Healthy communities - Understanding determinants of HIV

12 December 2014

Contents

1 Executive Summary

2 Introduction

The expiry date of the Millennium Development Goals (MDGs) is just around the corner, meanwhile the post-2015 agenda is being discussed intensively. In this context, it is important to assess the achievement of the MDGs and try to understand why some goals have not been reached.

Reducing HIV incidence is an important aim of the MDGs. 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). In most regions of the world this goal has been fulfilled: new HIV infections declined and the overall number of new HIV/AIDS infections per 100 adults (15-49 years old) decreased by 44 per cent between 2001 and 2012 (United Nations 2014). However, this trend cannot be observed in all 189 member states of the United Nations. On the contrary, HIV/AIDS prevalence has even increased in some countries.

---

3 Aim, Research Question and Hypotheses

This paper aims to provide evidence to assess why some countries struggle to achieve MDG 6A. We believe that one possible explanation for the failure of some interventions in reducing HIV/AIDS may lie in the lack of a full understanding of the determinants of the disease, which can in turn lead to ill-specified interventions and wrongly targeted campaigns.

The literature reviewed for this paper identifies a myriad of determinants of health. When it comes to specific diseases however, the existing literature only provides a narrow selection of potential determinants. The aim of this paper is to test whether social and community networks (often neglected by the literature on disease-specific determinants of health) significantly explain HIV/AIDS incidence rates at the country level. The hypothesis of this paper is that female school enrollment is a strong predictor of HIV incidence rates. By identifying variables that help to explain HIV/AIDS incidence, this paper will help move forward the discussion of the determinants of HIV/AIDS .

---

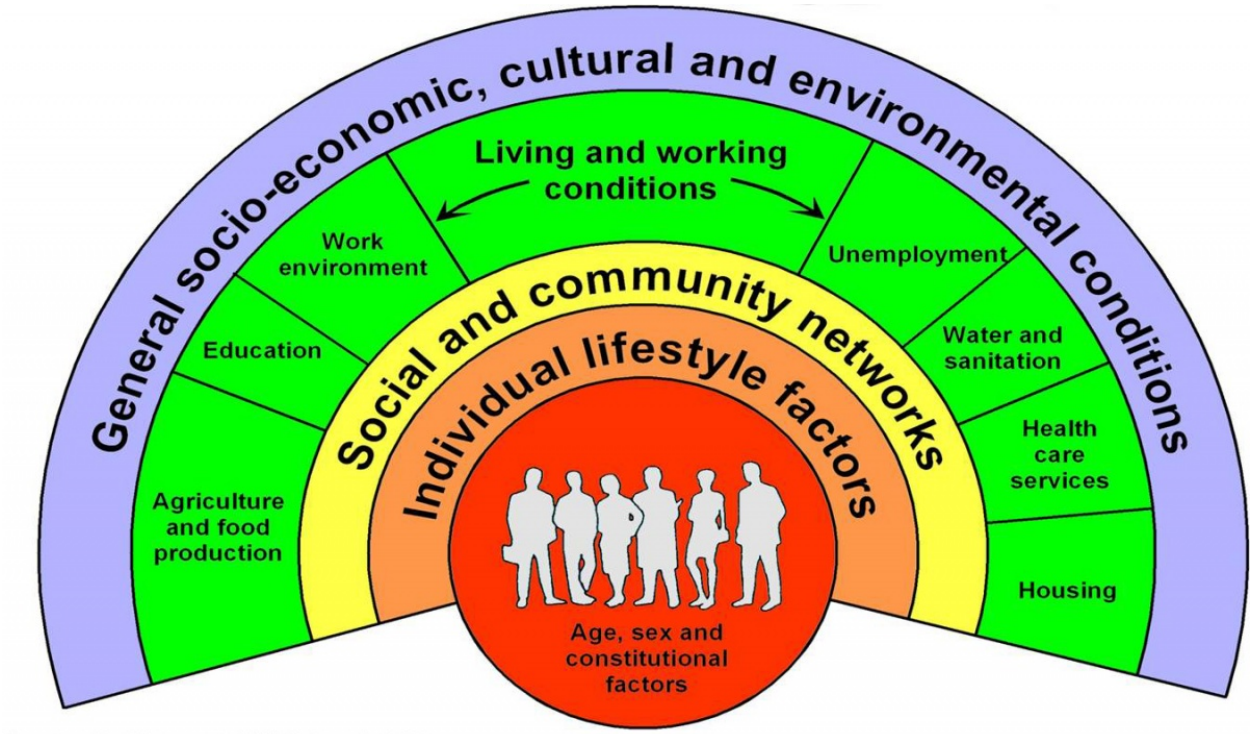
4 Literature Review

Hurrelmann (Hurrelmann 1989, 76) advocates an interdisciplinary framework for analysing what determines health outcomes. He considers it necessary to use a model that integrates all the aspects of the organism, individual and the environment.

