Books and Babies: An Analysis of the Relationship between Female Literacy Rate and Fertility Rate in South Asia (2000-2020)*

Jagpreet Singh

April 19, 2024

This paper utilizes data from the World Bank's Gender Statistics database to examine the relationship between the adult female literacy rate (% of females ages 15 and above) and fertility rate (total births per woman) in South Asia between 2000-2020. The analysis revealed a negative correlation between the two variables, indicating that as the literacy rate increased, the fertility rate decreased. These findings matter as they highlight the importance of investing in female education and empowerment to manage population growth and promote sustainable development. The insights can guide policymakers make decisions on education initiatives and reproductive health programs to further enhance socioeconomic progress in the region.

1 Introduction

According to the World Bank, the South Asian region comprises Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. It is the most populated region in the world with countries like India, Pakistan and Bangladesh which ranks among the top 10 in in terms of population size. Majority of these countries are also considered developing nations and play a major role in the world economy. In today's globalized world where population dynamics significantly influence a country's course it is necessary to analyze gender statistics indicators such as fertility rate and female literacy rate and how they relate to each other. By doing so we can gain valuable insights which can help make policy decisions on female education and empowerment that in-turn help manage population growth and promote sustainable development.

*Code and data are available at: https://github.com/Jagpreet5ingh/female_fertility_rate_south_asia.git

In this paper, we will examine the relationship between the fertility rate (total births per woman) and adult female literacy rate (% of females ages 15 and above) in South Asia through a linear regression analysis. The estimand here is how fertility rate and female literacy rate are related. Specifically, we will focus on South Asia, the most populated region in the world. We will draw data from the World Bank website. Based on the analysis, we found that there is a negative relationship between fertility rate and literacy rate of adult females.

In section 1, we discuss the source of data used in this paper, methodologies that follow it, and data terminology. In section 2, we present the results of our analysis, focusing on the trajectory of fertility rate and female literacy rate over the last two decades in the region of South Asia. In section 3, we will analyze the trend by establishing a linear regression model. In section 4 we will present the result of the model in a graph. Finally, in section 5 we will further discuss the findings of the result.

2 Data

2.1 Data Description and Methodology

The data used in this paper was obtained from the World Bank's Data Bank which is publicly available on the Data Bank website (DataBank 2024a). A considerable amount of the data in the World Bank's Data Bank is gathered and compiled by the World Bank itself, along with information from other reliable public and private sector sources, international organizations, national statistical offices, and specialized agencies of the United Nations (UN). The World Bank then compiles and examines these data to find patterns and provide guidance in policy matters.

The two data sets that are used in this paper are fertility rate (total births per woman)(DataBank 2024b) and literacy rate, adult females (% of females ages 15 and above) (DataBank 2024c). Both female literacy rate and fertility rate data was available for a total of 266 regions including individual countries and group of countries (for example - South Asia, Middle East etc.). While the literacy rate data was available from year 1960 to 2022, the fertility rate data consisted of data from 1960 to 2021. Not every country had data available for each year. The region of interest for this study is South Asia and the time period selected was 2000-2020. Both fertility rate and female literacy rate data was available for South Asia for the specified time period. The two indicators were merged to form a new data set with 21 rows that gives both fertility rate and female literacy rate in South Asia for any given year between 2000-2020.Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year. Female Adult literacy rate is the percentage of females ages 15 and above who can both read and write with understanding a short simple statement about their everyday life.

Table 1: A summary table of cleaned data

Year	fertility_rate_total	female_literacy_rate
2000	3.572224	44.92875
2001	3.515548	45.67490
2002	3.432268	46.99985
2003	3.339205	47.85153
2004	3.258046	48.89375
2005	3.166954	48.09562
2006	3.073746	48.75041
2007	2.999592	49.91730
2008	2.929633	50.99176
2009	2.877746	52.00314
2010	2.807602	55.80537
2011	2.742788	56.80440
2012	2.678348	57.57100
2013	2.621332	58.56601
2014	2.530008	59.59675
2015	2.498366	60.73805
2016	2.475185	62.34228
2017	2.406131	63.23561
2018	2.379341	63.78237
2019	2.320116	63.74677
2020	2.267098	64.60739

Table 1 shows the cleaned data with 21 observations from year 2000-2020. Here, fertility_rate_total and female_literacy_rate are the 2 variables of interest.

