Collaborative Research Project - Assignment 3

14 November 2014

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).

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

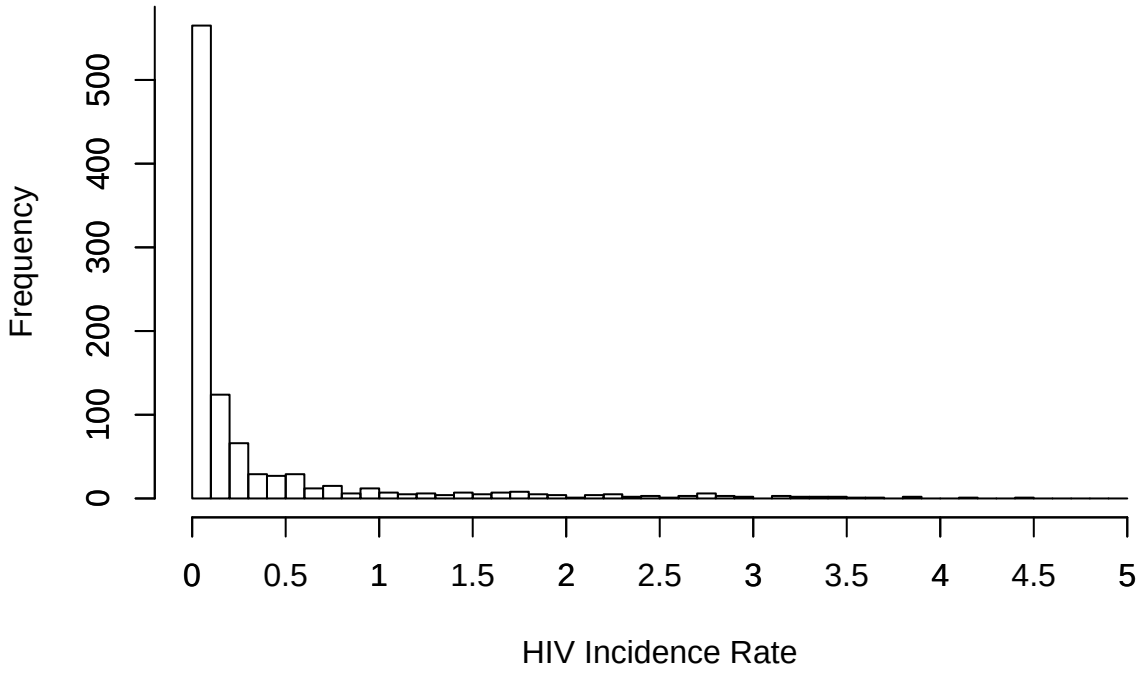
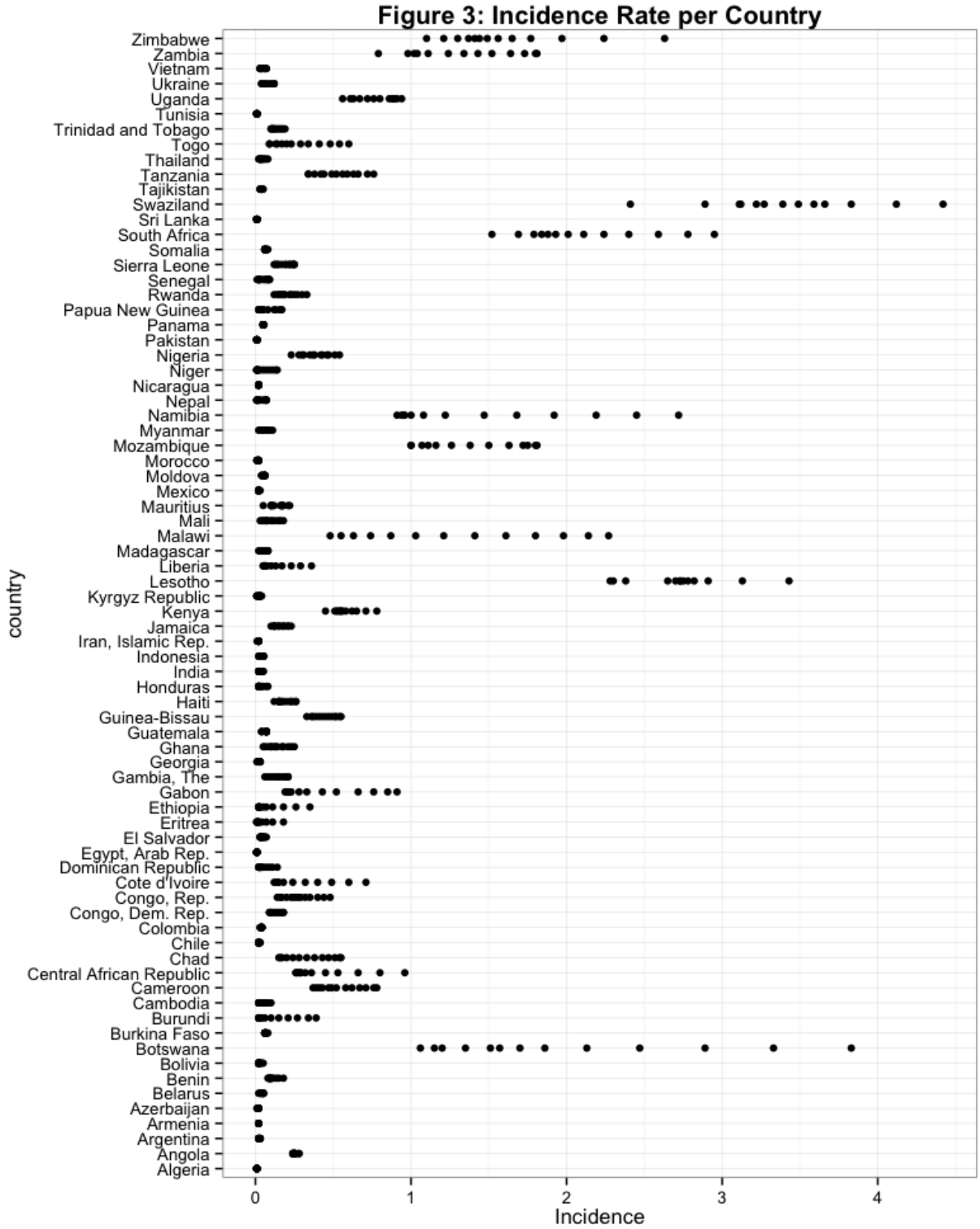


