference in the expected number of children compared to the reference group (ages 25-30). Overall, while literacy is a statistically significant predictor, age at marriage does not appear to have a meaningful effect on the number of children in this model.

Table 1: Coefficients and Confidence Intervals for the Poisson Model. This table shows the estimated coefficients, standard errors, z-values, p-values, and the corresponding 95% confidence intervals for the Poisson regression model. The variable literacy status is a statistically significant predictor, with illiterate women expected to have 12-23% more children than literate women, holding other factors constant. In contrast, the age at marriage categories do not show significant effects on the number of children.

	Estimate	Std. Error	z value	Pr(>|z|)	LowerCI	UpperCI
(Intercept)	-1.789	0.023	-77.858	0.000	0.160	0.175
literacyno	0.159	0.024	6.770	0.000	1.120	1.228
ageMarried0to15	0.036	0.081	0.448	0.654	0.882	1.210
ageMarried15to18	0.062	0.037	1.702	0.089	0.990	1.144
ageMarried18to20	0.048	0.031	1.557	0.120	0.988	1.115
ageMarried20to22	0.016	0.030	0.528	0.598	0.958	1.077
ageMarried22to25	-0.013	0.029	-0.468	0.640	0.933	1.044
ageMarried30toInf	0.008	0.060	0.136	0.891	0.894	1.133

Table 2 presents the means and variances of literacy and age at marriage. In several cases, the variance is at least twice as large as the mean, indicating potential overdispersion. To address this, a Negative Binomial model is applied, which allows the variance to exceed the mean by introducing an additional dispersion parameter. The model is specified as follows:

$$Y_i \sim \text{Negative Binomial}(\pi_i, \theta)$$

$$\log(\pi_i) = \beta_1 \cdot \text{literacy}_i + \beta_2 \cdot \text{ageMarried}_i + \log(O_i)$$

Where:

- Y_i represents the number of children in a family,
- π_i is the expected number of children per individual,
- θ is the dispersion parameter, allowing for overdispersion in the data,
- β_1 and β_2 are the coefficients for literacy and age at marriage, respectively,
- O_i is the offset term, defined as:

$$O_i = \frac{\max(1, \text{monthsSinceM}_i)}{12}$$

Table 2: Mean and Variance of Children by Literacy and Age at Marriage. The results indicate that in certain columns, the variance is significantly larger than the mean, suggesting the presence of overdispersion.

literacy	agemarried	Mean	Variance
yes	25to30	1.97	1.98
no	25to30	3.27	5.49
yes	0to15	2.79	2.64
no	0to15	4.46	2.27
yes	15to18	2.40	4.10
no	15to18	4.31	6.22
yes	18to20	2.15	2.98
no	18to20	4.87	12.98
yes	20to22	2.12	2.68
no	20to22	3.98	7.91
yes	22to25	1.97	1.98
no	22to25	3.92	7.05
yes	30toInf	1.42	1.79
no	30toInf	1.68	3.41

Table 3 presents the summary results of the Negative Binomial (NB) regression model, which was used to model the count of children. The predictor variable literacyno (indicating illiteracy) remains statistically significant in this model. The estimated coefficient for illiteracy is 0.148, with a standard error of 0.027, yielding a z-value of 5.570 and a p-value < 0.001. This p-value, being below the significance threshold of 0.05, strongly rejects the null hypothesis and indicates that being illiterate is associated with a higher expected number of children. Specifically, being illiterate increases the expected log count of children by 0.148, holding all else constant. The 95% confidence interval (CI) for this effect ranges from 1.100 to 1.221, suggesting that illiterate women are expected to have approximately 1.100 to 1.221 times more children than literate women, all else being equal.

In contrast, the effects of age at marriage on fertility are not statistically significant in this model. The coefficients for all categories of age at marriage (relative to the reference category of 25–30 years old) have p-values greater than 0.05, indicating that the timing of marriage, once literacy is accounted for, does not significantly impact the number of children. Additionally, the 95% confidence intervals for all categories include 1, further supporting this conclusion.

The estimated overdispersion measure is 0.2645, which is calculated as 1/sqrt(theta). This measure suggests that the variance in the data exceeds the mean, a condition known as overdispersion. This justifies the use of the Negative Binomial model, which includes an additional dispersion parameter to better account for this excess variability.

In conclusion, although evidence of overdispersion is present in the data, the results from both the Poisson and Negative Binomial models are similar. Both models suggest that literacy status is a significant predictor of fertility, while age at marriage is not. Although the standard error is slightly larger in the Negative Binomial model, and the confidence intervals are slightly wider, this can be attributed to the Negative Binomial model accounting for overdispersion. These results demonstrate that while both models yield similar findings, the Negative Binomial model provides a more accurate reflection of the data's variability due to its ability to handle overdispersion.

