Collaborative Research Project - Assignment 3

24 October 2014

Contents

1. Data Gathering and Cleaning

This section focuses on the process of gathering the data and cleaning the databases to prepare the variables for the data analysis.

The first step in this process was uploading the databases to R Studio. The first dataset consists of 29 World Development Indicators and it was downloaded from World Bank's website. These indicators represent the independent variables used for this research plus the population indicator that is used to filter small countries. Provided that the focus of this research is on country level data, all regional data was dropped. Further, 169 rows that contained only NA values were deleted.

After dropping empty rows, the data frame was alphabetically (ascending) ordered, rows were grouped by iso2c code and variables were renamed.

The dataset was further cleaned preparing the data for imputation using the AMELIA package. The imputation will be conducted however at a further stage of the research. This process requires that the panel is as balanced as possible, as it feeds from all variables to predict values for the missing observations. The next step was thus dropping variables for which more than 80% of the observations (552) were missing. In addition, countries with a population smaller than one million inhabitants were dropped from the database. 59 countries fell in that category: 46 islands, 5 European countries (Andorra, Liechtenstein, Luxemburg, Monaco and Montenegro), Bahrain, Bhutan, Belize, Djibouti, Equatorial Guinea, Guyana, Qatar and Suriname. Dropping these countries does not affect the research as the remaining database still contains a highly heterogeneous sample both in geographic and socio-economic terms. Furthermore, deleting these countries improves the dataset as most of these countries lacked information for most of the studied variables.

The second database used for this research was downloaded from UNAIDS' website and it provides information on HIV/AIDS incidence rates (as well as prevalence and deaths caused by HIV/AIDS). The data is publicly available. All columns except the country and the incidence rate were dropped. After renaming the variables, a unique identifier was created and missing values were recoded as NAs. Moreover, some observations in the database were not specific numbers; instead, it was indicated that for that year, prevalence was below a certain threshold (0.01%). In those cases, these observations were replaced by 0.009. The final step in the cleaning of the UNAIDS database consisted of deleting missing values for the dependent variable and deleting the regions with an iso2c equal to a country's iso2c (NA and ZA) to avoid problems in the merging process.

Once both databases were cleaned, the next step was to merge the datasets using the combination of iso2c and year as unique identifier. In the merging process, only observations that were present in both datasets were kept. It is worth noticing that UNAIDS' dataset included observations from 1990 to 2012 so all observation between 1990 and 1999 were dropped. Finally, unnecessary columns from the new database were eliminated.

2. Descriptive Statistics

The descriptive statistics part consists of the preparation of the variables for data analysis. Tables, plots and histograms are shown to understand the distribution of the variables. Table 1 shows the main descriptive statistics of all the variables that can be found in the cleaned dataset (number of observations, mean, standard deviation, and min and max values).

```
##
## Table 1: Descriptive statistics
## ================================================================================
## Statistic     N        Mean           St. Dev.           Min              Max
## --------------------------------------------------------------------------------
## X            988      494.5             285.4            1               988
## year         988     2,006.0             3.7           2,000            2,012
## GDP          950 62,121,352,248.0 161,136,361,501.0 482,341,824.0 1,389,049,411,478.0
## GDPpc        945     5,295.9           5,152.7          441.2          30,874.9
## Rural        988       57.7              19.5            8.7              91.8
## CO2          829       1.7                3.5           0.004            38.2
## HCexpend     962       5.9                2.2            1.8             18.4
## Water        982       75.5              17.9           23.5             99.8
## Sanitation   983       50.4              30.1            6.6             98.9
## Unemploym    988       8.7                6.2            0.6             38.7
## Primary      836      101.5              18.8           29.2             164.9
## HCexpendpc   962      124.3             165.4            2.4            1,103.4
## FemUnempl    988       10.2               8.1            0.3             53.7
## FemSchool    828       98.1              20.8           20.8             162.4
## LifeExpect   988       61.4               9.9           38.1             79.6
## DPT          988       79.9              18.0            19              99
## Measles      988       79.4              17.5            16              99
## Population   988   40,890,589.0     133,140,255.0     1,063,715      1,236,686,732
## Incidence    988       0.3                0.7           0.01             4.4
## Incidence2   912       0.4                0.7           0.01             4.4
## IncidenceDif 912      -0.02               0.1           -0.5             0.1
## DDif         912       0.9                0.3             0               1
## --------------------------------------------------------------------------------
```

The histogram of the dependent variable (Figure 1) shows that the incidence rates are not normally distributed but strongly skewed to the left and only few incidence rates are higher than 1.

