An analysis of Tanzanian's Education situation in 1996

A study of Tanzanian's education level obtained, using Tanzania's Demographic and Health Survey report, 1996

Xiao Bai Yichun Zhang

10/04/2022

Abstract

Tanzania demographic and health survey (TDHS) is designed to summarises basic information on fertility, mortality (infant, child, and maternal), contraceptive knowledge and use, and child bearing. We extracted page 38 of the report, which include information about Tanzanian's education level of male within different age groups, residences and regions. We plotted graphs to visualize the distribution of education level with comparison of different groups, and used statistical inference such as confidence interval and simple linear regression to study their relationship. We found that we are 95% confident that the population percentage of Tanzanians having no education is between 15.233 and 37.301, and the percentage of completing primary school and attending socondary and above education will positive affect median year of schooling of male.

Contents

Code and data are available at[1]

[1] https://github.com/XiaoBai-blip/paper4

1. Introduction

1.1 Background Information

It was in 1992 that the United Republic of Tanzania modified its Constitution to become a multiparty state, and it was in late 1995 that the country held its first multiparty general elections for president and parliament in more than three decades. CCM, the dominant party in Zimbabwe, maintained its grip on power by capturing 186 of the 232 seats up for election in the country's legislature. The minimum age for admission into contractual labour in vocations that have been authorized is set at 15 years. The legislation prevents a young person from working in any activity that is harmful to their health, is risky, or is otherwise improper for their age or experience. Industrial labour is permitted for young people between the ages of 12 and 15, but only between the hours of 6 a.m and 6 p.m., with a few exceptions, according to the law. The Ministry of Labor and Youth Development is in charge of enforcing the law, but the number of inspectors available is insufficient to keep up with the pace of change. According to reports, growing privatization has resulted in a decrease in the efficiency of government enforcement. Plantations that grow sisal, tea, tobacco, and coffee employ around 3,000 to 5,000 youngsters for seasonal work throughout the growing season. Children working on plantations often get fewer compensation than their adult peers, despite the fact that they may be doing similar tasks to their elders. It is especially dangerous and damaging to youngsters to work on sisal plantations. A sisal plantation had a child labour force that accounted for 30% of the total labour force, with barely half of the youngsters having finished basic education. They suffered from a high incidence of skin and respiratory ailments, were not given with protective clothes, and were deprived of proper nutrition and accommodation, among other things. Additional minors working in unlicensed gemstone mines range from 1,500 to 3,000 in number. Children labour with their parents in the informal economy, which is uncontrolled piecework manufacturing.

Tanzanian Primary School, which is taught in the students' native language of Kiswahili, is meant to be free, but the prices of necessary school uniforms, school supplies, and modest school overhead are considerably above the financial resources of many of the students. Students begin in standard one when they are seven years old and begin studying English in standard three when they are nine years old. Many pupils are unable to attend primary school due to the considerable distance they must go to school (the majority are far further away than the minimum 3 to 5 kilometers), duties at home, bad health, and insufficient money, among other factors (Tanzania, n.d.). Starting with standard 4, students must pass national tests in order to progress, and a passing score on the standard 7 exams determines where they will be put in secondary school. Kids with the highest test scores and financial aid may be admitted to boarding schools, which are often located far away from their homes, while students with lower test scores may be admitted to local day secondary schools, which are also located far away but less costly (Sana 2014).

The analysis will be conducted in R (R Core Team 2020), and the package we will use is tidyverse (Wickham et al. 2019). All graphs will be created using function ggplot2 (Wickham 2016). The packages knitr (Friendly et al. 2020) are also used to generate the R markdown report.

1.2 Our work

We focus on the education information gathered by Tanzania demographic and health survey (TDHS). With the 1996 Time Series of the United States' Demographic Health Survey (TDHS), researchers will be able to examine trends in fertility, child mortality, and other demographic indicators at the national level.We chosen one page of the report that include the percentage of no education, completed primary school, and completed secondary and above by different age groups, residence area and province. To have a general understanding of our data set, we made explanatory data analysis with our data. We break the data frame in to three components according to its different measure, and summarized two most important variable for our study into tables. The table shows that elder people tend to have a lower education rate, and younger generations are generally more educated. The other tables shows that people's education in Tanzania does not have a significant pattern regarding residence area. Further, we made some plots to visualize the distribution of education levels.

Regarding statistical inference and modelling, we used confidence interval, hypothesis test and simple linear regression. The estimated average percents is 25.64. We used bootstrap to estimated the sampling distribution about a given population. The graph is the result of bootstrap confidence interval for mean percentage of Tanzanian who have no education. The result shows that we are 95% confident that true mean of no education Tanzanians are between 15.233 and 37.301. As for linear regression, we are interested what factors affect the male population's median year of schooling. We assume that percentage of population who completed their primary education and who attended some secondary and higher education will drive the median year of schooling for men to be higher.

2. Data

2.1 Data Sources

The TDHS of 1996 is the third nationwide sample survey of this sort. The Tanzania Knowledge, Attitudes and Practices Survey (TKAPS) was conducted in 1994.The 1996 TDHS introduced questions on AIDS, maternal mortality, and female circumcision to the list of topics asked in the previous two surveys.

The TDHS sample consisted of 357 enumeration areas (EAs) from the 1991-92 TDHS (262 EAs in rural and 95 EAs in urban areas). First, wards branches were chosen, then EAs inside wards/branches. The chosen EAs were given lists of all homes, from which the third sample stage was drawn. The TDHS was meant to offer estimates for the whole nation, for urban and rural areas, and for regional groupings (zones). The sample will also offer estimates for each of the 20 mainland areas and two Zanzibar subgroups: Pemba Island and Ungaja. Men from every second home were interviewed in Dar es Salaam, Dodoma, Iringa, Kilimanjaro, Morogoro, and Shinyanga. The male's sample was intended to produce national and regional estimates. Unlike most previous DHS surveys, households in Tanzania were chosen at random from the household list for each ward (or branch). Due to the dispersed structure of houses, this selection technique was employed to minimize relocating difficulties. The 1996 TDHS aims to provide national-level data for calculating demographic rates, notably fertility and childhood mortality, as well as examining the direct and indirect variables that influence fertility levels and trends. Like earlier surveys, the 1996 TDHS included institutions and people. The survey was coordinated by the Planning Commission's Bureau of Statistics, with technical and logistical assistance from the Ministry of Health. USAID provided financial and technical assistance via Macro International Inc. The monies were utilized for field crew allowances, data processing, anthropometric equipment, questionnaire printing, field vehicle fuel and maintenance, and survey results distribution. This project was funded in part by the Tanzanian government.