One framework that shows the interaction between individual and environmental factors over time is the salutogenic model developed by Antonovsky in 1979. According to Hurrelmann, Antonovsky's model is a great contribution to interdisciplinary theory, but the downside is its complexity (Hurrelmann 1989).

A simpler and more common model on general determinants of health is the "rainbow model", developed by Dahlgren and Whitehead (Dahlgren and Whitehead 1991, 11). This model gives an overview of the main health determinants, reflecting the relationship between the individual, its environment and different health outcomes. Individuals are at the centre of the model with a set of fixed biological and genetical preconditions. Building upon these, four layers of influence on health can be identified: individual lifestyle factors, social and community networks, living and working conditions and general socio-economic, cultural and environmental conditions.

Figure 1: Main health determinants



Source: Dahlgren and Whitehead, 1991

The link between health, education and gender was extensively researched by Sachs and Dupas. While Sachs provides several examples of the interaction between these dimensions at the macro level, Dupas successfully tested the linkage between female education and health by conducting randomized control trials in Kenya. Among the findings of her research, it is worth mentioning the significant effect of risk reduction campaigns on HIV incidence. These campaigns were specifically targeting girls. (REF)

---

5 Data Sources

This paper utilises data from two different sources. For HIV/AIDS incidence rates at the country level, this research explored several databases such as UNAIDS, World Bank, Global Fund for AIDS, Tuberculosis and Malaria, WHO, the Institute for Health Metrics and Evaluation, PEPFA and the AIDS Data Hub. The database used for the data analysis was the one from UNAIDS given that it provided panel data for the period 2000-2012. All the other variables used in this research were obtained from the World Development Indicators (WDI) from the World Bank.

5.1 World Development Indicators (WDI)

The WDI database comprises 1342 indicators clustered in 10 thematic areas that range from health and education to infrastructure and public sector data. Information is available for 214 countries and dates back to 1960. All indicators are available for free at the World Bank website and can be downloaded as an Excel sheet, CSV, tabbed TXT or SDMX. In addition, there is a special R package WDI designed to download and use the data.

WDI have been used in a wide range of fields and HIV/AIDS research is not an exception. For examples of relevant literature that also make use of WDI please see (Haacker 2002), (Talbott 2007) and (Kalemli-Ozcan 2011). A list of all WDI indicators used for this research can be found in the Appendix.

5.2 UNAIDS Dataset

The UNAIDS dataset is available for free at the UNAIDS website (REF) and it has been put together by UNAIDS during the research for the UNAIDS GAP Report 2014 (REF). It provides data on HIV incidence, prevalence and HIV-related deaths for the period 1990-2013. The dataset aditionally provides uncertainty bounds (low and high estimated) for most of the variables. Further, there is information for all UN member states and there is also regionally aggregated data. Some of the indicators such as prevalence and incidence are disaggregated for the differents demographic groups, including data for adults, youth and children, male and female.

UNAIDS databases are often used in public health research. Some examples of papers using UNAIDS data on HIV incidence and prevalence are Letamo (REF), Bennell (REF) and Ferlay et al.().

The data is available as an Excel file.

---

6 Methodology and Data Analysis

The methodology consists of two statistical models. The first model has as dependent variable a dummy variable that takes the value of cero if HIV incidence rate is below the median and the value of one if the rate is above the median. This variable is regressed on N variables selected to represent each of Dahlgren and Whitehead's model.

Model 1:

$$I_{it} = \beta_0 + \beta_1 SE_{it} + \beta_2 WLC_{it} + \beta_3 SCN_{it} + \beta_4 ILF_{it} + \epsilon_{it}$$

Where I stands for HIV/AIDS incidence, SE stands for socioeconomic, cultural and environmental factors, WLC stands for working and living conditions, SCN stands for social and community networks and ILF stands for individual lifestyle factors.

