Investigating the Impact of Student-Teacher Ratio on Literacy Rates: A Multifaceted Analysis

Introduction

Education is fundamental to societal progress, with literacy rates being a key indicator of a society's educational health. This study investigates the relationship between the student-teacher ratio and literacy rates among adults. The student-teacher ratio is a critical indicator of educational quality, with lower ratios potentially leading to higher literacy due to more individualized attention.

This analysis extends beyond the student-teacher ratio to include factors like government expenditure on education, GDP per capita,primary education completion rate, and average years of schooling. These variables are considered to unravel the specific impact of student-teacher ratios from the broader socio-economic context.

The goal is to determine the extent to which the student-teacher ratio, among these factors, influences literacy rates.

Data

The dataset initially comprised 47 indicators across 272 countries or regions, spanning 20 years, resulting in a total of 12,784 rows. The indicators encompass a range of educational, and economic metrics. To address the challenge of numerous missing values in yearly data, we calculated the average of each indicator across the 20-year period for each country. This approach not only mitigated the issue of missing data but also provided a more consolidated and long-term perspective of each indicator at the country level.

After the cleaning and transformation process, our dataset was refined to encompass 13 indicators for 60 countries. This reduction in the number of indicators and countries was a necessary step to ensure data quality and reliability, focusing on the most relevant and consistently reported metrics.

For our analysis, we've selected key indicators from the World Bank data, grouped into two main categories:

Educational Indicators:

- Literacy Rate, Population 25-64 Years, Both Sexes (%): Reflects the adult literacy level, essential for understanding everyday life.
- Student-Teacher Ratio in Primary Education: Indicates the average number of students per teacher, a proxy for educational quality.
- Completion Rate, Primary Education, Both Sexes (%): Measures the effectiveness of primary education systems in retaining students.
- Average Years of Total Schooling, Age 25+, Total: Represents the average educational attainment of the adult population.

Economic Indicators:

- Government Expenditure on Education as % of GDP: Shows the government's financial commitment to education relative to the country's overall economic output.
- GDP per Capita (current US$): A key measure of economic performance and standard of living.

These indicators collectively provide insights into the relationship between student-teacher ratio, economic conditions, and educational outcomes, particularly literacy rates. The educational indicators focus on the quality and effectiveness of education systems, while the economic indicators reflect national priorities and capabilities concerning education.

Out[5]:

	avg_ed_years	prim_comp_rate	gdp_capita	gdp_capita_ppp	gov_exp	lit_rate	stu_teach_ratio
count	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000	61.000000
mean	6.132623	76.386557	4256.105246	8622.676557	4.022459	75.356393	34.117049
std	2.803604	22.480215	8872.239448	14981.295774	1.603432	22.886058	20.379016
min	1.180000	22.600000	230.060000	707.020000	1.350000	26.810000	11.340000
25%	3.810000	61.750000	783.090000	2329.790000	3.020000	59.120000	21.540000
50%	6.160000	83.840000	2239.150000	6063.530000	3.870000	78.790000	27.970000
75%	8.020000	96.950000	4735.060000	9977.220000	4.780000	94.920000	43.560000
max	11.580000	99.710000	67461.320000	116342.470000	10.250000	99.980000	135.610000

Analysis

We will construct several models to investigate the true correlation between the student-teacher ratio and literacy rates, examining how other educational and economic factors may influence this relationship. Our first and third models will employ single linear regression, focusing exclusively on these two variables. Subsequent models will incorporate multiple regressions with various "control variables" representing relevant educational and economic indicators. This multi-model strategy will allow for a more nuanced understanding of the underlying dynamics.

First, we will create a scatter plot to visualize the relationship between the student-teacher ratio and literacy rates.