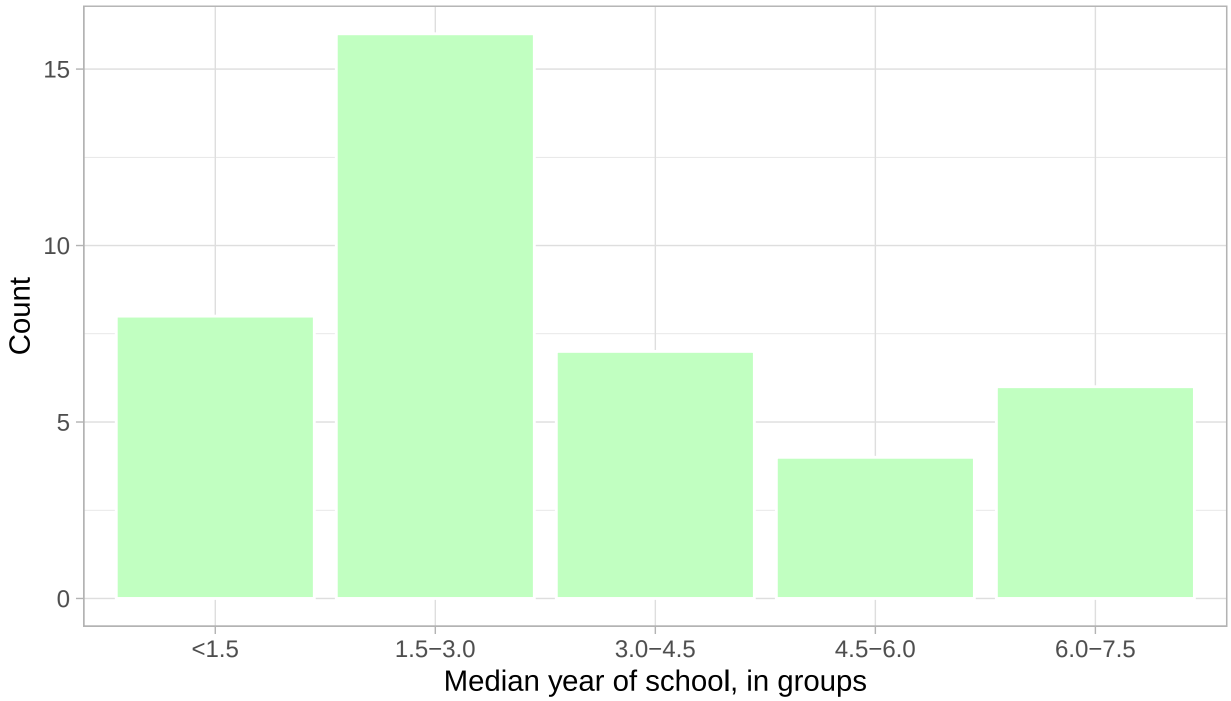
The primary survey field personnel was trained for three weeks in early July 1996 at the Vocational Training Institute (VETA) in Iringa. Guest professors from the UMATI, MCH professionals from the lringa regional hospital, and officials from the Tanzania Food and Nutrition Centre helped deliver the training. Trial interviews were held in adjacent communities and in Ifinga. Computer operators were also trained on the questions. There were simulated interviews between participants in the classroom and practice interviews with actual respondents in and around Iringa. For three days, supervisors and editors only discussed their tasks. The necessity of data quality was emphasized.

The supervisor was responsible for the well-being and safety of team members, completing given tasks, and maintaining data quality. The editor's tasks included interviewer performance and anthropocentric measurements of children and women. To guarantee accurate and thorough data collection, interviewers were closely supervised and completed questionnaires were edited. It started in late July and continued until late November. The household interview selected women and men for individual interviews. The home interview must be conducted by a different interviewer than the individual interview. This was done to decrease age-related errors, especially among women and men in the youngest and oldest age groups. The interviewers were allocated residences by the team supervisors. The field editors reviewed the completed household and individual questionnaires to verify they had all relevant items, followed the skip pattern guidelines, and were internally consistent. Before leaving the cluster, each team was told to finish revising and fixing any problems detected in the surveys. Supervisors were expected to guarantee that all chosen families and eligible women and men were questioned, and that interviewer and supervisor assignment forms were filled out accurately. The questionnaires and control sheets were sent to the Dar es Salaam headquarters for processing.

2.2 Data Cleaning and Data Overview

We extracted one page of the data, sliced them into different groups, and restored our data. The 1996 TDHS collected information on individual socioeconomic characteristics of all usualresidents and visitors who had spent the previous night preceding the survey interview. This was done by using a questionnaire which was completed for each household. A household was defined as a person or group of persons who live together and share a common source of food. The CSV file has three segments, the first segment studies the education level among different age groups. For the convenience of our study, we filtered out missing values. The missing values exists after the section split title. We omitted these missing values so that only meaningful numerical numbers are left. Also, after the web scraping, the data is not perfectly clean. We first clean the names using janitor's clean name function, and made all the numbers numeric. There are some values that does not match the page, so we also adjusted it.

Figure 1: The bar plot of median year of schooling



To have an overview of our data, table 1 shows the basic education information for people who are aged 20 and above.

Table 1: No education and primary incomplete percentage in different groups

background_characteristic	no_education	primary_incomplete
20-24	9.3	15.4
25-29	10.8	11.2
30-34	11.2	12.0
35-39	16.6	17.3
40-44	21.7	31.4
45-49	26.9	34.2
50-54	29.8	45.8
55-59	39.8	41.9
65+	64.7	27.5

As can be observed from this table, the percentage of Tanzanians with no education higher for older generations than young generations. For people who are 65 years old and above, 64.7% of them have no education, while for people who are 20-24 years old, there is only 9.3% of them has no education. This means the education generalization is increased, and more people are getting educated recently. Accordingly, the percentage of people who have not completed their primary school also shows this trend. For people who are younger, fewer of them have not completed their primary school.

