**PREDICTORS OF HEALTH INSURANCE COVERAGE AND REPRODUCTIVE OUTCOMES AMONG WOMEN IN SELECTED AFRICAN COUNTRIES**

***OLADOKUN, BLESSING ELIJAH***

***(31894290)***

**A PROJECT SUBMISSION FOR ADVANCED STATISTICAL MACHINE LEARNING (STAT 8730), UNIVERSITY OF NEBRASKA AT OMAHA.**

**CHAPTER 1: INTRODUCTION**

* 1. **Background of the problem**

Health insurance coverage is still a major challenge in many developing countries; this is evident in that the Millennium Development Goals (MDG) and Sustainable Development Goals (SDG) still emphasizes the Universal Health Coverage (UHC) as a global health priority (Chapman, 2016). Since the goal of UHC is that everyone has access to standard and basic healthcare services, when necessary, regardless of their financial status (Frenk, 2015); many African countries happen to fall in the category of people who have not been privileged to have readily accessible healthcare, either in that they are not remotely available to them, or they cannot afford it when it is available.

The reality in low- and middle-income countries (LMIC’s) is that the average resident survive on less than one-dollar per day (World Bank, 2021). Meanwhile, reproduction among women is still on the increase in these LMIC’s; hence, the average women in these countries are predictably lacking access to maternal healthcare, for financial reasons. Hence, why a study on the health insurance coverage in these countries is needed to understand how the problem can be solved to meet the people’s healthcare needs. The first step is knowing the distribution of the health insurance coverage in these countries; then, finding out factors and covariates that could predict whether a woman, given some background data, is likely to have an insurance or not.

Moving away from just having a health insurance coverage. The lack of health insurance coverage has in fact had a spiral effect on other reproductive outcomes. Owing to lack of access to healthcare from pregnancy stage, many of these women have lost their pregnancy or lost the child while at infant stage. The aftermath of not having access to healthcare does not stop at the point of women been deprived of reproductive healthcare, it leaves a mark on the livelihood of the kids they deliver; a good number of kids born to these women who lacked care die within their first year of birth, and those who survive beyond the first year may not survive till age of five. This study tries to trace this early child death to many pregnancy-related issues of the mothers; from access to ANC, timing to first attendance of ANC, number of ANC attended, history of pregnancy termination, delivery of child at a health facility, etc. The status of the child given-birth-to by these women was studied, to find out what background factors of the mothers (socio-demographics and pregnancy-related) can predict status of the child within their first year and after their first year of birth.

* 1. **Research questions**

For this project, we are much concerned about predictors and covariates. As a result, the broad aim of the project will be to develop predictive models that can determine three key outcomes: (a) Number of children born by a woman (b) If a woman has access to health insurance coverage (c) Status of a child delivered, within the first year and within five years. To that effect, the project will seek to provide answers to the following research questions:

1. What are the background factors of the women in African countries that can be used in a predictive model to determine the number of children they have had?
2. What are the significant predictors of health insurance coverage for women of reproductive health, in African countries?
3. Can the 1-year and 5-years outcome of a child, in terms of whether they are alive or dead, be determined based on the mother’s pregnancy-related factors?
   1. **Significance of the study**

At the end of the report writing, we would be able to pre-determine intrinsic factors that predispose women to lack of health insurance coverage. In addition, we would be able assess the impact of women’s lack of health insurance on the child’s status within the first year and within five years of birth.

Without neglecting other factors that could impact the child’s chances of survival up to age 5; the project will also attempt to quantify the contribution of other pregnancy- and delivery-related factors towards status of the child within the first year and within fifth year of birth.

* 1. **Structure of the report**

Excluding the title page, the report will consist of 7 major sections, including:

1. Introduction to the project: This will be a 2-page write-up on the background, significance, and objectives of the study.
2. Review of literature: This is a write-up not exceeding 2 pages to discuss what others have found out in terms of the study outcomes and objectives.
3. Method of data wrangling, cleaning, and analysis: This section will throw light on the details on the data, its source, how it was restructured and cleaned. The section will also reveal how the data will be analyzed to provide answer to each research questions.
4. Presentation of results and report: This section will show all results of the data analysis, using tables, charts, and graphics, as appropriate. The section will include report of the analyzed results.
5. Conclusion and Recommendations: This will be the concluding section of the report where the result will be briefly discussed, and recommendations will be made.
6. Appendix: The appendix will include all the R-code used in the entirety of the project (from data restructuring to cleaning to analysis to visualization).
7. References: This will include a list of all materials used in the course of project.