To study the impact of socioeconomic, cultural and environmental variables this model uses GDP per capita, the share of rural population and CO2 emissions (metric tons per capita). To operationalise working and living condition variables, healthcare expenditure, access to water and sanitation, umployment rates and primary school enrollment were selected. With regards to the social and community networks level, the model uses female school enrollment and the difference between female unemployment to total unemployment. Individual lifestyle factors were operationalised using immunization against measles and DPT and life expectancy.

This paper focus on the impact of social and community networks on HIV incidence rates. It is therefore important to explain the reasoning behind the selection of the variables used to operationalised this level. Female school enrollment was selected based on the premise that a higher female school enrollment rate would increase the chances of providing HIV prevention information campaigns to girls before they become sexually active. This in turn is expected to foster the use of condoms.

The difference of female unemployment to total unemployment was chosen in order to look at the relative position of women to men in the labour market. While female unemployment alone would be influenced by the dynamics of the labour market, discounting the total unemployment rate provides a better estimate of the relative position of women to men. It would have been desired to include a measure of female participation in the labour market. However, data on this indicator was not available.

Model 2 looks into more detail to the countries where HIV incidence rates are above the median and uses the same independent variables as model one. The dependent variable in model 2 is the logarithm of HIV incidences rates. In this case, given that the dependent variable is not a binomial variable but a contiuous one, model two is estimated using pooled OLS.

Model 2:

$$ln(I_{it}) = \beta_0 + \beta_1 SE_{it} + \beta_2 WLC_{it} + \beta_3 SCN_{it} + \beta_4 ILF_{it} + \epsilon_{it}$$

Where I stands for HIV/AIDS incidence, SE stands for socioeconomic, cultural and environmental factors, WLC stands for working and living conditions, SCN stands for social and community networks and ILF stands for individual lifestyle factors.

---

7 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 regionally aggregated 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. This process requires that the panel is as balanced as possible, as it feeds from all variables to predict values for the missing observations. A more detailed explanation of the imputation process will be providid in the inferential statistics section of this paper. In order to improve the balance of the panel, the next step consisted of 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 and WDI data covers the 2000-2012 period so all observation from the UNAIDS dataset between 1990 and 1999 were dropped. Finally, unnecessary columns from the new database were eliminated.

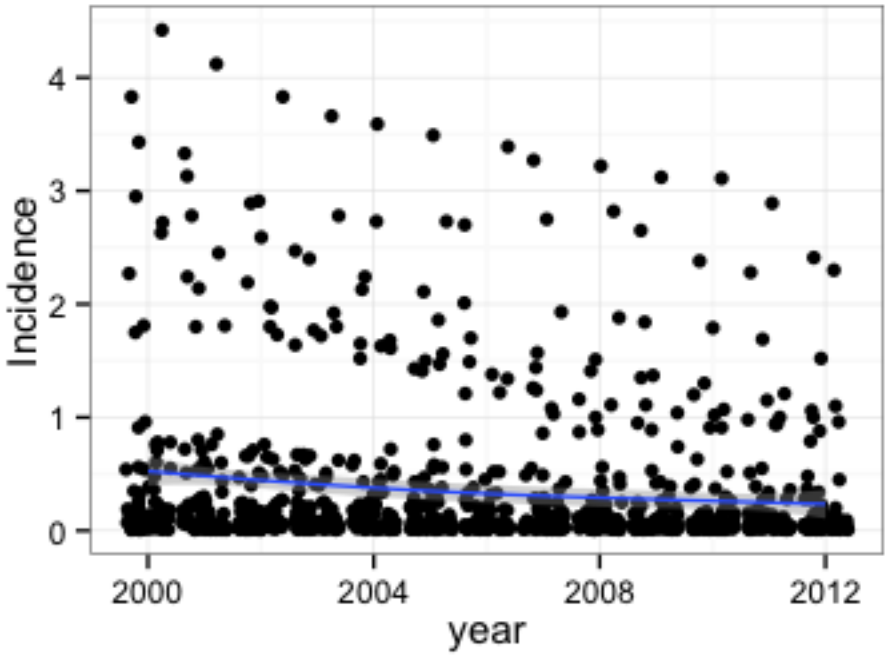
---

8 Descriptive Statistics

This section consists of the descriptive analysis of the main variables of interest. Tables, plots and histograms are shown to understand the distribution of the variables.