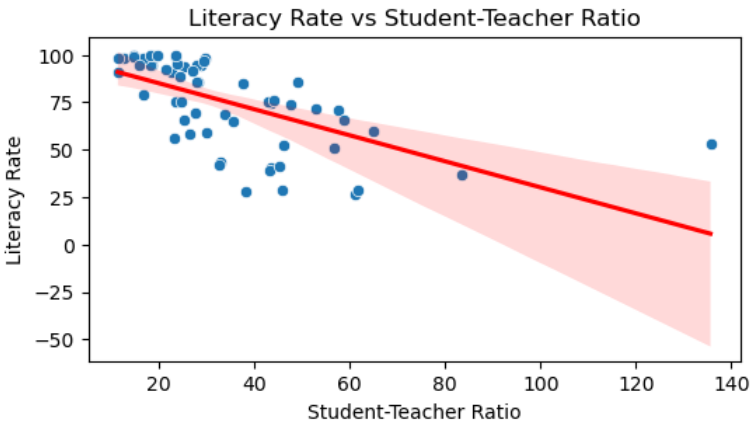

Fig. 1: Literacy Rate vs Ln-Student-Teacher Ratio

The plot (Fig. 1) reveals a negative correlation between the two variables, suggesting that an increase in the number of students per teacher could lead to a decrease in literacy rates. This may be because a teacher's capacity to effectively educate is limited by the number of students they must attend to.

In all our models standard errors are calculated using the Heteroskedasticity-Consistent (HC1) method. Heteroskedasticity occurs when the variance of the errors is not constant across all levels of the independent variable(s). The HC1 method adjusts for this by providing robust standard errors that are more reliable in the presence of heteroskedasticity.

Model 1

Level-level: $(literacy\_rate)^E = \alpha + \beta \times student\_teacher\_ratio$

In the linear regression model examining the relationship between `student-teacher ratio` and `literacy rates` (check **Appendix 1** first column), the intercept is found to be 98.761. This value represents the expected literacy rate when the student-teacher ratio is zero. While this scenario is hypothetical and not practically feasible, the high value of the intercept suggests a baseline literacy rate in an ideal educational setting with extremely low student-to-teacher ratios. The model's coefficient for the student-teacher ratio is -0.686, indicating that for each additional student per teacher, the literacy rate is expected to decrease by about 0.686 units (in our case percentages). This finding underscores the potential negative impact of higher student-teacher ratios on literacy rates.

Both the intercept and the coefficient of the student-teacher ratio are **statistically significant at 1%**, as evidenced by their p-values and confidence intervals. The p-values are sufficiently low to reject the null hypothesis that the coefficients are equal to zero, indicating a statistically significant relationship between student-teacher ratio and the literacy rate. Additionally, the confidence intervals for these coefficients do not include zero, further supporting their significance. The model's R-squared value of 0.373 implies that around 37.3% of the variation in literacy rates is explained by the student-teacher ratio, pointing to a moderate explanatory power but also suggesting the presence of other influential factors not included in the model. While the statistical analysis provides significant insights, it is important to interpret these results within the broader context of educational research, considering other variables that might also play a role in determining literacy rates.

Model 2

Level-level:
$(literacy\_rate)^E = \alpha + \beta_1 \times student\_teacher\_ratio + \beta_2 \times gov\_exp$

In the second model of our analysis, which includes both the student-teacher ratio and government expenditure on education as predictors of literacy rates (check **Appendix 1** second column), we observe some intriguing results. The model shows an R-squared value of 0.39, suggesting that about 39% of the variation in literacy rates can be explained by these two variables combined. This marks a slight improvement in explanatory power compared to the first model, which only included the student-teacher ratio.

The intercept in this model is 91.14 (signifant at 1%), which theoretically represents the expected literacy rate when both the student-teacher ratio and government expenditure on education are zero. Practically, this is an abstract baseline, as it's not feasible to have these conditions in reality. However, it does give us a starting point for understanding the model's predictions.

The coefficient for the student-teacher ratio (signifant at 1%) remains negative (-0.68) and is quite similar to that in the first model, reaffirming the inverse relationship between the student-teacher ratio and literacy rates. Interpretation is the same as in the first model (we just mention that other variables should remain constant). The consistency of this coefficient across both models strengthens the argument that higher student-teacher ratios, potentially indicative of less individual attention in classrooms, are associated with lower literacy rates.

The coefficient for government expenditure on education is 1.82, suggesting a positive relationship with literacy rates. However, its p-value of 0.08 indicates that this relationship is not statistically significant at conventional levels (like 0.05 or 0.01 that we accept for this study). This implies that while there seems to be a positive trend between government expenditure on education and literacy rates, we cannot confidently assert this as a statistically significant finding based on this model.