Figure 1: Incidence Rate

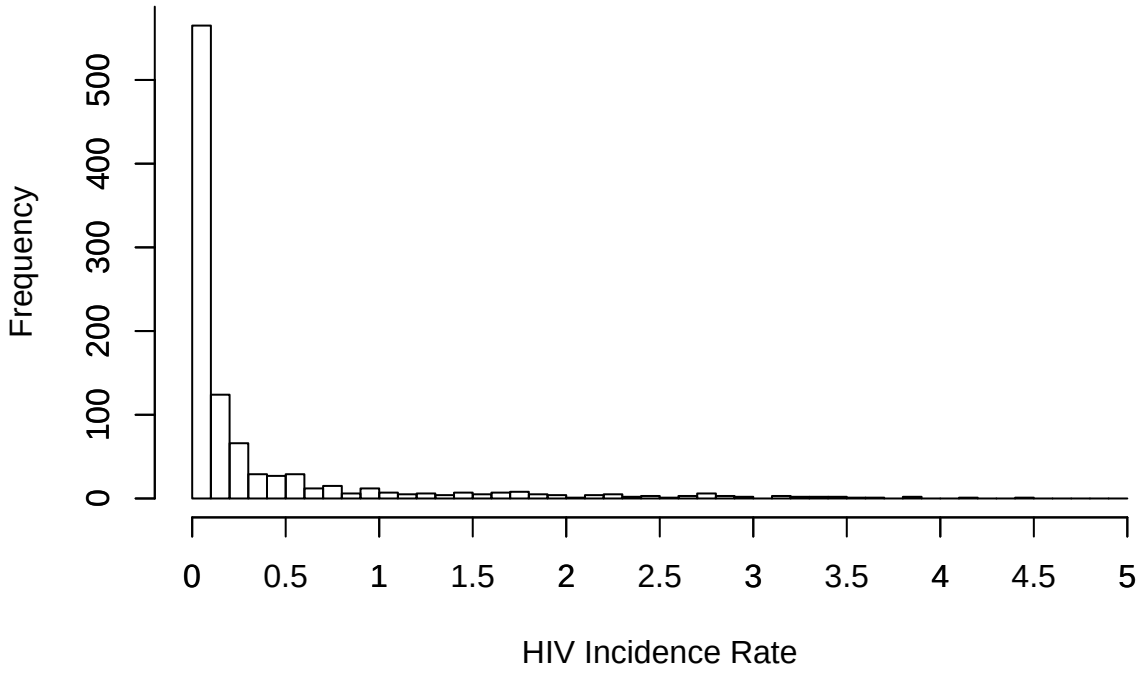


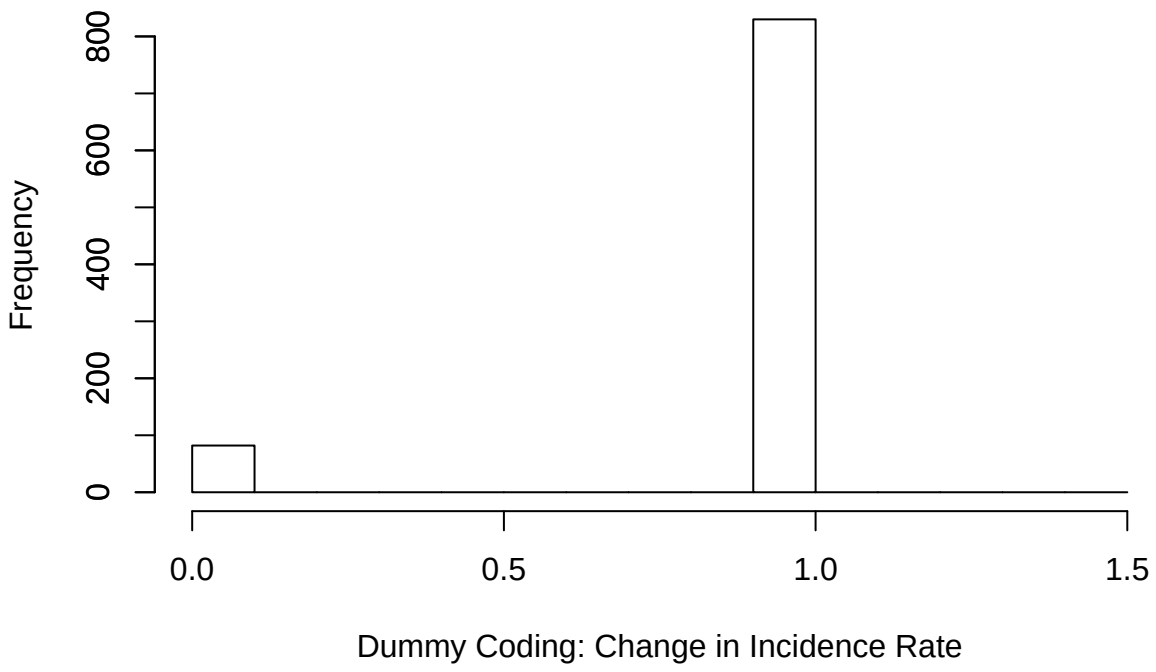
Figure 2 shows that in most countries of our dataset HIV/AIDS incidence rates decreased between the period of 2000 to 2015 (see Figure 2).

When plotting the incidence rates per country (Figure 3 in repository) the range of observations per country is shown and the outliers (countries with high incidence rates) can be identified.

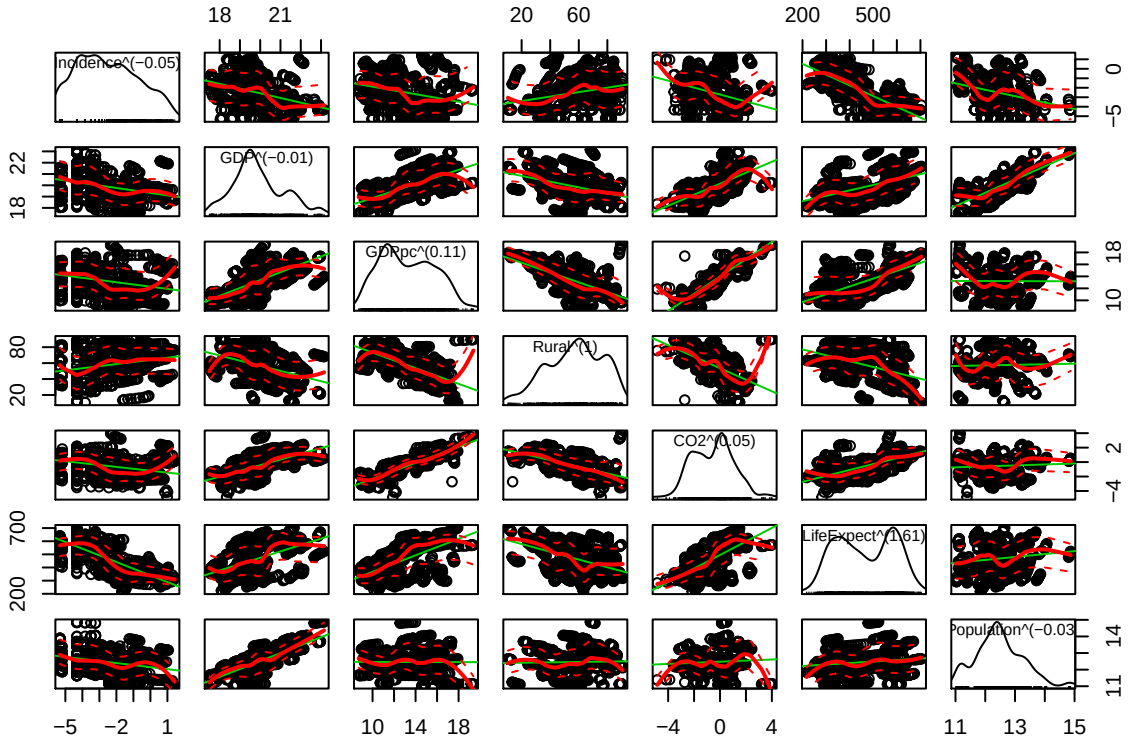
As the research question investigates why MDG 6.A is not being reached by some countries, the general HIV incidence rate is interesting, but also the decrease in the incidence rate from 2000 to 2015 is even more relevant. As stated in the research proposal Target 6.A of the MDGs specifies that countries should "have halted by 2015 and begun to reverse the spread of HIV/AIDS" [United Nations (2014)]. For this purpose, the dependent variable was lagged by one period and the difference between the lag and the current year was calculated (see Figure 4 in repository).

