

# An Analysis of effect of literacy and age of a marriage on family size using Portual data

Kyungrok Park

Yushen Chen

Feb 4, 2025

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Methods</b>	<b>1</b>
<b>3</b>	<b>Results</b>	<b>2</b>
<b>4</b>	<b>List of Figures and Tables</b>	<b>3</b>

## 1 Introduction

Fertility trends have been closely linked to the demographic transition in Portugal (Pimentel & Sousa Gomes, 2022). With the background of the population aging in Portugal, the Portuguese government started to pay more attention to the low fertility rate (Pimentel & Sousa Gomes, 2022). Since the 1970s, the birth rate in Portugal has decreased sharply and the impact of the COVID-19 crisis led to a historic decline in the fertility level (Pimentel & Sousa Gomes, 2022). Thus, it is essential to study the family size and the factors that influence the family size in Portugal to address the serious issue of population aging in Portugal. This research will analyze a fertility survey conducted in Portugal in 1979 which covers the national population except for Azores and Madeira, and focuses on ever-married women aged between 15 and 49.

Kebede et al. (2019) indicate that the fertility rates in sub-Saharan countries kept declining until 2000. This interruption of the decreasing birth rate seems to be connected to the disruptions in women's education (Kebede et al., 2019). Meanwhile, the decline in fertility in Indonesia and Uganda shows that age at marriage also affects the fertility rates. Ekoriano et al. (2023) state women in Indonesia who married before the age of 25 tend to have a longer reproductive period which could result in giving birth to more children compared with the women who married after 25 years old. Likewise, Ariho & Kabagenyi (2020) indicate that raising the age of first marriage can significantly decrease the birth rate in Uganda. Both the fertility data from Indonesia and Uganda suggest women delaying their marriage could lead them to have relatively fewer children, which results in a smaller family size. So, it is significant to consider women's literacy level and their marriage ages while studying the family size, as higher illiteracy rates often correlate with increased birth rates, and earlier marriage ages can also contribute to larger families.

Therefore, this research aims to examine how literacy and age at marriage influence family size in Portugal by using the Poisson regression and negative binomial regression. Additionally, it will explore the other variations, such as the impact of the number of sons and regional differences on family size

## 2 Methods

To examine the effects of literacy and age at marriage on family size, this study analyzes two types of regression models: Poisson regression and Negative Binomial regression. Although the temporal component

of the response variable, family size, is implicit, it can be interpreted as an accumulation over a woman's reproductive lifetime, making it suitable for modeling as count data within a Poisson framework.

Regarding the explanatory variables, the inclusion of the binary variable (literacy) and the categorical variable (age at marriage) is essential, as they are the primary focus of this research. Additionally, given that in Europe, fertility takes place later and is lower in cities than in rural areas (Riederer & Beaujouan, 2023), another categorical variable, regional population size, is incorporated as a key determinant of urban versus rural classification. Furthermore, in son-preferring societies, such as Korea, smaller families tend to have a higher sex ratio (number of males per 100 females) than larger families (Park, 1983). While it remains uncertain whether Portugal exhibits a strong preference for sons, this study introduces the proportion of sons among children as an additional explanatory variable. This inclusion aims not only to examine the impact of sex composition on family size but also to explore whether the data suggest any indirect indication of son preference in Portugal. Further justification and illustration of the response and explanatory variables will be discussed in the Introduction section.

In addition to the standard Poisson regression, a Poisson regression with an offset (months since marriage) is analyzed to account for the accumulation of family size over a woman's reproductive lifetime and differences in exposure time. Incorporating this offset allows the model to estimate birth rates rather than the total number of children. To address overdispersion, a Negative Binomial regression with an offset (months since marriage) is also examined. As a result, three main models are analyzed: Poisson regression without an offset, Poisson regression with an offset, and Negative Binomial regression with an offset.

Finally, the significance of predictors in each regression model will be assessed using confidence intervals and p-values, while residual plots will be used to evaluate the extent of variation in the response variable.

### 3 Results

To justify which regression model is appropriate for our analysis, the nature of variables are analyzed. From ??, each distribution highlights that every density plot of family size (number of children) for each condition of literacy and age of marriage has approximately a right-skewed distribution and it shows that responses (family size) can be reasonably modeled with a poisson distribution as the key variables, literacy status and age of marriage, are introduced. Furthermore, different means depending on literacy status for each plot and the fact that density of family size for illiterate females becomes greater than literate females after a certain family size (around 3~4) indicates a potential effect of literacy status of a female on family size. Additionally, different mean values for each age of marriage for same literacy status suggest a possible effect of age of marriage on family size.