The confidence intervals for the student-teacher ratio exclude zero, allowing us to reject the null hypothesis that its coefficient is zero and affirm its significance in the model. However, for government expenditure, the confidence interval includes zero, reflecting its statistical insignificance in this context.

Overall, the second model suggests a nuanced view where the student-teacher ratio continues to play a significant role in literacy rates, while the impact of government expenditure on education, although positive, lacks statistical certainty. As with any regression analysis, these findings should be interpreted in the context of the broader socio-economic and educational landscape, and they highlight the complexity of factors influencing literacy.

Our analysis aims to unravel the underlying dynamics, and even non-significant variables can yield insights. Thus, we have decided to retain non-significant variables in our model for their potential explanatory value.

Before constructing Model 3, which uses the natural logarithm of the student-teacher ratio, we created a LOESS function to visualize how literacy rates vary with changes in this transformed ratio. This can be seen in **Appendix 4** (**Fig. 2**). The LOESS curve is visually quite similar to a linear function, suggesting that the relationship could be linear in nature. However, we do not place significant emphasis on the exact functional form of our regression since our primary objective is to comprehend the average association between variables, not to make predictions.

Model 3

Level-log:
$(literacy\_rate)^E = \alpha + \beta \times \ln(student\_teacher\_ratio)$

Model 3 introduces (see **Appendix 2**) a transformative approach to our analysis by incorporating the natural logarithm of the student-teacher ratio as a predictor for literacy rates. This model marks a significant improvement in our ability to explain the variability in literacy rates, with an R-squared value of 0.49. This implies that about 49% of the variation in literacy rates can now be accounted for, a substantial increase from the previous models.

The key element in this model is the coefficient for the natural logarithm of the student-teacher ratio, which stands at -31.61. This coefficient suggests a strong, negative relationship between the student-teacher ratio and literacy rates. It indicates that as the student-teacher ratio increases by 1 percent, literacy rate decrease by 0.3 points (percentages), and this relationship is more pronounced and nuanced in the logarithmic scale than in a simple linear form.

The intercept of the model is 182.69. This value, in theory, represents the expected literacy rate when the logarithm of the student-teacher ratio is zero. While this is more of an abstract concept rather than a practical situation, it helps frame the context of the model's predictions.

The statistical significance of the coefficient of the logarithm of the student-teacher ratio is robust, as evidenced by a very low p-value (significant at 1%). This strong statistical significance, combined with the large negative coefficient, underscores the importance of the student-teacher ratio in understanding literacy rates, especially when considering the logarithmic transformation of the ratio. The confidence intervals, along with the p-values, allow us to reject the null hypothesis.

Models 4 and 5

4. $(literacy\_rate)^E = \alpha + \beta_1 \times \ln(student\_teacher\_ratio) + \beta_2 \times Gov\_expend\_on\_ed + \beta_3 \times GDP\_per\_capita\_PPP$
5. $(literacy\_rate)^E = \alpha + \beta_1 \times \ln(student\_teacher\_ratio) + \beta_2 \times Gov\_expend\_on\_ed + \beta_3 \times Prim\_ed\_comp\_rate + \beta_4 \times Avg\_ed\_$

Let's add 2 more models. One with including GDP per capita (PPP). And second with Primary education completion rate and average education years. In Models 4 and 5 (see **Appendix 3**), our analysis evolves to incorporate a broader range of variables, shedding light on the multifaceted nature of literacy rates.

