India Benefits Greatly if More of its Daughters are Sent to School*

A study to assess the importance of primary education using DHS 1992 data

Anshuman Agarwal Ishaan Bansal Saumya Bakshi Hailan Huang

11 April 2022

Abstract

This paper attempts to find the correlation between female enrollment in school and sexual, reproductive health, and health of children using data from the 1992 DHS survey across 25 states in India. When analyzing sexual and reproductive health we looked at the following factors – number of child marriages, use of contraceptives, utilization of antenatal care. When analyzing health of future generations, we looked at the following factors – percentage of children who are vaccinated and percentage of children that are underweight. Based on our analysis it is evident that there is a strong correlation between female education and sexual, reproductive health, and health of children.

Contents

*Code and data are available at: https://github.com/Saumya510/STA304-Paper4

References **15**

{r libraries, include = FALSE, echo=FALSE, message=FALSE, warning=FALSE} if (!require("tidyverse")) install.packages("tidyverse") if (!require("dplyr")) install.packages("dplyr") if (!require("janitor")) install.packages("janitor") if (!require("ggplot2")) install.packages("ggplot2") if (!require("knitr")) install.packages("knitr") if (!require("patchwork")) install.packages("patchwork") if (!require("here")) install.packages("here") if (!require("bookdown")) install.packages("bookdown") if (!require("pdftools")) install.packages("pdftools") if (!require("pointblank")) install.packages("pointblank") if (!require("car")) install.packages("car")

library(here) library(bookdown) library(tidyverse) library(dplyr) library(janitor) library(ggplot2) library(knitr) library(patchwork) library(pdftools) library(pointblank) library(car)

1 Introduction

India in the early 90s was a country which had recently experienced globalization and international trade for the first time since its independence from British colonialists. This meant that this time in the country's history saw lucrative economic growth. However, majority of the country was still dealing with cyclical poverty, which continues to pose severe implications on social ideology. Orthodox gender roles were rooted in the core of the Indian society, and conservative parents wanted to preserve these roles to prevent the 'bastardization' of the Indian culture from the western ideas of liberty and equality, citing their poverty as the key reason to do so.

The girl child was considered a liability to the parents, as a daughter was financial dead-weight to the household. Traditional gender roles dictated that a girl's job in society was to be married and to start her own family, thus she would not bring back the revenue spent on her upbringing by becoming a provider to the family in the future. Furthermore, the social responsibility of the father to get his daughter married and incur all the costs associated with the wedding (including an attractive dowry) set a dangerous precedent in society, to the extent where female infanticide became a pertinent problem for the country. This also disincentivized parents to invest in the education of their daughters, as educating their daughters seemed 'futile' because a girl doesn't need to be educated to be able to perform her role as a mother or a homemaker.

Past literature provides ample evidence which proves the need for further study into the benefits of female education. Analysis of DHS reports on Western and Central Africa provided statistical evidence which proves that increasing female education played a significant role in reducing child marriage (Male and Wodon 2018). Secondly, a 2006 DHS report on Uganda highlights the relationship between female education and contraceptive use – concluding that female education increases use of contraceptives and consequently, reduces fertility (Bbaale and Mpuga 2011). Thirdly, a study on Uganda concluded that female education accounted for 12% of the variation in utilization of antenatal care, a clear benefit of encouraging female education across India (Naigaga, Guttersrud, and Pettersen 2015). In addition, a different study on India concluded that children born in families with educated mothers were more likely to be fully vaccinated, representing another significant benefit of female education (Parashar 2005). Finally, a study on Ethiopia concluded that female education was a strong predictor of whether a child will be underweight, highlighting that educated females benefit health and welfare of future generations (Muche et al. 2021).

The aim of this paper is to study the benefits of increasing the number of young girls in school, on various aspects of their lives such as marriage and family planning, sexual and natal health, and the health of the future generations. The analysis uses a data-set from an era where the condition of women in the country was bad, and awareness about the benefits of female education was less known, thus the results of this study demonstrates the true extent of the benefits associated with the implementation of policies that aim to increase female representation in schools.

