Assignment 1

STA303: Methods of Data Analysis 2

Christoffer Tan (1008740445), Janis Joplin (1009715051)

Introduction

The number of children in a family varies widely from one household to another, making it crucial to understand the factors influencing family size. In Portugal—historically regarded as poorer by European standards—it is particularly important to consider both urban-rural differences and socioeconomic variables when examining childbearing decisions. Previous research points to multiple drivers of fertility: Najera et al. (1992) found that a family's decision to stop having children is associated with achieving a balanced gender composition, highlighting the influence of sociocultural preferences. Namboodiri (1970) showed that economic conditions shape reproductive behavior by imposing direct and indirect costs on parents, suggesting that wealth alone does not necessarily lead to larger families. Meanwhile, Zarate (1967) observed that rural areas in Mexico have higher family sizes, driven partly by earlier marriages and lower literacy rates, emphasizing the role of demographic and social structures.

Building on these findings, this study focuses on Portugal in 1979 and examines **how literacy and age at marriage affect family size**, while also assessing whether significant variability remains after accounting for these factors. We will use count-based generalized linear models, such as Poisson or Negative Binomial regressions, to analyze the number of children per family. The results will offer insights into whether the observed family sizes align with existing literature and clarify how literacy and age at marriage relate to fertility outcomes.

Methods

We are using data from a 1979 Portuguese fertility survey to explore how literacy and age at marriage might shape family size. Because previous research indicates that rural families might have more children, we consider three key predictors: mother's literacy status (literate vs. non-literate), age at marriage (under 20 vs. 20+), and region (rural vs. urban). Our response variable is the number of children, which is typically right-skewed, so we will begin with a **baseline Poisson model** that includes only literacy and age at marriage. To ensure completeness, we exclude observations with missing values in these key variables.

We then expand this to a **Poisson model with interaction term** to check whether the effect of literacy on family size might be different for mothers who married younger (under 20) versus older (20+). After that, we add rural/urban region as a potential confounding variable and account for how many years have passed since marriage with an offset, resulting in a **Poisson model with the interaction term, confounding, and offset variables**. We consider a predictor to be significant if the p-value is less than our chosen significance level, 0.05.

Finally, because Poisson models assume the mean and variance of the count data are the same, we will check for the presence of overdispersion in our data. If the variance is significantly higher than the mean, we will consider an **overdispersed model** using a Negative Binomial regression instead. Comparing these models will help us better understand the relationship between literacy, age at marriage, and region in relation to family size.

Results

To understand the distribution of key variables, we conducted an exploratory data analysis (EDA) in Figure 1, focusing on family size, literacy status, age at marriage, and residential region.