Table 2: No education and primary incomplete percentage in different residences

background_characteristic	no_education	primary_incomplete
Mainland	31.6	35.2
Total_urban	18.7	31.2
Dares_Salaam_city	14.6	26.3
Other_urban	20.6	33.5
Total_rural	34.8	36.2
Pemba	41.9	36.2
Unguja	26.6	33.2

No education population and people who haven't finish their primary education also differs in regions. Generally, total rural population have a higher no education and primary incomplete percentage, 31.6% and 35.2% respectively. Mainland has second highest no education and primary incomplete percentage.

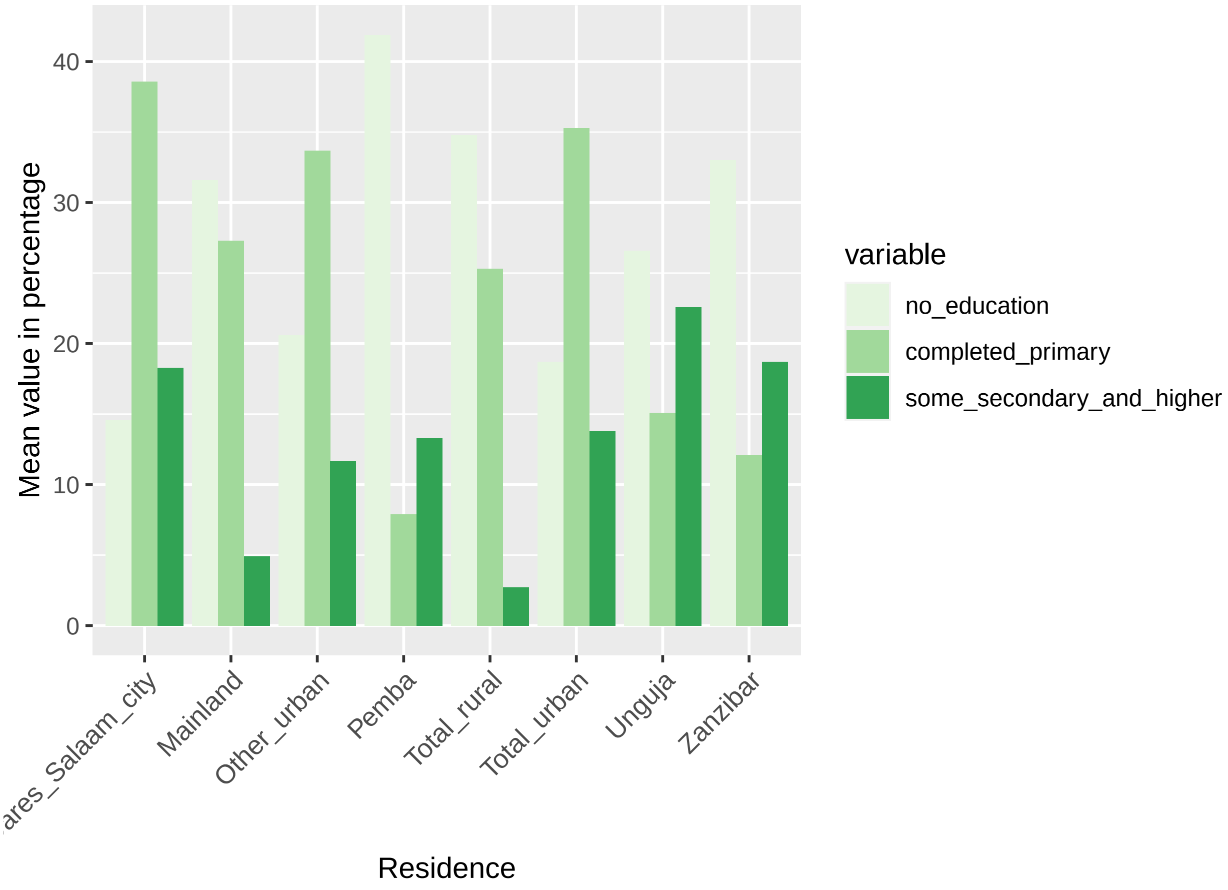
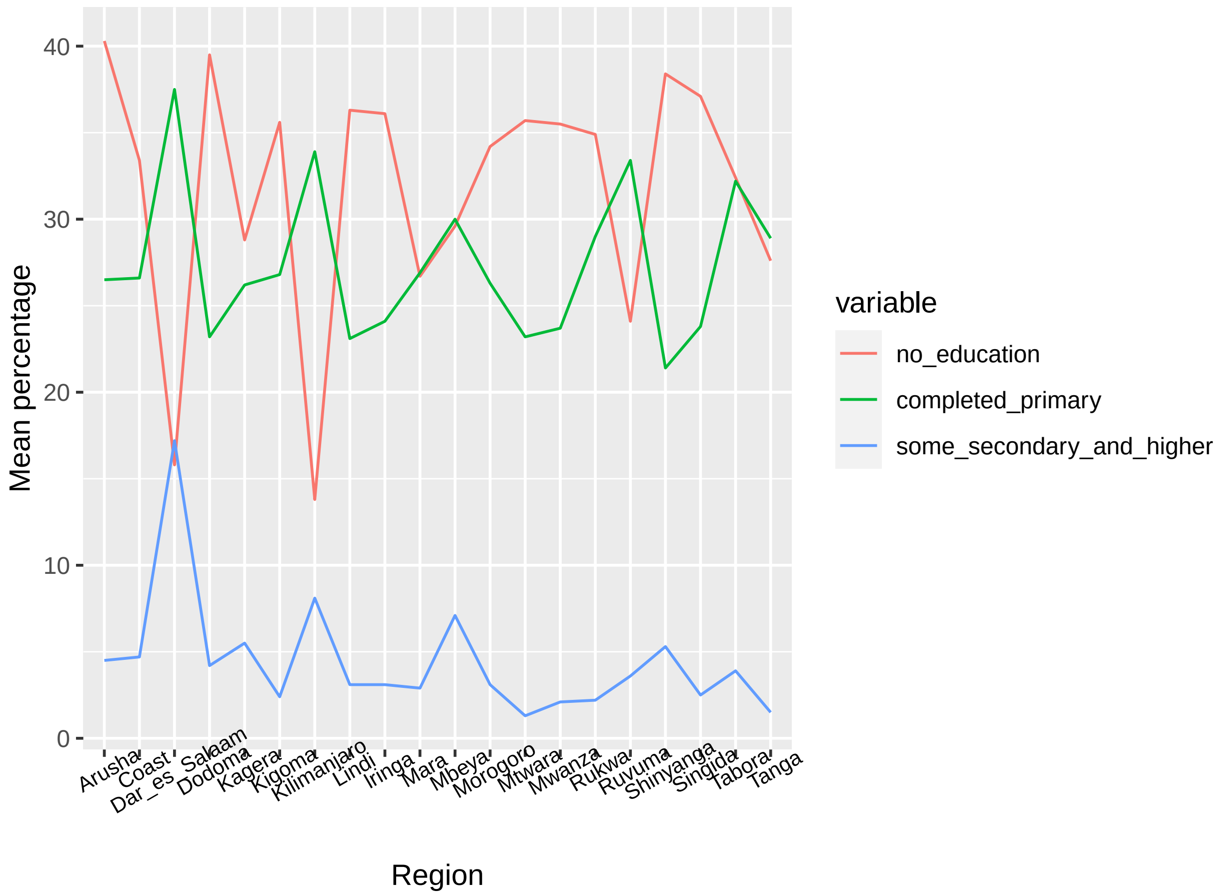