Figure 2 shows that in most countries of our dataset, HIV/AIDS incidence rates decreased between 2000 and 2012 (see Figure 2). The blue line in Figure 2 is however only representing the general downward trend of HIV/AIDS incidence over time in our sample. The black dots above the blue line show the dispersion of the observations around this trend. As it can be seen, there are severe outliers that still show a strongly higher HIV/AIDS incidence rate compared to the rest (mostly southern African countries).

Figure 2: HIV/AIDS Incidence Rate over Time

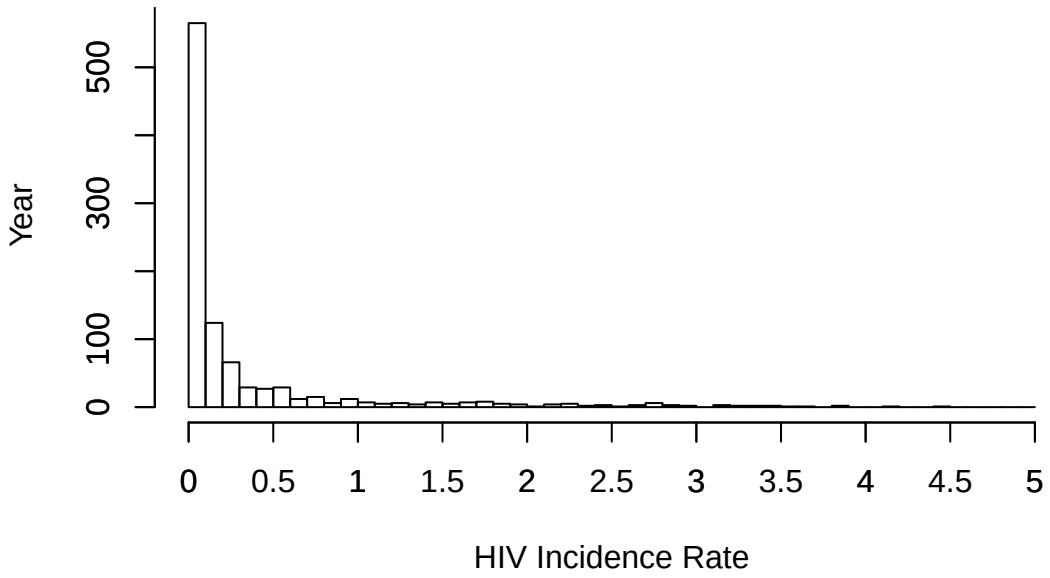


Looking at the incidence rate at the country level (Figures 3 and 4 in Appendix) it can be seen which countries are not following the general trend. While most countries have an HIV/AIDS incidence rate slightly above 0, some outliers (countries with high incidence rates) can be identified (Figure 3 in Appendix).

A similar pattern can be detected when analysing the direction of the change of the HIV/AIDS incidence rate compared to the previous year by country (Figure 4 in Appendix). To analyse the change of the incidence rate from one year to the following, the incidence variable was lagged by one year and a new variable, calculating the difference between the lagged and the original incidence rate was created.

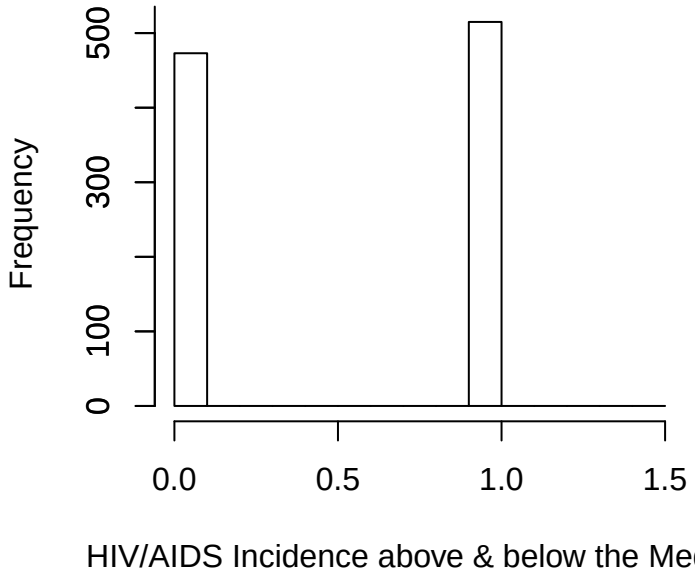
When looking at the distribution of the HIV/AIDS incidence variable in a histogram it becomes apparent that the incidence rate is indeed highly skewed to the left and only few incidence rates are higher than 1.