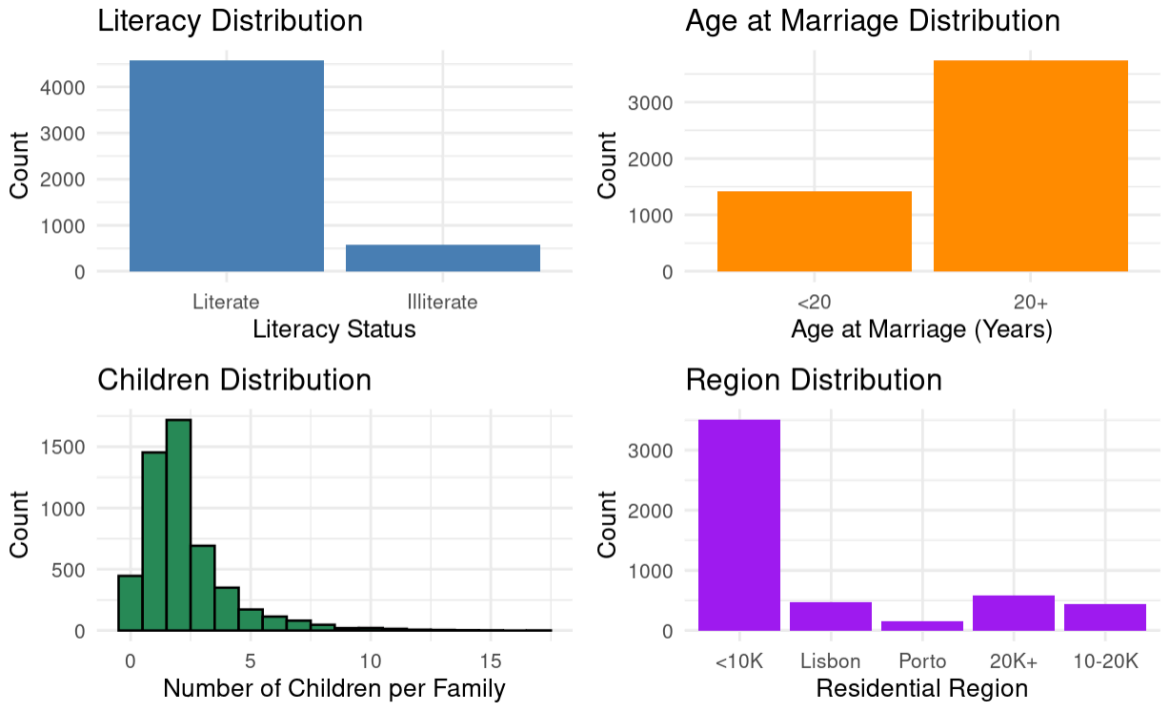

Figure 1: Distributions of literacy status, age at marriage, family size, and residential region. The number of children per family follows a right-skewed distribution, with most families having one to three children. Most individuals marry at 20 or older, while a smaller group marry younger. Literacy levels are high, with 85% of individuals being literate. Most of the sample resides in rural areas, where family sizes tend to be higher.

Baseline Poisson Model

We first fit a Poisson regression model to examine the effect of literacy and age at marriage on family size. The results indicate that illiterate individuals have 1.88 times more children than literate individuals ($\sigma = 0.02349, z = 26.789, p < 0.005, \alpha = 0.05$), and marrying before 20 increases the mean number of children by a factor of 1.16 ($\sigma = 0.0201, z = 7.582, p < 0.005, \alpha = 0.05$). However, this model assumes that the effect of age at marriage is the same for both literate and illiterate individuals.

To test whether early marriage affects family size differently based on literacy, we introduced an interaction term in the Poisson model. The results confirm that while illiterate individuals generally have more children, the effect of early marriage is even stronger for them ($\beta = 0.141, \sigma = 0.05, z = 2.824, p = 0.005, \alpha = 0.05$), confirming patterns shown in Figure 2.