Figure 2: Mean percent by residence

The bar plot shows the percent distribution of the male household population by three different levels of education attended according to different residence types. The x-axis shows eight main residences in Tanzania and the total male population is classified based on their highest education level. the lightest green represent percentage of no education population in Tanzania, and median light green bar represent the percentage of population in Tanzania who haven't finished their primary education, and the darkest green represent the percentage of Tanzanian who did some secondary education and higher. In other word, a darker green color represent a higher level of education. Total rural area, as observed, have a least number of people attended some secondary and higher education, and second largest population of no education. Dares Salaam city, on the other hand, has the most population who completed their primary school and second largest population of attending some secondary and higher education, compared to other residences.

Figure 3: Mean percent by region



The line graph shows the mean percentage of Tanzanians having education, completed primary or attended some secondary and higher education by region. The percentage of attending some secondary and higher is significantly lower than the population of having no education or completing their primary school only. In other word, the majority of people in Tanzania have not attend some secondary school or higher education. A significant peak in the blue line shows that Dares Salaam has a higher percent of people who attended some secondary and higher education.

3. Method and Model

Statistical methods are applied in this paper to help observing and interpreting the data in an alternative way. In this paper, we will use three statistical methods to explore deeply about our data. These methods are simple linear regression model, confidence interval and hypothesis test.

When conducting an estimate in statistics, there is always uncertainty around the estimate because the number is based on a sample of the population. Therefore, using the confidence interval method is also essential to a statistical research because it measures the degree of uncertainty or certainty in a sampling method. More specifically, it provides an approximate set of values that is likely to contain a true parameter that is uncertain. A true parameter can be a true mean, true proportion or standard deviation. Besides, each confidence interval has a percentage associated with it, called a confidence level. This percentage represents how confident we are that the results will capture the true population parameter, relying on the bond's luck together with your random sample. In surveys, confidence levels of 90%/95%/99% are frequently used.

Hypothesis testing refers to the procedures to accept or reject statistical hypotheses. We apply this method as we expected do a strict comparison with a pre-specified hypothesis and significance level. In common word, we know that the best way to determine whether a statistical hypothesis is true would be to examine the entire population. However, this is almost impossible. Therefore, we only examine a random sample from the population to see if the sample data is consistent with the statistical hypothesis. There are two types of hypothesis, null hypothesis and alternative hypothesis. Null hypothesis (H0) assumes that the difference between the chosen characteristics in a set of data is due to chance, and alternative hypothesis (Ha) is the opposite of null hypothesis.

In the following sections, we will explain more about how we conduct these methods to provide a better understanding of our dataset.

Table 3: Variable summary	Sample mean	Standard deviation
No education experience	25.644	17.81
Secondary or higher education experience	6.82	5.59

3.1 Confidence interval:

To analyze our data more deeply, we narrow down our topic to be more focusing on the average percentage of Tanzania communities within all background characteristics that had no education experience. We calculated the estimated mean percentage above using this dataset, which is 25.644. However, since the dataset is just one sample, there might be a problem about how we obtain a measure of precision and confidence about our estimate. Therefore, in order to describes the uncertainty surrounding an estimate, we will perform statistical inference and apply bootstrap method to get the confidence interval in this section.

Bootstrap is a statistical method that is used to estimate the sampling distribution about a given population. It creates multiple resamples (with replacement) from a single set of observations, and then computes the effect size of interest on each of these resamples. This bootstrap resamples of the effect size can then be used to determine the confidence interval. One type of bootstrap is empirical bootstrap, which samples from an estimator's sampling distribution without specifying the data distribution. In this paper, we will use empirical bootstrap. Besides, each confidence interval has a percentage associated with it, called a confidence level. More specificity, if we perform 95% confidence interval, 95% indicates that any such confidence interval will capture the population mean difference 95% of the time. Alternatively, it means that when repeating an experiment or survey over and over again, 95 percent of the time the results will match the results we get from a population. Moreover, with a 95 percent confidence interval, we have a 5 percent chance of being wrong. In addition, for a given dataset, increasing the confidence level of a confidence interval will only result in larger intervals (or at least not smaller). With the small sample, we expect to see that the 95% confidence interval is similar to the range of the data. But only a tiny fraction of the values in the large sample lie within the confidence interval. This is because the 95% confidence interval defines a range of values that you can be 95% certain contains the population mean. With large samples, we know that mean with much more precision than you do with a small sample, so the confidence interval is quite narrow when computed from a large sample. In our dataset, since the sample size is too small (n=43) and I think a wider confidence level might give an accurate result than a narrower one, I will use a relatively wider confidence level (95%) in this report. Before applying the bootstrap, there are some assumptions that need to be concerned. We assume that all samples are independent, and the parameter will be the true mean of percentage of people had no education experience.