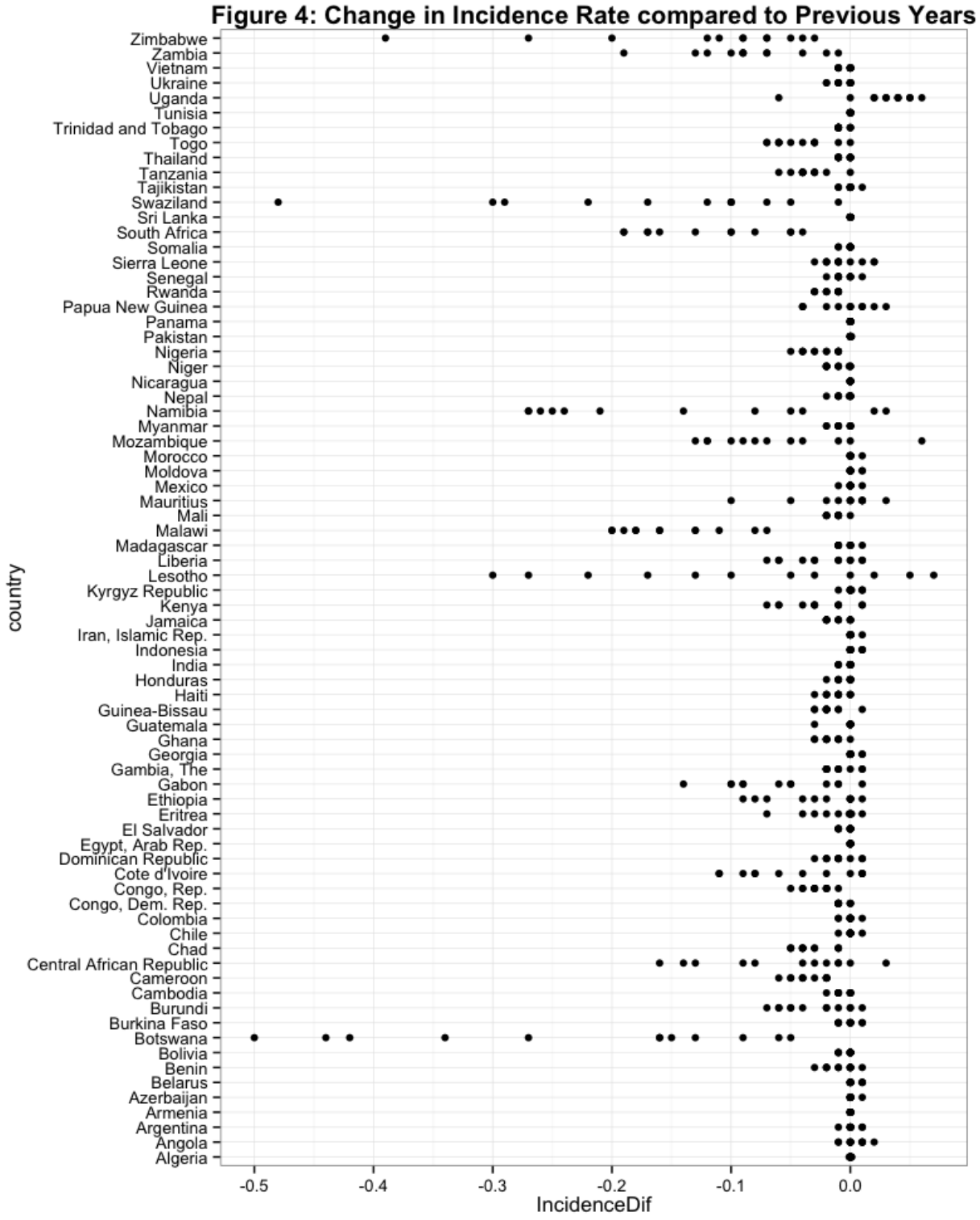
Figure 2 shows that in most countries of our dataset HIV/AIDS incidence rates decreased between the period of 2000 to 2015 (see Figure 2).

