Honey, I Shrunk the Nuclear Family: Analyzing the Size of Families over Time using GSS Data

DATASCI 203: Lab 2

Aman Kumar, Ayodele Oyewole, Assaph Aharoni, Gia Nguyen

2024-04-12

Contents

1 Importance and Context

The relationship between economic growth and population growth is at the center of significant debate among economists, sociologists, and politicians. Especially in high-income countries such as the US, low population growth can be associated with decreased GDP growth [1]. With immigration persisting as a political lightning rod, many are focused on studying the change in birth rates over time. Some commentators warn of dropping birth rates leading to a coming demographic crisis or "Great People Shortage" [2]. In this analysis, we attempt to contribute to this discourse by studying how the nuclear family size has changed over time in the US, as well as other factors. More specifically, we will be studying how the number of siblings in a family relates to the age of survey respondents, with age acting a generational proxy. Our modeling will also attempt to capture other family characteristics, such as household income and cultural context.

2 Data and Methodology

Our analysis relies on data from the 2022 General Social Survey (GSS), a nationally representative survey of adults in the United States that captures demographic, behavioral, and attitudinal data [3]. The 2022 GSS is an observational cross-sectional survey containing responses from 3,544 individuals 18 years or older. To operationalize the dependent variable, family size, we use the "sibs" field in GSS, which captures how many siblings respondents have had, including those who are deceased. Alternatively, one could have studied the number of children each respondent has had. However, such an analysis would require handling the fact that younger respondents may not have had kids yet. By looking at the results from the perspective of the respondent as a child, we employ the assumption that those who are 18 years or older are unlikely to have new siblings. Additionally, we used the "family16" field to restrict only to respondents whose parents were still together at the time they were 16 in order to avoid including outlier counts of siblings resulting from step-siblings. Second, we operationalized our main independent variable using the "age" field as a proxy for year of birth. Additionally, we attempt to control for other factors that may affect family size, such as income, location, and cultural context. To do so, we use the features shown in Table 1.

Table 1: Variables Analysis

Variables	Description	Data Manipulations	Data Scale
age	Respondent's age	Dropped 'no answer' and 'skipped on web' responses and categorized group '89 or older' as 89 for analysis	Ratio
incom16	Likert scale of perception of parent's income when respondent age 16	Dropped 'don't know,' 'no answer,' and 'skipped on web' responses and assigned numerical values for Likert scale	Ordinal
reg16	Respondent's region of residence at age 16	Dropped 'don't know' and 'skipped on web' responses and applied one-hot encoding	Binary after one-hot encoding
res16	Respondent's type of place lived in at age 16	Dropped 'don't know' responses and applied one-hot encoding	Binary after one-hot encoding
pasei10	Respondent's father's socioeconomic index	Dropped 'iap' responses	Interval
masei10	Respondent's mother's socioeconomic index	Dropped 'iap' responses	Interval
granborn	Number of grandparents born in US	Dropped 'don't know,' 'iap,' 'no answer,' and 'skipped on web' responses and recoded categorical values into numerical	Ratio
sibs	Number of siblings respondents have had (including deceased)	Dropped 'no answer' and 'skipped on web' responses	Ratio

After the data manipulations, Table 2 shows a summary of the final number of observations. We used this number of observations to partition our data into exploration and validation sets. To do so, we randomly selected 30% of the observations into the exploration subset in order to assist in model construction, and then validated the results on the remaining 70% of the observations. The results presented in this paper come from the validation subset of the data.

Table 2: Observations Analysis

Cause	Number of Samples Available for Analysis (after removal for cause)	Removed number samples for cause
Start	3,544	N/A
Insignificant responses	1,730	1,814
Family16 filter	1,428	302

To assess the correlation between the number of siblings in a family and the age of the respondent, we relied on an OLS regression. Figure 1 shows a scatter plot with number of siblings on the y-axis and age on the x-axis, as well as the OLS line.

