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 Sarma and Soni 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 birthrate 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 right-skewed, with most families having 2 to 3 children. The mean number of children is 2.26, the median is 2, and the standard deviation is 1.86, reflecting moderate variability. The histogram shows that the data is concentrated around 2 to 3 children, with fewer families having very high numbers of children.

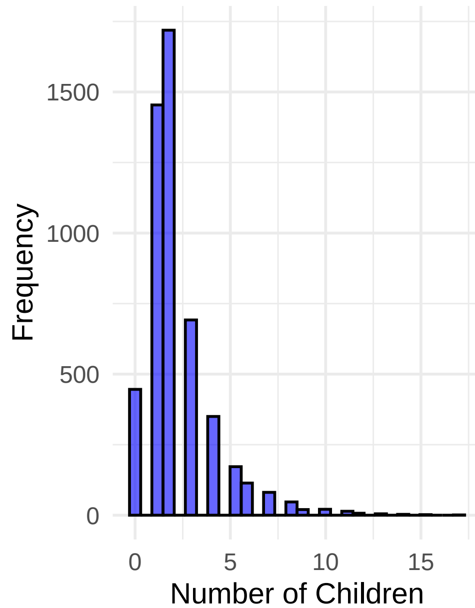
Figure 2 provides the statistical summary of the independent variables, "Age Married" and "Literacy." Among the 5,148 samples, the majority of individuals marry between ages 20 to 25. Specifically, 1,126 individuals married between 20 to 22, and 1,468 individuals married between 22 to 25. In contrast, 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 Married" follows an approximately bell-shaped pattern, slightly left-skewed, with the peak occurring in the 22 to 25 category.

For "Literacy," 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.

These statistical summaries provide key insights into the dataset, highlighting the central tendency, spread, and shape of each variable while contextualizing the findings within the sample population.

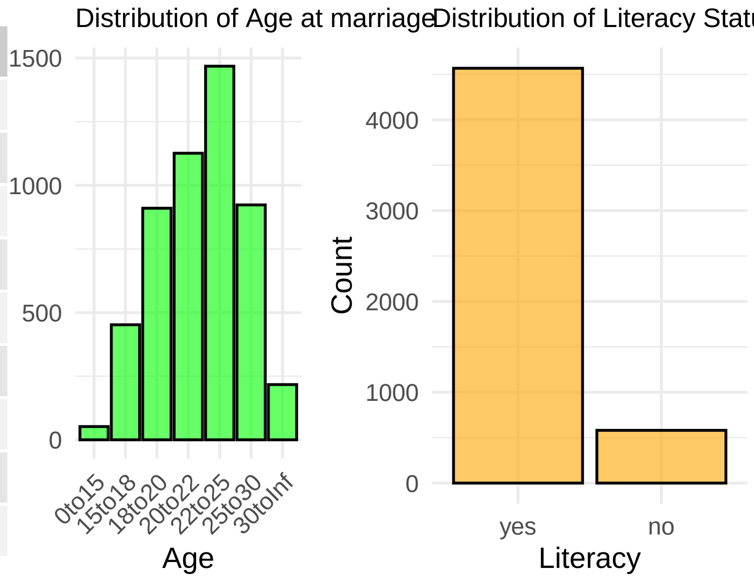
Figure1: Statistical Summary and Distribution of the Response Variable 'Children'. This figure provides an overview of the distribution of the number of children per family. The table on the left summarizes key statistics, including the mean, median, standard deviation, minimum, and maximum number of 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



re 2: Statistical Summary and Distribution of Predictor Variables 'Age at marriage' and 'Literacy stat
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.
bar plot for 'Age Married' illustrates a slightly left−skewed distribution, peaking in the 22 to 25 categ
The bar plot for 'Literacy Status' illustrates the proportion of literate and illiterate individuals, s
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 answer 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:

$$Y_i \sim \text{Poisson}\left(\exp\left(\text{offset}\left(\log\left(\max(1, \text{monthsSinceM}_i)/12\right)\right) + \beta_1 \cdot \text{literacy}_i + \beta_2 \cdot \text{ageMarried}_i\right)\right)$$

Where:

- children is the count of children,
- $\text{offset}(\log(\max(1, \text{monthsSinceM})/12))$ adjusts for the number of years married,
- literacy is a binary variable indicating whether the person is literate or not,
- ageMarried is the categorical variable representing the age at which the person married, with "25to30" as the reference category.

Figure 3 shows the summary of this model. According to the results, the variable `literacyno` (indicating illiteracy) has a significant effect on the number of children, with an estimate of 0.159 and a standard error of 0.024. The corresponding z-value is 6.770, and the p-value is less than 0.005, indicating strong evidence against the null hypothesis and suggesting that being illiterate (compared to being literate) is associated with a higher log count of children. Specifically, being illiterate increases the expected log count of children by 0.159, all else being equal. In contrast, the coefficients for the `ageMarried` categories (relative to the reference group of 25-30 years) do not show significant effects. For example, the coefficient for `ageMarried15to18` is 0.062 with a standard error of 0.037, yielding a z-value of 1.702 and a p-value of 0.089, which is above the standard significance level of 0.05, indicating that this result is not statistically significant. Similarly, other age categories, such as `ageMarried20to22` and `ageMarried30toInf`, show p-values of 0.598 and 0.891, respectively, suggesting no significant relationship between age at marriage and the number of children.

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: Table: Log-rate ratios and 95% confidence intervals for the Negative Binomial model relating school institution type, region, and interactions to the number of violent crimes.

	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

Figure 4 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. Since the Poisson model assumes equal mean and variance, this violation suggests the need for a more flexible model. 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:

$$Y_i \sim \text{NegBin}\left(\exp\left(\text{offset}\left(\log\left(\max(1, \text{monthsSinceM}_i)/12\right)\right) + \beta_1 \cdot \text{literacy}_i + \beta_2 \cdot \text{ageMarried}_i\right), \theta\right)$$

Where:

- children represents the number of children,
- $\text{offset}\left(\log\left(\max(1, \text{monthsSinceM})/12\right)\right)$ adjusts for the number of years married,
- literacy is a binary variable indicating whether the person is literate or not,
- ageMarried is the categorical variable representing the age at which the person married.

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

The summary of this negative binomial is shown in Figure 5. According to the model results, literacy has a statistically significant effect on the number of children. Specifically, the coefficient for ‘literacyno’ (indicating illiteracy) is 0.148 with a standard error of 0.027, yielding a z-value of 5.570. The associated p-value is less than 0.05, providing strong evidence against the null hypothesis and suggesting that being illiterate is associated with a higher expected number of children. This suggests that, holding all else constant, being illiterate increases the expected log count of children by 0.148.

In contrast, the effects of age at marriage do not appear to be statistically significant in this model. The coefficients for different ageMarried categories are estimated relative to the reference group 25–30 years old. with all levels in ageMarried has p value greater than 0.05, which suggest that in this population, the timing of marriage alone may not be a key determinant of fertility outcomes once other factors, such as education, are taken into account.

The estimated dispersion measure is 0.2645, indicating that the variance exceeds the mean, which violates the Poisson model's assumption of equidispersion. This justifies the use of a Negative Binomial model, which introduces an additional dispersion parameter to account for the excess variability.

Table 3: Table: Log-rate ratios and 95% confidence intervals for the Negative Binomial model relating school institution type, region, and interactions to the number of violent crimes.

	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

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.