Beside the regular poisson model for the family size, the research also includes years since marriage as an offset. (since the family size can be substantially determined by the years since marriage as a female with a longer marriage has a higher chance to have more children) Moreover, from @tbl-tbl2, as the variance-mean ratio of family size for most of conditions (except literate females with an age of marriage at 0-15 and illiterate females with an age of marriage at 0-15) are greater than 1, it is concluded that a negative binomial model is more suitable model than the poisson model since it can explain overdispersion of a poisson model.

As a result, from @fig-fig2, it shows that the poisson model without the offset have remarkably different estimates than the poisson model and negative binomial model with the offset as the literate female is set as a reference category. Alongside the estimates, the confidence interval with 95% significance level also shows a notable difference between the poisson model without the offset and rest of two models with the offset. Especially, most of age categories for poisson and negative binomial models with the offset include a value of zero which indicates no statistical significance of 95% confidence interval. Additionally, since most of estimates follow the same trend (the estimate decreases from poisson model without the offset to the poisson model with the offset, then increases from the poisson model with the offset to negative binomial model), it confirms that the extra variability in birth rates is accounted. Finally, seeing that the proportion of sons is not significant for all three models, it is excluded from the regression model.

To finalize which model should be used in this research based on the characteristics of data, the log-likelihood

test from @tbl-tbl3 is performed. Since the log-likelihood values for negative binomial model with the offset (-8909.044) is higher than the log-likelihood values for two poisson models(-9239.154, -8944.741), it is determined to analyze the negative binomial with the offset as our final regression model.

Finally, the interaction term between two main explanatory variables(age of marriage and literacy status) is not included in our final model. From @fig-fig3, it highlights that the distribution of the response variable appears similar across literacy groups within each age at marriage category, with no substantial variation in median, spread, or outliers that would indicate a strong interaction effect. This suggests that the relationship between age at marriage and the response variable does not significantly differ by literacy status. Therefore, including an interaction term between literacy and age at marriage in the regression model is unlikely to provide additional explanatory power.