**CHAPTER 2: DATA AND DATA PROCESSING PROCEDURES**

**2.1 Source of data**

The dataset for this project was obtained through DHS (<https://dhsprogram.com/>). The DHS has published the dataset for different countries in different years. The data was collected by the USAID funded program, DHS, through national surveys carried out in each of these countries. Worth of note is that the data was collected in different years in the different countries; but the target is the most recently collected data in each of the selected countries.

The DHS Kid’s Recode (KR) dataset for 11 African countries, including Angola, Egypt, Ethiopia, Gabon, Ghana, Kenya, Mali, Nigeria, South Africa, Zambia and Zimbabwe, were pooled into a single datafile for this project. The combined dataset across the 11 countries has record of 137,229 women and their most recent kids (under 5-year-old).

Graphical user interface, application, Word

Description automatically generated**2.2 Overview of raw data**

Figure : Type of variables in raw dataset

Graphical user interface, application

Description automatically generated

Figure : List of variables in raw dataset

**Table

Description automatically generated**

Figure : Overview of the raw data (Top rows)

**Table

Description automatically generated**

Figure : Overview of the raw data (Bottom rows)

**2.3 Study variables**

The combined dataset across the selected countries has 28 distinct variables. This includes the variables measuring the outcomes, and the variables considered as predictors (factors and covariates). Below is a full list of the variables in the dataset:

1. Country,
2. Year (when survey was conducted),
3. Age (of women),
4. Region (where women lived- Urban or Rural),
5. Highest level of education attained (for women and partner)
6. Wealth index class,
7. Marital status of women (as of time of survey)
8. Age at first sex
9. Index child is twin (Single or multiple)
10. Sex of index child (Male or Female)
11. Preceding birth interval (in month)
12. Duration of pregnancy
13. Currently pregnant
14. History of pregnancy termination
15. Modern contraceptive use
16. Child delivery in health facility (Place of delivery)
17. Delivery by CS
18. Size of child at birth
19. Timing of first ANC check
20. Number of ANC visits during pregnancy
21. Received ante-natal care in health facility
22. Mother smokes
23. Number of children ever born
24. Number of living children
25. Coverage by health insurance
26. Status of last child (dead or alive)
27. If dead, age at death
28. If alive, age as of time of survey

**2.31 Dependent variables**

From the above list, three variables will be considered outcome variables for this study. They are: (a) Number of children ever born, (b) Health insurance coverage (c) Status of last child.

**2.32 Predictors/Independent variables**

**Table 1: List of predictors and outcome**

|  |  |  |
| --- | --- | --- |
| **Outcome** | **Predictors/Co-variates** | **Comment(s)** |
| **Number of children ever born** | Age group of women  Region  Educational level of women  Wealth index class  Marital Status  Age at first sexual intercourse  History of pregnancy termination  Smoking status of women | Outcome is categorized as follows:  0–1 child, 2–3 children, 4+ children  9 categorical predictors |
| **Health insurance coverage** | Age group of women  Region  Educational level of women  Educational level of their partners  Wealth index class  Marital Status  Currently pregnant  History of pregnancy termination  **Timing of ANC visit**  **Number of ANC visits during pregnancy**  Modern contraceptive use  Child delivery in a health facility  Delivery of index child by Caesarean Section  Received Ante-natal care in health facilities | Outcome is categorized as:  Yes- covered by insurance  No- not covered by insurance  13 categorical predictors |
| **Status of index child** | Age group of women  Region  Educational level of women  Wealth index class  Marital Status  Index child is single or multiple  Sex of index child  Preceding birth interval (in months)  Duration of pregnancy (in months)  History of pregnancy termination  Child delivery in health facility  Delivery of index child by CS  Received Ante-natal care in health facilities  Size of child at birth (in kg)  Smoking status of women  Ever lost a child previously  Health insurance coverage | Outcome is categorized as:  Alive – <= 1 year old  Alive – Older than a year  Dead – Within first year of birth  Dead – After first year of birth  10 categorical predictors and 3 numerical predictors  Numerical predictors will be categorized using standard classification system. |

**2.4 Data restructuring and wrangling**

The DHS default data coding included lots of redundant code names and values that makes the data not usable in its raw state. Many of the variables were coded to include categories such as: “Don’t know”, “Inconsistent”, and “Missing. Some of the variables had a mix of both numeric and texts, while some variables were repeated in the dataset with different variable names (E.g., v025 and v102 – both contained region of the women).

