Research Project #2:

Internet Usage on Literacy Rates Among Youth

Maya Jimenez

Department of Economics: Loyola Marymount University

ECON 3300: Econometrics

Professor Jain

December 11, 2023

My research question is centered around the effect of how internet usage affects literacy rates among youth. The relationship between internet usage and literacy rates among youth is important because it represents the role that internet usage plays in the education system, specifically in shaping literacy outcomes during early education. Being literate is a big contributor to one's social mobility and has an impact on economic growth within society. With the internet becoming a part of everyday life, we have been more exposed to information and educational resources than ever before, which is why we expect there to be a causal link between the two. Policymakers can implement the findings from this research to allocate more resources for groups with limited access to technology and integrate it into education systems.

The independent variable in this research is individuals using the internet (% of population)[1]. Internet users can be defined as individuals who have used the internet (from any location) in the last 3 months. The internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV, etc). The Global Partnership on Measuring ICT for Development plays a significant role in standardizing information and communications technology (ICT)- related data for developing countries with limited statistics. The data surrounding this variable has typically been collected through telecommunication operators, who gather information on subscriptions, providing a general idea of access to telecommunication services. However, the penetration rate metric- the share of households with access to telecommunications, has provided us with more information recently through the annual household and business surveys given to a random sample of households, which tend to include questions on accessibility to the internet, types of devices used, etc., and reveal the weighted averages.

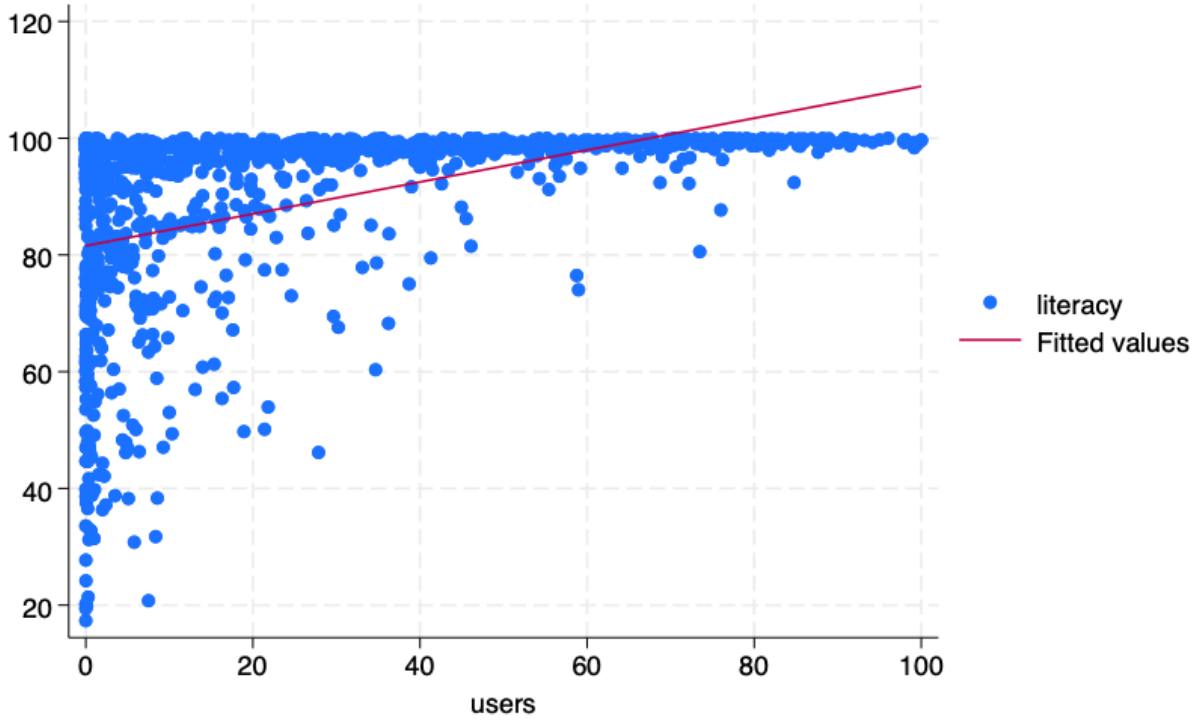
[1] https://data.worldbank.org/indicator/IT.NET.USER.ZS