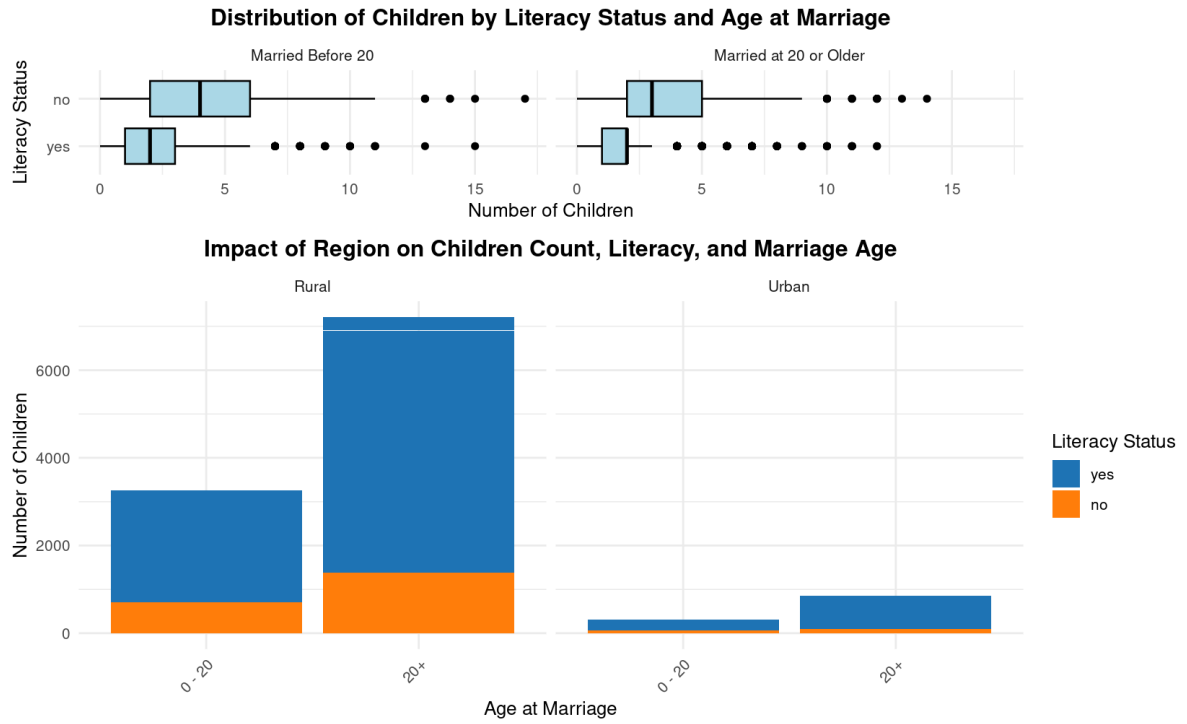

Figure 2: Distributions of the number of children in the family by literacy status, age at marriage, and residential region. The number of children per family varies by literacy and marriage age, with illiterate individuals and those married before 20 generally having more children. The boxplots reveal right-skewed distributions, with higher variability in the number of children among illiterate individuals. The bar chart highlights a stark contrast between rural and urban areas, where rural families, particularly those with illiterate members, tend to have significantly larger family sizes. This underscores the combined impact of literacy, marriage age, and regional differences on family sizes.

Adjusting for Marital Duration

Since longer marriages provide more opportunities for childbearing, we included months since marriage as an offset variable to model family size relative to time. To align birth rates with yearly intervals, we scaled months since marriage by dividing by 12. After this adjustment, the effects of literacy and age at marriage decreased but remained significant, while the interaction term became non-significant ($z = -0.162, p = 0.871, \alpha = 0.05$), suggesting that the previously observed interaction was due to differences in marital duration rather than a direct effect of literacy on the relationship between age at marriage and fertility.

Accounting for Regional Differences

Given that rural families tend to have more children and illiteracy is more prevalent in rural areas, we tested whether region (urban vs. rural) confounds the observed relationships. Figure 2 shows that illiterate individuals are more common in rural areas, and rural families tend to have more children overall. After adding region as a predictor, the effects of literacy and age at marriage on family size slightly decreased but remained significant. Illiterate individuals have 1.17 times more children than literate individuals ($\sigma = 0.029, z = 5.374, p < 0.001, \alpha = 0.05$), and marrying before 20 increases family size by a factor of 1.05 ($\sigma = 0.023, z = 2.243, p = 0.025, \alpha = 0.05$). Living in an urban area reduces the mean number of children

by a factor of 0.84 ($\sigma = 0.031, z = -5.661, p < 0.001, \alpha = 0.05$), confirming regional differences in fertility.

Assessing Overdispersion

To test whether the Poisson model was appropriate, we examined the mean and variance of family size across groups defined by literacy, region, and age at marriage (Table 1).