## 4 List of Figures and Tables

Figure 1: Density of Number of Children by Age at Marriage and Literacy

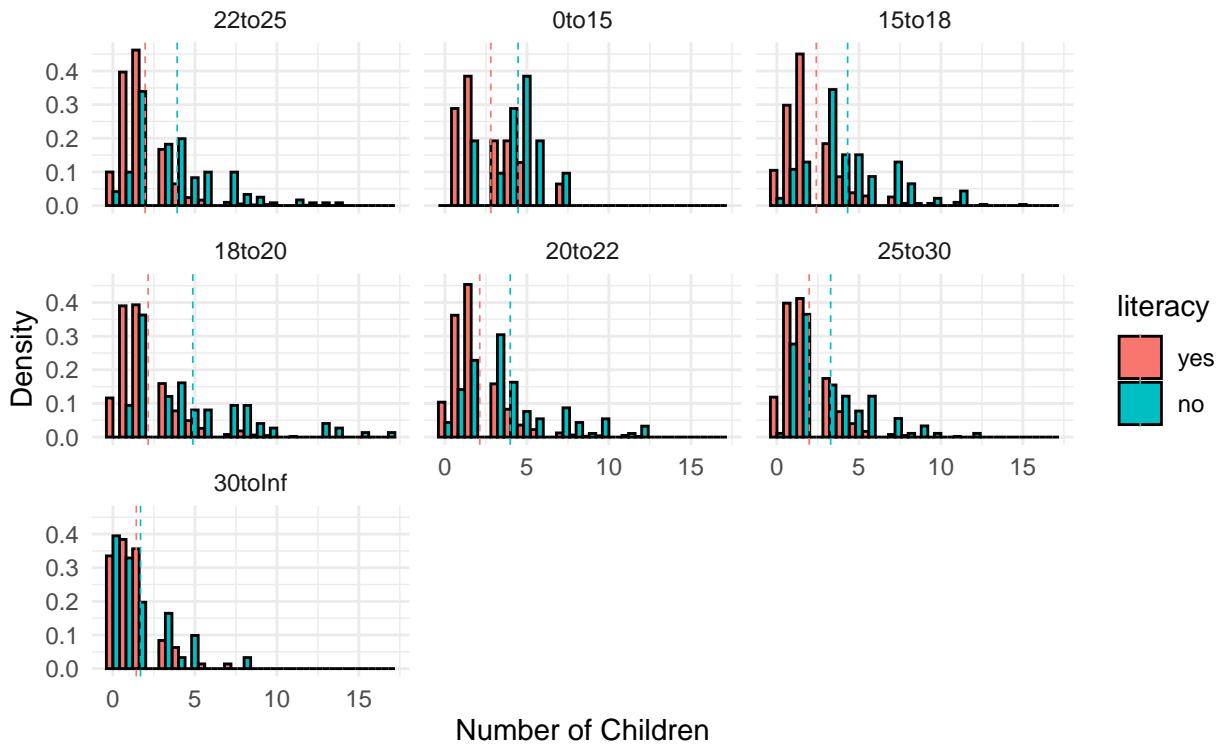
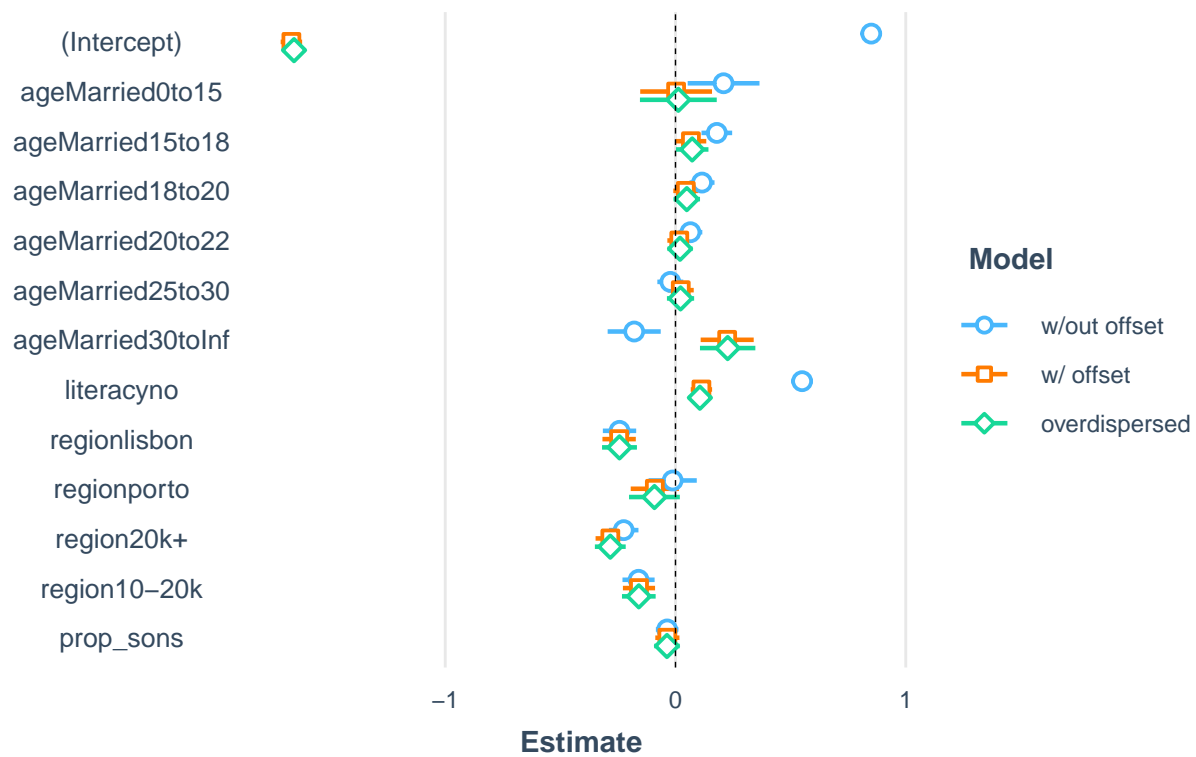


Figure 1: A Density plot which shows the distribution of Family size based on the Literacy Status and Age of Marriage. Notably, the density of family size for each plot has approximately a right-skewed distribution. Furthermore, although the centre of the distribution is approximately same (around the family size of 2-3), the density of family size for illiterate females becomes greater than literate females after a family size of 3, and it suggests a potential literacy effect on number of family size (children)

