**Department of Statistics, The Chinese University of Hong Kong**

**STAT5102 Regression in Practice (Term 1, 2018-19)**

**Group Project**

**Generalised Linear Model for Estimating Mortality Rate Using Development Indicators**

**Date: 19 December 2018**

**Group members (Name, Student ID):**

CHAN, Yiu Fung Frankie, 1155010561

LI Wai Lan, 1155118254

WONG, Tsz Wing, 1004666311

WONG, Yiu Chung, 1155017920

**Objective**

In year 2015, the United Nations adopted the 2030 Agenda for Sustainable Development and 17 Sustainable Development Goals (SDG)[[1]](#footnote-1). One of the important goals is ensuring healthy lives and promoting the well-being including the target to reduce the under-five mortality to at least as low as 25 per 1,000 live births by 2030. According to the 2017 report on “Levels and Trends in Child Mortality” produced by the UN Inter-agency Group for Child Mortality Estimation[[2]](#footnote-2), 7,000 children died under the age of five globally every day. If the SDG 2030 target can be met, 10 million lives of children under five will be saved between period from 2017 to 2030.

To contribute to the global efforts in improving child mortality, this study aims to develop a generalised linear model for estimating mortality rate using development indicators. We investigate the dataset obtained from The World Bank: World Development Indicators. A Generalised Linear Model is used to perform a prediction analysis on the response variable: Mortality rate, using a selected set of predictor variables.

**Import Data**

The World Bank data set contains relevant data from 214 countries and jurisdictions for the year 2010, covering the following:

| No. | Variables Labels | Variables Meaning |
| --- | --- | --- |
|  | Year | Year |
|  | YearCode | Year Code |
|  | Country Name | Country Name |
|  | Country Code | Country Code |
|  | ARI treatment (% of children und | ARI treatment (% of children under 5 taken to a health provider) |
|  | Adjusted savings: education expe | Adjusted savings: education expenditure (% of GNI) |
|  | Adolescent fertility rate (birth | Adolescent fertility rate (births per 1000 women ages 15-19) |
|  | Age dependency ratio (% of worki | Age dependency ratio (% of working age population) |
|  | Age dependency ratio, young (% o | Age dependency ratio, young (% of working age population) |
|  | Agricultural land (% of land are | Agricultural land (% of land area) |
|  | Birth rate, crude (per 1,000 peo | Birth rate, crude (per 1,000 people) |
|  | CPIA gender equality rating (1=l | CPIA gender equality rating (1=low to 6=high) |
|  | Central government debt, total ( | Central government debt, total (% of GDP) |
|  | Children with fever receiving an | Children with fever receiving antimalarial drugs (% of children under age 5 with fever) |
|  | Fertility rate, total (births pe | Fertility rate, total (births per woman) |
|  | GDP (constant 2005 US$) | GDP (constant 2005 US$) |
|  | GDP growth (annual %) | GDP growth (annual %) |
|  | GDP per capita, PPP (constant 20 | GDP per capita, PPP (constant 2005 international $) |
|  | GINI index | GINI index |
|  | Health expenditure per capita (c | Health expenditure per capita (current US$) |
|  | Health expenditure, public (% of | Health expenditure, public (% of GDP) |
|  | Health expenditure, public \_0001 | Health expenditure, public \_0001 (% of government expenditure) |
|  | Income share held by lowest 20% | Income share held by lowest 20% |
|  | Inflation, GDP deflator (annual | Inflation, GDP deflator (annual %) |
|  | Life expectancy at birth, female | Life expectancy at birth, female (years) |
|  | Life expectancy at birth, male ( | Life expectancy at birth, male (years) |
|  | Life expectancy at birth, total | Life expectancy at birth, total (years) |
|  | Mortality rate, infant (per 1,00 | Mortality rate, infant (per 1,000 live births) |
|  | Mortality rate, under-5 (per 1,0 | Mortality rate, under-5 (per 1,000 live births) |
|  | Mortality rate, under-5, female | Mortality rate, under-5, female (per 1,000) |
|  | Mortality rate, under-5, male (p | Mortality rate, under-5, male (per 1,000) |
|  | Population growth (annual %) | Population growth (annual %) |
|  | Immunization, DPT (% of children | Immunization, DPT (% of children ages 12-23 months) |
|  | Immunization, measles (% of chil | Immunization, measles (% of children ages 12-23 months) |
|  | Physicians (per 1,000 people) | Physicians (per 1,000 people) |
|  | Women's share of population ages | Women's share of population ages 15+ living with HIV (%) |

**Preliminary Data Wrangling**

We first remove the columns of Year, Year Code, Country Name and Country Code which are only for naming of each case of data.

We also noted that the variable "Age dependency ratio (% of working-age population)" includes people who are below 15 or above 64, while the variable "Age dependency ratio, young (% of working-age population)" only includes people below 15. To separate these two, we will subtract the later from the former. Then, the original "Age dependency ratio (% of working-age population)" is renamed to "Age dependency ratio (% of working-age population)".