2 Data

{r, include = FALSE, echo=FALSE, message=FALSE, warning=FALSE} data <- read_csv(here::here("inputs/data/clean_da

2.1 Statistical Software

This report analyzes data collected by the 1992 Demographic and Health Survey (International Institute for Population Sciences Bombay) to investigate overall national family health across India from 1992-93. The statistical analysis in this report, including preparation of the data set after appropriate cleaning, reformatting and visualization of analyzed data, was done using `R Statistical Programming Language` (R Core Team 2020). In addition to the base features included in `R`, various other publicly available packaged were utilized for the analysis. For the preparation of the cleaned dataset, the packages `tidyverse` (Wickham et al. 2019) were used to manipulate the raw data which was collected from the DHS survey. The cleaned dataset was then analyzed using `dplyr` (Wickham et al. 2021) and `pointblank` (Iannone and Vargas 2022) and visualized using `ggplot2` (Wickham 2016) and `car` (Fox and Weisberg 2019). This report was constructed in R Markdown format and formatted using the following packages – `knitr` (Xie 2021) and

`bookdown` (Xie 2020). Finally, to maintain a completely reproducible workflow, `here` (Müller 2020) was used to index the location of the files necessary to create this report.

The raw data used for analysis in this paper was taken from the publicly available DHS final report. In order to create the required dataset, one full-page table containing the variables pertaining to the topic of this paper was taken from the DHS program final report using the `pdftools` (Ooms 2022) package. Screenshots of the data table for the DHS final report are available in the appendix.

2.2 Information on the 1992 DHS

Published in 1995, the DHS report aimed to compile data pertaining to Indian National Family Health and factors effecting health of individuals across India. The National Family Health Survey is a vital component of the project to strengthen survey research capabilities of the population research centers in India.

The survey was conducted with the key objective of providing state and national level estimates of fertility, infant and child mortality, practice of family planning, maternal and child healthcare, and utilization of services provided for mothers and children. The data collected through the survey is to be utilized by policymakers to allow for more accurate and applicable welfare policies to be put in place based on the key health issues facing Indians outlined in the results of the survey. In addition, a key objective of the National Family Health Survey was to provide high quality data to academics and researchers for undertaking empirical research on various population and health topics.

The National Family Health Survey covered 24 states of India, comprising of 99 percent of the Indian population. In total, 89,777 participants responses were recorded along with 88,562 households using uniform questionnaires to maintain data integrity. The data collection was carried out on a state-by-state basis from April 1992 up to September 1992.

Response and non-response bias are a key source of concern in the National Family Health Survey. There exists a strong possibility that respondents may choose to avoid responding to sensitive and personal questions which the survey is forced to ask to collect accurate and meaningful data. This could exist due to the culture of social stigma omnipresent in India surrounding women's health, childbirth, sexually transmitted disease, etc. To combat this factor, as opposed to surveys being sent through mail or electronically, researchers personally went to each household to interview respondents. In addition, in a majority of the interviews, there was a majority of female researchers as the main source of data for this survey was female respondents who were more comfortable responding to other females as opposed to males. The non-response bias could have been taken care of by providing incentives for participation, however due to funding issues throughout the survey, providing incentives was not feasible due to the scale of the study and the large sample size.

Additionally, an important weakness of the survey is the possibility of the presence of the conformity bias. Due to the social stigma surrounding women's' health and sexually transmitted disease in India, respondents may have felt the need to tailor their answers to conform to societal expectations which in some areas of India enforce silence on these topics.

2.3 Methodology

The aim of this paper is to create a holistic view on the state of Female sexual and reproductive health and its correlation with proportion of female attendance in primary education using the 1992 DHS report on National Family Health. Our analysis utilizes a dataset from 1992 as it provides us with the most meaningful data for our paper, while remaining relevant enough to warrant further research into the plight of women. This study attempts to categorize the true impact of female education on sexual and reproductive health along with health of future generations.

Our key variable used throughout this study is the percentage of women attending school. We believe this would be a more meaningful relationship to study as opposed to the relationship with female illiteracy rates. While illiteracy rates may show increased correlation with our response variables, the purpose of this paper is to highlight the importance of female education and provide empirical evidence to encourage policymakers

to incentivize female attendance in schools across India. Female education is more easily influenced by government policies as opposed to female illiteracy rates.

Our key response variables include – female contraceptive use, antenatal care, and health of infants. While there were other variables available to study in the datasheet, we believed these to be the most meaningful for the purpose of our paper. Use of female contraceptives will be shown to be directly correlated with female education, representing sexual health benefits of female school attendance. While there were other variables available which could also be used to demonstrate increased sexual health – percent of women using sterilization – they included data from male respondents as well which would take away from the accuracy of our results and conclusions.

