The Impact of Literacy, Marriage Age, and Region on Family Size in Portugal

Shivraj Sambus, Liam Wall

10 February 2025

Introduction

From 1979 to 1980, The Demographic and Health Surveys Program conducted a fertility survey in Portugal to better understand the factors contributing to the birth rate across the country. The questionnaire consisted of basic questions such as pregnancy history and contraceptive use. Our report explores the impacts that literacy, age of marriage, and region of residence has on the size of a given family in Portugal. We will conduct this analysis by using a negative binomial general linear model and 95% confidence intervals to prove that these factors are significant predictors of family size. An accurate estimation of family size is essential to determining the size, age, fertility trends of a given population. Using this information, one can better understand the population's growth in the future and assess what areas of the local healthcare and education systems need to be improved (Lutz, 2006, p.25). This is a highly important part of improving the infrastructure and the quality of life in a given country. Next, we will present summaries of three research papers on fertility rates in Portugal and the rest of Europe.

In 2014, Dr. Maria Testa conducted an analysis of the relationship between European women's level of education and intended family size, with respect to their individual-level and country-level demographics. She found that the relationship is positive on both levels and that highly educated women do not necessarily have fewer children. With regards to Portugal, however, Testa's findings show that, on average, family sizes are smaller among highly educated women (Testa, 2014. p.40). In 2022, Dr. Dulce Pimentel and Dr. Cristina Gomes conducted an analysis of the regional fertility patterns in Portugal over the last 20 years, specifically considering economic and health crises such as the 2008 financial crisis and the COVID-19 pandemic, respectively. They found that the fertility rate in Portugal is declining, one of the lowest in Europe, and that regions with a higher population density tend to have a higher fertility rate, especially those occupied by younger generations. An important result of their study is that rural areas with low population densities were awarded the highest amount of government funding to improve sexual education and fertility rates. This implies that those that have multiple children and live in rural areas may continue to have children without facing excessive financial constraints (Pimentel, Gomes, 2022). In 2008, Dr. Analia Torres conducted an analysis of the relationships between gender and labor in the European Union. She found that the rate of female employment is high in Portugal and that most employed women, on average, work nearly the same full-time hours as men do but do not earn as high of an income. An important part of her results is that gender inequality negatively affects employed women and, in many cases, they end up taking more responsibility over childcare and housework. This implies that highly educated women may not have as large of a family size as their lesser educated counterparts (Torres, 2008, p. 36,51).

An assumption of our model is that those in Portugal with a lower level of literacy tend to have a larger family size due to the lack of sexual education resources. This aligns with the implications of Dr. Pimentel and Dr. Gomes' analyses since those in rural areas have access to government funding to continue having children while gaining an understanding of child care. Another assumption is that those in rural areas are generally less wealthy and tend to have larger family sizes because they need their children to assist them in generating the household income through manual labor. This aligns with Dr. Testa's analyses because she mentioned that, in Portugal, highly educated women tend to have smaller family sizes than those who are

less educated. Our last assumption is that younger married couples tend to have larger family sizes because they remain fertile for longer than their older counterparts. This does not align closely with Dr. Torres' analyses because of the unequal division of labor among Portuguese couples, leading to insufficient time to raise a large number of children.

Methods

When choosing a generalized linear model (GLM) to best fit and model the relationship between literacy, age of marriage and children, we have to carefully consider the characteristics of our data. We aim to predict the size of a family (children) given a person's age at marriage, the number years since marriage, the region they reside and whether they are literate or not (literacy). Literacy and age of marriage will be included as predictors in our GLM because the main goal of this paper is to examine the effect of these predictors on the size of a family. The number of years since marriage will be included in our GLM as an offset. Number of years since marriage is an offset because it allows the GLM to model the number of children while adjusting for how long someone has been married. This is an important inclusion because we already incorporate the age of marriage as a predictor yet we want the GLM to consider that recent married couples may have a larger family in future whereas lontime couples may not have another child. Additionally, we include region as a confounding predictor because it is well known that families in rural areas are larger. Lastly, we make sure to relevel the predictors so that the most common variable is the reference for our model.