Figure 5: Distribution of HIV/AIDS Incidence Rate



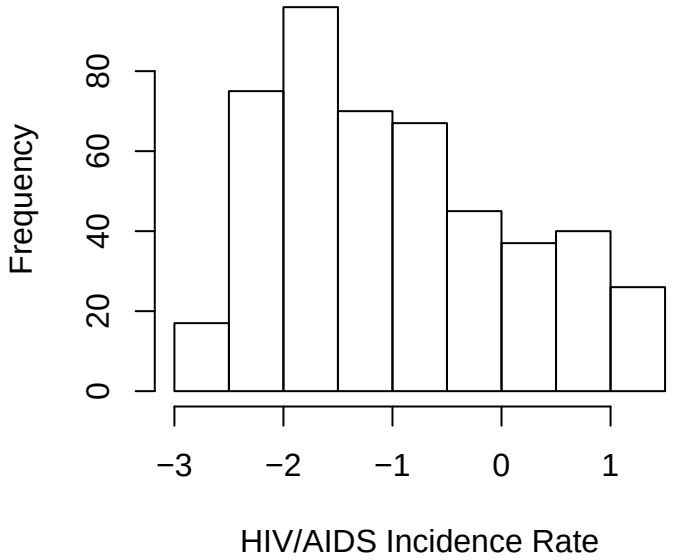
In order to deal with the skewness of this variable, a dummy variable, differentiating between countries with a high HIV incidence and countries with a low HIV incidence was created for the first empirical model. The most accurate measure of the central tendency of a skewed distribution is the median. Therefore, the dependent variable was coded as 1 if the HIV incidence rate is below the median and a value of 0 was assigned to countries with an HIV incidence rate above the median.

Figure 6: Dummy Variable for High and Low HIV/AIDS Incidence



For the second empirical model the sample was reduced and focused only on countries with a high HIV/AIDS incidence rate in order to explore whether the determinants identified in Model 1 hold true when zooming in on the more problematic cases. As the dependent variable was still skewed to the left after restricting the sample to those countries above the mean a log transformation was applied to approach a normal distribution.

Figure 6: Logged HIV/AIDS Incidence Rate for Countries lying above the Median



Scatterplots were created for all independent variables by the levels of Dahlgrehn's model, in order to see whether further relevant variables are skewed and to broader explore the relationship of the selected determinants for each level (see Appendix). Due to high skewness in most of the variables and to ensure a better comparability of the variable units, all but one of the independent variables in the sample were logged.