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.



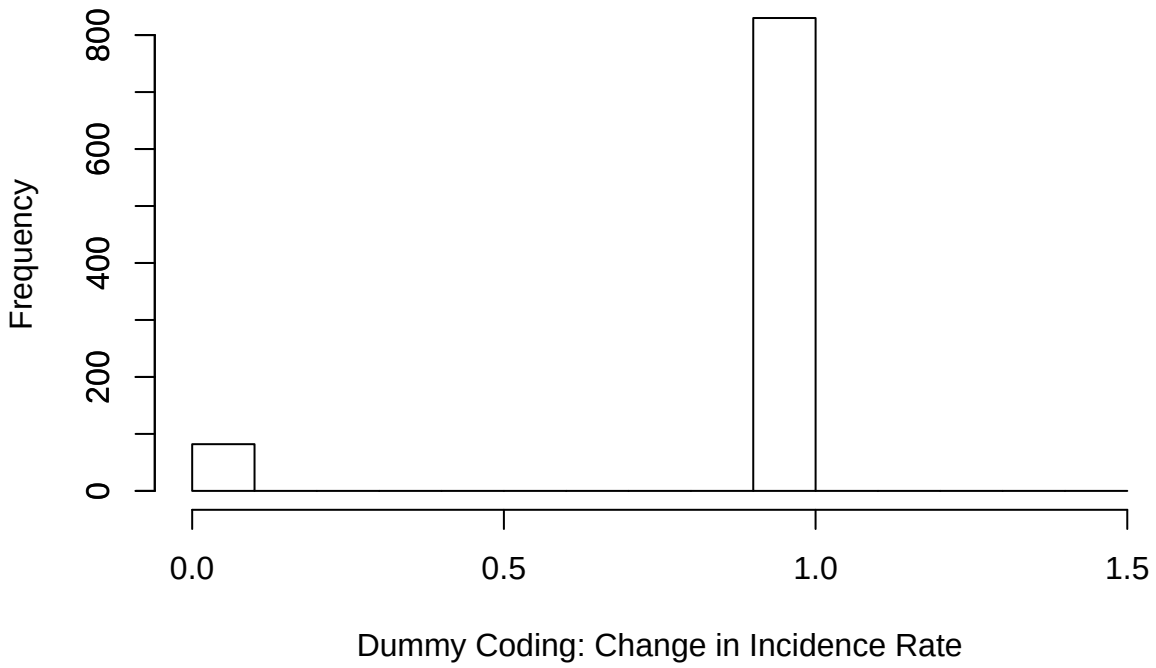
As the research question investigates why MDG 6.A is not being reached by some countries, the general HIV incidence rate is interesting, but also the decrease in the incidence rate from 2000 to 2015 is even more relevant. As stated in the research proposal Target 6.A of the MDGs specifies that countries 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).