3.2 linear Regression Model:

The multiple linear regression (MLR) is used to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. The general form for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

Where β_i represents the coefficients need to estimated, $x_i, i = 1...p$ is the predictor variables, y is the response variable, ϵ is the error term. Mathematically, β_0 measures where the line intercept y-axis, and $\beta_1 - \beta_p$ measures the slope of the line. More practically, β_o is the value of y when x equals the zero, and $\beta_1 - \beta_p$ is the average change of y when x increase by 1.

4 Results

In general, educational attainment is greater in urban areas than it is in rural areas, according to the data. For women, the percentage of those without a formal education in urban regions (25 percent) is lower than that in rural areas (46 percent); for males, the proportion of those without a formal education in urban areas is 19 percent, compared to 35 percent in rural areas. Women and men in urban areas are more likely than those in rural areas to have completed elementary and secondary school. Both men and girls with no education are in greater proportion in Zanzibar than on the mainland, which is a result of a combination of factors. Zanzibar, on the other hand, has the greatest percentage of the population with a secondary or higher level of education. This is owing to the fact that obligatory primary education includes three years of secondary school as part of the curriculum. The Dodoma, Arusha, Lindi, Mtwara, lringa, Singida, Kigoma, Shinyanga, and Mwanza areas have the greatest percentage of women with no education (over 40 percent) and males with no education (above 35 percent). The areas of Dar es Salaam and Kilimanjaro have the lowest shares of male and female respondents who have no formal education, respectively (below 20 and 25 percent, respectively).

4.1 Confidence interval

The graph is the result of bootstrap confidence interval for mean percentage of Tanzania that did not have any education experience. Values between the 2 red lines are in the 95% interval. We rounded our result to three significant digits (refer to the table). We are 95% confident that true mean is between 15.233 and 37.301. The confidence interval is meaningful because it is between 0 and 1, and both number (15.233 and 37.301) is close to and bounded around the sample mean we calculated above. Specifically, we are 95% sure that the true average of percentage of Tanzania communities that did not have any education experiences is between 15.233 and 37.301.

Table 4: Confidence interval	2.5%	97.5%	CI
value	15.233	37.301	(15.233, 37.301)

4.2 Linear Regression Model:

Since we are interested in identifying factors that may affect calculating male population's median year of schooling, we conduct a regression based on these potential factors. We selected two factors, each representing the mean value of total male population that had completed primary education and those who had some secondary or higher education experiences respectively.

Our assumption is that these two variables might have a positive linear relationship with our response variable, median year of schooling. This is because a relatively large proportion of population in Tanzania had difficulty affording tuition fee, so they may chose to not getting education. Similarly, since proportion that completed primary education and secondary or higher education looks equal, we need to perform regression to check which one is more significant in contributing the number of year of school. However, as by intuition we may think of population of total population that have higher educational experience should be lower relative to population that purely attended primary education, we may assume that variable "completed primary" is likely to have stronger correlation with response variable.

In summary, independent variables we chose that may be influential to the dependent variable is listed below:

1. Completed primary education
2. Secondary or higher education

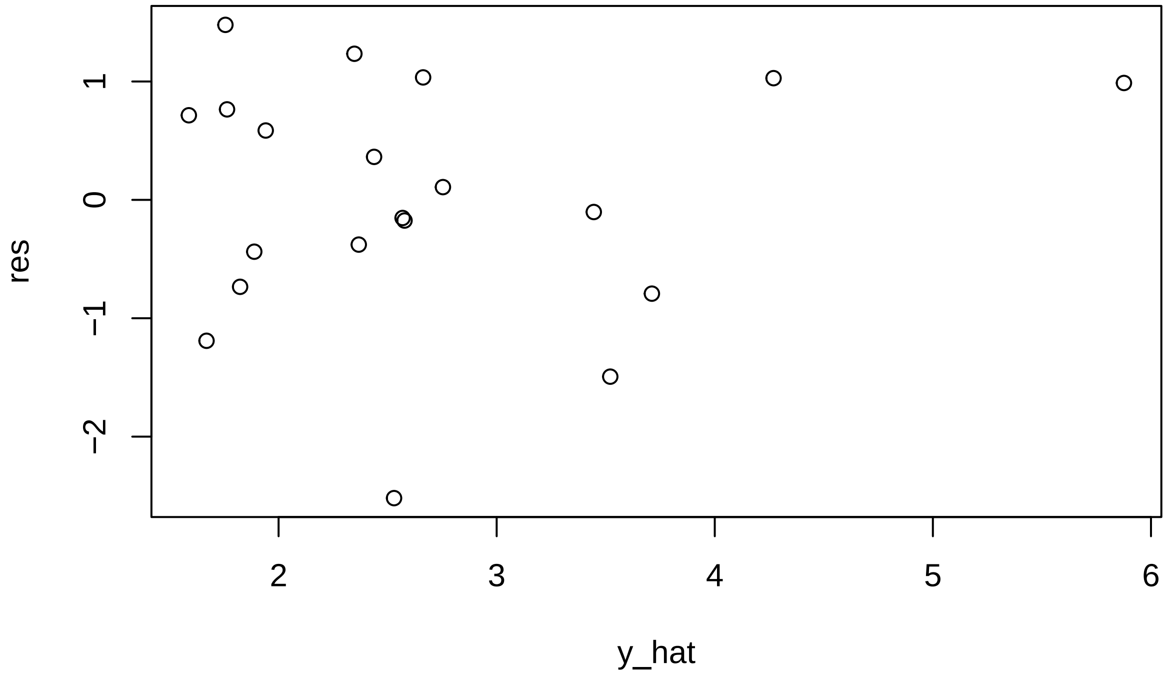
Then we conduct a multiple regression based on these explanatory variables to see if there exist some sort of relationship with median year of schooling.