The data was cleaned and analyzed using the `R` programming language(R Core Team 2022) and the following packages were used: `tidyverse` (Wickham et al. 2019), `janitor` (Firke 2023),`dplyr` (Wickham et al. 2023), `ggplot2` (Wickham 2016), `here` (Müller 2020), `kableExtra` (Zhu 2021), `knitr` (Xie 2014), `modelsummary` (Arel-Bundock 2022), `readr` (Wickham, Hester, and Bryan 2024), `tibble` (Wickham, Müller, and Hester 2021) and `arrow` (Richardson et al. 2024)

2.2 Data Visualization

2.2.1 Trend of Female Literacy Rate in South Asia (2000-2020)

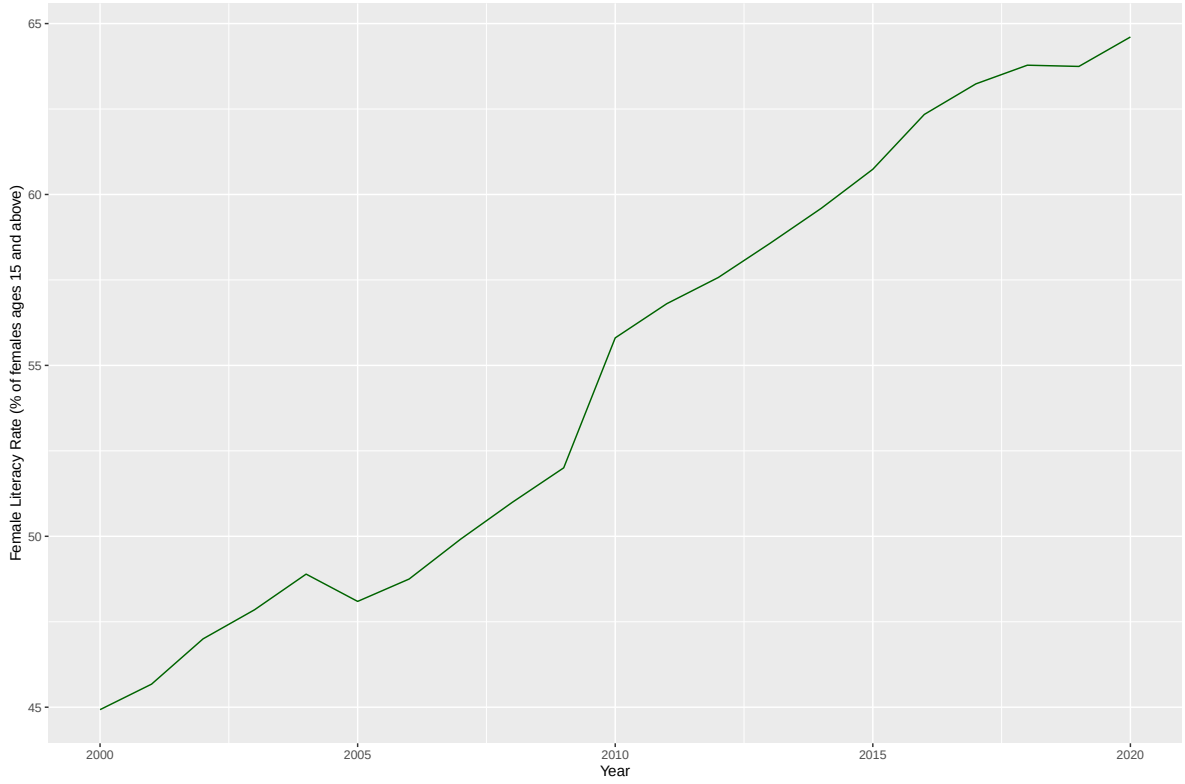

Figure 1: Literacy rate, adult females (% of females ages 15 and above) in South Asia (2000-2020)