Secondly, we chose Antenatal care to be our key indicator of maternal health and welfare. Utilization of antenatal care services will be shown to be directly correlated with female education, representing improved health of future generations as a result of female school attendance. At first glance percentage of births delivered in a health facility would be a better indicator of health of future generations. However, births delivered in a health facility does not necessarily mandate presence of a specialist in natal care. By analyzing utilization of antenatal care, we are demonstrating the availability or lack of natal care specialists in rural parts of India. This is in conjunction with the purpose of our paper which is to demonstrate the multitude of advantages of promoting female education and highlight areas which require government policy intervention.

List of Variables used for analysis - Percent attending school (females aged 6-14): Variable containing information on female school attendance - Percent of women aged 20-24 married before age 18: Variable containing information on female child marriages - Percent of women using any contraceptive method: Variable containing information of married women, aged 13-49, using contraceptive methods - Unmet need for family planning: Variable containing information on percentage of married women not using family planning, regardless of their intentions to bear a child - Mothers receiving antenatal care: Variable containing information on female utilization of antenatal care - Percent of children fully immunized (age 12-23 months): Variable containing information on percentage of children receiving the BCG, Measles, DPT, and Polio vaccine - Percentage of living children under four years of age who are under-weight: Variable containing information on underweight children assessed by 'weight-for-age'.

3 Results

The study is structured to see the effect of an increase in the percentage of women (aged 6-14) attending school on different aspects of a woman's life. The first category includes aspects for marriage, sexual health, and family planning, and the second category includes aspects regarding the health of a mother and her children. The dataset which is used to conduct the analysis is a cross-sectional dataset which contains a range of historical information for 25 states and 1 observation for the whole of India, bringing the total sample size to 26. The analysis uses scatterplots to visualize the association between percentage of girls aged 6-14 attending school with various variables which represent the chosen aspects of a woman's life. All the figures used to get the following results can be found in the Appendix.

The first association which was looked at was between female school attendance of a state and its child marriage rates. (Figure **??**) shows how much of the geographical variation in the percentage of women who are married under 18 is associated with the geographical variation in the percentage of women (aged 6-14) who are enrolled in school. The trend line seen in the figure represents the linear regression function with child marriage rates as the response and percent attending school as the predictor. As seen in (Figure **??**), there seems to be a linear and monotonically decreasing association between child marriage rates of a state and the percent attending school. Majority of the data points show evidence of a linear relation and fall inside the shaded region surrounding the regression function, which maps the confidence interval of the slope.

The next association that was analysed was between female attendance of a state and the percentage of women in that state that use any form of contraceptives. (Figure **??**) uses a scatter plot to demonstrate the extent to which these factors are related. Like (Figure **??**), a linear trend line and the confidence interval band is added to the figure. As seen in (Figure **??**), there seems to be a monotonically increasing association between the percent of women attending school and the percent of women using any contraceptive methods.

The confidence interval band for the slope of the linear trend line is much larger compared to the (Figure **??**). The scatter plot also shows a very faint signs of a linear pattern in the data points, and the function seems to be suffering from the presence of outliers.

Extending the context to see the effect of female school attendance on women's sexual health in the country, the association between percentage of girls attending school in a state and the percentage of females with unmet family planning needs which is defined as married females in India who reported that they do not want to have kids in 1992 or wanted to wait at least 2 years before having another child, but have not reported using any means of family planning. (Figure **??**) uses a scatter plot to demonstrate the extent to which these factors are related. Like the previous figures, a linear trend line and the confidence interval band is also added to the figure. As seen in (Figure **??**), the downward sloping trend line indicates that a higher percentage of girls attending school is associated with a lower percentage of women with unmet family planning needs. Like in the section above, this association too seems to be influenced by the presence of outliers.

The study next wishes to assess the effect of varying female attendance on maternal health and the nourishment of the next generation. To examine the effect on maternal health, the association between female attendance rates and the percentage of women receiving antenatal care was seen. (Figure **??**) below plots the respective variables like done previously. (Figure **??**) shows a positive association between female attendance in school and percent of mothers receiving antenatal care. The scatter plot also shows some evidence of a linear relationship, however due to the small sample size, outliers seem to have significant influence.