The new model we built based on these variables is:

$$Median year \hat{o} f schooling = \hat{\beta}_0 + \hat{\beta}_1 * Primary + \hat{\beta}_2 * Secondary$$

Before we run the result, we need to test the validation of this regression, and this is based on the statistical theory about assumptions of linear regression model. The linear regression model has four assumptions: linearity, uncorrelated errors, constant variance and normality. As we saw when we were deriving the unbiasedness and the covariance of our estimator, we use the assumptions many times to obtain our results. When all the model assumptions are satisfied, we can then be sure that the estimators will behave in a nice way and have all these lovely properties. However, if even one assumption is violated, this can have a large impact on how we can use our estimates.

We can use residual plot to determine whether there are violations of model assumptions. Residual plots allow us to visually inspect the model assumptions. Moreover, we work with residual plots because the data can sometimes be too noisy to see model violations clearly. There are three main types of residual scatter plots that we use: residuals versus predictor plots, residuals versus fitted values plots, and normal qq plots. Both residuals versus predictor and residuals versus fitted value plots can be used to assess whether our first three assumptions hold. We can check by observing from the residuals plot and if there is no discernible pattern seen in the residual's plots, then the assumptions hold. In other words, to satisfy the assumptions, residual verses fitted predictors plot should not have any pattern or large clusters of residuals. Even thought we have a very small size of data, our plot which can be seen below seems not violate this assumption. This means the result of our regression might be as accurate as we expected.



To check the normality with residual plot, we do so by using a QQ plot. Normality is verified by using a QQ plot which computes quantiles from the residuals and plot them against the standard Normal quantities. We expect to see a straight diagonal string of points in the plot with minimal deviations at the ends. Surprisingly, our plot generally satisfies this assumption as points are all distributed along the diagonal. It looks like follows a bimodal pattern. This means our regression model does not violate normality assumption.

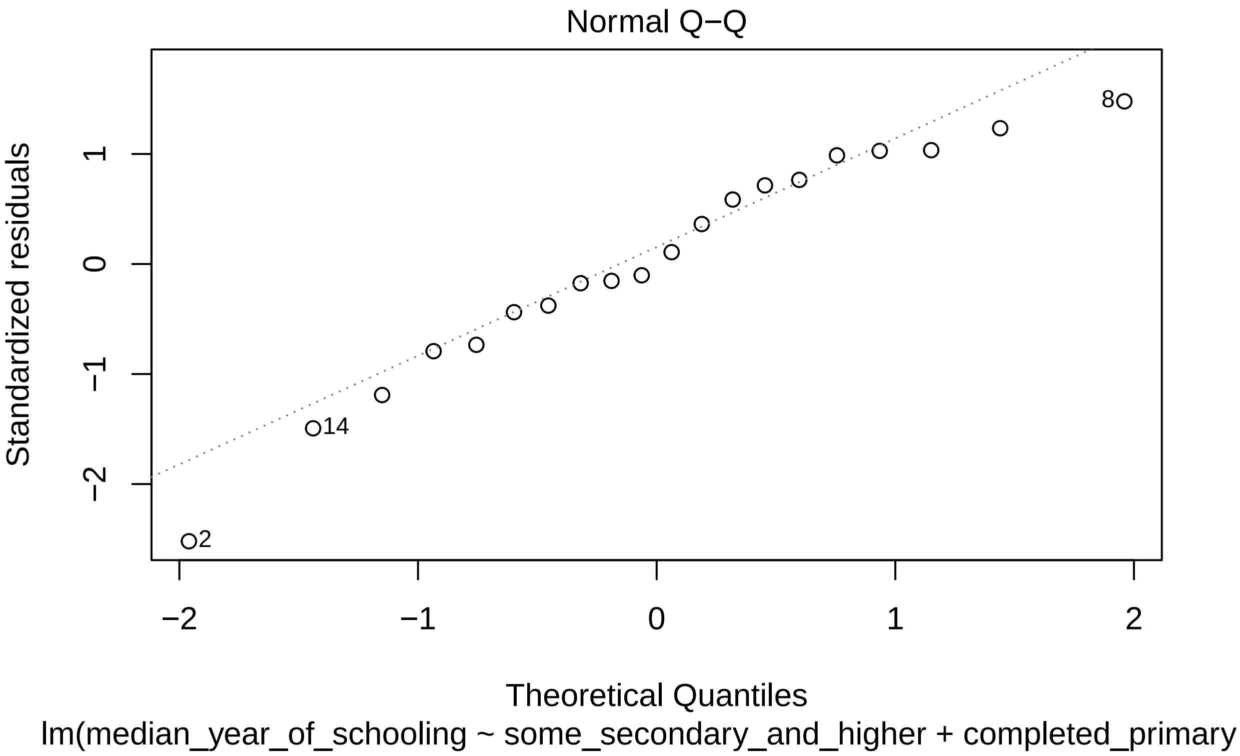


Table 5: Regression result	Estimate	Std. Error	t value
Intercept	-2.834	0.786	-3.603
completed_primary	0.185	0.032	5.746
some_secondary_and_higher	0.103	0.039	2.624

P-values and coefficients in regression analysis work together to be used for showing which relationships in the model are statistically significant and the nature of those relationships. The p-values for the coefficients indicates whether these relationships are statistically significant. The sign of a regression coefficient can tell us whether there is a positive or negative correlation between each independent variable and the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase.

Our regression output can be seen in table 3, where all the coefficients of our predictors and their p value are listed. Based on the table we can see that both two predictors are statistically significant because their p-values are less than the usual significance level of $\alpha = 0.05$.

The coefficients describe the mathematical relationship between each independent variable and the dependent variable. According to the table, $\beta_1 = 0.185$ represents that as one unit increase in mean percentage of male population completed primary education, the average percentage of median year of schooling will increase around 0.185. Similarly, as one unit increase in mean percentage of male population that have had secondary education, the average percentage of median year of schooling will increase around 0.103.

5. Discussion

5.1 Findings

In this paper, we discovered that Tanzania has a generally low education level. Many people have not completed their primary school and most people does not have secondary education experience. We also see that the popularity of education is getting better gradually, more young generations are receiving education. By doing confidence interval, the result indicates that we are 95 percent certain that the real mean of Tanzanians without a secondary education is between 15.233 and 37.301. In terms of linear regression, we are interested in the variables that influence the median year of education for males. We anticipate that the proportion of the population that finished elementary education and attended some secondary and higher education will result in a higher median year of schooling for males.