In beginning our statistical analyses we will consider two models: a poisson model and a negative binomial model. The response variable in our data, number of children, is a classic example of count data (taking non-negative integer values: 0, 1, 2,...) which aligns exactly with the Poisson distributions each model follows. We did think about other models like skew-normal and Gamma, however we are assuming the underlying distribution of the number of children has a true distribution that exists. Therefore, because skew-normal may take non-negative values and because Gamma is non-zero, we choose to not consider them. It should be noted that the two chosen models follow different assumptions. The Poisson distribution assumes that our data has an equivalent mean and variance whereas the negative binomial distribution assumes our data has a greater variance than mean, or overdispersion. Thus, in order to decide on the best model for our question, we will examine the mean and variance of our data across several features, as well as the outputs of each of the fitted models. Specifically, if we find that the variances are greater than the means of the number of children among different age groups at marriage and among those who are literate, we will suspect using the negative binomial model is better. Also, we will look at the dispersion term given to us when we fit a negative binomial model to our data from which we can find the relative standard deviation of our data. This relative standard deviation tells us the amount of variation that exists in our data compared to its mean. As the Poisson distribution has equal mean and variance, a non-zero relative standard deviation would be further evidence that a negative binomial model is better due to over dispersion.

Results

First, we examined the distribution of the number of children and the number of children given age of marriage and literacy. As can be seen in the three side by side plots below, each distribution of the number of children, alone and conditioned on our predictors of literacy and age of marriage, are heavily right skewed and evident of a Poisson distribution. Beyond showing us that data looks like it follows a Poisson distribution, we can also see how illiterate women display greater numbers of children and that the distribution of children among women married at different ages is relatively similar (aside from '30toInf').