To assess the relationship between female attendance and the health of the future generations, information about immunization and nourishment were picked. (Figure **??**) plots the data available for 25 states and the national average as a scatter plot. A linear trend line representing the linear regression function is added to the plot. As seen from [Fig], there seems to be a positive association between female attendance in school and percent of children aged between 12-23 months who are fully immunized as seen by the upward sloping regression function. There is some evidence of linearity, and this plot seems to be less influenced by the presence of outliers. Most data points in the plot show signs of a positive linear association, however due to small sample size, sampling error cannot be ruled out. Following the same methodology from the sections above, (Figure **??**) visualizes the association between female attendance in school and the percent of children under 4 who are classified as under-weight. As seen from the (Figure **??**), there is an evident association between female attendance in school and the percentage of children who can be classified as under-weight. States with higher female attendance, on average, have children who are relatively healthy and nourished. The scatter plot shows evidence of a linear relationship; however, some elements of the plot show non-linear patterns, which could be indicative of the insufficiency of a linear model or the presence of unexplained heterogeneity.

4 Discussion

The results reported in the section above are only to be interpreted in the context of a state-level comparison, and causality (even asymptotic causality) should not be assumed due to the observational nature of the data-set, and the small sample size being used here.

4.1 Benefits of Investing in Increasing Female Education

Child Marriage was a social evil rampant throughout India (with 54.2% of women aged between 20-24 at the time of data collection, being married before the age of 18), and while the government of India has made it illegal for women below 21 years of age to be married, the practice of child marriage is still present in some of the rural parts of the country. According to the results obtained from looking at (Figure **??**), on average, states with a high female attendance rate is associated with low child marriage rates. The reason why this result is seen could be due to the endogeneity issue of the urbanity and poverty of states. The states which are majorly urban and on average richer, would have low child marriage rates and have relatively higher female attendance in school. However, this association cannot be ignored as past research has been successful in finding an association like the one found in this study. The urbanity of the states will pose a

limitation to the analysis that will follow, the main take away would be about the importance of including this information in the models, therefore using a multi-variate model would improve the ability of this study to make accurate inferences.

In India, back in the 90s, the biggest responsibility imposed on women by society was to become a parent and a homemaker while her husband provides financially for the family. Thus, even the influence of western ideologies could not destigmatize intercourse for pleasure and the use of contraceptives, especially for women for whom even the conversation of intercourse outside the bounds of reproduction was considered impure. (In 1992, only 40.6% of married women aged between 13-49 reported using contraceptives.) This way of thinking was extremely detrimental to the sexual health of married women, and it would not be logically incorrect to attribute this phenomenon to non-existent sexual education or awareness about the existence of female contraceptives in general. According to the results obtained from looking at (Figure **??**), we see an upward sloping linear trend line, however the inherent validity of using simple linear regression must be questioned. The Gauss-Markov assumptions which form the basis of the OLS framework seem to be violated. A very faint linear pattern is noticed, and if residual plots are consulted it will not be surprising to see other violations. Even though it makes logical sense that higher levels of education correlate with increased awareness about the availability and benefits of using contraceptives, this might not be a direct association. Rural areas of India might suffer from low availability of female contraceptives (a probable cause of the geographical variation seen) and increasing education levels will not have an immediate effect in reversing that. However, in the long run, higher awareness could correspond to a higher demand for contraceptives and thus females might advocate for a supply that meets their demand. Lastly, just increasing levels of educations might not have the expected effects until the relevant topics such as sexual education are being taught in schools, but the study seems to find evidence suggesting that increasing female attendance might be the first step towards achieving this.

Alongside economic growth, India was also experiencing concerning population growth in the 90s. Due to the taboo surrounding the use of female contraceptives and non-reproductive intercourse, combined with the poor presence of family planning resources; out of all the married females in India who reported that they do not want to have kids in 1992 or wanted to wait at least 2 years before having another child, almost 20% had family planning needs that were unmet. According to results seen from (Figure **??**), if more females were educated at a younger age about the existence of family planning or about the family planning resources that are available to them, more married females would actively try to get their needs met. Furthermore, inferences made using the linear regression framework would be misleading as non-linear patterns are seen along with a very wide confidence interval area. Thus, even though an association is visible, it might not be direct and heavily influenced by the presence of confounding variables like urbanity, which are left out from the model. Past literature establishes that increased knowledge about the existence of family planning resources leads to a fall in the number of women who have unmet needs, however like with contraceptives, increasing female attendance might be the necessary first step, and the effects of policies might only be visible in the long run.