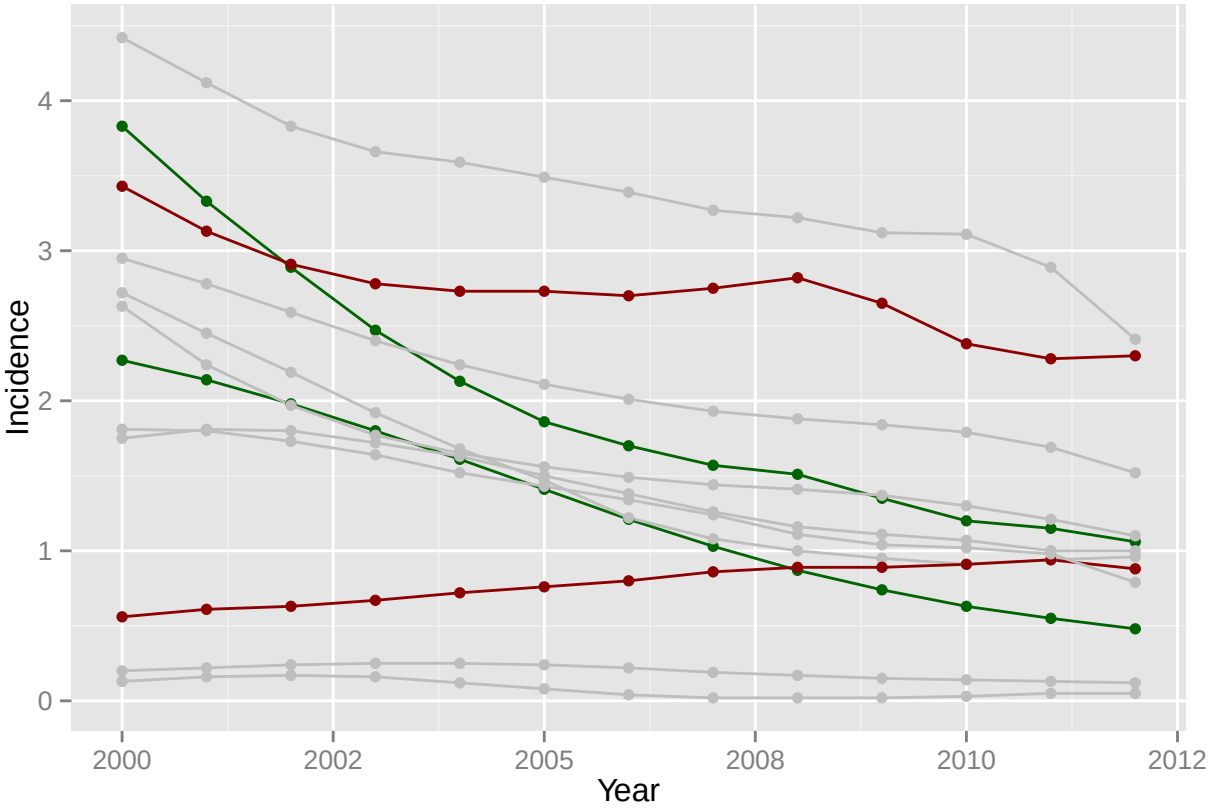
The only variable that was not logged is a variable comparing the share of female unemployment to total unemployment. The distribution of this variable comes already close to a normal distribution, so a log transformation is not of utter importance. Further, it has some values of zero, which would make it impossible to log the variable without further transforming it.

9 Case Studies - Botswana, Lesotho, Uganda and Malawi

In the following, four case studies are presented to develop a better understanding of best practices and bad practices among countries with high changes in HIV/AIDS incidence rates between 2000 and 2012. The cases were selected by the "extreme case selection method" on the basis of peculiar positive or negative changes in incidence rates over time (see Figure 4 in Appendix) (Gerring 2008).

Figure 8 shows the four case studies that were selected: two best practices (green colour), namely Botswana and Malawi and two bad practices (red colour), Uganda and Lesotho.

Figure 8: Interesting Cases for Change in HIV Incidence Rates



In Figure 9 the case studies are plotted individually, showing the development of the HIV/AIDS incidence over time for each of the four case studies. Botswana and Malawi show a steep decrease in HIV incidence rates from 2000 to 2012. Malawi is even reaching to an HIV/AIDS incidence rate that is lower than 0.5 in 2012. The HIV/AIDS incidence rate of Uganda in contrast increases continuously between 2000 and 2012. The HIV/AIDS incidence rate of Lesotho is slowly decreasing over time, however there were to backlashes one in 2009 and one in 2011.