3 Case Studies - Botswana, Lesotho, Uganda & Malawi

4 Inferential Statistics

```
## 'data.frame':    988 obs. of  24 variables:
##  $ X           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ iso2c       : Factor w/ 75 levels "AM","AO","AR",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year        : int  2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 ...
##  $ country     : Factor w/ 76 levels "Algeria","Angola",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ GDP         : num  2.75e+09 3.02e+09 3.42e+09 3.90e+09 4.30e+09 ...
##  $ GDPpc       : num  2919 3214 3654 4182 4635 ...
##  $ Rural       : num  35.3 35.6 35.7 35.8 35.8 ...
##  $ CO2         : num  1.127 1.158 0.999 1.129 1.205 ...
##  $ HCexpend    : num  6.25 5.94 5.4 5.56 5.5 ...
##  $ Water       : num  92.6 93.1 93.7 94.3 94.9 95.5 96.1 96.7 97.3 98 ...
##  $ Sanitation  : num  88.9 89 89.2 89.3 89.4 89.6 89.7 89.8 90 90.1 ...
##  $ Unemploym   : num  24.7 35.9 27.8 28.6 32.3 ...
##  $ Primary     : num  98.5 102.2 95.6 94.3 94.6 ...
##  $ HCexpendpc  : num  38.9 41.1 42.1 51.4 65 ...
```

```
##  $ FemUnempl   : num  29 40.1 33.8 35.2 36.3 ...
##  $ FemSchool   : num  98.3 103.1 96.5 96 97.4 ...
##  $ LifeExpect  : num  71.3 71.8 72.2 72.6 73 ...
##  $ DPT         : int  93 94 94 94 91 90 87 88 89 93 ...
##  $ Measles     : int  92 93 91 94 92 94 92 92 94 96 ...
##  $ Population  : num  3076098 3059960 3047002 3036032 3025652 ...
##  $ Incidence   : num  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 ...
##  $ Incidence2  : num  NA 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 ...
##  $ IncidenceDif: num  NA 0 0 0 0 0 0 0 0 0 ...
##  $ DDif        : num  NA 1 1 1 1 1 1 1 1 1 ...
```

Table 1: Variance Inflation Factors 1

	VIF
GDP	3.09
GDPpc	11.78
Rural	2.49
CO2	3.20
HCexpend	1.61
Primary	47.65
Water	3.88
Sanitation	4.41
Unemploym	12.07
HCexpendpc	7.00
FemUnempl	11.52
FemSchool	53.08
LifeExpect	3.54
DPT	8.82
Measles	9.17
Population	2.93

install.packages(“knitr”) library(knitr)

```
##      GDPpc      Rural        CO2   HCexpend      Water Sanitation
##     11.650      2.346      3.135      1.594      3.815      4.394
##  Unemploym HCexpendpc  FemUnempl  FemSchool LifeExpect        DPT
##     11.123      6.551     11.032      1.370      3.384      8.664
##    Measles Population
##      9.009      1.190
```

Table 2: Variance Inflation Factors 2

	VIF
GDPpc	11.65
Rural	2.35
CO2	3.14
HCexpend	1.59
Water	3.81
Sanitation	4.39
Unemploym	11.12
HCexpendpc	6.55

	VIF
FemUnempl	11.03
FemSchool	1.37
LifeExpect	3.38
DPT	8.66
Measles	9.01
Population	1.19

As can be seen from the Scatterplots (see appendix) 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.

