## **Part1: Project Overview**

### **Step 1: Establish Research Objectives**

* Primary Objective: Investigate whether wealthier and more educated counties exhibit lower fertility rates between 2017 and 2021.
* Secondary Objective: Examine the impact of family planning participation on fertility rates.

### **Step 2: Data Collection**

I found the majority of the dataset on the California Open Data Website.

* Primary Data Sources:
  + Final Birth Data (2000-2021): [Final Birth by year by county](https://data.ca.gov/dataset/live-birth-profiles-by-county/resource/360ea8da-7bf9-437a-bdc5-328ec5451cea)
  + County Revenue per Capita (2003-2021)：[County revenue per capita](https://data.ca.gov/dataset/county-revenues-per-capita)
  + Educational Attainment for Adults Aged 25+ in the U.S., States, and Counties (1970-2021): [Educational attainment for adults age 25 and older in the U.S., States, and counties](https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/)

### **Step 3: Data Preparation**

* Data Type Standardization: Convert all data to appropriate numerical types.
* Temporal Alignment: Restrict analysis to data from 2017-2021.
* Data Integration: Merge datasets based on the common 'county name' column.
* Data Forecasting: Project family plan data for 2020 and 2021 using historical averages.

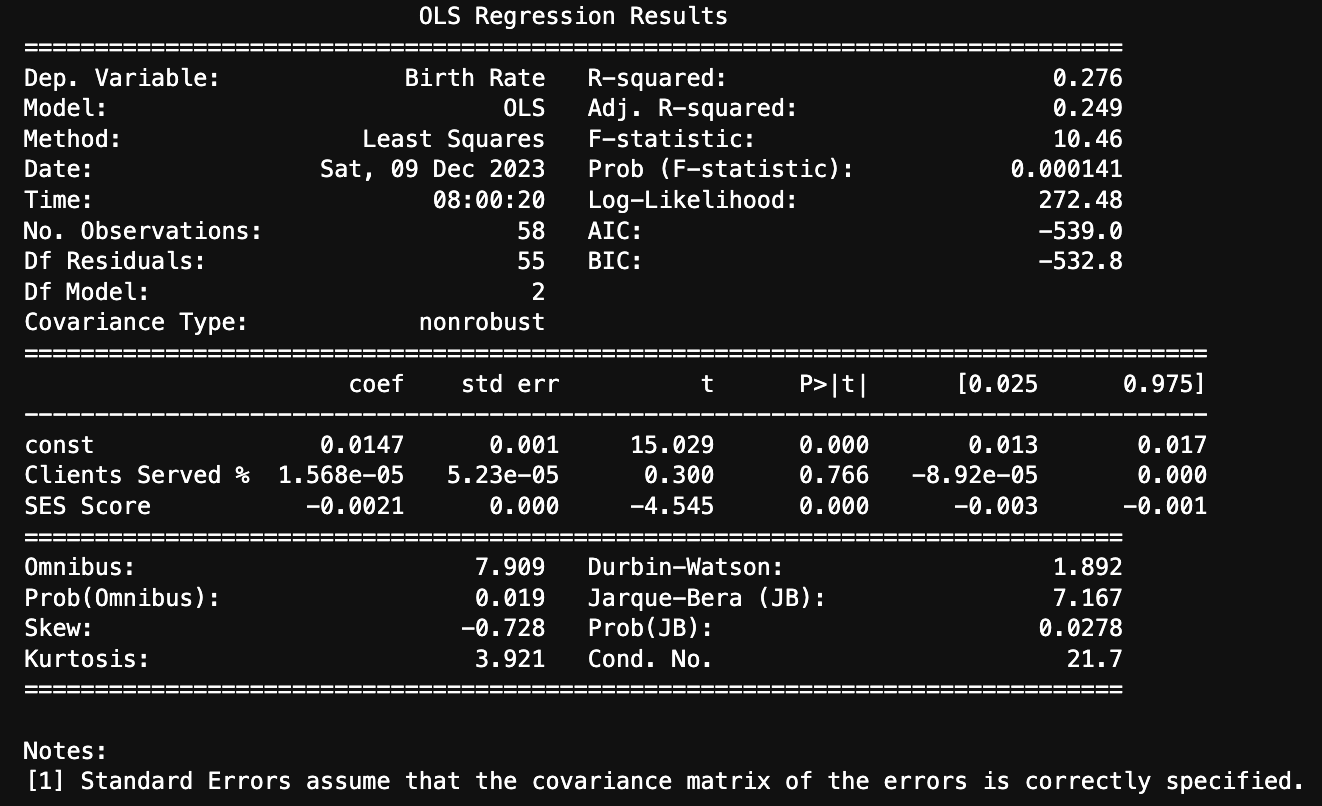
### **Step 4: Data Consolidation**

* Column Standardization: Harmonize column names across datasets for 'county'.
* Dataset Merging: Create a composite dataset with county name, revenue per capita, education level, and family planning participation.

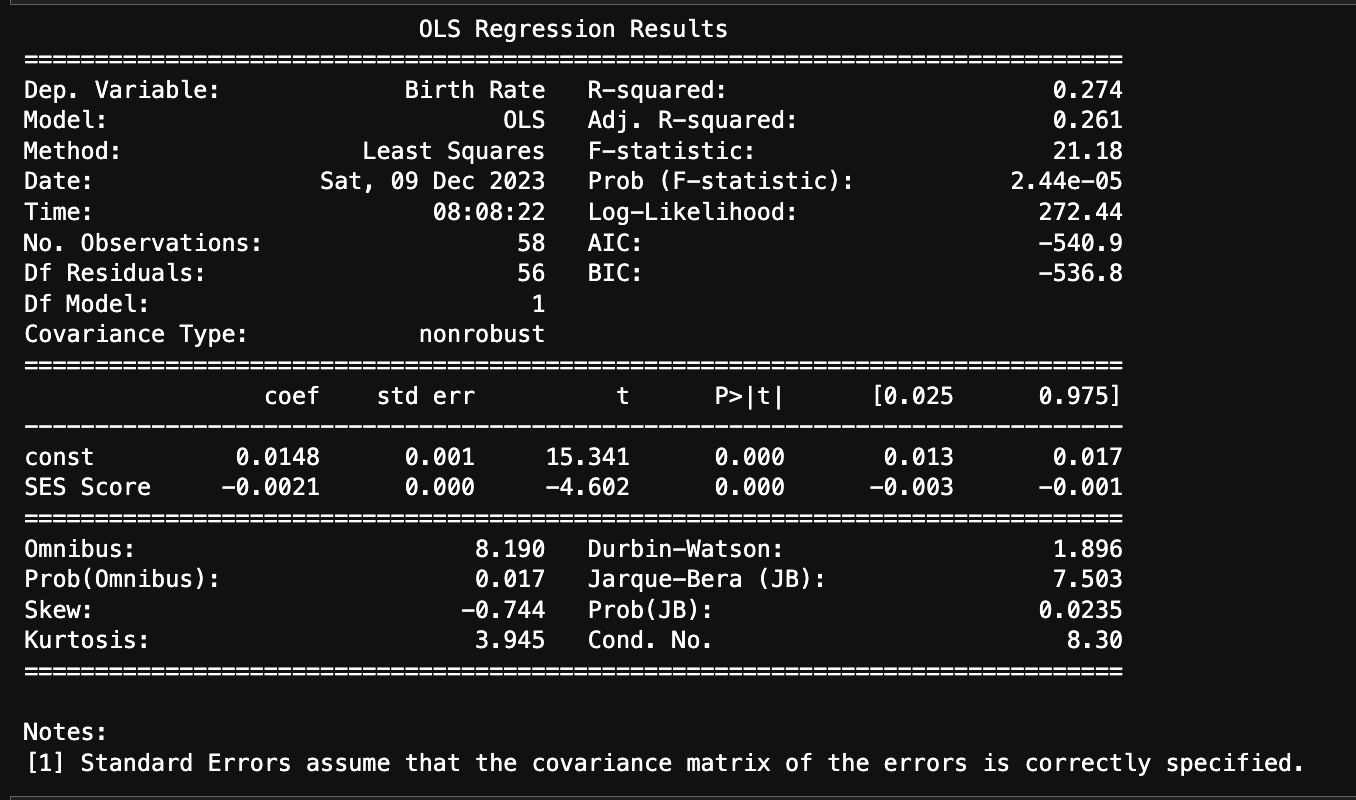
### **Step 5: Data Annotation and Classification**