Below is an overview of the default coding for some variables (that needed to be restructured), as used by DHS:

Graphical user interface, application, Word

Description automatically generated

Figure : Unique values of selected variables

Graphical user interface, application, Word

Description automatically generated

Figure : Unique values of selected variables

Graphical user interface, application, Word

Description automatically generated

Figure : Unique values of selected variables

Table

Description automatically generatedBelow is a screenshot of the R-codes used in restructuring the affected variables:

Figure : R-code used in data restructuring

**2.5 Variable transformation and creation of new variables**

Many of the variables in the dataset were transformed to a new variable to make them ready for analysis. Below is a list of concerned variables and what they were transformed into:

1. Contraceptive methods: This variable originally had numerous categories, but was transformed into 3 categories- No method, non-modern methods, and modern methods
2. Marital status: This was recoded into 2 categories of “married” and “unmarried”. The unmarried category included women who were “single/never married”, “divorced”, “widowed”, or “separated”.
3. Twin child status: This was recoded into “Single birth” and “Multiple birth”.
4. Place of child delivery was recoded into “Health facility” – which included those who had their deliveries in any form of hospital/clinic or health center, whether government- or private-owned.
5. Duration of pregnancy was recategorized into 3 groups: Pre-term (those who delivered before 9 months), term (those who delivered at 9 months), and post-term (those with delivery after 9 months).
6. Preceding birth interval, which was reported in months, was classified into 4 groups namely: Less than 12 months, 12 – 23 months, 2 – 5 years, and more than 5 years.
7. Timing of first ANC attendance: First trimester (within first 3 months), Second trimester (within 3 – 6 months), Third trimester (6 – 9 months).
8. Number of ANC visits was initially reported in days, but was recategorized as “No ANC visit”, “Insufficient ANC visit”- which includes those who attended ANC not up to 8 times during the pregnancy, “Sufficient ANC visit” – which includes those who had ANC attendance at least 8 times.
9. Age at first sexual intercourse was originally reported in years but recoded into 4 categories: Below 18-years, 18 – 24 years, 25 – 29 years, and 30+ years.
10. Total children ever born was reported as counts by DHS but was recorded into 3 categories, for the purpose of this project: 0-1 child, 2-3 children, 4+ children.

In some other cases, new variables created from a combination of two or more variables from the initial dataset. Three variables were newly created from the dataset, which includes – Child status, Smoking status of mother, and ANC facility.

1. Child status: This was created from a combination of 3 variables, “Child is alive”, if ALIVE – “Age of child at time of survey”, and if DEAD – “Age of child at time of death”. The resulting categories were: “Alive: <= 1 year old”, “Alive: > 1 1 year old”, “Dead: Within 1st year”, and “Dead: After 1st year”.
2. Smoking status: This was created from a list of 6 variables that asked the women if they smoked any of cigarettes, tobacco pipe, kreteks, cigars or cheroots, and water pipe. The resulting variable had a binary category: “Yes”- for those who smoked any of the above listed option, and “No”- for those who did not smoke any from the list.
3. ANC facility: The new variable was created from 6 variables in the dataset. Each of the 6 variable asked if the women visited any of the facilities including – government hospital, government health center, government health post, other government health facility, private hospital/clinic, and other private health facility.

To carry-out the aforementioned data restructuring procedures, some functions from the *dplyr* package in R were used, these were *gsub, mutate,* and *case\_when*. Below is a screenshot of how the lines of code were used in R.

Table

Description automatically generated

Figure : R-codes used for variable transformation and new variable creation

**2.6 Procedures for dealing with missing data**

Following restructure of the dataset, transformation of variables and creation of new variables, the entire dataset was screened for redundant variables and missing values among needed variables.

21 variables from the original dataset were removed; 6 of which were variables duplicated in the dataset by DHS, and the other 15 were original variables that have been transformed into a new variable.

The *is.na* function in R was used along with other functions to detect and identify columns in the dataset that had at missing values. Then then entire dataset was disaggregated into 3 three subsets of data using the *subset* function in R. Each subset data represents the dataset to be used for each of the 3 objectives of this project; hence, each subset has ONLY variables that are essential to each of the outcome, as listed on Table 1. Within each of the 3-subset dataset, the *drop\_na* function in the *tidyr* package was used to remove any row that had at least one missing value. Below is a screenshot of the R codes used for the data Graphical user interface, text, application, Word, email

Description automatically generatedcleaning:

After cleaning the individual dataset for each of the specific objectives, some observations were lost and there were non-uniform number of observations (rows) for each of the subset. For subset 1: There were 14 variables and 120,653 observations. For subset 2: There were 17 variables and 70,382 observations. For subset 3: We had 16 variables and 69,151 observations.

