How Economic Development and Educational Attainment Impact Female Fertility Rate: A Regression Approach

Amanda Yang, Caitlin Wei, Max Jiang, Angela Xie

2024-12-09

Group Member Emory IDs

Amanda Yang: 2512275 Caitlin Wei: 2551023 Max Jiang: 2547409 Angela Xie: 2515217

Introduction

This research examines the relationship between economic development, educational attainment, and female fertility rates, focusing on adolescent fertility and total fertility rate as indicators of reproductive decisions and societal attitudes. Using World Bank data, the study employs multivariable regression to analyze how GDP per capita and school enrollment levels influence fertility rates across four regions: Eastern and Southern Africa, North America, East Asia & Pacific, and the European Union. The findings aim to guide policymakers on addressing demographic trends, labor markets, and public health while promoting gender equity and societal development.

Our analysis reveals a relatively strong inverse relationship between GDP per capita and adolescent fertility rates, with higher economic development linked to lower adolescent fertility. Additionally, educational attainment, particularly at the primary and secondary levels, emerged as a significant factor in reducing fertility rates, highlighting the need for targeted investments in education to address demographic challenges.

Data Description

We selected six variables from the World Bank open database: Adolescent fertility rate (births per 1,000 women ages 15-19), Fertility rate, total (births per woman), School enrollment, tertiary (% gross), School enrollment, secondary (% gross), School enrollment, primary (% gross), and GDP per capita (current US$).

Data Analysis

Data Cleaning

We used both SQL and Python in our data cleaning process. We began by converting the CSV file into a Dask DataFrame, which allowed us to manage large datasets efficiently and in a distributed fashion. After loading the data, we examined the DataFrame's columns and their data types to understand the structure and identify any potential irregularities. Whenever we encountered column names that included spaces or special characters, we renamed them to more SQL-friendly versions without changing their fundamental meaning, ensuring consistent naming conventions throughout the dataset. We also addressed missing values by imputing them with appropriate statistical measures, where we filled in the missing values with the average of two most recent years' values. We also dropped out columns that will not be used later for data analysis to further clarify the dataset. Additionally, we verified that all numeric columns were properly recognized as numerical data types, converting strings to numeric types when necessary. Once these steps were complete, the data was free from easily fixable inconsistencies, the column names were standardized, and the dataset was prepared for SQL-based operations. At that point, we registered the cleaned and preprocessed DataFrame as a SQL table within Dask-SQL, ensuring that each column was accurately represented in the query context.

Summary Statistics

Summary Statistics By Region

Region	Africa Eastern and Southern	East Asia & Pacific	European Union	North America
Obs	50	50	50	50
Avg_Fertility	123.659179	27.250583	18.515741	39.815799
Min_Fertility	94.688181	20.101836	9.066463	14.375896
Max_Fertility	152.504673	39.598045	36.541219	57.855551
Avg_GDP	989.952961	5103.408136	20890.010973	35286.638374
Prim_Enroll	86.117031	108.536252	102.800165	100.923277

Region	Africa Eastern and Southern	East Asia & Pacific	European Union	North America
Sec_Enroll	30.448847	64.249117	98.799087	95.774863
Ter_Enroll	4.896809	21.045803	47.37911	73.226633

Based on the summary statistics above, we can see that Africa Eastern and Southern generally have higher adolescent fertility, lower gdp per capital and education enrollment compared to other regions.

GDP Growth by Region

	Region	avg_gdp_per_capita	gdp_growth
0	North America	35286.638374	71466.380836
1	European Union	20890.010973	37143.177577
2	East Asia & Pacific	5103.408136	12529.099105
3	Africa Eastern and Southern	989.952961	1336.021219

The findings highlight significant economic disparities, with North America and the European Union leading in GDP per capita and growth, while Africa Eastern and Southern lags far behind.