Table 3: Coefficients and Confidence Intervals for the Poisson Model. This table shows the estimated coefficients, standard errors, z-values, p-values, and the corresponding 95% confidence intervals for the Poisson regression model. The variable literacy status is a statistically significant predictor, with illiterate women expected to have 10-23% more children than literate women, holding other factors constant. In contrast, the age at marriage categories do not show significant effects on the number of children.

	Estimate	Std. Error	z value	Pr(>|z|)	LowerCI	UpperCI
(Intercept)	-1.772	0.025	-71.292	0.000	0.162	0.178
literacyno	0.148	0.027	5.570	0.000	1.100	1.221
ageMarried0to15	0.057	0.090	0.631	0.528	0.883	1.263
ageMarried15to18	0.073	0.040	1.810	0.070	0.993	1.165
ageMarried18to20	0.059	0.034	1.730	0.084	0.992	1.134
ageMarried20to22	0.025	0.033	0.754	0.451	0.961	1.093
ageMarried22to25	-0.011	0.031	-0.339	0.735	0.931	1.052
ageMarried30toInf	0.011	0.064	0.165	0.869	0.889	1.145

Conclusion

According to the result, the negative binomial model has been chosen as the fitted model, which be present as:

$$\begin{aligned}\log(\text{Children/Year Since Married}) = &-1.772 + 0.148 \cdot \text{literacyno} \\ &+ 0.057 \cdot \text{ageMarried0to15} + 0.073 \cdot \text{ageMarried15to18} \\ &+ 0.059 \cdot \text{ageMarried18to20} + 0.025 \cdot \text{ageMarried20to22} \\ &- 0.011 \cdot \text{ageMarried22to25} + 0.011 \cdot \text{ageMarried30toInf}\end{aligned}$$

Among the predictors, literacy shows a statistically significant effect. The p-value for the "literacyno" variable < 0.001, indicating strong evidence against the null hypothesis. The estimated coefficient for literacy is 0.148, and exponentiating this coefficient (eˆ0.148 = 1.16) shows that illiterate women aged 25 to 30, on average, experience a 16% higher birth rate compared to literate women, holding all other variables constant. The 95% confidence interval (1.100 to 1.221) suggests that the true effect is consistently positive, as it does not include 1. In contrast, age at marriage does not appear to significantly influence the number of children, with p-values greater than 0.05 across all categories. This suggests that, within this dataset, the timing of marriage does not have a meaningful effect on fertility rates, once literacy is accounted for.

These findings diverge from some literature that suggests the timing of marriage plays a significant role in fertility rates. For example, Talwar's research suggests that delaying marriage generally leads to lower fertility (Talwar 1967). However, this study's findings align with the understanding that the relationship between marriage age and fertility is context-dependent. As this study focuses on Portugal, where cultural and socioeconomic factors may differ significantly from those in other countries like India, it is reasonable to observe these discrepancies. Furthermore, the significant effect of literacy is consistent with previous research, such as (Saurabh, Sarkar, and Pandey 2013), which found that illiterate women tend to have higher fertility rates.

This study provides valuable insights into the relationship between literacy and fertility in Portugal. The findings suggest that literacy is a significant predictor of fertility, while age at marriage does not significantly influence birth rates in this context. The results have important implications for policymakers, as they highlight the potential role of women's education in addressing population growth. By improving educational opportunities for women, policymakers could foster greater control over fertility rates, ultimately contributing to sustainable population management.

References

Demographic and Health Surveys (DHS) Program. 2024. "World Fertility Survey (WFS) Data." https://wfs.dhsprogram.com/.

Fauser, B. C. J. M., G. D. Adamson, J. Boivin, G. M. Chambers, C. de Geyter, S. Dyer, M. C. Inhorn, et al. 2024. "Declining Global Fertility Rates and the Implications for Family Planning and Family Building: An IFFS Consensus Document Based on a Narrative Review of the Literature." *Human Reproduction Update* 30 (2): 153–73. https://doi.org/10.1093/humupd/dmad028.

Saurabh, S., S. Sarkar, and D. K. Pandey. 2013. "Female Literacy Rate Is a Better Predictor of Birth Rate and Infant Mortality Rate in India." *Journal of Family Medicine and Primary Care* 2 (4): 349–53. https://doi.org/10.4103/2249-4863.123889.

Song, J.-E., J.-A. Ahn, S.-K. Lee, and E. H. Roh. 2018. "Factors Related to Low Birth Rate Among Married Women in Korea." *PLOS ONE* 13 (3): e0194597. https://doi.org/10.1371/journal.pone.0194597.

Talwar, P. P. 1967. "A Note on Changes in Age at Marriage of Females and Their Effect on the Birth Rate in India." *Eugenics Quarterly* 14 (4): 291–95. https://doi.org/10.1080/19485565.1967.9987739.