Figure 1 represents South Asia's female literacy rate over the span of 21 years i.e. 2000-2020. The literacy rate is represented as the percentage of females ages 15 and above in South Asia who can both read and write with understanding a short simple statement about their everyday life. The graph depicts an overall upward trend in the female literacy rate over the two decades. The female literacy rate in year 2000 was approximately 44.9287 % which grew consistently over the period of 21 years to reach 64.6073 % as of 2020. There were few minor fluctuations in the opposite direction but no drastic change was observed in the upward trend.The figure suggests that more and more females of ages 15 and above in South Asia acquired literacy over the years.There could be various social, economic and political changes that may have influenced the literacy rate over this time period.

2.2.2 Trend of Fertility Rate in South Asia (2000-2020)

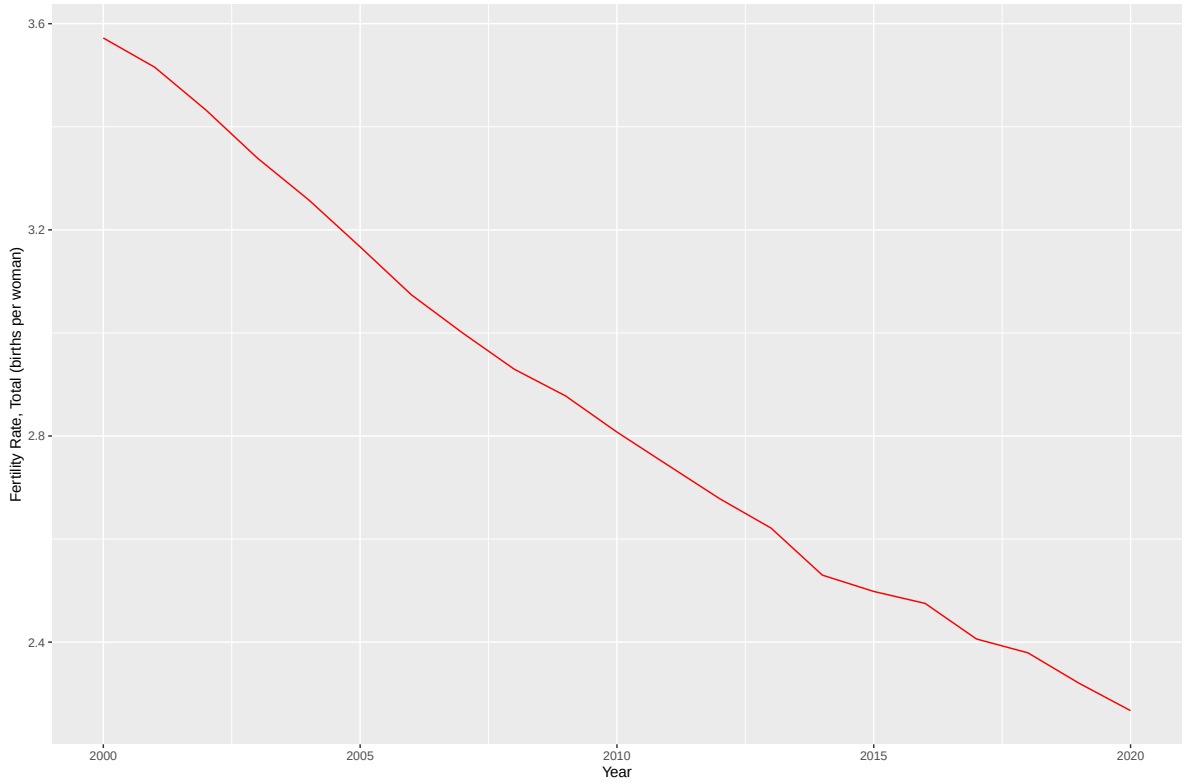

Figure 2: Fertility Rate, Total (births per woman) in South Asia (2000-2020)