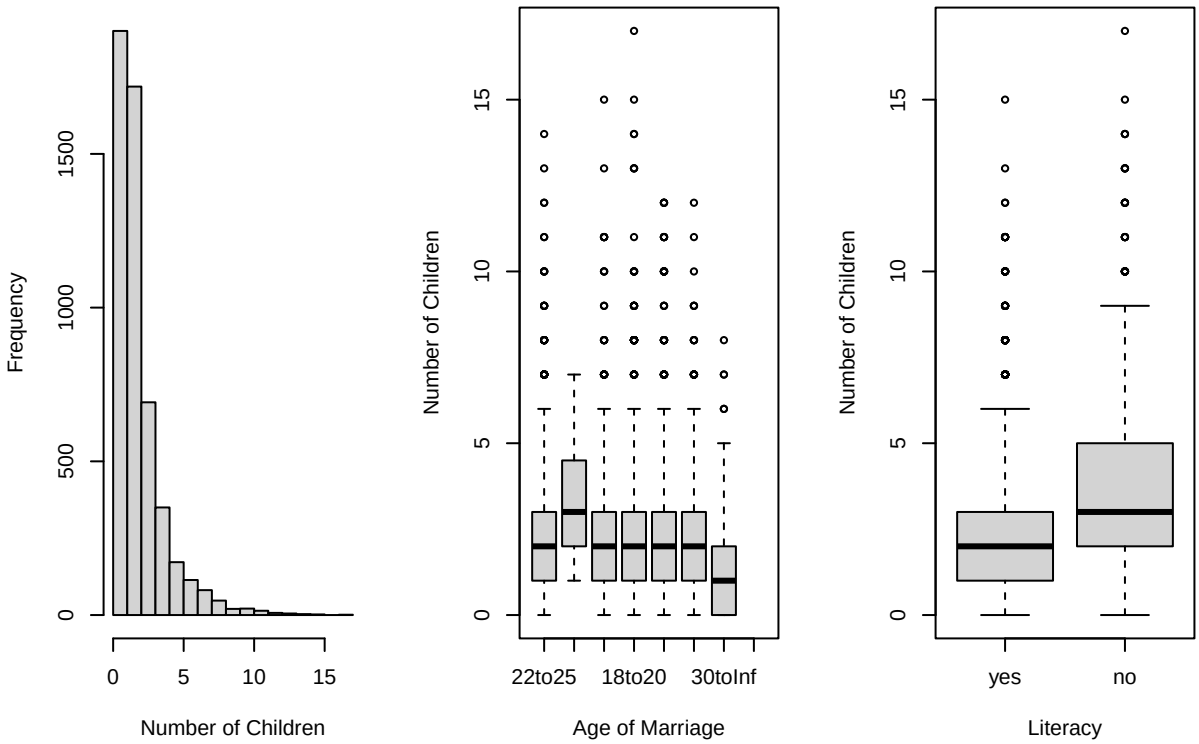

Table 1 below shows that our data centers around 2.26 children, and given each predictor this varies slightly. The highest average number of children is present among illiterate women at 3.864 children while the lowest average number of children is in women who married at age 30 or higher, with 1.47 children on average. Further, we see the average number of children is less than the variance for every category in our predictors. In categories with a relatively large number of observations, the means and variances are closer, but never approximately equal. For example, there are 4567 literate women in the data set and the variance of children among literate women is only 0.47 greater than the mean. In contrast, there are only 581 illiterate women and the variance of the number of children among illiterate women is 4.081 higher than the mean. Thus, based on the statistical summaries of the variables in our data, we have evidence for overdispersion and suspect that the negative binomial model will perform better.

Table 1: Table showing the overdisperssion present in the response variable, children, given each predictor. We observe that the mean of number of children is less than the variance of the number of children in each category of age married and also in each category of literacy.

Group	Mean	Variance	Difference	Count
22to25	2.167	2.847	0.680	1468
0to15	3.212	3.033	0.179	52
15to18	2.644	4.771	2.127	452
18to20	2.426	4.674	2.248	910
20to22	2.313	3.522	1.209	1126
25to30	2.127	2.584	0.457	923
30toInf	1.470	2.065	0.595	217
yes	2.056	2.526	0.470	4567
no	3.864	7.945	4.081	581

In order to finally determine which model performs best with our data in order to answer our question, we

will look at the dispersion term in our fitted negative binomial model. When fitting the negative binomial data, as discussed in the Methods section, we are predicting the number of children based on literacy, age of marriage, region and offsetting by the number of years married. The 'glmmTMB' package (glmmTMB, 2017) turns a dispersion parameter, sigma. We can then find the inverse square root of sigma to get the relative standard deviation of our data when compared to its average. As we can see from the model results in Table 2, the last row gives us the estimate and corresponding 95% confidence interval for the relative standard deviation. A relative standard deviation of 0.245 means that the number of children per women in our data varies by 24.5% compared to the mean of the entire population. This suggests a level of dispersion beyond what is expected under a Poisson distribution (which would have a relative standard deviation of 0.0). Based on these findings, we concluded that our final model for analyzing the effects of literacy and age of marriage on family size should follow a negative binomial distribution.

Table 2: The negative binomial model's estimated coefficients as well as their corresponding 95% confidence interval. The last row shows the calculated relative standard deviation and its corresponding 95% confidence interval.

	Estimate	2.5 %	97.5 %
(Intercept)	-1.712	-1.753	-1.671
literacyno	0.119	0.067	0.171
ageMarried0to15	0.056	-0.118	0.230
ageMarried15to18	0.083	0.010	0.156
ageMarried18to20	0.065	0.006	0.125
ageMarried20to22	0.029	-0.027	0.086
ageMarried25to30	0.016	-0.045	0.077
ageMarried30toInf	0.028	-0.094	0.150
regionlisbon	-0.276	-0.354	-0.199
regionporto	-0.097	-0.211	0.017
region20k+	-0.294	-0.363	-0.226
region10-20k	-0.161	-0.237	-0.086
SD	0.245	0.214	0.279

Finally, we can examine and interpret the model's estimated coefficients. The negative binomial model has a log link so therefore the given coefficients are relative log rates. To help with interpretation, we also provided the natural rates, the exponentiated coefficient. All the relevant information from the negative binomial model can be found in Table 3 including the log rates, natural rates, standard errors, and p-value.

Table 3: This table shows the estimated coefficients on the log scale, on the natural scale, and their corresponding standard error and p-value.