Literacy	Age Married	Area Type	Mean	Variance
Literate	20 +	Rural	2.03	2.29
Illiterate	20 +	Rural	3.63	7.12
Literate	0 – 20	Rural	2.30	3.47
Illiterate	0 – 20	Rural	4.64	10.25
Literate	20 +	Urban	1.72	1.52
Illiterate	20 +	Urban	2.77	4.36
Literate	0 – 20	Urban	1.84	2.19
Illiterate	0 – 20	Urban	4.67	3.70

Table 1: Mean and variance of family size across groups defined by literacy status, age at marriage, and residential region. In all groups, the variance exceeds the mean, indicating overdispersion in the data, which suggests that family sizes vary due to cultural, economic, or personal factors, making the Poisson assumption of equal mean and variance inappropriate.

To account for this, we fit a Negative Binomial model, which introduces a dispersion parameter to capture extra variability in family size.

Variable	Estimate (exp)	Std. Error	2.5% CI	97.5% CI
Intercept	0.174	0.014	0.169	0.179
Illiterate	1.155	0.032	1.085	1.231
Married Before 20	1.059	0.025	1.009	1.112
Urban Area	0.840	0.033	0.786	0.897
Illiterate × Married Before 20	0.993	0.057	0.888	1.110

Table 2: Final model estimates from the Negative Binomial regression, showing exponentiated coefficients (interpreted as multiplicative effects on family size), standard errors, and 95% confidence intervals. Illiteracy and early marriage are associated with larger family sizes, while urban residence corresponds to fewer children. The non-significant interaction term suggests that the effect of early marriage on family size does not significantly differ based on literacy status.

The Negative Binomial model provides the best fit, producing wider confidence intervals than the Poisson models, reflecting its ability to capture greater variation in family sizes.

The dispersion parameter of this model ($\sigma = 0.26, 95\%$ CI : $[0.23, 0.29]$) confirms that family sizes differ across individuals, possibly due to cultural or economic influences. Given these findings, the Negative Binomial model is preferred, as it corrects for overdispersion while still accounting for regional influences on fertility.