**Missing Data**

When examining the data set, we find that it contains some obviously problematic data. For instance, some variables contain more than 90% missing data. Here, we remove variables with more than 5% missing data, and remove cases with more than 5 missing points. These cut offs are set objectively deeming that any variables or cases with missing data more than the cut off would undermine the usefulness. By now, there are 184 cases and 19 variables remain.

| No. | | Variables Labels | | Variables Meaning |
| --- | --- | --- | --- | --- |
|  | Adolescent fertility rate (birth | | Adolescent fertility rate (birth per 1000 women ages 15-19) | |
|  | Age dependency ratio, old | | Age dependency ratio, old (% of working age population) | |
|  | Age dependency ratio, young (% o | | Age dependency ratio, young (% of working age population) | |
|  | Agricultural land (% of land are | | Agricultural land (% of land area) | |
|  | Birth rate, crude (per 1,000 peo | | Birth rate, crude (per 1,000 people) | |
|  | Fertility rate, total (births pe | | Fertility rate, total (births per woman) | |
|  | Health expenditure per capita (c | | Health expenditure per capita (current US$) | |
|  | Health expenditure, public (% of | | Health expenditure, public (% of GDP) | |
|  | Health expenditure, public \_0001 | | Health expenditure, public \_0001 (% of government expenditure) | |
|  | Life expectancy at birth, female | | Life expectancy at birth, female (years) | |
|  | Life expectancy at birth, male ( | | Life expectancy at birth, male (years) | |
|  | Life expectancy at birth, total | | Life expectancy at birth, total (years) | |
|  | Mortality rate, infant (per 1,00 | | Mortality rate, infant (per 1,000 live births) | |
|  | Mortality rate, under-5 (per 1,0 | | Mortality rate, under-5 (per 1,000 live births) | |
|  | Mortality rate, under-5, female | | Mortality rate, under-5, female (per 1,000) | |
|  | Mortality rate, under-5, male (p | | Mortality rate, under-5, male (per 1,000) | |
|  | Population growth (annual %) | | Population growth (annual %) | |
|  | Immunization, DPT (% of children | | Immunization, DPT DPT (% of children ages 12-23 months) | |
|  | Immunization, measles (% of chil | | Immunization, measles (% of children ages 12-23 months) | |

For the remaining missing data, we use Multiple Imputation by Chained Equations (MICE) to impute them assuming that the missing data are missing at random. In gist, MICE operation imputes missing data through a series of regression models on the basis that each missing data is conditional upon the other observed variables in the data.[[3]](#footnote-3)

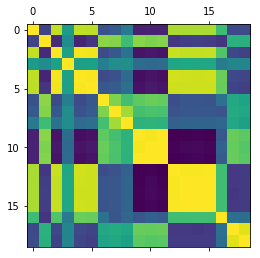
We have considered single imputation, as such mean imputation, as an alternative. However, the subsequent analysis will be biased as since mean value would be treated as true data rather than imputed data, reducing the reliability of conclusion. Maximum likelihood estimate was also considered but rejected, due to its limitation of application to longitudinal or structured equations models.

There are merits in using MICE. First, the imputation of missing data takes into account given relations in the observed data by including other observed data in the imputation model (Schafer and Graham, 2002). It is also flexible and yielding accurate standard errors (Greenland and Frankle, 1995).

The imputed data set after using MICE is enclosed as **Annex [1]**.

**Correlation Heatmap**

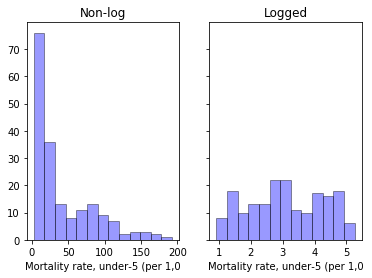
A correlation heatmap (*Figure.1*) is produced for the 19 variables based on the data set with missing data imputed using MICE. Based on the visualisation, it seems that there exist multicollinearity between variables. We will examine this in the later sections.



*(Figure 1, Correlation Heatmap)*

**Plotting distribution of numeric variables**

We also evaluate the distribution of the dependent variable “Mortality rate, under-5 (per 1,000 live births)” as shown in *Figure 2*. Since the dependent variable is a count data, the non-log distribution seems follows poison distribution. As the logged distribution shows improvement, this gives an indication for transformation which will be discussed in the later section.



*(Figure 2: Distribution plots for “Mortality rate, under-5 (per 1,000 live births)”)*

1. United Nations Sustainable Development Goals 2030 (<https://www.un.org/sustainabledevelopment/>) [↑](#footnote-ref-1)
2. Levels and Trends in Child Mortality, Report 2017, by the UN Inter-agency Group for Child Mortality Estimation (<https://www.unicef.org/publications/files/Child_Mortality_Report_2017.pdf>) [↑](#footnote-ref-2)
3. “Multiple Imputation by Chained Equations: What is it and how does it work?”, Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf, March 2011 [↑](#footnote-ref-3)