Figure 2 represents South Asia's fertility rate over the span of 21 years i.e. 2000-2020. The fertility rate is represented as the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year. The graph depicts a strictly downward trend in the fertility rate over the past two decades. The fertility rate in year 2000 was 3.5722 children per woman which declined over the period of 21 years to 2.2670 children per woman as of 2020.The figure suggests that females in South Asia are choosing to have lesser and lesser number of children in recent times.There could be various social, cultural, economic and political changes that may have influenced this decline in fertility rate.

3 Model

Based on the preliminary analysis, we observed a negative relationship between fertility rate, total (births per woman) and literacy rate, adult females (% of females ages 15 and above). This indicates a potential linear regression.

Below is the equation of the linear regression model:

$$Yij = \beta 0 + \beta 1 Xij + \varepsilon ij$$

where:

- Y is the Fertility Rate
- X is the Female Literacy Rate
- Y_{ij} is the Fertility Rate for observation j in year i.
- X_{ij} is the Female Literacy Rate for observation j in year i.
- β_0 is the intercept/constant term, which represents the expected value of Fertility Rate when the Female Literacy Rate is equal to zero.
- β_1 is the slope coefficient or the estimated change in Fertility Rate for a one-unit increase in the Female Literacy Rate.
- ϵ_{ij} is the error term or the deviation of the actual value of Fertility Rate from the predicted value based on the regression equation.

This linear regression model aims to estimate the values of 0 and 1 such that the model fits the data well, and predicts the expected value of the Fertility Rate for different values of the Female Literacy Rate. The statistical significance of 1 can be assessed using a t-test, which tests whether the estimated coefficients are significantly different from zero. If the p-value of the t-test is less than the selected level of significance, we can conclude that there is a significant relationship between the Female Literacy Rate and the Fertility Rate.

4 Results

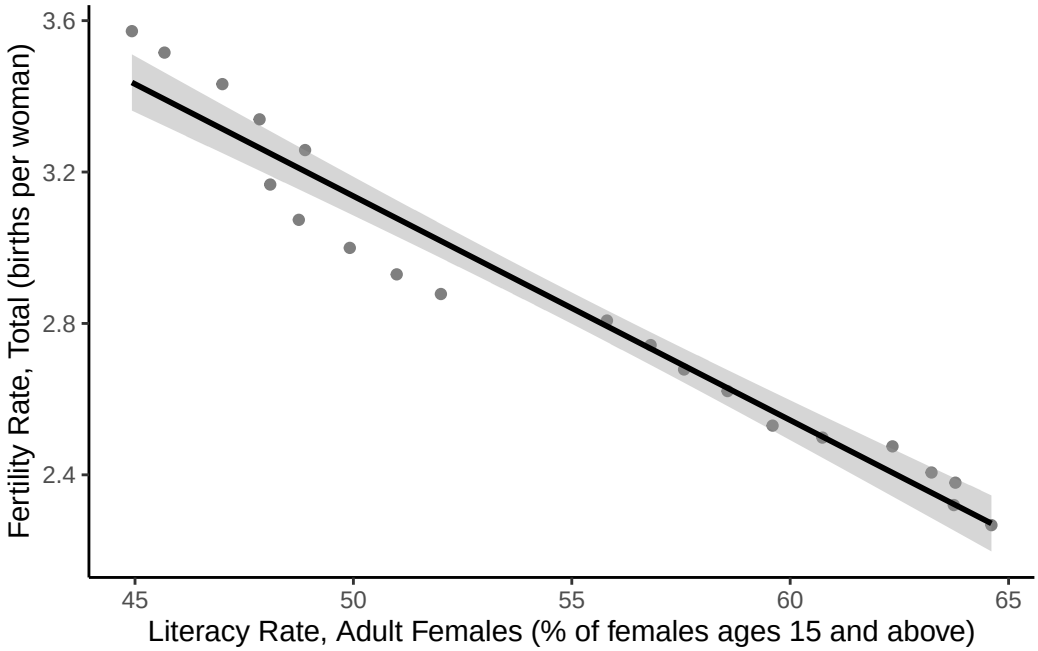

Figure 3: The model shows the relationship between Female Literacy Rate and Fertility Rate in South Asia (2000-2020)

