The likelihood of childbearing based on educational, socioeconomic, and interpersonal factors

J. Abigail Joseph
Department of Electrical Engineering and Computer Science
Florida Atlantic University College of Engineering and Computer Science
Boca Raton, United States
jemimahjosep2021@fau.edu

I. INTRODUCTION

There have long been reports about the declining birth rate in the United States. According to the White House, in 2023, the CDC reported that the birth rate had fallen to its lowest level in history[1]. There have been multiple reasons discovered by researchers as to why the US birth rate has been declining for decades, such as social and economic disparities, and the increased age at which many Americans are choosing to become parents, if at all[2]. However, there have been several countries, particularly in the EU, that have managed to stabilize and increase their falling birth rates through family-oriented policies such as monetary incentives for having children, increased parental leave, and better childcare supplied by the government[3]. It is crucial to understand a country's most prevalent issues that influence its citizens' decision to have children. This report will study how several factors impact the likelihood of a person in the US having children. The most impactful factors will enable researchers to propose effective policies that will provide more support for Americans to have children. For example, currently, the US only offers twelve weeks of unpaid parental leave to new parents[4], one of the worst parental leave policies among countries in the developed world. There have been no US policies that directly address the falling birth rate and encourage Americans to have more children[5].

For this project, a logistic regression model will be used to examine five factors that may increase or decrease the likelihood of an individual having children.

II. MAIN BODY

A. Motivation

As stated previously, five factors were examined based on their relative impact on a person's likelihood of having children, using a logistic regression model. Logistic regression is ideal for this type of analysis because it can make predictions based on multiple variables. It is also ideal because with a research question like this, multiple of the variables could be confounders; this is accounted for as well in the analysis.

B. Design Framework

This study utilized the 2013-2015 Male Respondent dataset from the CDC's NSFG data. According to the CDC, "The National Survey of Family Growth (NSFG) gathers information on pregnancy and births, marriage and cohabitation, infertility, use of contraception, family life, and general and reproductive health.[6]"

First, the dataset was read into the R Studio environment. The structure of the dataset was evaluated to ensure that it was imported properly, and then it was saved as an RData file so that the dataset could be reloaded easily at any time. Next, the dataset was filtered to extract the target variables that would be used in the logistic regression model. The target variables were recoded into binary categories as necessary, and the dataset was cleaned via filtering to eliminate any missing or unknown values. The following were the variables used in the logistic regression model:

- MARSTAT – respondent's marital status

1. Married
2. Non-married cohabitating
3. Widowed
4. Divorced/annulled
5. Separated
6. Never been married
7. Refused (to answer)
8. Unknown

- HAVEDEG – whether a respondent has a college or university degree

1. Yes
2. No
3. Other (missing/recoded values)

- COVER12 – whether a respondent had health insurance coverage (in the last 12 months)

1. Yes
2. No
3. Unknown

- ORIENT – respondent's sexual orientation

1. Heterosexual (straight)
2. Homosexual (gay)
3. Bisexual
4. Other (missing or N/A values)

- EARN – a respondent's income/earnings during their last period of employment

1	Under $96 (weekly)/Under $417 (monthly)/Under $5,000 (yearly)
2	$96-$143 (weekly)/$417-624 (monthly)/$5,000-7,499 (yearly)
3	$144-191 (weekly)/$625-832 (monthly)/$7,500-9,999 (yearly)
4	$192-239 (weekly)/$833-1,041 (monthly)/$10,000-12,499 (yearly)
5	$240-288 (weekly)/$1,042-1,249 (monthly)/$12,500-14,999 (yearly)
6	$289-384 (weekly)/$1,250-1,666 (monthly)/$15,000-19,999 (yearly)
7	$385-480 (weekly)/$1,667-2,082 (monthly)/$20,000-24,999 (yearly)
8	$481-576 (weekly)/$2,083-2,499 (monthly)/$25,000-29,999 (yearly)
9	$577-672 (weekly)/$2,500-2,916 (monthly)/$30,000-34,999 (yearly)
10	$673-768 (weekly)/$2,917-3,332 (monthly)/$35,000-39,999 (yearly)
11	$769-961 (weekly)/$3,333-4,166 (monthly)/$40,000-49,999 (yearly)
12	$962-1,153 (weekly)/$4,167-4,999 (monthly)/$50,000-59,999 (yearly)
13	$1,154-1,441 (weekly)/$5,000-6,249 (monthly)/$60,000-74,999 (yearly)
14	$1,442-1,922 (weekly)/$6,250-8,332 (monthly)/$75,000-99,999 (yearly)
15	$1,923 or more (weekly)/$8,333 or more (monthly)/$100,000 or more (yearly)
97	Not ascertained
98	Refused
99	Don't know