In 1992, only 62.3% of expecting mothers received antenatal care. Low accessibility of antenatal care in terms of number of hospitals available with neonatal specialists, especially in rural areas can be an issue, areas which also see low female attendance in school, giving rise to any associations we see in (Figure **??**) above. However, another reason we can be witnessing the upward sloping trend line, and some evidence of positive linear association, is that with increasing education increases the overall level of awareness about the importance and existence of antenatal care. Some elements of the figure like existence of discomforting non-linear patters suggest that we take the results of this figure with a pinch of salt, as even though linear regression seems to have found an association, making definitive inferences can be misleading. Furthermore, education might only affect a part of the variation that arises due to the exogenous decision of women to not opt for antenatal care (due to misinformation about the importance or existence of such care). Other factors such as availability of this care in the rural parts of the country seems to hold more logical significance than education in this case.

Educating women at an early age should not only benefit them, but also benefit the next generation. When the data was collected, only 35.4% of children between the ages of 12-23 months were immunized against illnesses like measles and polio. Even today, third world nations are battling to eradicate these diseases and it

is common knowledge that vaccination is the way to do that. Research in the past have confirmed the positive association between the level of education of mothers and immunization rates of infants, thus we don't see any results that conflict those findings obtained from (Figure **??**). However, the limitation of heterogeneity and lack of robustness to outliers still applies to the findings here. That being said, this association might be the most direct non-causal association found by this study. Logically, if vaccine availability and coverage are controlled for, immunization becomes a personal choice based on the persons knowledge in the subject which can be increased with education.

Malnutrition is a serious concern in countries like India where due to large levels of poverty and rapidly increasing population, proper nutrition is a scarce commodity. This is why, when the data was collected it was found that a shocking 53.4% of all children under 4 years of age fell into the classification of being under-weight. (Figure **??**) tells us that on average states with higher female attendance report fewer cases of children under 4 being under-weight. While malnutrition is a causal effect of extreme poverty, the national average being more than half suggests that other factors indirectly contribute towards this phenomenon. Thus, education's association can be explained as a long run impact. Educating a generation will lead to better living conditions for the next generation, which might mean lower poverty and thus lower rates of children being under-weight. The statistical limitations of using a simple linear regression function to model the association still applies, and many confounding factors need to be controlled to see the marginal effect of increasing female attendance in schools on the health of the next generation.

4.2 Conclusion

After extensive statistical analysis detailed above, we found a strong correlation between women's' education and sexual, reproductive and infant health. Using the datasheet, containing the variable which collects number of women enrolled in school, we were able to show a strong correlation between education and health of women and their children. Firstly, we see a strong correlation between women's education and child marriage – states with higher women literacy rates also report lesser child marriages. However, this result can also be attributed to the fact that states with higher rates of women's education are likely to be the larger cities in India which had already enacted strict rules forbidding child marriage in 1992. The actual effect of women's education on number of child marriages, considering the demographic element, is further analyzed in the discussion section above.

Subsequently, we see a strong positive correlation between women's education and use of contraceptives. This is a strong indicator of improved women's health as a result of higher literacy rate. However, similar to the child marriage case, this can be attributed to larger cities being more progressive as opposed to a direct correlation between literacy rates and use of contraceptives.

A big issue facing India in 1992 was the lack of pre-natal care which increased the health risks facing women and infants. However, our analysis shows that increased women's education is positively correlated with increased pre-natal care. As a result, we see reduced infant mortality rates in states with higher women's literacy rates. However, similar to the case above, this correlation can be attributed to increased availability of pre-natal care in larger cities which were more progressive and actively encouraged women's education.

In summary, based on our analysis of the 1992 DHS report on India, we have found a strong correlation between women's education and sexual, reproductive, and infant health. While there may be other geographical and demographic variables influencing this conclusion, it is evident that women's education plays a vital role and should be the focus of policymaker's decisions in the coming years.