* Define an SES score by integrating revenue and education data.
  + I get the upper bound and lower bound of revenue per capita, and assign label based on upper/lower bound -> 0 means low, 1 means mid , 2 means high
  + Education attainment is quantified using a weighted scoring system (1-4)
    - Percent of adults with less than a high school diploma: 1
    - Percent of adults with a high school diploma only: 2
    - Percent of adults completing some college or associate's degree: 3
    - Percent of adults with a bachelor's degree or higher: 4
  + with a combined SES score assigned to each record.
    - I combine revenue per capita score and Education attainment score to generate a new SES(social-economic status) score
* Client served %: the percent of clients served by the Family Planning, Access, Care, and Treatment (Family PACT) Program. I use this variable from dataset directly.
* Birth Rate: I use the dataset birth amount and the population of this county, and manually calculate the birth rate.

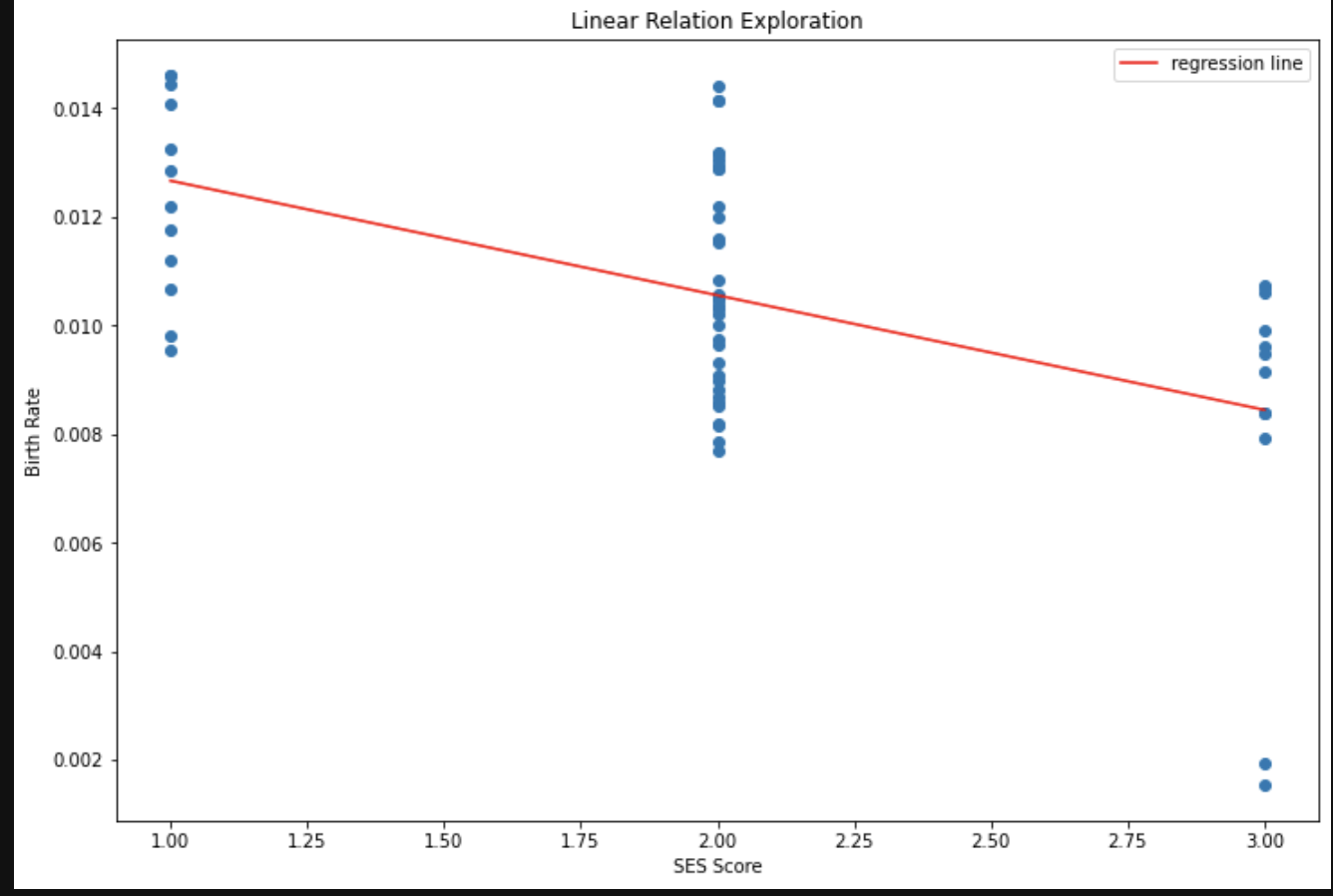
### **Step 6: Data Analysis**



* OLS Regression Analysis: Fit a model with SES score and client service percentage as predictors and birth rate as the response variable.
* Model Evaluation: The model's R-squared value is 0.276, and the F-statistic is significant (p < 0.0001), indicating a strong model. The SES score's coefficient is significantly negative (-0.0021, p < 0.0001), suggesting a decrease in birth rate with increasing SES. In contrast, the client service percentage was not significant (p = 0.766) and was excluded in subsequent modeling.
* Adjustment: I remove the client served factor to run the regression model again.



### **Step 7: Data Visualization**



* Visualization of Findings: The negative slope of the regression line is consistent with the coefficient for the SES Score reported in the regression output (-0.0021). This confirms that there is a statistically significant negative association between SES Score and Birth Rate as indicated by the regression analysis.The distribution of points suggests that there could be other factors affecting the Birth Rate since the data points show some spread around the regression line. This spread indicates the degree of variance in the Birth Rate that the SES Score doesn't account for.