Figure 1: Values for 'EARN' variable

These variables were chosen because they are generally known to have at least some part in whether a person has children. It has been studied whether higher education makes a person more or less likely to have children, particularly women (although this factor should not be examined on its own). Factors such as a low income or lack of health insurance may discourage people in those situations from having children at all, or having multiple children. There is a belief that poor people have more children, but some studies have found this to be false[9]; much like education, poverty on its own is not a reliable predictor of fertility. Sexual orientation is an interesting and relevant factor. Although people who identify as lgbtq+ in the United States often have smaller families and less children in general than non-lgbtq+ people, lgbtq+ inidividuals are more likely to foster and adopt children than 'straight' people[10].

Each of the variables for the logistic regression model were recoded into binary variables for ease of use in the logistic regression model; any missing values were put into an 'Unknown' category:

- MARSTAT_BINARY

1. Married/Cohabitating (1, 2)
2. Other (3, 4, 5, 6)
3. Unknown (8, 9)

- HAVEDEG_BINARY

1. Yes (1)
2. No (5)
3. Other (missing or N/A values)

- EARN_BINARY

1. "HighIncome" (>= 8)
2. "LowIncome" (< 8)
3. Unknown (single or other values)

- LIVTOGWF_BINARY

1. Yes (1)
2. No (5)
3. Unknown (all other values)

- COVER12_BINARY

1. Yes (1)
2. No (5)
3. Unknown (all other values)

- ORIENT_BINARY

1. Heterosexual (1)
2. Other (2, 3)
3. Unknown (all other values)

Additionally, a new variable to identify whether a person has children was created; this variable was used in the baseline model. When the dataset was correctly filtered and contained no missing values, beginning with one of the variables, each variable was implemented into the logistic regression model one by one so that eventually, a 'full' model was built that used all of the variables. This was done so that there was an accurate picture of how each individual variable impacts the model, and variables that significantly impacted the model could be identified. This step-by-step approach also aided in assessing multicollinearity and confounding variables.

After the model was fully built, the Akaike Information Criterion (AIC) was calculated for all of the models. By comparing the AIC of each successive model with the previous model, it could be determined how much each variable improved the model as it was added. This was an additional method to assess how good the model became at predicting the outcome (the likelihood of an individual having children) based on the variables that were added.

After AIC was calculated, each variable was assessed for confounding. This was done by performing a chi-square test, in which each variable was assessed for significance when it was added to the model on its own, and whether the significance of the variable changed when it was included in a model with the other variables. Next, the performance of the 'full' model was measured via comparison with the baseline model. Lastly, in order to provide results for the research aim, which explores which factors impact the likelihood of having children the most, the odds ratios of each variable were calculated. The results of this calculation and the other calculations and tests will be discussed further in the next section.