Data Visualization

Figure 1 illustrates trends in adolescent fertility rates (births per 1,000 women aged 15-19) across regions from 1974 to 2019: 1. Africa Eastern and Southern consistently shows the highest fertility rates, but there has been a steady decline over time. 2. North America has experienced fluctuations but maintains moderate levels of adolescent fertility compared to other regions. 3. East Asia & Pacific and the European Union have the lowest adolescent fertility rates, showing significant declines and stabilizing at minimal levels over the years.

Figure 2 shows the relationship between tertiary school enrollment and adolescent fertility rate. The graph shows an obvious negative correlation between tertiary school enrollment and adolescent fertility rate; and compared to the previous graph of secondary education, the slope is steeper.

Figure 3 shows an obvious negative correlation between GDP per capita and adolescent fertility rate; and compared to the previous graphs, the slope is the steepest.

Figure 4 highlights a positive correlation between tertiary school enrollment (% gross) and GDP per capita. Regions such as North America and the European Union exhibit high GDP per capita alongside higher tertiary enrollment rates, indicating the potential influence of

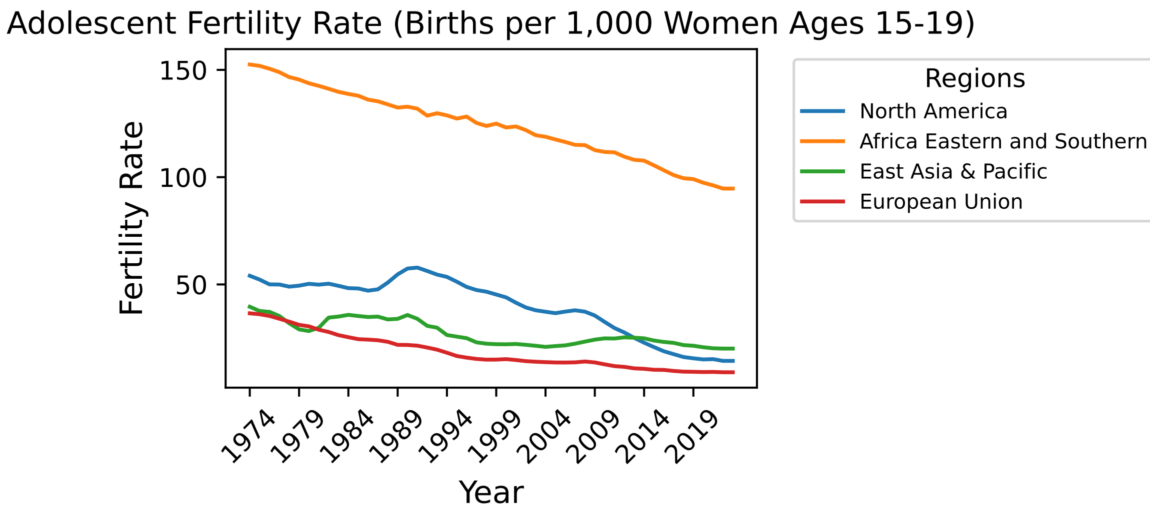


Figure 1: Line Plot showing the adolescent fertility rate from 1974 to 2019

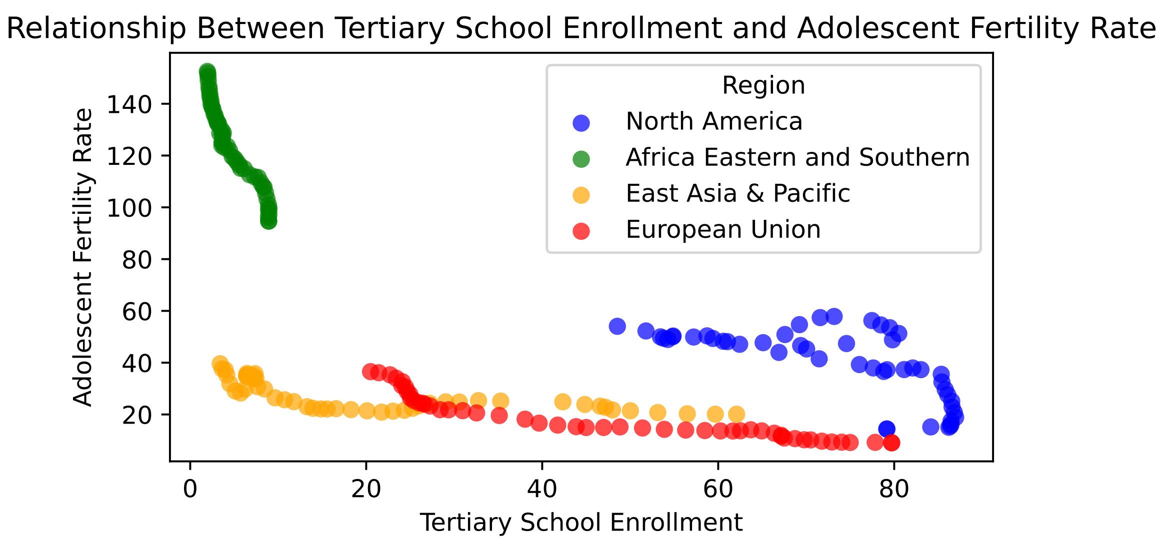


Figure 2: Scatterplot showing the Relationship Between Tertiary School Enrollment and Adolescent Fertility Rate

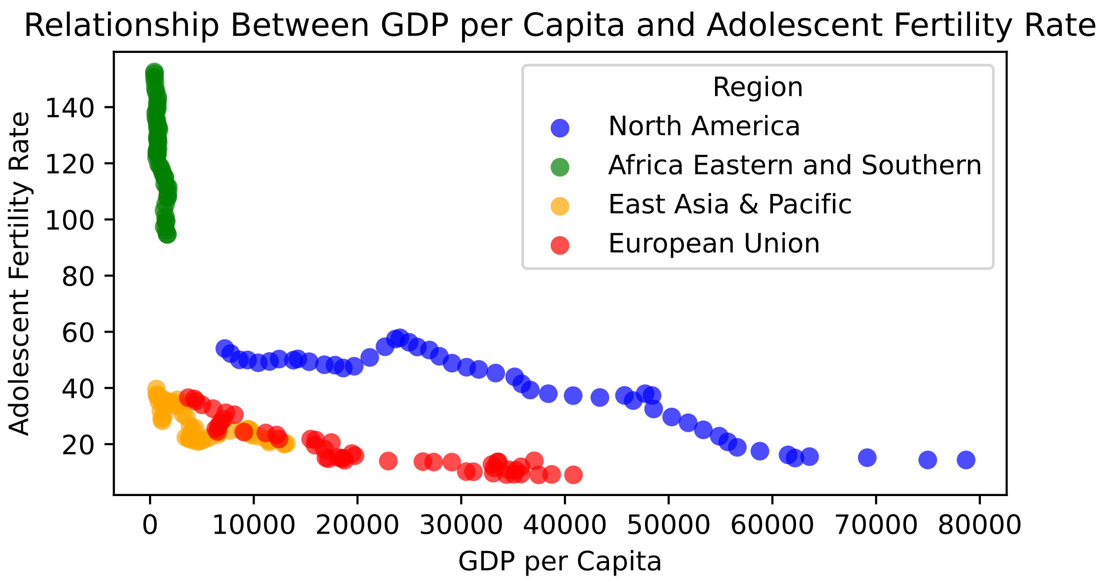


Figure 3: Scatterplot showing the Relationship Between GDP per Capita and Adolescent Fertility Rate