The dependent variable in my research is literacy rate, youth total (% of people ages 15-24)[2], which is the percentage of people ages 15-24 who can both read and write with an understanding of a short simple statement about their everyday life. More specifically, it is the portion of the population that has completed a primary level education and acquired literacy along with numerical skills such as the ability to make simple arithmetic calculations. Data on literacy is based on information collected annually from the national censuses and household surveys that represent a large sample of the population, this data is compiled by the UNESCO Institute for Statistics. The population of interest is those 15-24, in order to reflect the outcomes over the past decade. The Global Age-Specific Literacy Projection Model is a process that UNESCO uses to estimate literacy rates for countries that have limited information through comparison of countries with similar characteristics and demographics.

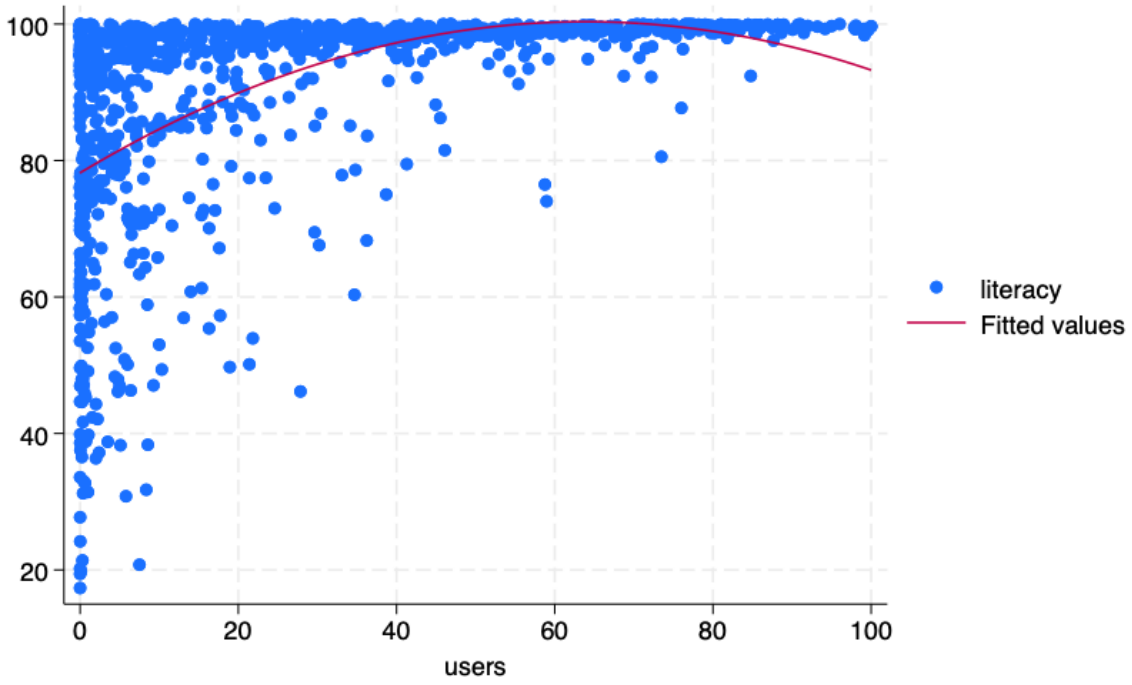
The cleaned variables in my data set are *users* and *literacy*, *users* representing the portion of the population that has used the internet in the past three months, focusing only on countries, and *literacy* representing the literacy among youth (ages 15-24), after the years of 1990. The process of cleaning my dataset involved converting my data from a numbers file into a CSV file. In order for Stata to properly interpret my CSV file, I deleted the first three rows for both of my CSV files which had unnecessary information. Once imported into Stata, in order to clean up the variable names, I dropped variables *countrycode*, *indicatorname*, and *indicatorcode* which were unnecessary for my analysis. In order to rename my variables, I used a loop to iterate the variables for v5- v68, which changed the variable names to *users*, *literacy,* and *year* to better suit my data and research question. Additionally, reshaping the data so it's oriented in a vertical (long view) format as opposed to horizontal (wide view), was important for allowing me to merge both