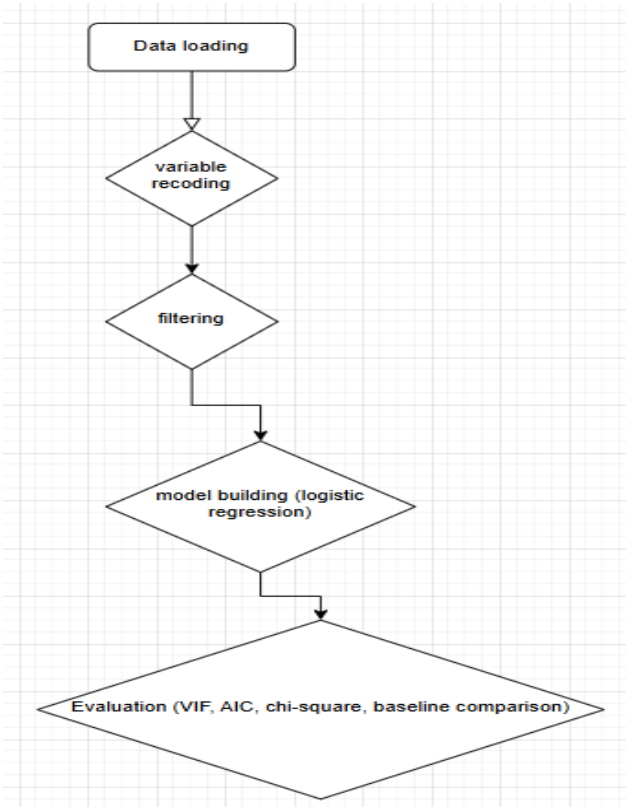

Figure 2: design framework flowchart

III. Experiments

a. To fully evaluate the performance of the 'full' logistic regression model, a baseline model was used for comparison. This baseline model is a majority class classifier. The accuracy of the 'full' model was assessed against the baseline accuracy. For programming and to complete all of the analysis, R and R Studio were used. Additionally, the libraries SAScii and pROC were used. The original dataset from the CDC's website was in .sas file format, so SAScii was used to read the data into R Studio. Along with the dataset, a setup file was also read into R Studio to ensure that the data was set up and formatted correctly. The pROC library was used to evaluate the Area Under the Curve (AUC) of the 'full' model. The Receiver Operating Characteristic (ROC) curve was calculated for the model, and from that, AUC of the model was determined. Additionally, the ROC calculation was used to create a visual for the model performance.

b. As stated previously, the dataset used was the 2013-2015 Male Respondent dataset from the CDC's National Survey of Family Growth (NSFG) data[7]. The dataset contains 2,940 columns (variables) and 4,506 rows (observations); this is quite an extensive dataset. Here is additional information about the 2013-2015 dataset respondents according to the CDC: 10,205 respondents aged 15-44 (5,699 women and 4,506 men)[8].

c. As stated previously, the baseline model served as the benchmark by which the performance of the 'full' logistic regression model was compared. A new, binary variable, 'HasChildren', which reflects the majority class in the dataset (men who do have children), was created. This variable was created because there was no variable in the original dataset which represented solely men who have children. Variables in the initial dataset that indicated whether men had children were a bit too specific (e.g. whether men had children with their current wife/partner or a previous wife/partner). The 'HasChildren' variable was more general and the most ideal for the analysis. The baseline classifier classified all observations as '1' ('1' being that a man has children and '0' being that he does not) if more than 50% of the observations in 'HasChildren' had a value of '1', the majority class.

The 'HasChildren' variable was created from the variable 'HHKIDTYP' in the dataset. The 'HHKIDTYP' variable indicates whether the respondent has biological or non-biological children, '0' being no children, '1' being biological children, and '2' being non-biological children. Based on this, if a respondent answered '1' or '2', 'HasChildren', a binary variable, was set to 1. Otherwise, it was set to 0.

The accuracy of the baseline model was determined based on its predictions compared to the actual values of 'HasChildren' in the dataset, the proportion of correct predictions made by the baseline model. Then, after the 'full' logistic regression model was fitted with all the variables, the accuracy of the 'full' model was compared to the baseline accuracy.

d. Results

Baseline Model

- Accuracy: 66.9%

AIC of logistic regression models after adding each (binary) variable

- Model 1 (MARSTAT) - 1919
- Model 2 (MARSTAT + HAVEDEG) – 1921, no significant improvement
- Model 3 (MARSTAT + HAVEDEG + EARN) - 1883.6, significant improvement
- Model 4 (MARSTAT + HAVEDEG + EARN + COVER12) - 1882.8

'Full' Model (Model 5)

- AIC – 1877.4
- AUC – 0.857
- Accuracy – 80.7%

Significant Predictors (odds ratios from the 'full' model)

- The 'Other' category of MARSTAT had an odds ratio of 0.041, indicating a strong negative association with having children. Recall that the 'Other' category of MARSTAT_BINARY represented those respondents in the dataset who were *not* cohabitating with their current wife/partner, including those who were widowed, separated or divorced, cohabitating *but not married* to their partner, etc.). From this, it is observed that people who did not get married *and* cohabitate with the individual that they were married to were much less likely to have children than those who did marry and cohabitate. Even if an individual was married, if they separated from their partner, the likelihood of them having a child decreased significantly.