**Table

Description automatically generated2.7 Overview of cleaned data**

Figure : Overview of Subset 1 Data

**Table

Description automatically generated**

Figure : Overview of Subset 2 Data

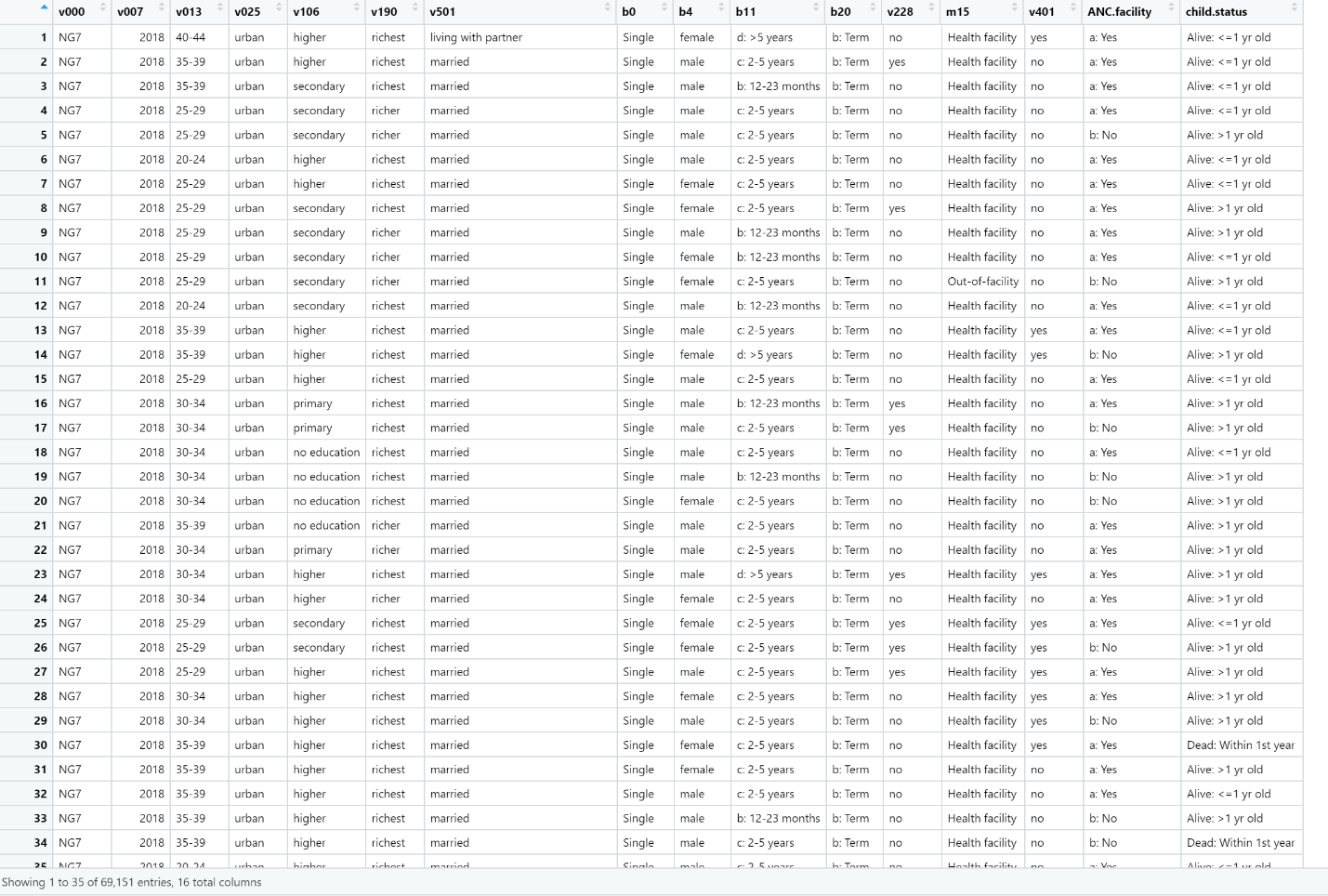


Figure : Overview of Subset 3 Data