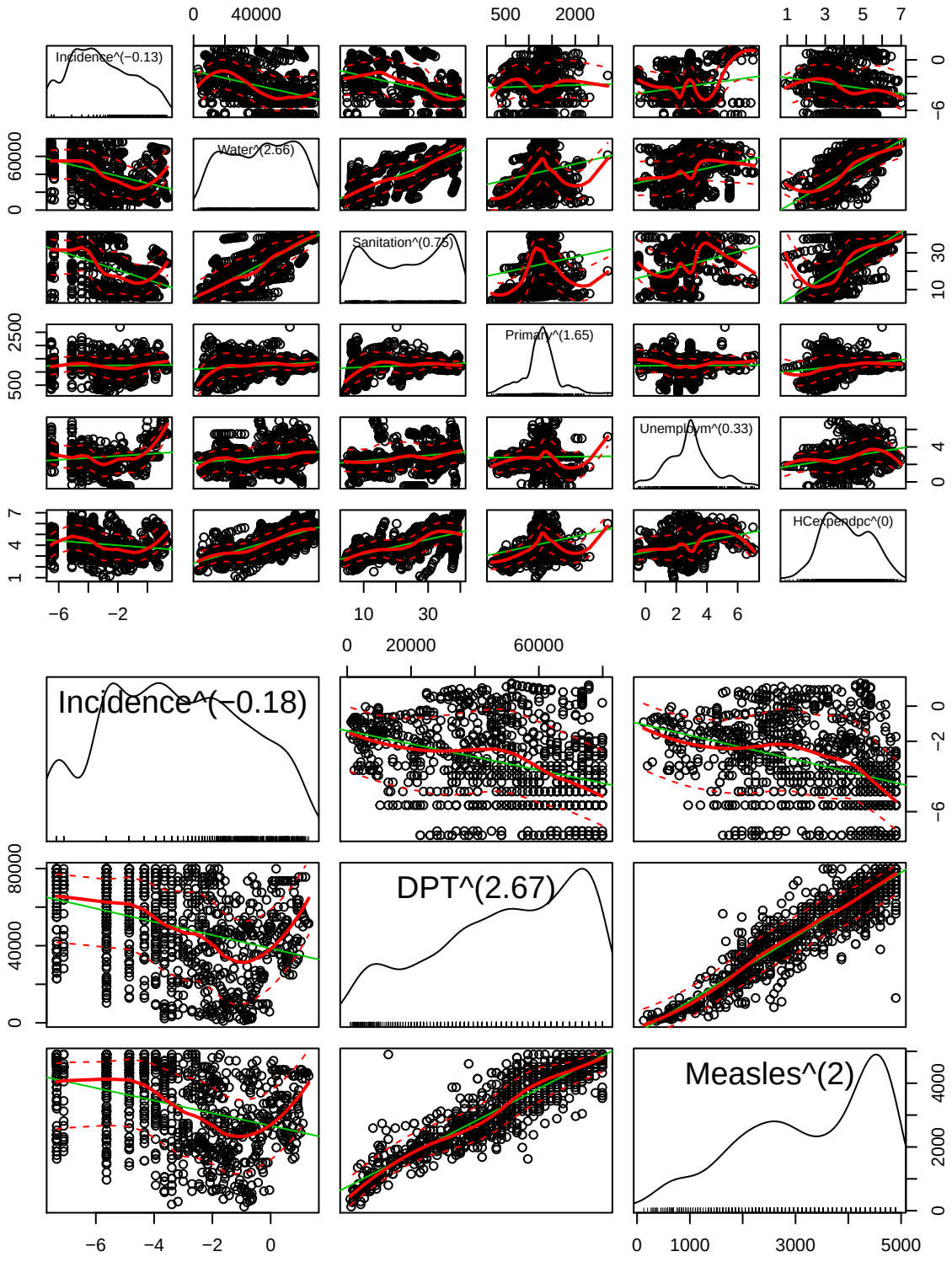
Further, a dummy variable was created assigning a value of zero for those observations where the incidence rate decreased compared to the previous year or stayed the same (countries reaching MDG 6.A) and a value of one was assigned to those observations where the incidence rate increased (countries not reaching MDG 6.A). Figure 4 shows the direction of the change in the incidence rate compared to the previous year by country.