- Another significant predictor was EARN_BINARY, with an odds ratio of 0.426. Individuals who earned lower income were much less likely to have children. Refer to Figure 1, which shows the different incomes in each numerical category for the EARN variable. Those who earned less than category 8 ($25,000 - $29,000) annually were classified as low income in the model.
- The final significant predictor was ORIENT_BINARY, with an odds ratio of 0.372. Those who identified with a sexual orientation categorized as 'Other' (i.e. not straight/heterosexual) were less likely to have children.

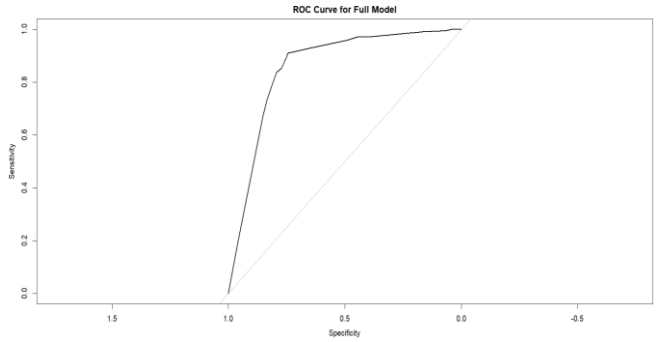

Figure 3: ROC curve for the 'full' model

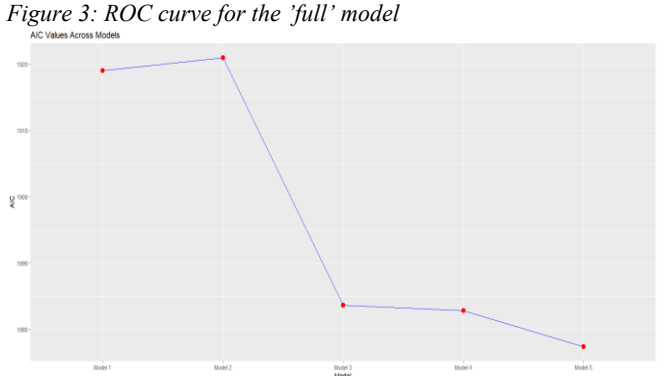

Figure 4: AIC values across models. Notice the significant drop from Model 2 to Model 3

e. The baseline model had an accuracy of 66.90%, while the full logistic regression model had an accuracy of 80.75%. The full model significantly outperformed the baseline model by around 14%, and the AUC score of the full model (0.857) indicated that it was able to predict the likelihood of an individual having children with strong accuracy.
The full logistic regression model performed well for several reasons. The variables chosen for the model are generally known to have strong impacts on fertility and childbearing. The filtering and recoding process accounted for any missing or unknown values in the dataset; after filtering, the dimensions of the dataset changed. It is crucial that all five models were trained on the exact same dataset with the exact same dimensions, so filtering and recoding had to be done very carefully, especially considering that the initial dataset was quite large. The odds ratio and confidence interval calculations ensured that each predictor had a meaningful contribution to the model overall.
None of the variables were found to be confounders. Adding each new variable to the model did not change the relationship of the primary variable and the outcome variable. The coefficients of each variable remained relatively unchanged across the five models. The chi-square tests were significant ($p < 0.001$); each variable increased the predictive accuracy of the model without affecting the other predictors and the outcome.

Consider the following case study:
A heterosexual male is separated from his partner. He is an electrician educated via trade school, and he made $27,000 the previous year. He also has no health coverage. Based on the model, this man has a low probability of having children:

- Marital status: the man has a 4% chance of having children compared to a man who is married and cohabitating
- Income: the likelihood of the man having children based on his income is 43% of the chance of a man who is high income
- Education level: the man has a slightly lower chance of having children based on his lack of a college or university degree; however, this number is not statistically significant.

Other variables, such as insurance coverage status were not significant. This was a general finding of the model.