How to cite this model in Zelig: Kosuke Imai, Gary King, and Olivia Lau. 2014. "logit: Logistic Regression for Dichotomous Dependent Variables" in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software," http://gking.harvard.edu/zelig

Model: logit Number of multiply imputed data sets: 5

Combined results:

Call: glm(formula = formula, weights = w, family = binomial(link = "logit"), model = F, data = data)

Coefficients: Value Std. Error t-stat p-value (Intercept) -39.2895 6.9307 -5.6689 1.545e-08 lGDPpc 0.4144 0.3471 1.1938 2.333e-01 lRural -2.5682 0.5490 -4.6776 3.183e-06 lCO2 -0.6242 0.2031 -3.0734 2.269e-03 lHCexpend 0.9005 0.4390 2.0511 4.468e-02 lWater -2.3599 0.8513 -2.7720 5.659e-03 lSanitation 0.8936 0.2807 3.1830 1.459e-03 lLifeExpect 19.5374 1.6989 11.5003 1.702e-30 lDPT -0.6119 1.0441 -0.5861 5.581e-01 lMeasles 1.6430 1.2264 1.3397 1.821e-01 Inverse 1.8616 0.2588 7.1946 6.305e-13 lFemSchool -5.8963 0.7479 -7.8838 1.846e-11

For combined results from datasets i to j, use summary(x, subset = i:j). For separate results, use print(summary(x), subset = i:j).

How to cite this model in Zelig: Kosuke Imai, Gary King, and Olivia Lau. 2014. "logit: Logistic Regression for Dichotomous Dependent Variables" in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software," http://gking.harvard.edu/zelig

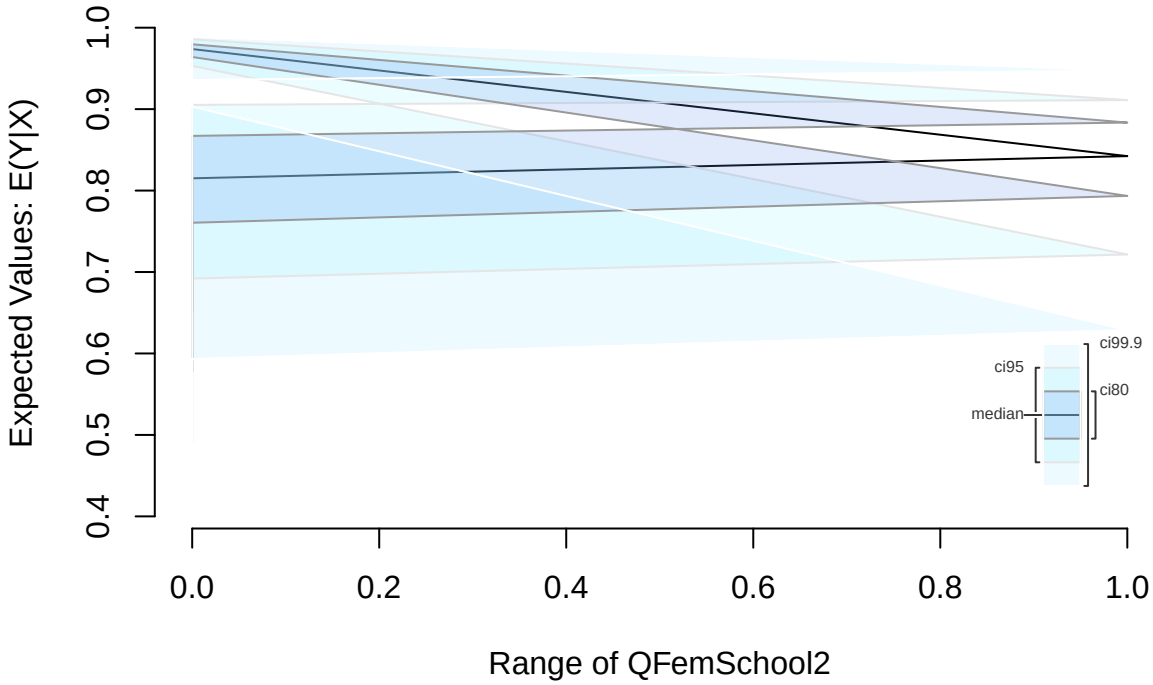
Model: logit Number of multiply imputed data sets: 5