[2] https://data.worldbank.org/indicator/SE.ADT.1524.LT.ZS

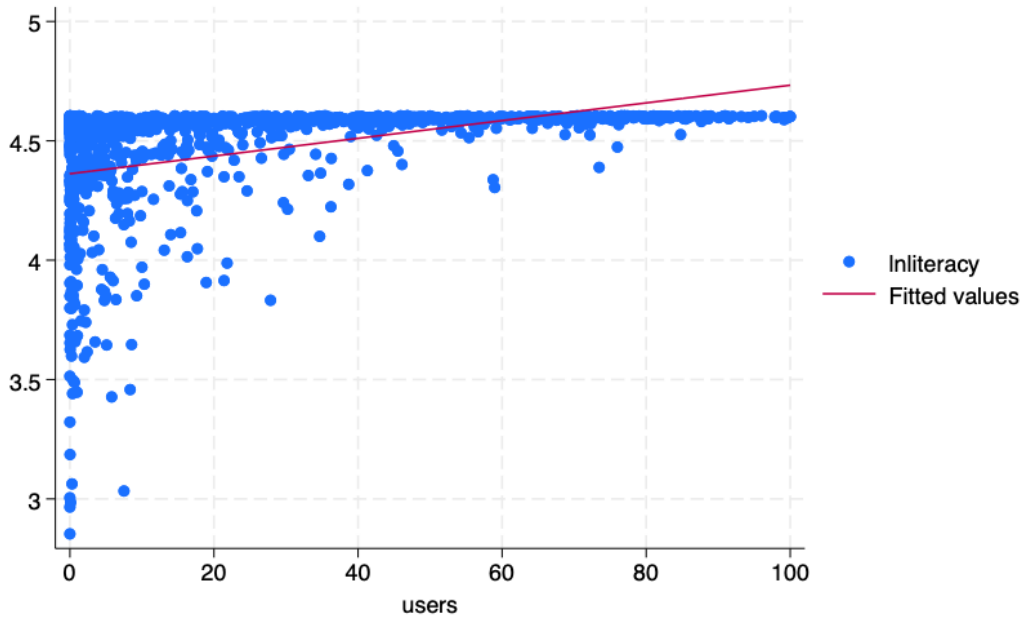
data sets; so they can be viewed side by side. Then, I got rid of all of the regions within the world to only focus on countries in the years 1990 and beyond. The sample mean for variable *users* was 25.78154, with a range of [0,100] and a total of 6,378 observations. The variable *literacy* had a sample mean of 89.12348, with a range of [17.34738, 100] and a total of 1,052 observations.



For my first naive regression using OLS with standard robust errors, I regressed literacy rates on the number of internet users, $literacy_{it=}\ \beta_{0(81.568850)} + \beta_{1users(.2732577)} + u_{it,}$ and estimated that a one percentage point increase in the number of internet users would result in an increase of 27.32 percentage points in literacy rates among youth. The effect of internet users on youth literacy is statistically significant at the 10%, 5%, and 1% levels because the p-value is 0. Compared to the average literacy rate of 89.12348, the effect of a 1 percentage point increase in internet users on literacy rates among youth is a 30.63% change, which I consider to be economically significant/large.



In my second polynomial regression using OLS with standard robust errors, I regressed literacy rates on users and users squared, $literacy_{it=}\ \beta_{0(78.19004)} + \beta_{1\ users(.6940459)} + \beta_{2\ users\ squared(-.0054336)} + u_{it,}$ and estimated that as the internet users increase, literacy rates also increase and become larger in magnitude with diminishing returns. The effect of internet users squared on youth literacy is statistically significant at the 10%, 5%, and 1% levels due to a p-value of 0. Compared to the average literacy rate of 89.12348, the effect of a 1 percentage point increase in internet users on literacy rates is a 77. 86% change, which I consider to be economically significant/large.