	Estimate	Estimate..Natural.Scale.	Std..Error	z.value	Pr. . . z..
(Intercept)	-1.712	0.181	0.021	-81.839	0.000
literacyno	0.119	1.126	0.026	4.494	0.000
ageMarried0to15	0.056	1.058	0.089	0.631	0.528
ageMarried15to18	0.083	1.086	0.037	2.216	0.027
ageMarried18to20	0.065	1.067	0.030	2.155	0.031
ageMarried20to22	0.029	1.030	0.029	1.024	0.306
ageMarried25to30	0.016	1.016	0.031	0.525	0.600
ageMarried30toInf	0.028	1.029	0.062	0.454	0.650
regionlisbon	-0.276	0.758	0.040	-6.982	0.000
regionporto	-0.097	0.908	0.058	-1.668	0.095

	Estimate	Estimate..Natural.Scale.	Std..Error	z.value	Pr...z..
region20k+	-0.294	0.745	0.035	-8.402	0.000
region10-20k	-0.161	0.851	0.038	-4.203	0.000

Conclusion

To interpret the results of the fitted negative binomial model we will reference Table 3. Illiteracy was found to have a significant effect on the number of children born. With a p-value < 0.001, we can reject the null hypothesis. Further, a natural rate of 1.126 means that women who are illiterate had 12.6% more children than literate women. The effect of age of marriage had mixed results with the categories of ages 15 to 18 exhibiting a p-value < 0.03 and ages 18 to 20 a p-value < 0.04, both indicating significant effect. Women married at ages 15 to 18 had 8.6% more children than those married at ages 22 to 25. Women married at ages 18 to 20 had 6.7% more children than those married at ages 22 to 25. The other categories for age at marriage did not have a significant effect on the number of children. Lastly, we can see regional differences did have an effect the number of children. Compared to women residing in rural areas, every other region showed less children per woman. Those in Lisbon had 0.758 more children with a p-value $<$ 0.001, corresponding to 24.2% less children. Women in cities with a population of 20,000 or more had 25.6% less children than women in rural areas with a p-value < 0.001. Similarly, women in cities with 10,000–20,000 people had 14.9% fewer children than those in rural areas with a p-value < 0.001. However, the effect for Porto was marginally non-significant with an estimate of 9.2% less children than women in rural areas with a p-value > 0.09.

Overall, the results suggest that literacy, age at marriage, and regional differences significantly influence the size of families in Portugal. Women with lower literacy and those who married at younger ages tend to have more children, while those in urban areas, particularly Lisbon and other large cities, have less children - all of which is consistent with the research mentioned in the Introduction. Any shortcomings in this paper or inconsistencies with current research may be due to the small and outdated nature of the samples in our data set. These findings highlight the importance of socio-demographic factors in shaping the future of a country.

References

1. Torres, A. (2008). Women, Gender, and Work: The Portuguese Case in the Context of the European Union, International Journal of Sociology, 38:4, 36-56, https://www.tandfonline.com/doi/epdf/10.2753/IJS0020-7659380402?needAccess=true
2. Pimentel D., Gomes C. (2022). Beyond the crisis: fertility variations and the family policies in the Portuguese municipalities. Space, populations, societies. OpenEdition Journals http://journals.openedition.org/eps/12990
3. Testa, M. (2014). On the positive correlation between education and fertility intentions in Europe: Individual- and country-level evidence. Advances in Life Course Research. Elsevier. https://www.sciencedirect.com/science/article/pii/S1040260814000069?ref=pdf_download&fr=RR-2&rr=90ffa28488e4aa96
4. Lutz, W. (2006), Fertility rates and future population trends: will Europe's birth rate recover or continue to decline?. International Journal of Andrology. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1365-2605.2005.00639.x
5. R Core Team. (2023) R: A Language and Environment for Statistical Computinghttps://www.R-project.org/
6. Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill

Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani. (2019) Welcome to the {tidyverse}. doi: 10.21105/joss.01686

7. Mollie E. Brooks and Kasper Kristensen and Koen J. {van Benthem} and Arni Magnusson and Casper W. Berg and Anders Nielsen and Hans J. Skaug and Martin Maechler and Benjamin M. Bolker (2017) {glmmTMB} Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. doi: 10.32614/RJ-2017-066