Combined results:

Call: glm(formula = formula, weights = w, family = binomial(link = "logit"), model = F, data = data)

Coefficients: Value Std. Error t-stat p-value (Intercept) -59.8572 7.1530 -8.3681 6.123e-17 lGDPpc 0.4011 0.3480 1.1528 2.496e-01 lRural -2.1136 0.5050 -4.1851 2.984e-05 lCO2 -0.5999 0.2091 -2.8688 4.585e-03 lHCexpend 0.8629 0.3898 2.2134 2.778e-02 lWater -3.1306 0.7885 -3.9703 7.227e-05 lSanitation 0.9771 0.2769 3.5290 4.240e-04 lLifeExpect 19.1798 1.6264 11.7925 2.181e-31 lDPT -1.0271 1.0329 -0.9944 3.206e-01 lMeasles 1.5999 1.1417 1.4014 1.616e-01 Inverse 1.8117 0.2547 7.1126 1.265e-12 QFemSchool2 -2.0895 0.4408 -4.7397 8.934e-06 QFemSchool3 -2.0315 0.4658 -4.3614 8.637e-05 QFemSchool4 -3.1151 0.4315 -7.2197 1.623e-11

For combined results from datasets i to j, use summary(x, subset = i:j). For separate results, use print(summary(x), subset = i:j).



The test for variance inflation factors showed that in our first logistic regression model six variables showed high multicollinearity and had a higher 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.

5 Preparing the dataset for Model 2

6 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.

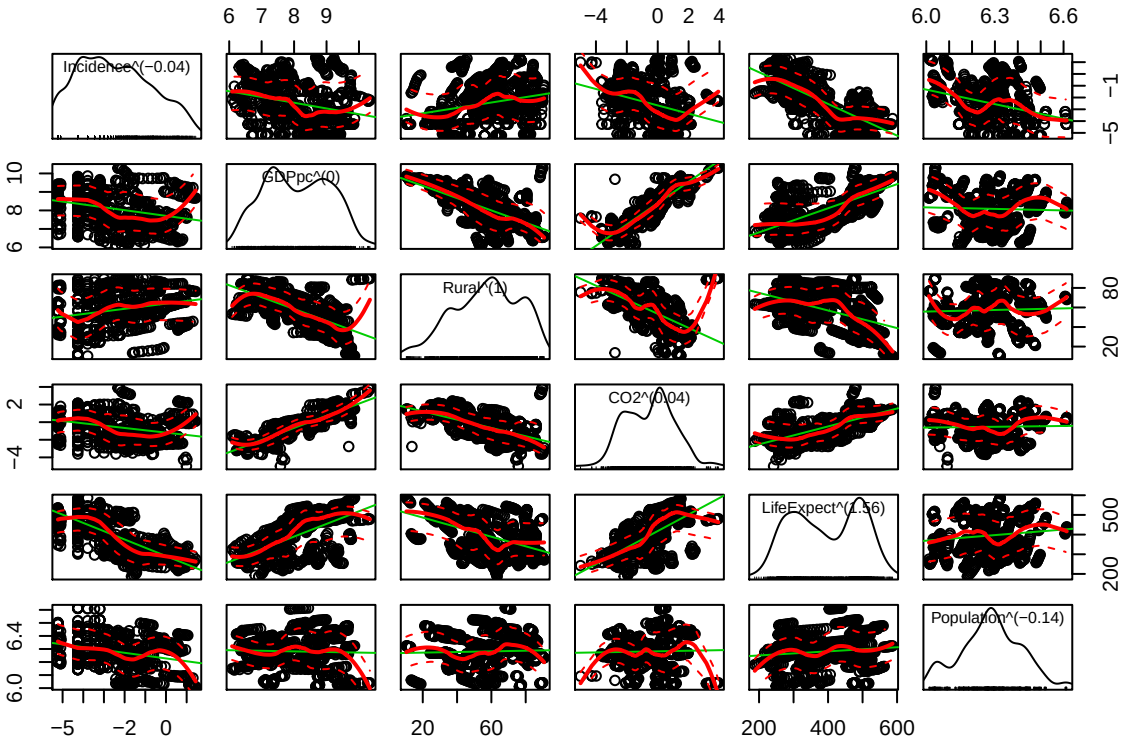
The selection of these variables was not arbitrary but followed instead the theoretical framework guiding this research, i.e. Dahlgren's model. Two levels of Dahlgren's model (Social and Community Networks and Individual Lifestyle Factors) ended up underrepresented after dropping these variables. To 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.