Figure 5: Dummy Variable



Scatterplots were used for each category of Dahlgrehn's model in order to see whether the variables are skewed or multicollinear (Figures 5, 6 & 7).




Incidence^(−0.13)
Water^(2.66)
Sanitation^(0.75)
Primary^(1.65)
Unemploym^(0.33)
HCexpendpc^(0)
Incidence^(−0.18)
DPT^(2.67)
Measles^(2)

3. Inferential Statistics

As can be seen from the Scatterplots in the descriptive statistics most of the variables are not normally distributed. Further, the variables all have different scales. Therefore, the independent variables were logged for enabling comparisons in the data analysis part.

As the dependent variable is coded as a dummy, being 1 if MDG 6.A is reached and 0 for countries that are not reaching MDG 6.A logistic regressions are used for the inferential statistics part. As Odds and Odds ratios are difficult to present to a broad audience predicted probabilities are calculated after running the logistic regressions. The interpretation of the results will mainly focus on the predicted probabilities.

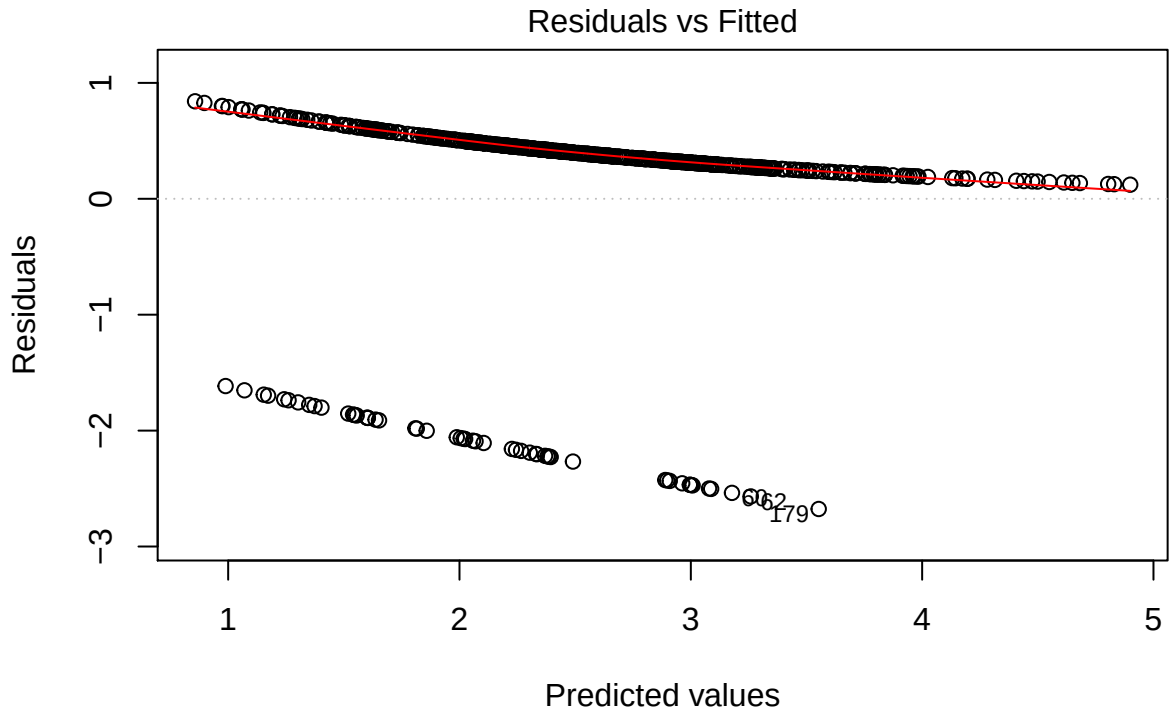
The test for variance inflation factors showed that in our first logistic regression model six variables showed high multicollinearity and had a heigher variance inflation than the threshold of 10. We tested the multicollinearity between the variables and found that there was high multicollinearity between the GDP and GDP per capital, Unemployment and Female unemployment, Primary education and female schooling. Therefore, we excluded one of these multicollinear variables for each group based on their explanatory strength for our research question, namely unemployment, primary education and GDP.

```
##        lGDP      lGDPpc      lRural        lCO2   lHCexpend      lWater
##    2.324163   23.743894    2.208177    7.068360    2.642353    3.346136
## lSanitation  lUnemploym    lPrimary lHCexpendpc     Inverse  lFemSchool
##    3.554170   26.087093   44.287369   12.980265   24.852079   47.902670
## lLifeExpect        lDPT    lMeasles
##    3.132095    7.817427    8.464431
```