This study brings the education and poverty situation of underdeveloped countries to us. Literatures show that child labour is widely used in Tanzania, and extreme poverty and disease, as well as low popularity of schools in Tanzania, made the education level greatly under the world average education percentage.

5.2 Limitation and Next steps

Constructing a proper model to illustrate our findings is challenging. Our research may have overlooked the problem of variable bias. One's education, sexual orientation, and religious beliefs may all have a significant impact on one's education, as can one's financial situation and whether or not one has children. By deleting these elements from our model, we can see whether the variables we added were overstated in their importance. This is due to the fact that the predictor variables are no longer distributed independently of the error term, resulting in inaccurate conclusions.

Even though we have a large survey population, our data that we selected from one page is limited, therefore, the linear model we fit should be expanded with more variables. Our knowledge of statistical modelling is limited, and there should be more possible studies we can do to fully use this dataset. This survey is conducted in 1996, which is far from today, making this paper less sustainable for today's development and less suggestive. The dataset also have its ambiguities, and we are not sure if we separated the same group of people intro different parts. There should be lots more information to be provided together with this dataset.

When seeking to identify independent variables that are statistically significant, more severe statistical approaches should be applied. When an excessive number of independent variables is included, the issue of omitted variable bias is aggravated by the problem of multi-collinearity, which further increases the number of independent variables. There is a trade-off between omitted variable bias and multicollinearity, since multicollinearity increases our estimations' standard error. To collect meaningful sets off independent variables for our investigation, it may be essential to utilize either the Akaike or Bayesian information criteria. The mean duration between visits is the most critical metric to watch since it is unaffected by other variables.

A sample survey's estimates are influenced by two categories of mistakes: nonsampling errors and sampling errors. Nonsampling errors are the most common form of error. When mistakes are made in the implementation of data collecting and data processing, such as failure to discover and interview the proper household, misinterpretation of the questions by either the interviewer or the responder, and data entry errors, nonsampling errors occur as a consequence. Nonsampling errors are hard to eliminate and difficult to quantify statistically, despite the fact that extensive attempts were taken throughout the development of the 1996 TDHS to reduce this sort of mistake.

We can collaborate with seniors and experts, using a more recent dataset, and conduct more complex and meaningful statistical studies that can be used to suggestive papers.

6. Appendix

Extract of the questions from Gebru (2021)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to enable analysis of Australian politicians. We were unable to find a publicly available dataset in a structured format that had the biographical and political information on Australian politicians that was needed for modelling.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - DHS program (Demographic and Health survey) conducted researches and surveys and create datasets.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The DHS Program has earned a worldwide reputation for collecting and disseminating accurate, nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria, and nutrition.
    - The DHS Program is funded by the U.S. Agency for International Development (USAID). Contributions from other donors, as well as funds from participating countries, also support surveys. The project is implemented by ICF, since September 2013
4. *Any other comments?*
    - The DHS Program believes that the ultimate purpose of collecting data is its use in policy formation, program planning, and monitoring and evaluation, aims at fostering and reinforcing host country ownership of data collection, analysis, presentation, and use; coordinates with key stakeholders on data collection and dissemination.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - no_education: Tanzanians who have no education experience.
    - Age groups: Tanzanians' education level divided into different age groups.
    - Residence: parts of country
    - primary_incomplete: Tanzanians who have not completed their primary education.
    - Primary complete: Tanzanians who have completed their primary but have no further education.
    - Some secondary and higher: Tanzanians who have attended some secondary education or above.
2. *How many instances are there in total (of each type, if appropriate)?* -A total sample of 8,900 households were selected with the objective to have 9,000 completed interviews of women 15 to 49 years old. A total of 8,141 households were occupied and in 7,969 households, interviews were completed. In those households interviewed, 8,501 women 15 to 49 years old were identified and 8,120 were completed interviews.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

- The sample size is all participants of the survey, and we extracted one section of the survey result but the sample is not reduced. The initial large data set should be the data information collected directly from 1996 DHS survey in Tanzania.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - This data mostly contain data frames with characters summarized. The data are mostly numeric and characteristic. There are nine variables in total, eight of them are numerical representing male population's mean percentage of education level. The only one categorical variable represents background characteristics. These characteristics are group by different age group, different residence of Tanzania and different regions.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - Yes. The whole dataset describes the average education level among male population. Columns are divided by education level (eg: no education, primary incomplete, primary complete, etc. . .), and divided into different age groups/residences/regions.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - NO. There is no missing information in the cleaned dataset. In the raw data , they indeed had some missing values in some variables. However, we removed those missing values in the data cleaning process, and the cleaned data now should contain values for most of the observations.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - Each respondent's interview and biomarker data files are identified only by a series of numbers, including enumeration area (EA) number, household number, and individual number. After data processing, questionnaire cover sheets containing these identifier numbers are destroyed, and EA and household numbers are randomly reassigned.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - Yes. This page is divided into different sections, age groups, residence and regions. We can analyze data according to different sections, but in the paper we group all three sections together and provide a general overview for the data as a whole.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - NO, there is not. Our data is extracted from the report, and is clean and structure. There should be sources of noise and redundancies in the original survey data obtained, but they are not in our scale of study.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - Yes. DHS website and the entire survey result. It has guanranteed access. a) yes, b) yes, c) yes.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

- NO. confidential information has already been hiden when the information is provided. The dataset retrived contain no confidential information.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - Yes. At a first glance of the data, we found out that the mean percentage of male population that had no educational experience is extremely higher than other two education levels, that is, completed primary and secondary or higher education. This caused a bit of anxiety as if one variable contains values that are significantly larger than others, it may be difficult to perform statistically inference and make comparison with other variables that have similar characteristics.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes. The dataset classifies the total population and collected the mean percentage of education levels by age, neighbor, gender, etc. The age section contains 13 different age groups such as 10-14, 20-24 etc,.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - NO, we can not. Confidential information are hiden in the report.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - NO. By checking the questionnaire in the Appendix, we can see that some sensitive information, and ethics and income are asked, but the raw data is hiden from the report, and our study is based on the dataframe provided in the report.