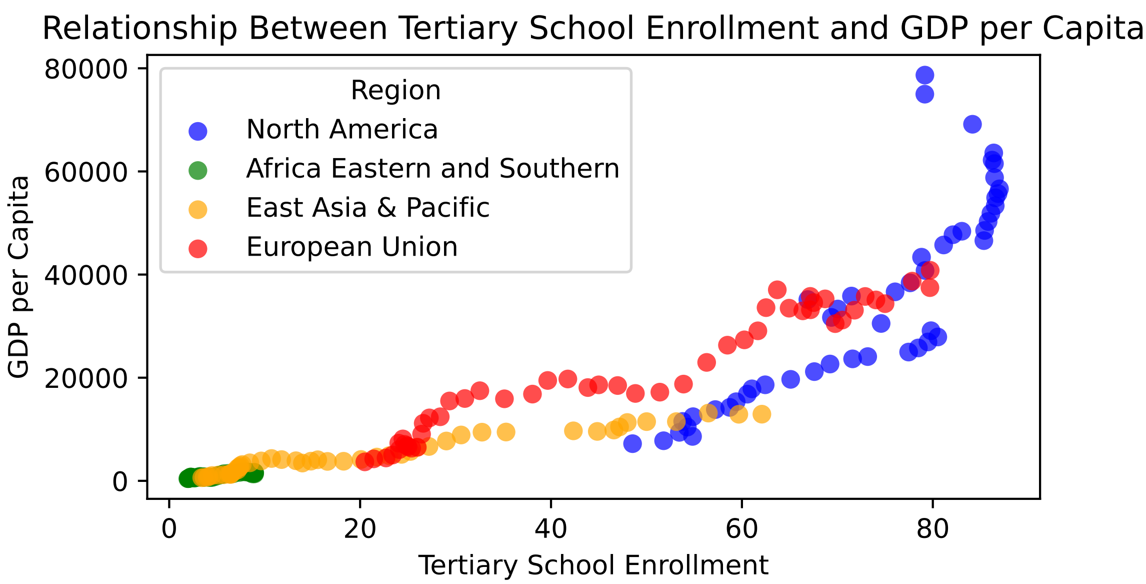


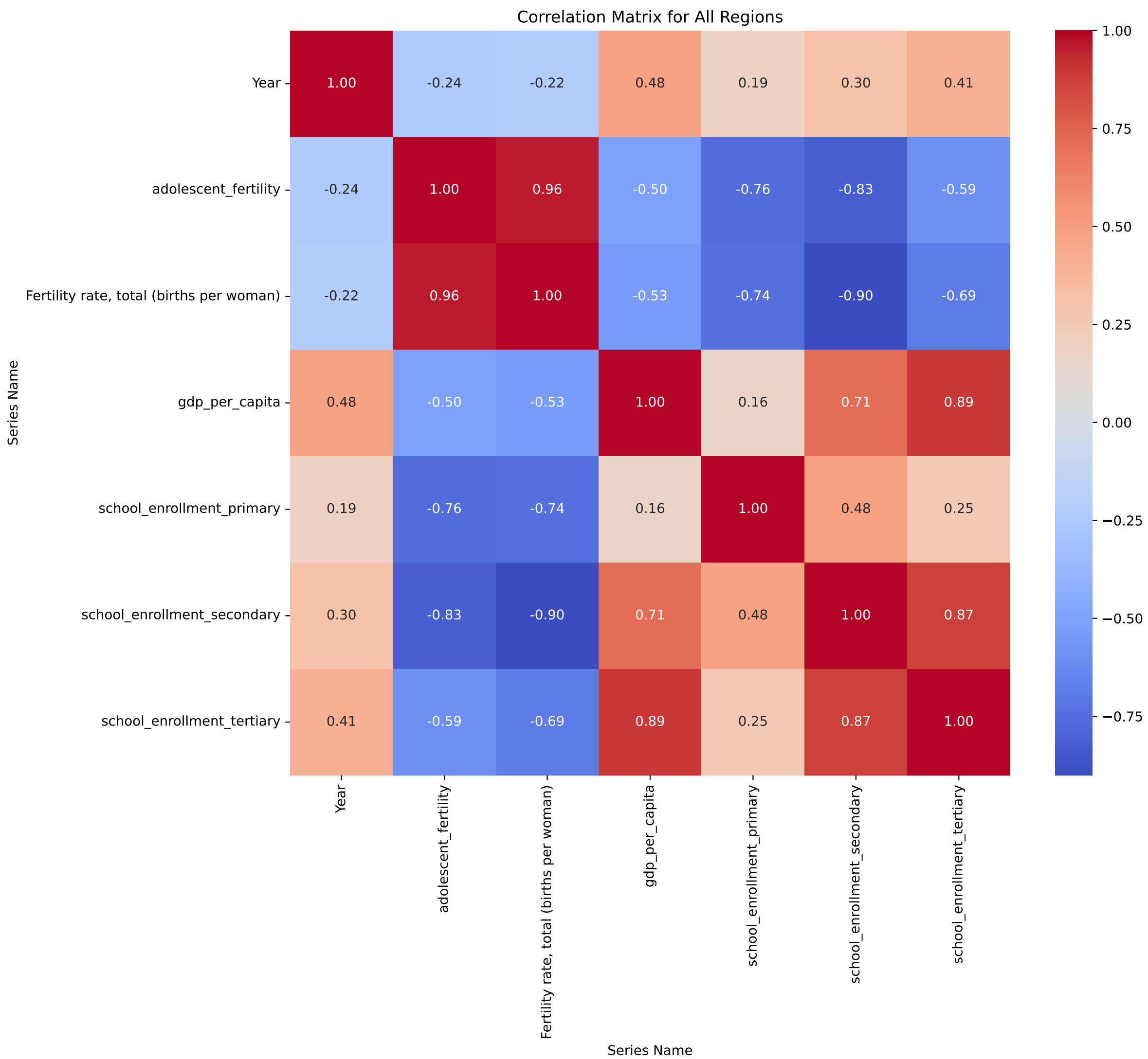
Figure 4: Scatterplot showing the Relationship Between Tertiary School Enrollment and GDP per Capita

advanced education on economic prosperity. Conversely, regions like Africa show lower GDP per capita and tertiary enrollment, suggesting gaps in higher education access and economic outcomes.

Regression Analysis

Correlation Analysis

In this part, we generate the correlation matrices for all variables in four regions.



Findings:

1. The Role of Education: In all regions, higher secondary and tertiary school enrollments are strongly negatively correlated with adolescent fertility, which highlights the importance of advanced education in reducing teenage fertility rates.
2. Economic Influence: GDP per capita consistently shows a strong negative correlation with adolescent fertility rates, with wealthier regions tending to have lower adolescent fertility.
3. Regional Differences: Africa shows the strongest link between education and reduced fertility, while North America and East Asia show weaker correlations, indicating other factors may be at play.

Regression Analysis

OLS Regression before standardizing variables

For the regression analysis, we initially use Ordinary Least Squares (OLS) as our baseline model. The independent variables include GDP per capita, region, year, primary school enrollment, and secondary school enrollment. To avoid perfect multicollinearity, we exclude tertiary school enrollment and use n-1 categories for the year and region variables.

The OLS regression results reveal challenges in interpreting the model due to the inconsistent scaling of the independent variables, which makes it difficult to compare the effects of different regressors. For instance, most of the coefficients for the region variable are below -100. Additionally, the model exhibits strong multicollinearity.

Standardizing and refining dataset

To address these problems, we standardize variables using StandardScaler library to make sure each variable has a (0,1) distribution. There might be too many dummy variables as well. So we create a different variable called year_grouped that classify specific years into decades.