	MARSTAT_BINARY	HAVEDEG_BINARY	EARN_BINARY	COVER12_BINARY	ORIENT_BINARY	Predicted_Prob
1	Married/Cohabiting	Yes	HighIncome	Yes	Heterosexual	0.61059904
2	Other	No	LowIncome	No	Heterosexual	0.04218908

Figure 5: table of case study results

IV. Conclusion.

This study aimed to explore several factors that influence the likelihood of an individual having children. These predictors included marital status, college/university education, annual income, insurance coverage, and sexual orientation. A dataset was filtered and the variables recoded in order to be compatible with the logistic regression model used. Each variable was added to the model one-by-one in order to assess their impact on the model and general significance to the research aim. It was found that individuals who are no married and cohabitating with their wife/partner had a significantly lower chance of having children; marital status was a significant predictor in this model. Additionally, income and sexual orientation both had a notable impact on having children, with those classified as lower income less likely to have children, and those who identified as anything other than straight/heterosexual also less likely to have children. Variables such as education level and health insurance coverage had no notable impact on the likelihood of having children.

The full logistic regression outperformed the baseline model significantly, and was found to have strong predictive power.

From this study, it is observed that social and economic demographic factors influence the likelihood of an individual having children the most. Therefore, policies and measures to

stabilize and increase the birth rate in the United States should focus on those areas.

References

[1] Council of Economic Advisers, "Issue Brief: A First Principles Look at Historically Low U.S. Fertility and its Macroeconomic Implications," May 23, 2024. [Online]. Available: https://www.whitehouse.gov/cea/written-materials/2024/05/23/issue-brief-a-first-principles-look-at-historically-low-u-s-fertility-and-its-macroeconomic-implications/#:~:text=Recently%2C%20the%20CDC%20released%20provisional,onset%20of%20the%20Great%20Recession. [Accessed: Dec. 5, 2024].

[2] The Week, "Why is the U.S. fertility rate declining?" 2023. [Online]. Available: https://theweek.com/science/us-fertility-rate-declining-2023. [Accessed: Dec. 5, 2024].

[3] VoxUkraine, "Few of us and even fewer of you: What policies for increasing birth rates really work?" [Online]. Available: https://voxukraine.org/en/few-of-us-and-even-fewer-of-you-what-policies-for-increasing-birth-rates-really-work#:~:text=In%20Sweden%2C%20between%201983%20and,in%20the%20country%20since%201974. [Accessed: Dec. 5, 2024].

[4] UIndy Reflector, "The U.S. needs paid parental leave for everyone," Dec. 15, 2021. [Online]. Available: https://reflector.uindy.edu/2021/12/15/the-u-s-needs-paid-parental-leave-for-everyone/#:~:text=At%20the%20national%20level%2C%20parents%20are%20only,through%20the%20Family%20and%20Medical%20Leave%20Act. [Accessed: Dec. 5, 2024].

[5] Axios, "U.S. birth rate hits a new low: Policy solutions," Sep. 9, 2024. [Online]. Available: https://www.axios.com/2024/09/09/us-birth-rate-low-policy-solutions. [Accessed: Dec. 5, 2024].

[6] Centers for Disease Control and Prevention, "National Survey of Family Growth," [Online]. Available: https://www.cdc.gov/nchs/nsfg/index.htm. [Accessed: Dec. 5, 2024].

[7] Centers for Disease Control and Prevention, "NSFG 2013-2015 Public Use File," [Online]. Available: https://www.cdc.gov/nchs/nsfg/nsfg_2013_2015_puf.htm. [Accessed: Dec. 5, 2024].

[8] Centers for Disease Control and Prevention, "About the National Survey of Family Growth," [Online]. Available: https://www.cdc.gov/nchs/nsfg/about_nsfg.htm. [Accessed: Dec. 5, 2024].

[9] Institute for Family Studies, "More Money, More Babies: What's the Relationship Between Income and Fertility?" [Online]. Available: https://ifstudies.org/blog/more-money-more-babies-whats-the-relationship-between-income-fertility. [Accessed: Dec. 5, 2024].

[10] The Williams Institute, UCLA School of Law, "LGBT Parenting in the United States," [Online]. Available: https://williamsinstitute.law.ucla.edu/publications/lgbt-parenting-us/. [Accessed: Dec. 5, 2024].