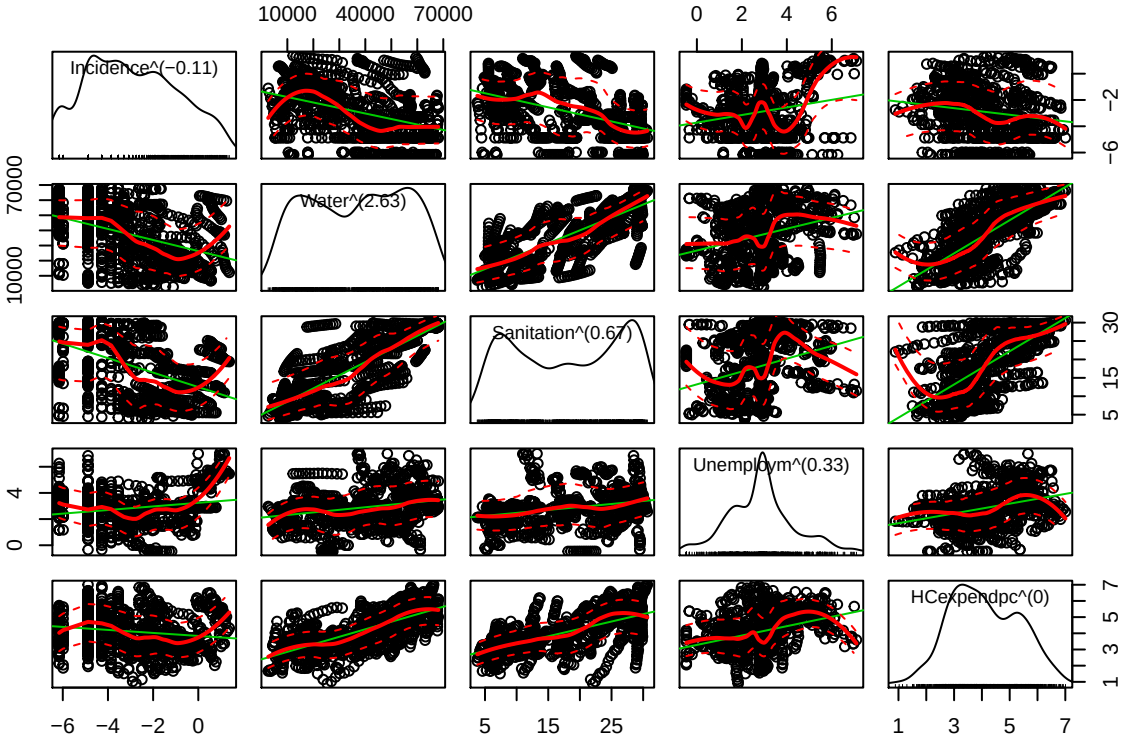
Another shortcoming faced at this stage was the integration of figures from the descriptive statistics into the final report. A transitory solution was to save those pictures in a subfolder of the repository.

7 Appendix

Scatterplot of variables for socio-economic, cultural and environmental conditions



Scatterplot of variables for living and working conditions



Scatterplot of variables for individual lifestyle factors