My third regression is a log-linear regression regressing ln(literacy) on users by implementing OLS with standard robust errors. The equation $\ln(\text{literacy}_{it}) = \beta_{0(4.31247)} + \beta_{1\ \text{users}(.0037094)} + u_{it,}$ predicts that if the number of internet users increases by one percentage point, then the literacy rate for youth is predicted to increase by 37.09 percentage points. The effect of internet users on youth literacy rates is statistically significant at the 10%, 5%, and 1% levels because the p-value is 0. Compared to the average literacy rate of 89.12348, there is a .00416% change which I consider to be economically insignificant/small.

For my fourth regression, I implemented a fixed-effects regression model with clustered standard error and interactions, $\text{literacy}_{it} = \beta_{0(43.65782)} + \beta_{1\ \text{users}(-.1738125)} + u_{it.}$ Holding constant all factors that vary across countries but are constant over time, if internet users increase by one percentage point, then the effect of internet users on literacy rates is expected to decrease by 17.38 percentage points. The effect of the number of internet users on literacy rates is statistically significant at the 10%, 5%, and 1% levels, suggesting that changes in the number of internet users are highly correlated with literacy rates among youth. Compared to the average literacy

rate of 89.12348, there is a 0.195% change which is considered to be economically insignificant/small.

		Dependent Variable:	Literacy Rates among youth	
	(1) Literacy Rates	(2) Literacy Rates	(3)Ln(Literacy Rates)	(4) (Literacy Rates)
Internet users (X_1)	.2732577*** (.0150261)	.6940459*** (.0508953)	.0037094*** (.0002416)	-.1738125*** (.0406295)
Internet users squared (X_2)		-.0054336*** (.0005077)		
Constant/Intercept	81.56885***	78.19004***	4.362157***	43.65782***
Country Fixed Effects	no	no	no	yes
Sample Size (n)	997	997	997	997
Adjusted R^2	.21275	.26293602	.16749629	.91914933

Each column represents a separate regression. The entries in the first 4 rows are estimated regression coefficients, with standard errors below them in parentheses. In columns (1), (2), and (3), robust standard errors are reported. In column (4) standard error is reported at the country

level. The asterisks indicate whether the coefficient is statistically significant at the 10% level (*), 5% level (**), or the 1% level (***).

Of the regressions I implemented, the regression that best fits the data and captures the causal effect of internet users on literacy rates among youth is the 4th regression, which is the fixed-effects regression with clustered standard error and interactions. This is because fixed effects allow us to hold constant all factors that vary across countries but are constant over time. This regression fits the data the best because of its increased adjusted r-squared from .16749629 to .91914933 and it remains statistically significant at the 10%, 5%, and 1% levels. One omitted variable that I could not control for that could result in omitted variable bias is education because education affects how much one can read, write, and speak. It is also directly correlated with the portion of internet users because those who are more educated are more likely to have more internet access. If the true effect of education on internet users and literacy rates among youth is positive, then the literacy rates would be overestimated. Measurement error as a threat to internal validity is a concern in this context because there is a potential random measurement error in responses that are nonsystematic due to interpretation issues that could arise in survey questions. This can lead respondents to underreport their internet usage, resulting in an underestimation of the relationship between internet users and literacy rates. There is also a concern with sample selection error because there are certain third-world or war-stricken countries that may have no responses available leading to bias in the portion of internet users. This sample would not be representative of the broader population, it would result in an underestimation of internet users. Simultaneous causality is also a concern in this context since both causal channels are present. The bidirectional relationship between internet users and literacy rates would cause an overestimation because of the influence of internet users on literacy rates and literacy rates on internet users. The external validity is affected by the limited data for most countries prior to

1995, impacting the generalizability of the study to post-1995. The absence of data during this time period reflects the lack of technological advancement that occurred. Educational and technological policymakers from developing countries with missing data would especially be interested in the findings of this research to ensure there is equitable access to technological resources.