STA303 Assignment 1

Rebecca Li & Ebrahim Mohamud

Introduction

Family size is a powerful reflection of a country's social and economic landscape, making it a crucial area of study. Family size differentials influence birth rates, fertility rates, and population growth—key drivers of economic stability and social welfare. Analyzing the factors that shape family size provides valuable insight into demographic shifts leading to effective policy implementation.

While fertility and personal preferences are well-established factors behind the number of children that a couple has, the influences behind family size span beyond personal or biological reasons, often tracing to socio-economic and time factors. Freedman (1963) argues that high birth rates in underdeveloped countries can be attributed to social factors such as the age of marriage and societal norms. In underdeveloped countries, children are used as sources of labor and are less likely to survive to adulthood (Freedman, 1963). These findings are relevant to the context of the present report, which analyzes data from the underdeveloped country of Portugal to investigate the primary factors behind family size, allowing for more accurate projections of future demographic trends.

Freedman (1963) also highlights the importance of age of marriage on the number of children had, citing that underdeveloped areas witness high birth rates to compensate for high mortality rates prompting people to marry and having children at a younger age. Furthermore, related studies of undeveloped countries demonstrate that literacy also play a significant role. In a study of 178 males in Chakwal city, Punjab, Mahmood et al. (2016) found that 19.2% of uneducated men were in favour of smaller families compared to a 77.0% of educated men. Together, these two factors played a primary role in motivating our research question.

Finally, internal factors such as differences between regions also have significant effects. Gutkind (1962) conducted an observational study of married African women in rural and urban areas, concluding that urban populations tend to have smaller families than rural populations, attributing this phenomenon to high cost of urban living. These results encouraged us to consider an additional variable of *region* in our model. The present report expands on the results from the aforementioned studies, using them to guide our analysis of fertility data from a 1979 study of Portuguese families in order to answer the motivating research question: "How do literacy, region, and age of a marriage affect family size?". Our analysis employs a Generalized Linear Regression Model (GLM) to explain variations in family size based on the aforementioned variables. We will justify our model choice, interpret the results, and discuss potential limitations, leading to a comprehensive answer to our research question.

Methods

Prior to our analysis, we performed simple data wrangling by redefining the age brackets to ensure each level contained a roughly equal number of observations. This adjustment helped prevent sparsely populated categories, reducing the risk of unstable estimates and enhancing the reliability of our analysis. To answer our research question, we required a functional form for the GLM being employed. Since the number of children is a count variable distributed over a unit of time, an appropriate model for this data is a Poisson model. We conducted a preliminary exploratory data analysis by visualizing the data through a histogram and observed whether the data satisfied the properties of a Poisson model.

Additionally, we compared the mean and variance of each age group within the *age married* variable to assess whether they were equate; an assumption behind the Poisson model. Overdispersion is present when they don't equate, therefore, we decided to create two models: a Poisson regression model featuring *age married*, *literacy*, and *region* as predictor variables and a Negative Binomial model which features the same predictors while accounting for overdispersion.

As previously discussed, we included *region* in our model because of the surrounding support from the literature. Additionally, we included an offset in both models to account for variation in exposure times. The offset ensures that each observation yields an estimated rate with a standard time unit of one year, thereby maintaining a consistent interpretation of the intensity of births across our data.

We conclude our analysis by comparing the two models based on their estimated coefficients, p-values, confidence intervals, and standard errors. This evaluation allows us to determine which model better captures the relationship between the predictors and the number of children. The results from the better model will be interpreted, providing meaningful insights that address our research question.

Results

Our dataset captured key demographic details from 5,148 observations of married couples from 1979 Portugal. Since we are interested in how age at marriage, literacy, and region contribute to family size, we plotted a histogram visualizing our data with respect to our response (the number of children). The histogram (Figure 1) exhibits a heavy right skew with a mean of 2.26 and a standard deviation (spread) of 1.86. To further support our GLM choice, we plotted a Poisson fitted line with parameters based on the global mean, and group means from the sub-categories within *age married* to determine if they matched the data. This initial assessment illustrates that a Poisson model is a reasonable fit both overall and across the different age levels.