Figure 3 is a depiction of a linear regression model showing the relationship between fertility rate, total (births per woman) and literacy rate, adult females (% of females ages 15 and above). The female literacy rate is given by the x-axis and the fertility rate is given by the y-axis. The solid black line indicates the trend of this relationship as estimated by the linear regression model. The confidence interval for the regression line is given by the grey shaded area and also represents the uncertainty around the estimated relationship. Analyzing the regression line, we can observe that there exists a negative relationship between fertility rate, total (births per woman) and literacy rate, adult females (% of females ages 15 and above). As the literacy rate increases the total number of children a woman gives birth decreases. The negative sloping regression line indicating a negative correlation and the confidence interval giving the range of the uncertainty, we can say that a true regression lies in this area.

Table 2: T-Value and P-Value of Regression

	t_value	p_value
(Intercept)	36.95598	0
female_literacy_rate	-19.81133	0

Table 3: Summary of the linear regression model

	Coefficients	Standard_Error
(Intercept)	6.097	0.165
female_literacy_rate	-0.059	0.003

Metric	Value
Num.Obs.	21.00000
R2	0.95400
R2 Adj.	0.95100
AIC	-37.28422
BIC	-34.15066
Log.Lik.	21.64211
RMSE	0.09000

Table 2 and Table 3 provides us with a vital summary of the regression model that further helps us to understand the relationship between fertility rate, total (births per woman) and literacy rate, adult females (% of females ages 15 and above). Together, these summary statistics tells us the following about our regression model:

- The t-value for the female literacy rate is approximately -19.811, which means that the estimate is -19.811 standard deviations away from 0. The negative sign of the t-value indicates a negative relationship between female literacy rate and fertility rate in South Asia.
- The t-value for the intercept is 36.95598, indicating that the estimate is significantly different from zero.
- The p-value for both female literacy rate and the intercept is 0 which means that the coefficient for female literacy rate and the intercept is statistically significant at common alpha levels.
- The coefficient for the intercept is approximately 6.097, suggesting that when the female literacy rate is 0, the fertility rate is predicted to be around 6.097 births per woman. However, interpreting the intercept as meaningful may not be appropriate as a female literacy rate of 0 is not feasible and is outside the range of the data.
- The coefficient for female literacy rate is approximately -0.059, with a standard error of 0.003.This suggests that for each one-unit increase in the female literacy rate, the fertility rate is expected to decrease by 0.059 births per woman. The negative coefficient indicates a negative relationship between female literacy rate and fertility rate which means as female literacy rate increases, the fertility rate will decrease.

- The R-squared value of 0.954 represents that approximately 95.4% of the variability in the fertility rate is explained by the model.This high R-squared value suggests a strong fit of the model to the data.

The linear regression model suggests a statistically significant negative relationship between female literacy rate and fertility rate which means as female literacy rate increases, the fertility rate will decrease. However, the intercept is not statistically significant. While the model has a high R-squared value, indicating a strong fit to the data, the practical significance of the intercept may be questionable.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

DataBank. 2024a. "Data Bank Home Page." *The World Bank*. https://databank.worldbank.org/home.

———. 2024b. "Fertility Rate, Total(births Per Woman)." *The World Bank*. https://data.worldbank.org/indicator/SP.DYN.TFRT.IN.

———. 2024c. "Literacy Rate, Adult Female (." *The World Bank*. https://data.worldbank.org/indicator/SE.ADT.LITR.FE.ZS.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Kirill Müller, and James Hester. 2021. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.