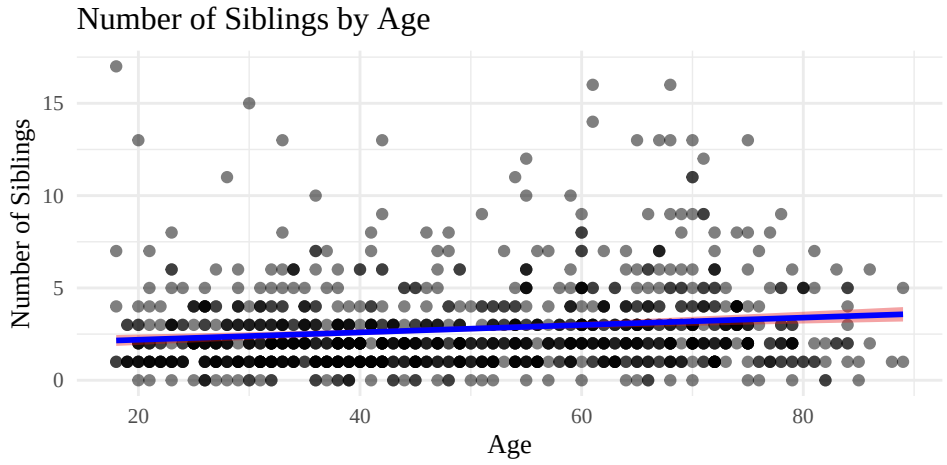

Figure 1: Number of Siblings by Age

Because our dataset contains responses from 1,428 individuals, we are able to assess and leverage the large-sample assumptions for regression. Use of the GSS data allows us to assume that observations are IID. However, while the data fulfills the criteria of no perfect collinearity, the distribution of number of siblings is heavy-tailed. To make sure a unique BLP exists, we explored transformations of the dependent variable to reduce the impact of the heavy tails. The log transformation could help achieve that, but because we have many observations with a sibling count of zero, we decided in the end to use the square root of the number of siblings.

3 Results

In Model 1 ($\sqrt{siblings}$ regressed on age), there is a statistically significant relationship between age and number of siblings. To assess the robustness of this result, we controlled for the various features described in the previous section in Model 2. Given that the socioeconomic indices of the parents are related to one another, we decided to use an interaction term.

Table 3: Relationship Between Number of Siblings and Various Socioeconomic Factors

	sqrt(sibs)	
	Base Model	Additional Model
	(1)	(2)
Constant	1.233^{***} (1.111, 1.355)	1.975^{***} (1.703, 2.247)
incom16		-0.052^{**} (-0.100, -0.003)
masei10		-0.009^{***} (-0.013, -0.004)
pasei10		-0.008^{***} (-0.012, -0.003)
age	0.006^{***} (0.004, 0.008)	0.004^{***} (0.002, 0.007)
reg16foreign		0.184^{**} (0.022, 0.346)
masei10:pasei10		0.0001^{**} (0.00002, 0.0002)
N	997	997
R^2	0.025	0.083
Adjusted R^2	0.024	0.077
Residual Std. Error	0.658 (df = 995)	0.640 (df = 990)
F Statistic	25.677^{***} (df = 1; 995)	14.906^{***} (df = 6; 990)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

We believe that the age coefficient is lower in Model 2 because a portion of age's effect on family size is explained by other covariates. Seeing as there is a square root transformation on the dependent variable, the coefficient of age can be interpreted by partially differentiating with respect to age, giving us $\frac{\partial sibs}{\partial age} = 0.008\sqrt{sibs}$. Meaning that if the average person has two siblings, another person with 5 more years of age would have an average of 2.06 siblings (assuming that the parents of both people have similar socioeconomic indices, perceived income, and grew up within the US). Similarly, the fact that a person grew up in the US versus overseas has an impact on the number of siblings they have. For two people of the same age whose parents have similar socioeconomic indices and perceived income, if the person in the US had an average of two siblings, the person who grew up overseas would have an average of 2.52 siblings.

4 Discussion

Our analysis suggests areas for future research. First, studying family size from the respondent's childhood perspective limits insights into how family size trends have evolved over the past 18 years; future work could aim to address this gap. Second, our reliance on perceived variables, like perceived parents' income when respondent was 16 years old, could introduce recall bias; objective data could reduce such biases. Lastly, the presence of fat tails in our model's residuals, as indicated by a QQ plot (see attached code), suggests that exploring further transformations could refine our findings.

5 References

[1] Peterson, E. W. F. (2017). The Role of Population in Economic Growth. Sage Open, 7(4). https://doi.org/10.1177/2158244017736094

[2] North, A. (2023, November 27). You can't even pay people to have more kids. Vox. https://www.vox.com/23971366/declining-birth-rate-fertility-babies-children

[3] The General Social Survey (1972). GSS. https://gss.norc.org/About-The-GSS