Model 4, with an R-squared of 0.50, integrates the natural logarithm of the student-teacher ratio, government expenditure, and GDP per capita PPP. The model explains about 50% of the variance in literacy rates. Interestingly, the ln(student-teacher ratio) coefficient stands at -31.42 (statistically significant at 1 %), indicating a significant negative relationship, on average 1 percent more students per teacher means 0.31% lower literacy rate when other variables are constant. This suggests that while the student-teacher ratio remains a vital factor, its impact is moderated when economic variables are considered. The government expenditure, with a coefficient of 1.543, hints at a positive influence on literacy rates, but its lack of statistical significance casts doubt on its direct impact. GDP per capita PPP, though included, shows a negligible and statistically insignificant effect. The constant is significant, but coefficient is not interpretable

Model 5 marks a significant leap in explanatory power, with an R-squared of 0.82. By replacing GDP per capita PPP with primary completion rate and average years of schooling, the model captures a more comprehensive picture, accounting for nearly 82% of the variance in literacy rates. Here, the influence of the ln(student-teacher ratio) is diminished (a coefficient of -4.93, signifacnt at 5%), highlighting that its effect on literacy rates is less dominant when considering direct educational factors. In contrast, the primary education completion rate and average years of schooling emerge as strong, positive correlated variables, with coefficients of 0.490 and 3.024(both significant at 1%), respectively. One point increase in these variables mean respectiveley around 0.5 and 3 percent higher literacy rate, when all other variables are constant. These findings underscore the critical importance of educational attainment and completion in fostering literacy.

The contrast between Models 4 and 5 is striking. While Model 4 suggests a reduction of the student-teacher ratio's impact by economic factors, Model 5 vividly illustrates the overpowering influence of educational achievement variables. The consistent, yet reduced significance of the student-teacher ratio across the models reinforces its role in literacy development, but also points to the complexity beyond mere classroom dynamics. The role of government expenditure remains ambiguous, with its persistent non-significance suggesting that the link between funding and literacy might be more intricate than direct financial inputs.

Overall, these models emphasize the need for a comprehensive view of education policy, where quality and outcomes, represented by completion rates and schooling years, play a pivotal role alongside quantitative measures like student-teacher ratios. The complex array of factors influencing literacy rates highlights the complexity of educational ecosystems and the need for nuanced policy interventions.

The high R-squared of our model prompted us to check the condition number, which turned out to be very high, indicating potential multicollinearity issues. The condition number exceeds 1000 (**see Appendix 6**).

Turning to the correlation matrix in **Appendix 5** (**Fig. 3**), we observe a very strong correlation between literacy rate and average schooling years, at 0.85. Similarly, the correlation between literacy rate and primary education completion rate is also very strong, at 0.88. These findings, combined with the high condition number, lead us to conclude that including these two variables in the model could cause multicollinearity, resulting in inaccurate coefficients for our variables. Therefore, we should consider removing them from Model 5 and revert to Model 3.

Conlusion

As we conclude our analysis, we've explored the impact of various educational and economic factors on literacy rates, progressing from simple linear models to more complex ones. Throughout our analysis, the condition number has been a consistent concern, especially in Model 5, where it exceeded 1,000, signaling potential multicollinearity. Nevertheless, Model 5's adjusted R-squared of 0.82 demonstrates substantial explanatory power, indicating a significant association of our included variables with literacy rates.

From a causal standpoint, our models indicate a negative correlation between the student-teacher ratio and literacy rates. The high number of students per teacher logically affects literacy rates, though our analysis does not account for factors such as the availability of water, electricity, and equipment. Therefore, it's likely that other factors contribute to changes in literacy rates. Model 3 provides statistically significant coefficients, allowing us to generalize our results to similar countries included in our analysis.

Assessing the external validity of our models is more challenging; we cannot confidently apply our findings to developed countries. Our dataset primarily includes developing countries from Asia and Africa, which may limit external validity. However, as our indicators are aggregated averages over 20 years, we can potentially expect similar results in the near future for the same countries.

In conclusion, it's evident that literacy is shaped by a complex array of factors. Therefore, effective educational policy should consider not only quantitative measures like student-teacher ratios but also the quality of educational outcomes. Future research could build upon this study by examining these relationships over time, better addressing the directionality and causality of factors influencing literacy rates.