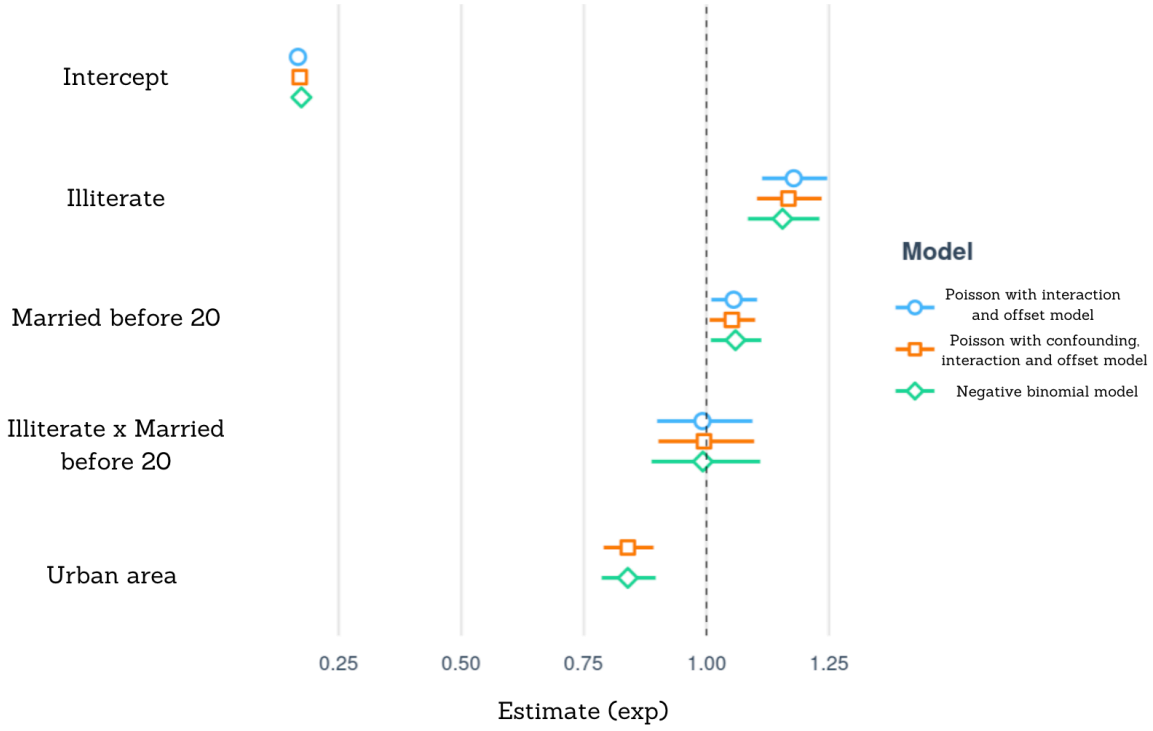

Figure 3: Exponentiated coefficient estimates and 95% confidence intervals from three models: a Poisson model without confounding, a Poisson model including region as a confounder, and a Negative Binomial model accounting for overdispersion. The wider confidence intervals in the Negative Binomial model reflect its ability to capture unobserved variability in family sizes, suggesting that the Poisson models may have underestimated dispersion in the data. Observe that literacy, age at marriage, and residential region are all significant predictors of family size.

Conclusion

Our final Negative Binomial model confirms that literacy and age at marriage both play a significant role in determining family size, even after accounting for regional differences and overdispersion. Illiterate women tend to have 1.15 times more children than literate women ($\sigma = 0.032, z = 4.482, p < 0.001, \alpha = 0.05, 95\%$ CI : $[1.085, 1.231]$), and marrying before 20 increases family size by a factor of $1.06 (\sigma = 0.025, z = 2.327, p = 0.02, \alpha = 0.05, 95\%$ CI : $[1.009, 1.112])$. Regional differences also persist, with women in urban areas having 16% fewer children than those in rural areas ($\sigma = 0.033, z = -5.223, p < 0.001, \alpha = 0.05, 95\%$ CI : $[0.786, 0.897]$). The interaction between literacy and early marriage was not significant ($z = -0.128, p = 0.898, \alpha = 0.05$) in the final model, suggesting that early marriage has a similar effect on fertility regardless of literacy status once we account for marital duration and regional factors.

These findings align with previous research showing that education and rural-urban differences influence fertility patterns. Najera et al. (1992) emphasized the role of social factors in family planning, which may explain why illiterate women have more children, possibly due to limited access to education and contraception. Similarly, Zarate (1967) found that rural areas tend to have higher family sizes, which is consistent with our findings. However, our results diverge slightly from Namboodiri (1970), who argued that economic constraints primarily drive fertility. While economic conditions likely matter, literacy remains a strong predictor of family size even after controlling for regional differences, suggesting that education has an independent effect on fertility choices.

Overall, our study highlights the importance of literacy and age at marriage in shaping fertility trends and family size. It also reinforces the need to consider regional differences, as rural families still tend to have more children. These insights could be valuable for policymakers focused on family planning and education initiatives, particularly in areas with high birth rates and lower literacy levels. Future studies could further investigate how cultural and economic factors interact with literacy and marriage age to influence fertility decisions.

Bibliography

- Nájera, C., Aparisi, M. L., & Gómez, F. (1992). *Sex ratio and factors influencing family size in a human population from Spain*. Behavior Genetics, 22(5), 531–543. https://doi.org/10.1007/BF01074306
- Namboodiri, N. K. (1970). *On the relation between economic status and family size preferences when status differentials in contraceptive instrumentalities are eliminated*. Population Studies, 24(2), 233–239. https://doi.org/10.1080/00324728.1970.10406126
- Zarate, A. O. (1967). *Some factors associated with urban-rural fertility differentials in Mexico*. Population Studies, 21(3), 283–293. https://doi.org/10.1080/00324728.1967.10406104