4.3 Limitations and Scope for future analysis

While we were able to use the data to reach meaningful conclusions, there were some limitations to the data obtained from the DHS report that potentially reduced the accuracy of our conclusions and the ability to generalize our predictions to the Indian population. Key sources of limitations include, but are not limited to, the small sample size, the use of simple linear regression modelling to see associations, the fact that many important variables were missing from the data set such as information about the urbanity and poverty level

of the states. As a note to researchers planning to work in this area, using multi-variate regression models with robust standard errors would prove to be better for statistical inference.

Furthermore, response and non-response bias are a key source of concern in the National Family Health Survey. There exists a strong possibility that respondents may choose to avoid responding to sensitive and personal questions which the survey is forced to ask to collect accurate and meaningful data. This could exist due to the culture of social stigma omnipresent in India surrounding women's health, childbirth, sexually transmitted disease, etc.

Finally, an important weakness of the survey is the possibility of the presence of the conformity bias. Due to the social stigma surrounding women's' health and sexually transmitted disease in India, respondents may have felt the need to tailor their answers to conform to societal expectations which in some areas of India enforce silence on these topics.

Appendix

A Datasheet

A.1 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 analyze societal welfare and women's health across India. In addition, the dataset served the purpose of strengthening research capabilities of population research centres across India.

2. Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

- Dataset created by Ministry of Health and Family Welfare

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 government of India

A.2 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.

- Instances represent the different states and territories across India.

2. How many instances are there in total (of each type, if appropriate)?

- Total of 26 Instances

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).

- Dataset takes into account all states of India, as of 1992. Sample represents 99% of Indian population

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.

- Each instances has 23 variables and 1 defining variable

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

- Each instance has a label which provides information regarding which state that instance is providing data on

6. 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.

- No relationships between individual instances present

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

- Errors arise as a result of biases – Non-Response bias and confirmation bias

8. 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.

- The dataset does not contain any confidential information

9. 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.

- The dataset is comprised of Women and infants

10. 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.

- It is not possible to identify individuals based on information in dataset

11. 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.

- The data provides some health data (sexual and reproductive health of women in India)

A.3 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.

- Data associated with each instance was acquired through interviews across all states

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?

- Manual Human Curation

3. 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.

- April 1992 – September 1993

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

- Data collected from DHS (Demographic and Health Surveys)

5. 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.

- Data was collected through voluntary interviews

A.4 Processing/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 miss-

ing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

- The dataset was obtained from a PDF report and converted into a usable dataset using R pdftools

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.

- The raw data was saved – https://github.com/Saumya510/STA304-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.

- R was used

A.5 Uses

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

- The dataset has not been used for other tasks

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.

- https://github.com/Saumya510/STA304-Paper4

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

- Any research into the state of women and children in 1992-93 India

4. Are there tasks for which the dataset should not be used? If so, please provide a description.

- The dataset would not be useful for any research not looking into women and children in 1992-93 India

A.6 Distribution

1. 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.

- The dataset is openly available for personal use

2. How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

- Dataset will be distributed using GitHub: https://github.com/Saumya510/STA304-Paper4

3. 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.

- No restrictions have been imposed

4. 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 such control or restriction is applicable

A.7 Maintenance

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

- Anshuman Agarwal
- Saumya Bakshi

- Ishaan Bansal
- Hailan Huang

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

- Owner can be contacted using GitHub

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

- No erratum available

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)?

- Currently no plans to update dataset, if in the future it were to be updated the changes would be reflected on GitHub: https://github.com/Saumya510/STA304-Paper4

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.

- No applicable limits as respondents voluntarily provided responses

6. 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.

- Currently no mechanisms to contribute or build on the dataset

B DHS Fact Sheet

{r, warning=FALSE, echo=FALSE, out.width="75%", fig.cap="Screen Grab of Tables used as the dataset"}
include_graphics(here::here("outputs/figures/1.jpeg")) include_graphics(here::here("outputs/figures/2.jpeg"))

C Figures

{r one, fig.cap="A Scatterplot showing the Relationship between the Percentage of Females Attending Schools and Percentage of Women Married before Age 18", echo = FALSE, message = FALSE, warning = FALSE} data %>% ggplot(mapping = aes(x = `Percent attending school (females age 6-14)`, y = `Percent women age 20-24 married before age 18`)) + geom_point() + geom_smooth(method = lm, colour = "red") + theme_minimal() + labs(x = "Percentage of Females Attending Schools", y = "Percentage of Women Married Before Age 18")