### **Step8: Problems**

In the research process, identifying datasets that align with the study's variables and timeframe proved challenging. Initial datasets were abundant but often failed to coincide with the dependent variable's time period or lacked a common "county name" identifier. Ultimately, I narrowed the independent variables to revenue, educational attainment, and family planning participation due to these constraints. Although the combined Socioeconomic Status (SES) variable revealed a significant correlation with birth rates, the regression model's explanatory power was limited, as indicated by the moderate R-squared value. Visual analysis further highlighted the likelihood of additional, unaccounted factors influencing birth rates. This suggests a need for further data exploration to capture the complexity of the factors at play.

## **Part2: Reflection on Class Materials**

When I investigate the interplay between socioeconomic status (SES) in different counties, family planning, and fertility rates, I find the material on categorization highly relevant. The project's foundation is built upon creating meaningful and analytical categories that can help interpret complex social data. Through the lens of classical and cognitive categorization theories, I can better understand and structure the diverse variables.

Classical categorization, which relies on clear, distinct attributes, initially guides the project's data collection and preprocessing. Here, uniform data types and the consolidation of the time period are essential to create comparable categories. And this also reflects on the single property categories.

And the classification of counties by SES level, combining revenue and education attainment, follows a traditional categorical approach. Each category has precise boundaries, much like the classical view where objects either belong to a category or they don't based on specific shared properties.