Before committing to a Poisson model, we needed to ensure all the Poisson assumptions were met within our data. Therefore, we proceeded to compare the means and variances of each age group within the *age married* variable, which demonstrated general positive differences between the variances and means, suggesting possible overdispersion in the data (Table 1). Therefore, our analysis compares the results between a Poisson regression model against a Negative Binomial model to determine which more appropriately answers our research question.

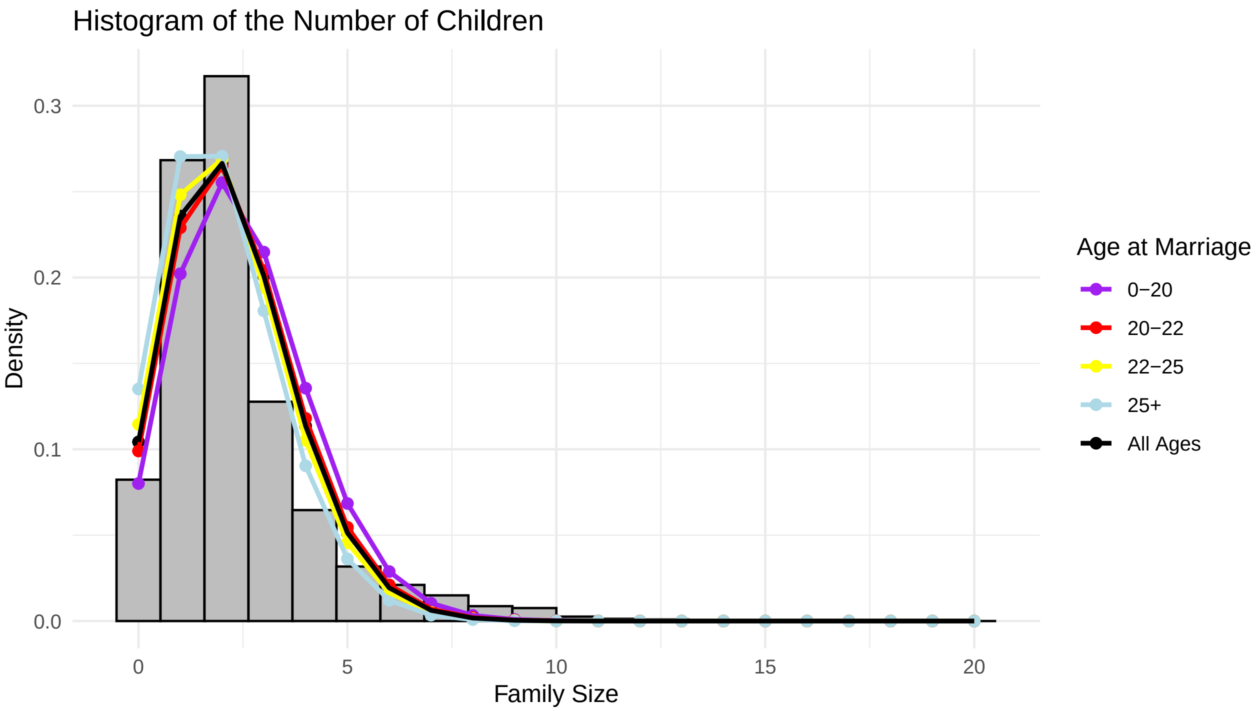

Figure 1: Comparison of the density distributions and fitted Poisson models for the number of children (in the 1979 Portugal dataset) across age of marriage categories (0–20, 20–22, 22–25, and 25+ years), with an additional model using the overall dataset mean as the Poisson parameter. Each line represents the fit for a specific age category using the category-specific mean number of children, while the black line corresponds to the model fitted to the entire dataset. The histogram reflects the observed family size distribution, providing context for model fit and differences across age groups.

Table 1: Summary statistics for the number of children overall and across the age marriage levels, showing mean, variance, and the variance-mean gaps

Age Married	Mean Num Children	Variance Num Children	Variance-Mean Gap
0to20	2.52475247524752	4.66711511915523	2.1423626439077
20to22	2.31261101243339	3.52174304322084	1.20913203078745
22to25	2.16689373297003	2.84670145563895	0.679807722668925
25+	2.00175438596491	2.54960183602113	0.547847450056221
All Levels	2.26048951048951	3.46370382095076	1.20321431046125

After fitting the Poisson and Negative Binomial models, we summarized the coefficient estimates and confidence intervals on the natural scale, along with the standard errors, test statistics, and p-values, in Figures 2 and 3, respectively. Comparing the two tables, we notice that both models have similar coefficient estimates, with differences occurring in the hundredths decimal place. Both models had similarly low standard errors (with slightly larger standard errors for the Negative Binomial model), and both models indicate that the variables *Age Married 20-22*, *Age Married 25+*, and *Region Porto* are insignificant in explaining family size (at a significance level of 0.05).