{r two, fig.cap="A Scatterplot showing the Relationship between the Percentage of Females Attending Schools and Percentage of Women Using any Contraceptive Method", echo = FALSE, message = FALSE, warning = FALSE} data %>% ggplot(mapping = aes(x = `Percent attending school (females age 6-14)`, y = `Percent of women using any contraceptive method`)) + geom_point() + geom_smooth(method = lm, colour = "red") + theme_minimal() + labs(x = "Percentage of Females Attending Schools", y = "Percentage of Women Using any Contraceptive Method")

{r three, fig.cap="A Scatterplot showing the Relationship between the Percentage of Females Attending Schools and Unmet Need for Family Planning", echo = FALSE, message = FALSE, warning = FALSE} data %>% ggplot(mapping = aes(x = `Percent attending school (females age 6-14)`, y = `Unmet need for family planning`)) + geom_point() + geom_smooth(method = lm, colour = "red") + theme_minimal() + labs(x = "Percentage of Females Attending Schools", y = "Unmet Need for Family Planning")

{r four, fig.cap="A Scatterplot showing the Relationship between the Percentage of Females Attending Schools and Unmet Need for Family Planning", echo = FALSE, message = FALSE, warning = FALSE} data %>% ggplot(mapping = aes(x = `Percent attending school (females age 6-14)`, y = `Mothers receiving antenatal care (last 4 years)`)) + geom_point() + geom_smooth(method = lm, colour = "red") + theme_minimal() + labs(x = "Percentage of Females Attending Schools", y = "Percentage of Mothers Receiving Antenatal Care")

{r five, fig.cap="A Scatterplot showing the Relationship between the Percentage of Females Attending Schools and Percentage of Children Fully Immunized", echo = FALSE, message = FALSE, warning = FALSE} data %>% ggplot(mapping = aes(x = `Percent attending school (females age 6-14)`, y = `Percent of children fully immunized (age 12-23 months)`)) + geom_point() + geom_smooth(method = lm, colour = "red") + theme_minimal() + labs(x = "Percentage of Females Attending Schools", y = "Percentage of Children Fully Immunized")

{r six, fig.cap="A Scatterplot showing the Relationship between the Percentage of Females Attending Schools and Percentage of Children Underweight", echo = FALSE, message = FALSE, warning = FALSE} data %>% ggplot(mapping = aes(x = `Percent attending school (females age 6-14)`, y = `Percent of children under 4 years of age underweight`)) + geom_point() + geom_smooth(method = lm, colour = "red") + theme_minimal() + labs(x = "Percentage of Females Attending Schools", y = "Percentage of Children Underweight")

References

Bbaale, Edward, and Paul Mpuga. 2011. "Female Education, Contraceptive Use, and Fertility: Evidence from Uganda." *Consilience*, no. 6: 20–47.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Iannone, Richard, and Mauricio Vargas. 2022. *Pointblank: Data Validation and Organization of Metadata for Local and Remote Tables*. https://CRAN.R-project.org/package=pointblank.

Male, Chata, and Quentin Wodon. 2018. "Girls' Education and Child Marriage in West and Central Africa: Trends, Impacts, Costs, and Solutions." In *Forum for Social Economics*, 47:262–74. 2. Taylor & Francis.

Muche, Amare, Mequannent Sharew Melaku, Erkihun Tadesse Amsalu, and Metadel Adane. 2021. "Using Geographically Weighted Regression Analysis to Cluster Under-Nutrition and Its Predictors Among Under-Five Children in Ethiopia: Evidence from Demographic and Health Survey." *PloS One* 16 (5): e0248156.

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

Naigaga, Mpolampola DAS, Øystein Guttersrud, and Kjell S Pettersen. 2015. "Measuring Maternal Health Literacy in Adolescents Attending Antenatal Care in a Developing Country–the Impact of Selected Demographic Characteristics." *Journal of Clinical Nursing* 24 (17-18): 2402–9.

Ooms, Jeroen. 2022. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents*. https://CRAN.R-project.org/package=pdftools.

Parashar, Sangeeta. 2005. "Moving Beyond the Mother-Child Dyad: Women's Education, Child Immunization, and the Importance of Context in Rural India." *Social Science & Medicine* 61 (5): 989–1000.

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

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

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

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2020. *Bookdown: Authoring Books and Technical Documents with RMarkdown*. https://github.com/rstudio/bookdown.

———. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r. R Package Version 1.31*.