**Figure 2: Comparison of Regression Models**



**Figure 3: Distribution of Birth Rate by Literacy and Age at Marriage**

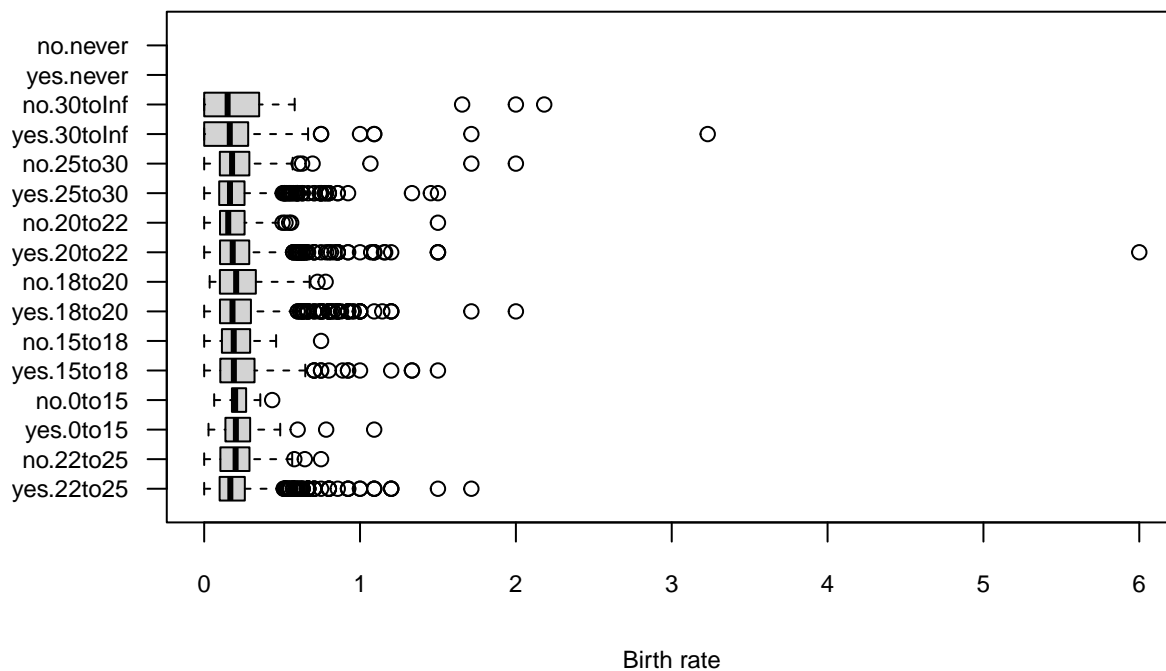


Table 1: Mean and Variance of children by Literacy Status and Age at Marriage

literacy	ageMarried	mean_family_size	var_family_size	vmr
yes	22to25	1.965831	1.976795	1.0055771
yes	0to15	2.794872	2.641026	0.9449541
yes	15to18	2.398477	4.102898	1.7106264
yes	18to20	2.148103	2.984166	1.3892101
yes	20to22	2.122651	2.676031	1.2607024
yes	25to30	1.966667	1.975402	1.0044416
yes	30toInf	1.424581	1.785010	1.2530073
no	22to25	3.920530	7.046976	1.7974550
no	0to15	4.461538	2.269231	0.5086207
no	15to18	4.310345	6.217786	1.4425263
no	18to20	4.870968	12.983170	2.6654189
no	20to22	3.982609	7.911976	1.9866314
no	25to30	3.274336	5.486568	1.6756274
no	30toInf	1.684211	3.411095	2.0253378

*Note:*

hello

Table 2: Log-Likelihood Comparison of Models

Model	Log_Likelihood
Poisson (No Offset)	-9239.154
Poisson (With Offset)	-8944.741
Negative Binomial (With Offset)	-8909.044

Table 3: Exponentiated Model Coefficients (Incident Rate Ratios)

Predictor	IRR (Exp Coeff)	Std. Error	z-Statistic	p-Value	2.5% CI (IRR)	97.5% CI (IRR)
<b>(Intercept)</b>	0.1805386	0.0209168	-81.8388728	0.0000000	0.1732869	0.1880938
<b>literacy</b> no	1.1260636	0.0264199	4.4938836	0.0000070	1.0692377	1.1859096
<b>ageMarried</b> 0to15	1.0576504	0.0888719	0.6306815	0.5282488	0.8885753	1.2588965
<b>ageMarried</b> 15to18	1.0862735	0.0373453	2.2158919	0.0266989	1.0096035	1.1687659
<b>ageMarried</b> 18to20	1.0674649	0.0303020	2.1545296	0.0311987	1.0059131	1.1327829
<b>ageMarried</b> 20to22	1.0298949	0.0287769	1.0236232	0.3060133	0.9734148	1.0896521
<b>ageMarried</b> 25to30	1.0164150	0.0310174	0.5249219	0.5996375	0.9564648	1.0801229
<b>ageMarried</b> 30toInf	1.0286826	0.0623122	0.4538264	0.6499538	0.9104187	1.1623090
<b>region</b> lisbon	0.7584540	0.0396004	-6.9815663	0.0000000	0.7018128	0.8196665
<b>region</b> porto	0.9077321	0.0580393	-1.6679382	0.0953280	0.8101297	1.0170934
<b>region</b> 20k+	0.7449670	0.0350392	-8.4024557	0.0000000	0.6955232	0.7979257
<b>region</b> 10-20k	0.8509668	0.0383936	-4.2033644	0.0000263	0.7892815	0.9174730