Table 2: Summary of exponentiated coefficients, 95% confidence intervals, standard errors, test statistics, and p-values from the fitted Poisson models. Results highlight the significance of literacy and regional factors, while age at marriage categories (20–22 and 25+) and Region: Porto are identified as statistically insignificant predictors.

	2.5 %	97.5 %	Estimate	Standard Error	Test Statistic	P_Value
Intercept	0.183	0.196	0.189	0.003	-90.239	0.000
No Literacy	1.086	1.191	1.138	0.027	5.455	0.000
Age Married: 20-22	0.913	1.011	0.961	0.025	-1.550	0.121
Age Married: 22-25	0.894	0.984	0.938	0.023	-2.613	0.009
Age Married: 25+	0.908	1.009	0.957	0.026	-1.628	0.103
Region: Lisbon	0.704	0.814	0.758	0.028	-7.512	0.000
Region: Porto	0.818	1.009	0.910	0.049	-1.765	0.078
Region: Population 20k+	0.698	0.794	0.745	0.024	-9.021	0.000
Region: Population 10-20K	0.793	0.912	0.851	0.030	-4.527	0.000

Table 3: Summary of exponentiated coefficients, 95% confidence intervals, standard errors, test statistics, and p-values from the fitted Negative Binomial models. Results highlight the significance of literacy and regional factors, while age at marriage categories (20–22 and 25+) and Region: Porto are identified as statistically insignificant predictors.

	2.5 %	97.5 %	Estimate	Standard Error	Test Statistic	P-Value
Intercept	0.186	0.202	0.194	0.004	-80.196	0.000
No Literacy	1.069	1.186	1.126	0.030	4.503	0.000
Age Married: 20-22	0.908	1.014	0.960	0.027	-1.457	0.145
Age Married: 22-25	0.884	0.982	0.932	0.025	-2.649	0.008
Age Married: 25+	0.896	1.005	0.949	0.028	-1.805	0.071
Region: Lisbon	0.702	0.820	0.759	0.030	-6.977	0.000
Region: Porto	0.810	1.017	0.908	0.053	-1.671	0.095
Region: Population 20k+	0.695	0.798	0.745	0.026	-8.406	0.000
Region: Population 10-20K	0.790	0.918	0.851	0.033	-4.191	0.000

Figure 2 compares the exponentiated coefficients (rate ratios) with confidence intervals for Poisson and Negative Binomial models. Both models yield similar coefficient estimates and trends. However, the Negative Binomial model generally displays slightly wider confidence intervals, reflecting its ability to account for overdispersion. For the *Age Married 25+* category, the Negative Binomial model's confidence interval covers less of the line at 1.0 compared to the Poisson model, indicating improved precision. Also the model estimated an overdispersion parameter of 24.4% (CI 21.4%, 27.8%) indicating that birth rates vary by about 24.4% between families. Therefore, we believe a Negative Binomial model is better at explaining variation in family size and it serves as our model choice.