16. *Any other comments?*

    - NO.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - This data is reported by survey responses, and summarized by characteristics. Yes, the data link to the survey reponses.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The absolute probability of selecting an EA (product of the probability of selecting a ward/branch and the conditional probability of selecting an EA within a ward/branch

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- The sample for the 1996 TDHS was selected from the same primary sampling units used in the 1991- 92 TDHS. The sample frame for the 1991-92 survey was based on the list of enumeration areas from the 1988 Population Census; therefore, this census is also implicitly a frame for the 1996 TDHS. The list of census enumeration areas for the 1996 TDHS survey was stratified by each of the 20 regions (for the mainland) and within each region by urban and rural areas. In total, 357 EAs were selected, 95 in the urban area and 262 in the rural. Table A1 shows the sample distribution of EAs

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - A total sample of 8,900 households were selected with the objective to have 9,000 completed interviews of women 15 to 49 years old. A total of 8,141 households were occupied and in 7,969 households, interviews were completed. In those households interviewed, 8,501 women 15 to 49 years old were identified and 8,120 were completed interviews.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data is collected from 1995 to 1996. It matches the timeframe of the data associated with the instances.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - Procedures and questionnaires for standard DHS surveys have been reviewed and approved by ICF Institutional Review Board (IRB). Additionally, country-specific DHS survey protocols are reviewed by the ICF IRB and typically by an IRB in the host country. ICF IRB ensures that the survey complies with the U.S. Department of Health and Human Services regulations for the protection of human subjects (45 CFR 46), while the host country IRB ensures that the survey complies with laws and norms of the nation.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Individauls are interviewed directly or they complete the questionnaire on their own. The data is directly collected.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - The informed consent statement emphasizes that participation is voluntary; that the respondent may refuse to answer any question, decline any biomarker test, or terminate participation at any time; and that the respondent's identity and information will be kept strictly confidential.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Before each interview or biomarker test is conducted, an informed consent statement is read to the respondent, who may accept or decline to participate. A parent or guardian must provide consent prior to participation by a child or adolescent. DHS informed consent statements provide details regarding:The purpose of the interview/test; The expected duration of the interview; Interview/test procedures; Potential risks to the respondent; Potential benefits to the respondent Contact information for a person who can provide the respondent with more information about the interview/test.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - NO, there is not.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.* -The programme has the potential of doubling the current national contraceptive prevalence rate.
12. *Any other comments?*

    - NO

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the dataset was extracted from the pdf about Tanzania, so we have done lots of cleaning process for making the data readable. For example, we remove all the meaningless characteristics and change the name of some variables. We also used some packages such as tidyverse for analyzing our data. We also remove all the missing values and mutate new variables.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Yes, raw data can be found under input folder of my Github. It can be found at this link: https://github.com/XiaoBai-blip/paper4.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - We only use packages from R to help preprocessing the data such as pdftools.

4. *Any other comments?*

    - No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The dataset has not been used by other people or organizations except me. Since the dataset was extracted from a pdf document reporting the basic life-being condition of Tanzania. This dataset contains only 48 observations, which may lead to inaccurate results so no other people has used this before.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - Code and data are available at: https://github.com/XiaoBai-blip/paper4

3. *What (other) tasks could the dataset be used for?*

    - The dataset could be used to compare education level of male population to female population. Also, since the dataset is collected in 1996, people can use it to make comparison with data collected in another year.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - Since the data was collected based on different background characteristics, consumers should avoid reusing these data to calculate the mean percentage. For example, calculating the mean percent based on different age group might gives you the same result of calculating the mean percent based on region.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*
    - No. **Distribution**
7. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - No, because the dataset contains few variables and observations that are not good for used by third parties.
8. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset will be distributed throught my Github under input folder.It can be found here: https://github.com/XiaoBai-blip/paper4
9. *When will the dataset be distributed?*
    - The dataset will be distributed on 10th April, 2022.
10. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - Yes. DHS has the license and copyright.
    - License to DHS. To the extent Provider-owned works are incorporated into Work Product, Provider grants to DHS a perpetual, non-exclusive, paid-up, world-wide license in the use, reproduction, publication and distribution of such Provider-owned works when included within the Work Product. Provider shall not copyright Work Product without DHS' prior written consent.
11. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - Yes, by Government, Statistics Act.
12. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - NO
13. *Any other comments?*
    - NO

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - The dataset is supported by Family Planning Unit in the Ministry of Health in proving logistical support particularly female interviewers and transport. Maternal and Child Health Unit in the Iringa Regional Hospital also provided support in training services to the interviewers.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - GitHub link: https://github.com/DHSProgram
   - Website page: https://www.dhsprogram.com/
   - Phone number: +1 (301) 407-6500
   - Email Adress: info@dhsprogram.com
   - Code: codeshare@DHSProgram.com

3. *Is there an erratum? If so, please provide a link or other access point.*

   - No

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - This given dataset is conducted in 1995-1996, and the report is set to be final. The GitHub of retrieving data is https://github.com/DHSProgram, and feedback will be collected. Questions are accepted: codeshare@DHSProgram.com.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - The DHS Program is authorized to distribute, at no cost, unrestricted survey data files for legitimate academic research. Registration is required for access to data.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - The report is set to be final and is already published. However, the data and errors can be reported using email.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - There is no such extension. All risks, information and consent are slowly and clearly asked by two experimenters. Respondents can quite at any time, but after the final consent and the data collection period is ended, respondents have no access any more.

8. *Any other comments?*

   - NO

7. Reference

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2020. *Lahman: Sean 'Lahman' Baseball Database*. https://CRAN.R-project.org/package=Lahman.

Gebru, Jamie Morgenstern, Timnit. 2021. "Datasheets for Datasets." Communications of the ACM."

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sana, Asante. 2014. "Tanzania Education System." https://www.asantesanaforeducation.com/tanzania-education-system-.

Tanzania, United Republic of. n.d. "Education Has Played a Vital Role in Tanzania's Development Since Independence." https://www.unicef.org/tanzania/what-we-do/education.

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.