STA303 Report

Rebecca Li & Ebrahim Mohamud

2025-02-07

Introduction

Family size is one of the most critical and statistically relevant markers of a countries social and economics condition. Measuring the change in family size serves as an good indicator for birth rates, fertility rates, and population growth which have broad social and economic consequences. Understanding the various factors which explain changes in family size is therefore an essential statistical measure.

Infertility and personal preferences are well-established factors behind the number of children that a couple has. However, the influences behind family size span beyond personal or biological reasons, often tracing to socioeconomic factors and time influences. Impacts such as the age of marriage and literacy demonstrate a strong relationship to family size, particularly in the context of developing countries, where children are utilized as a source of labour and less likely to survive to adulthood compared to their developed country counterparts (Freedman, 1963). These findings are relevant to the context of the present report, which analyzes data from the underdeveloped country of Portugal in order to investigate the primary factors behind family size, allowing for more accurate projections and predictions of future population trends.

Further inquiring into the influences on family size, Freedman (1963) highlights the significance of marrying and having children at a younger age, which serves as a protective measure in case of a parent or child dying, relaying the importance of age of marriage in the number of children. Furthermore, related studies of undeveloped countries have demonstrated that literacy can also play a significant role. In a study of 178 males in Chakwal city, Punjab, Mahmood et al. (2016) found that a minority of uneducated men were in favour of smaller families (19.2%) compared to a majority of educated men (77.0%). Mahmood et al. also established that one leading motivation behind larger families in uneducated males was the desire for a son. Together, these two factors played a primary role in motivating our research question.

Previously, we discussed the implications of a country's developmental stage, however, there may also be within-country factors to consider. In particular, the differences between urban and rural regions have demonstrated significant impacts on family size. 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 factors such as high costs of urban living. The significance of these results encouraged us to consider an additional variable of "region". 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?". Specifically, our analysis focuses on using a

generalized linear regression model to explain variation in family size through the aforementioned variables. Our analysis will involve justifying our model choice, fitting our model, interpreting the results, and outlining the possible limitations present in our findings.

Methods

To answer our research question, we began by plotting the response variable, children, in a histogram to examine its distribution. Since the number of children is a count variable distributed over a unit of time, the only appropriate model for this data is a Poisson model. To establish further support for our model choice, we fit a Poisson fitted line with a chosen parameter based on the global mean, as well as the means from the various level within the "age married" predictor (the main predictor in our model) to determine if they matched the data. This initial visual assessment is captured in Figure 1 below and illustrates that a Poisson model is a reasonable fit for our data.

Before committing to a Poisson model, we need to ensure all the Poisson assumptions are met within our data. Therefore, we proceeded by checking whether the mean and variance of each age group within the "age married" variable equated or whether they didn't. As Figure 2 illustrates, there is a positive gap between the variances and means suggesting possible overdispersion in the data. 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 except accounts for overdispersion. We conclude our analysis by first interpreting then comparing the two model estimates and confidence intervals to measure which model more appropriately serves to answer our research question.

We chose to include "Region" in addition to the "age married" and "literacy" variables specific to the research question because of the surrounding literature by Peter Gutkind which emphasized the explanitory power of the "Region" variable. Specifically that rural families produce significantly more children that their urban counterparts. Additionally we included an offset in both models to account for variation in exposure times. The offset ensure 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. Adding the offset is essential to our research question since meaningfully estimating the birth intensity across each family fully specifies our regression model.

Results

Our dataset includes 5148 observations of married couples from Portugal in the year 1979 with information such as the age of marriage, literacy status, the number of children, the months since married, etc. Our research question specifically focus on how much the first two variables explain variation in the number of children a family produces. Visualizing our data as in Figure 1, we witness a large right skew with a mean of 2.26 and a standard deviation of 1.86. Graphically, the Poisson model appears to be a reasonable fit, both overall and across the different categories of the age at marriage predictor variable. To improve balance across categories, we redefined the age brackets to ensure that each level contains a roughly equal number of observations. This adjustment was made to avoid sparsely populated categories, which could lead to unstable estimates and reduced reliability in our analysis.

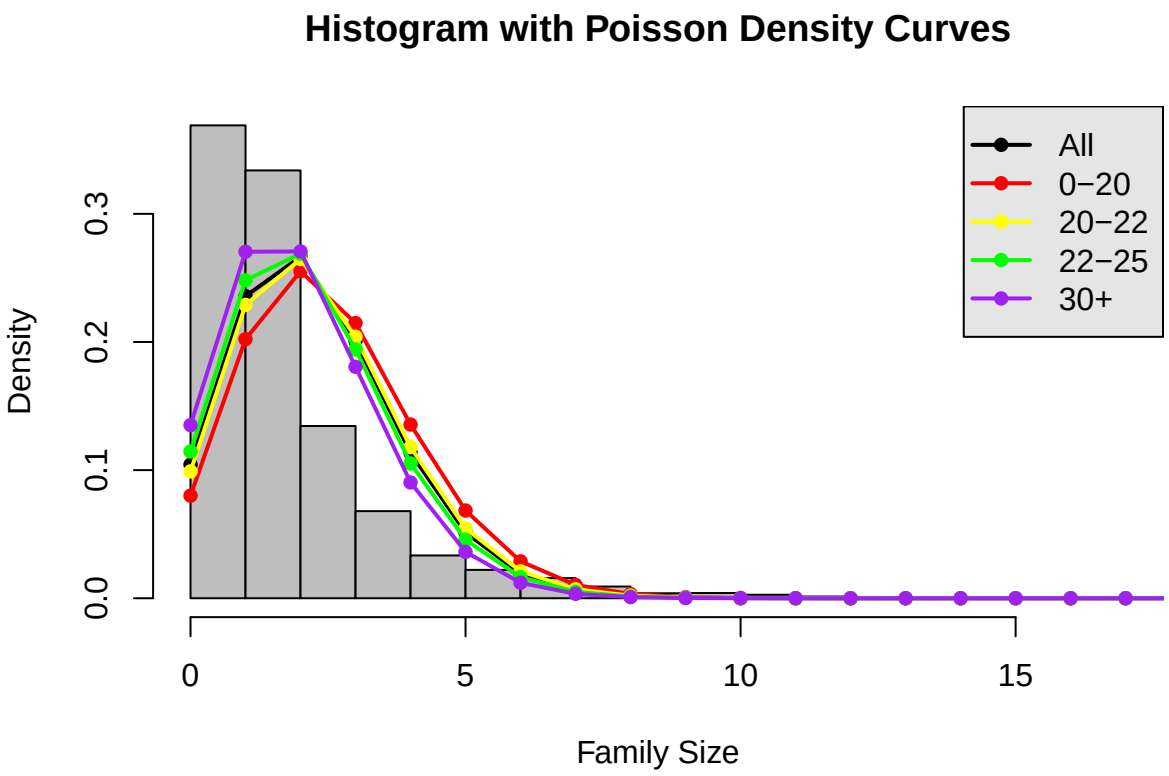

Figure 1: Histogram with Poisson Density Curves

While the Poisson model appears to visually match the data, a critical Poisson assumption is violated, namely that the mean and variance of each category dont equate. The table below illustrates this violation well as a the variance-mean gap is positive, suggesting potential overdispersion and leading us to favour a Negative Binomial model over a Poisson.

Age Married	Mean Num Children	Variance Num Children	Variance-Mean Gap
0to20	2.525	4.667	2.142
20to22	2.313	3.522	1.209
22to25	2.167	2.847	0.680
30+	2.002	2.550	0.548

To truly compare the two models we must compare their model estimates and . . .