Appendix 1

Out[7]:

	Dependent variable: lit_rate	
	(1)	(2)
Studenent-Teacher Ratio	-0.686***	-0.677***
	(0.210)	(0.207)
Government Expenditures on Education		1.817*
		(1.036)
Constant	98.761***	91.142***
	(6.604)	(8.673)
Observations	61	61
R^2	0.373	0.389
Adjusted R^2	0.363	0.368
Residual Std. Error	18.273 (df=59)	18.191 (df=58)
F Statistic	10.699*** (df=1; 59)	8.354*** (df=2; 58)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix 2

Out[8]:

OLS Regression Results

Dep. Variable:	lit_rate	**R-squared:**	0.491
Model:	OLS	**Adj. R-squared:**	0.482
Method:	Least Squares	**F-statistic:**	58.42
Date:	Fri, 22 Dec 2023	**Prob (F-statistic):**	2.20e-10
Time:	20:47:45	**Log-Likelihood:**	-256.42
No. Observations:	61	**AIC:**	516.8
Df Residuals:	59	**BIC:**	521.1
Df Model:	1		
Covariance Type:	HC1		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	182.6920	13.252	13.786	0.000	156.719	208.665
ln_stu_teach_ratio	-31.6139	4.136	-7.643	0.000	-39.721	-23.507

Omnibus:	5.007	**Durbin-Watson:**	2.265
Prob(Omnibus):	0.082	**Jarque-Bera (JB):**	4.488
Skew:	-0.584	**Prob(JB):**	0.106
Kurtosis:	2.368	**Cond. No.**	25.4

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Appendix 3

Out[9]:

	Dependent variable: lit_rate	
	(1)	(2)
ln(student-teacher ratio)	-31.419***	-4.931**
	(4.722)	(2.334)
gov_exp	1.543	0.922
	(0.970)	(0.678)
gdp_capita_ppp	-0.000	
	(0.000)	
prim_comp_rate		0.490***
		(0.115)
avg_ed_years		3.024***
		(0.875)
Constant	175.948***	32.443***
	(16.925)	(11.106)
Observations	61	61
R^2	0.503	0.822
Adjusted R^2	0.477	0.809
Residual Std. Error	16.558 (df=57)	9.992 (df=56)
F Statistic	20.386*** (df=3; 57)	59.922*** (df=4; 56)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix 4