Comparison between Ridge and OLS Regression after revising dataset

To address problems about multicollinearity, we used Ridge regression to compare it with OLS using the refined dataset. We fine tuned the ridge regression by trying out different alphas. For ridge regression, the r squared is 0.958, and MSE is 0.042. For OLS regression, R squared is 0.987 and MSE is 0.013. Overall, ridge is better since it address multicollinearity and remain in a high r squared.

Run Ridge Model to Check Rubustness

In order to check robustness, we tried out different ridge models by adding gap_per_capital, school_enrollment, region and Year_grouped variables accordingly.

```
R^2: 0.225, MSE: 0.775, Best Alpha: 100.0
Coefficients:
         Variable  Coefficient
0  gdp_per_capita    -0.335722

================================================================================

R^2: 0.851, MSE: 0.149, Best Alpha: 1.0
Coefficients:
                      Variable  Coefficient
0               gdp_per_capita    -0.010159
1    school_enrollment_primary    -0.471573
2  school_enrollment_secondary    -0.587996

================================================================================

R^2: 0.952, MSE: 0.048, Best Alpha: 10.0
Coefficients:
                      Variable  Coefficient
0               gdp_per_capita    -0.186083
1    school_enrollment_primary    -0.316172
2  school_enrollment_secondary    -0.509877
3   Region_East Asia & Pacific    -0.656009
4        Region_European Union    -0.328958
5         Region_North America     0.117550

================================================================================

R^2: 0.958, MSE: 0.042, Best Alpha: 10.0
Coefficients:
                       Variable  Coefficient
0                gdp_per_capita    -0.217385
1     school_enrollment_primary    -0.322671
2   school_enrollment_secondary    -0.516964
3    Region_East Asia & Pacific    -0.648517
4         Region_European Union    -0.294195
5          Region_North America     0.171980
```

```
6             Year_grouped_1980     0.000768
7             Year_grouped_1990    -0.009741
8             Year_grouped_2000     0.076593
9             Year_grouped_2010     0.137006
10            Year_grouped_2020     0.067865

==========================================================================
```

After trying different ridge models, the coefficients for regressors do not change significantly.

1. Economic and Educational Factors:

- gdp_per_capita (-0.217): A 1-standard-deviation increase in GDP per capita is associated with a 0.217 standard deviation decrease in adolescent fertility. This aligns with the expectation that economic development reduces fertility rates, particularly among adolescents.

- school_enrollment_primary (-0.323) and school_enrollment_secondary (-0.517): Higher enrollment rates in primary and secondary education are strongly associated with lower adolescent fertility rates. Secondary education has a more effect, indicating its critical role in delaying childbearing.

2. Regional Effects:

- Region_East Asia & Pacific (-0.649) and Region_European Union (-0.294): Adolescent fertility is significantly lower in these regions compared to Africa Eastern and Southern.
- Region_North America (0.172): This positive coefficient indicates slightly higher fertility in North America compared to Africa Eastern and Southern.

3. Temporal Trends

- Year Dummies (Year_grouped_1980 to Year_grouped_2020): The coefficients for year groups are relatively small, indicating gradual changes in fertility over decades. Later years (e.g., 2010, 2020) show positive coefficients compared to earlier years, suggesting slight increases in adolescent fertility over time, though the effects are minimal.

Discussion

The findings highlight the importance of economic and educational interventions in reducing adolescent fertility, particularly in high-fertility regions. However, some limitations should be considered. The use of grouped year variables may oversimplify temporal trends, potentially overlooking finer changes. Multicollinearity among predictors, especially between education and regional variables, may affect coefficient stability despite using Ridge regression. Additionally, excluding factors like cultural influences or healthcare access limits the analysis.

Future research could address these gaps by incorporating additional predictors, exploring interaction effects, and employing alternative modeling techniques to better understand the complex factors influencing adolescent fertility. Despite these constraints, the study offers valuable insights for policymakers, emphasizing the need for targeted investments in education and economic development to address adolescent fertility challenges effectively.