We tested the variance inflation factors of the new logistic regression model and all remaining variables passed the test.

```
##        lGDP      lRural        lCO2   lHCexpend      lWater lSanitation
##    1.956140    1.971864    3.384596    1.230200    3.017148    3.307859
##  lFemUnempl  lFemSchool lLifeExpect        lDPT    lMeasles
##    1.368089    1.380339    2.509395    8.095830    9.059972

##      lGDPpc      lRural        lCO2   lHCexpend      lWater lSanitation
##    6.802097    1.860802    6.493329    1.217211    3.118498    3.135771
##  lFemUnempl  lFemSchool lLifeExpect        lDPT    lMeasles
##    1.386350    1.370772    2.479665    7.807571    8.507410
```



Residual vs. Fitted Plot

Confidence Intervals

```
## Waiting for profiling to be done...
```

```
##                    2.5 %     97.5 %
## (Intercept) -14.2519359 15.9336922
## lGDPpc       -0.4288777  1.1120998
## lRural       -1.4501349  0.6481201
## lCO2         -0.4791310  0.4184275
## lHCexpend    -1.1496653  0.8892495
## lWater       -0.9627660  3.1179108
## lSanitation  -1.8687353 -0.3378941
## lFemUnempl   -0.4515607  0.2978412
## lFemSchool   -2.6601865  0.6642645
## lLifeExpect  -1.9389062  3.4995207
## lDPT         -0.9412698  5.2886166
## lMeasles     -5.5504639  1.3556460
```