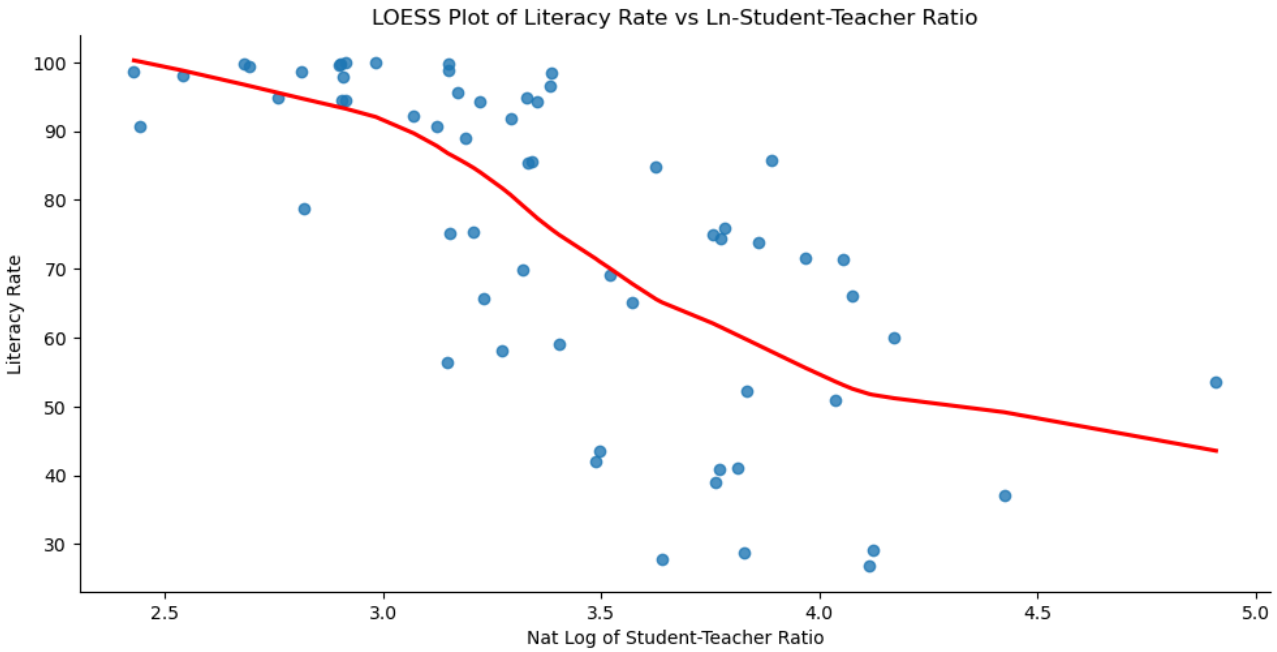

Fig. 2: Loess Plot of Literacy Rate vs Ln-Student-Teacher Ratio

Appendix 5

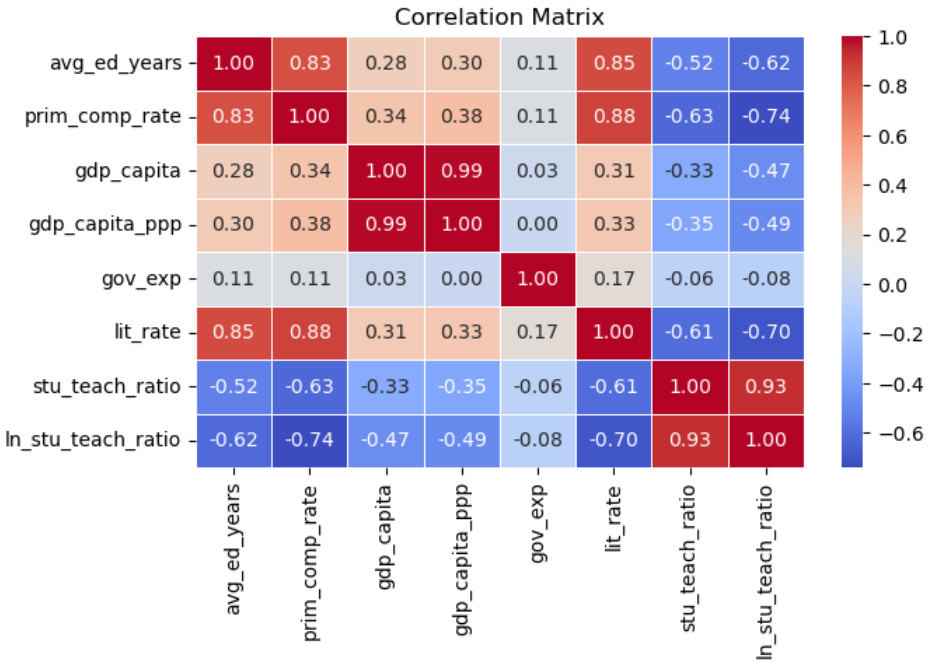

Fig. 3: Correlation matrix of numerical variables

Appendix 6

Out[12]:

OLS Regression Results

Dep. Variable:	lit_rate	**R-squared:**	0.822
Model:	OLS	**Adj. R-squared:**	0.809
Method:	Least Squares	**F-statistic:**	59.92
Date:	Fri, 22 Dec 2023	**Prob (F-statistic):**	1.38e-19
Time:	20:47:47	**Log-Likelihood:**	-224.36
No. Observations:	61	**AIC:**	458.7
Df Residuals:	56	**BIC:**	469.3
Df Model:	4		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	32.4426	11.106	2.921	0.003	10.676	54.209
ln_stu_teach_ratio	-4.9310	2.334	-2.113	0.035	-9.506	-0.356
gov_exp	0.9219	0.678	1.360	0.174	-0.407	2.251
prim_comp_rate	0.4896	0.115	4.257	0.000	0.264	0.715
avg_ed_years	3.0244	0.875	3.456	0.001	1.309	4.739

Omnibus:	0.294	**Durbin-Watson:**	2.079
Prob(Omnibus):	0.863	**Jarque-Bera (JB):**	0.051
Skew:	-0.065	**Prob(JB):**	0.975
Kurtosis:	3.058	**Cond. No.**	1.18e+03

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 1.18e+03. This might indicate that there are strong multicollinearity or other numerical problems.