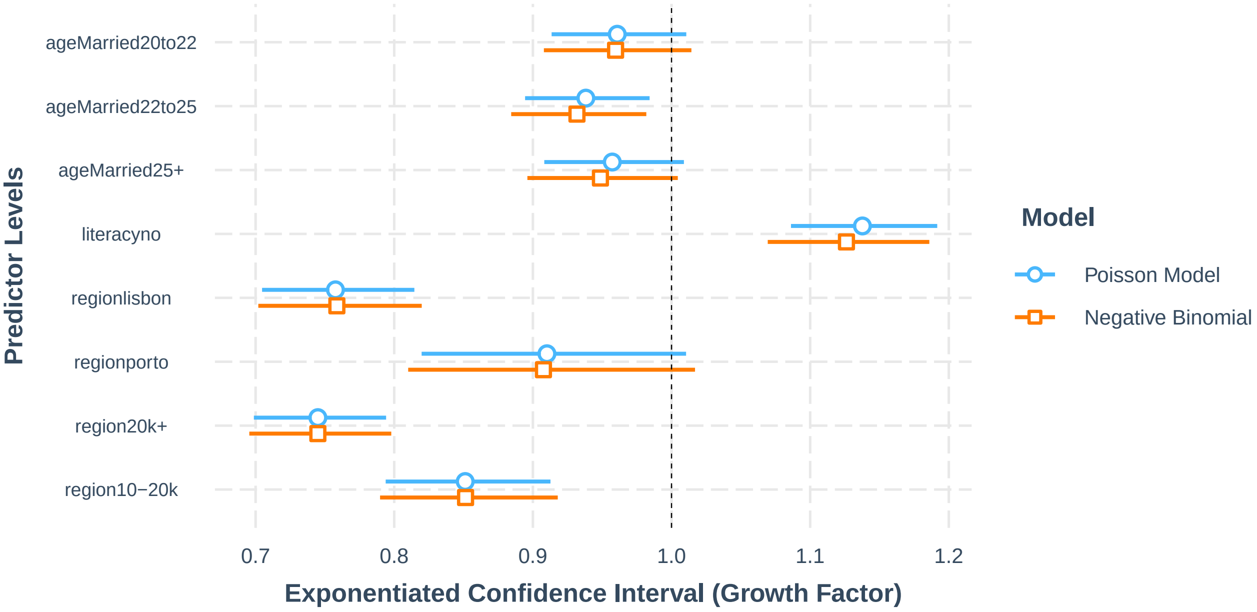

Figure 2: Comparison of exponentiated coefficients (growth factors) and 95% confidence intervals for predictor levels across Poisson and Negative Binomial models. The plot highlights differences in effect sizes and overdispersion adjustments for key predictors, including age at marriage, literacy, and region.

Conclusion

Overall, the results from our data analysis align with ideas from pre-existing literature and demonstrate a positive progression in selecting an appropriate model to fit the data. Both our original Poisson model and the Negative Binomial model demonstrated generally significant p-values for variables such as literacy and region which is supported by the results of Gutkind's and Mahmood et al. 's studies. However, contrary to our previous literature review, some variable levels did not prove to be as significant as initially thought, such as *Age Married 20-22*, *Age Married 25+*, and *Region Porto*.

In particular, the variable *literacy* under the Negative Binomial model had an estimated coefficient of 1.126 on the natural scale (Figure 3). Since this represents a growth factor; the birth rate of uneducated couples compared to their educated counterparts increases by a factor of 1.126, holding all else constant. This trend matches Mahmood et al.'s (2016) findings that 77% of educated males favoured smaller families compared to only 19.2% of uneducated males ($p < 0.001$). Additionally, the variable at level *Age Married 22-25* has an estimated coefficient 0.932 (Figure 3) which is interpreted as a decrease in birth rates by a factor of 0.932 for couples married at an age between 22 to 25 compared to couples married at the age between 0 to 20, holding all else constant. This matches the findings from Freedman (1963) which also found an inverse relationship between age of marriage and the birth rates.

Our final model results indicate that marrying later as well as higher literacy status is generally associated with having fewer children, with indications of overdispersion and the need for an offset within the data. Our Negative Binomial model seems to be an appropriate fit for the 1979 Portugal fertility data, demonstrating a significant relationship between literacy, region, and family size, with room for further investigation of the age of a marriage, in order to deeper understand the population behaviors and trends.

References:

Freedman, R. (1963). Norms for family size in underdeveloped areas. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, **159**(974), 220–245. https://doi.org/10.1098/rspb.1963.0074

Gutkind, Peter C. W. (1962). African Urban Family Life: Comment on and Analysis of Some Rural-Urban Differences. *Cahiers d'Études Africaines*, **3**(10), 149–217. http://www.jstor.org/stable/4390830

Mahmood, H., Khan, Z., & Masood, S. (2016). Effects of male literacy on family size: A cross sectional study conducted in Chakwal city. *Journal of the Pakistan Medical Association*, **66**(4), 399–403