LOGISTIC REGRESSION

The rest of this section focuses on logistic regressions using only those variables where no clear multicollinearity was observed. The results of the logistic regressions were exported with Stargazer. Table 2 shows the logistic regression results. The results show that out of all the tested variables in the model only the variable Sanitation has a significant impact on changes in HIV/Aids incidence rates.

```
##
## Table 2: Logistic Regression Results
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                     DDif
##                           (1)       (2)       (3)
## ------------------------------------------------
## lGDPpc                    0.328     0.216     0.249
##                          (0.411)   (0.417)   (0.424)
##
## lRural                   -0.406    -0.218    -0.265
##                          (0.548)   (0.530)   (0.543)
##
## lCO2                     -0.114    -0.051     0.022
##                          (0.239)   (0.233)   (0.231)
##
## lHCexpend                -0.076    -0.109    -0.111
##                          (0.524)   (0.530)   (0.541)
##
## lWater                    0.930     1.316     1.084
##                          (1.024)   (1.041)   (1.038)
##
## lSanitation             -1.055*** -1.113*** -1.150***
##                          (0.406)   (0.400)   (0.408)
##
## Inverse                   0.128               0.363
##                          (0.193)             (0.385)
##
## as.factor(QFemSchool)2   -0.127
##                          (0.524)
##
## as.factor(QFemSchool)3    0.502
##                          (0.601)
##
## as.factor(QFemSchool)4   -0.764
##                          (0.482)
##
## as.factor(QInverse)2               -0.523
##                                    (0.458)
##
## as.factor(QInverse)3                0.187
##                                    (0.526)
##
## as.factor(QInverse)4               -0.504
##                                    (0.500)
##
## lFemSchool                         -1.022    -0.940
```

```
##                                     (0.831)   (0.848)
##
## lLifeExpect                 1.363     1.718     1.309
##                            (1.429)   (1.424)   (1.416)
##
## lDPT                        1.929     2.102     2.312
##                            (1.602)   (1.569)   (1.601)
##
## lMeasles                   -1.858    -1.783    -1.686
##                            (1.798)   (1.764)   (1.771)
##
## as.factor(QInteraction)2                       -0.631
##                                                (0.522)
##
## as.factor(QInteraction)3                        0.333
##                                                (0.753)
##
## as.factor(QInteraction)4                       -1.080
##                                                (0.958)
##
## Constant                   -4.129    -3.363    -1.387
##                            (8.019)   (7.912)   (8.322)
##
## ------------------------------------------------------
## Observations                 613       613       613
## Log Likelihood            -171.200  -172.826  -170.898
## Akaike Inf. Crit.          370.400   373.652   371.796
## ======================================================
## Note:                       *p<0.1; **p<0.05; ***p<0.01
```

PREDICTED PROBABILITIES

All three predicted probability models fix the values of all independent variables except one to the mean values of Uganda, an interesting case given the failure of this country to contain the spread of HIV/AIDS. The first regression looks at the impact of female school enrolment on the dependent variable.

As it can be observed, the probability of successfully halting or reversing the spread of HIV/AIDS does not change substantially when female school enrolment increases. In fact, the probability of being successful at combating the disease falls between the first and the second quintile and between the third and the forth.

```
##
## Predicted Probabilities Model 1
## ======================
##   1     2     3     4
## ----------------------
## 0.902 0.890 0.938 0.810
## ----------------------
```

The second regression (Predicted Probabilities Model 2) focuses on the single impact of female unemployment in halting and reversing the spread of HIV/AIDS. The predicted probabilities of this model show a similar pattern as in the first model, although the discrepancy in the probabilities of being successfully are larger in this model.

```
##
```

```
## Predicted Probabilities Model 2
## ========================
##   1     2     3     4
## ------------------------
## 0.872 0.802 0.892 0.805
## ------------------------
```

Finally, the third model (Predicted Probabilities Model 3) integrates the interaction between both variables. As with model 2 and 3, there is a discontinuous reduction in the probability of successfully halting and reversing the spread of HIV/AIDS and the probability decreases between quintile one and two and between quintile three and four.

```
##
## Predicted Probabilities Model 3
## ========================
##   1     2     3     4
## ------------------------
## 0.903 0.832 0.929 0.760
## ------------------------
```

All three models show inconsistent results with the predictions of this paper. The disproportionate number of cases of successful halt and reverse of HIV/AIDS in the sample might be responsible for the inconsistency of the results.

4. Limitations

The paper had to make some compromises regarding its original aim as outlined in the first research proposal. Due to the significant amount of missing values and the presence of multicollinearity, a considerable number of variables had to be dropped and could ultimately not be integrated in the logistic regression models.

The selection of these variables was not arbitrary but followed instead the theoretical framework guiding this research, i.e. Dahlgren's model. Two levels of Dahlgren's model (Social and Community Networks and Individual Lifestyle Factors) ended up underrepresented after dropping these variables. To deal with this limitation, the research will only use the theoretical framework as an instrument to guide the selection of variables but will not utilise the findings to test the validity of the model.

In terms of the data used to run the regressions, the relative high number of countries that have already halted or reversed the spread of HIV/AIDS in our sample can lead to biased results. In the next stage of the research, the effect of excluding those countries that only halted the spread will be explored.

Another shortcoming faced at this stage was the integration of figures from the descriptive statistics into the final report. A transitory solution was to save those pictures in a subfolder of the repository.