Figure 9: HIV/AIDS Incidence in Selected Countries

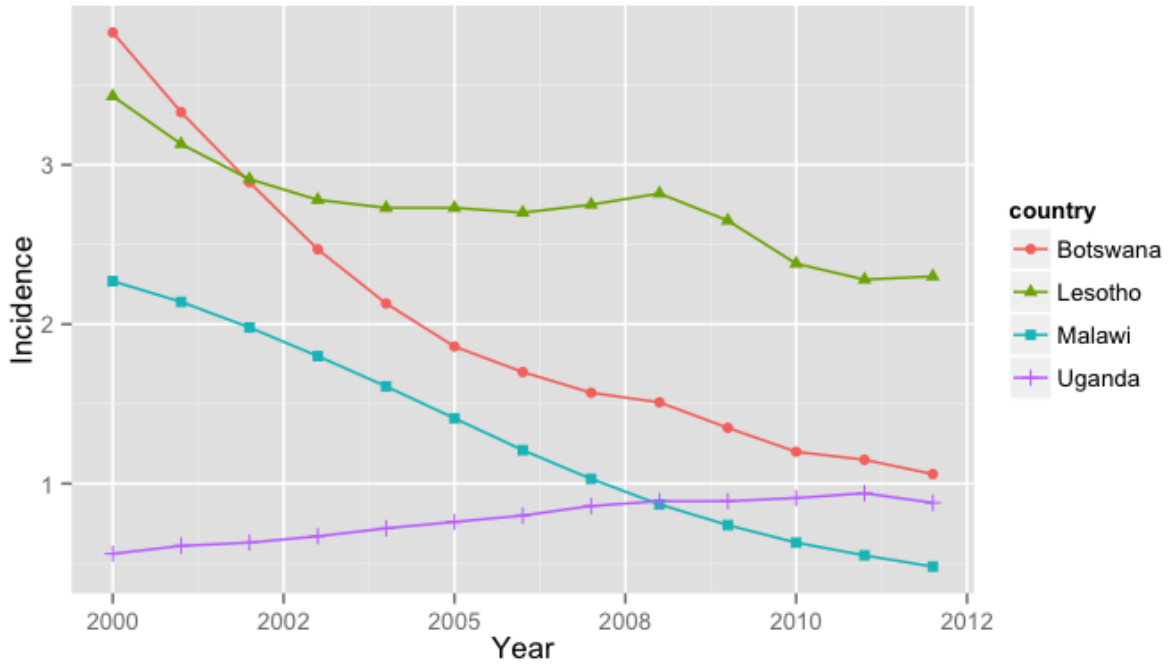
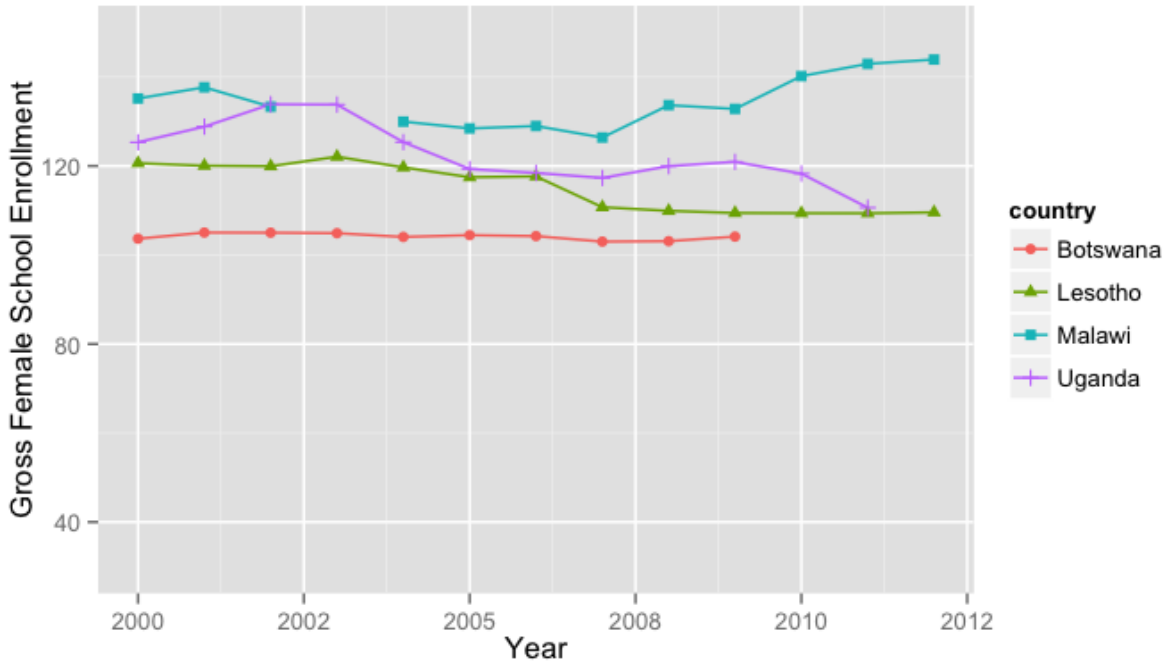
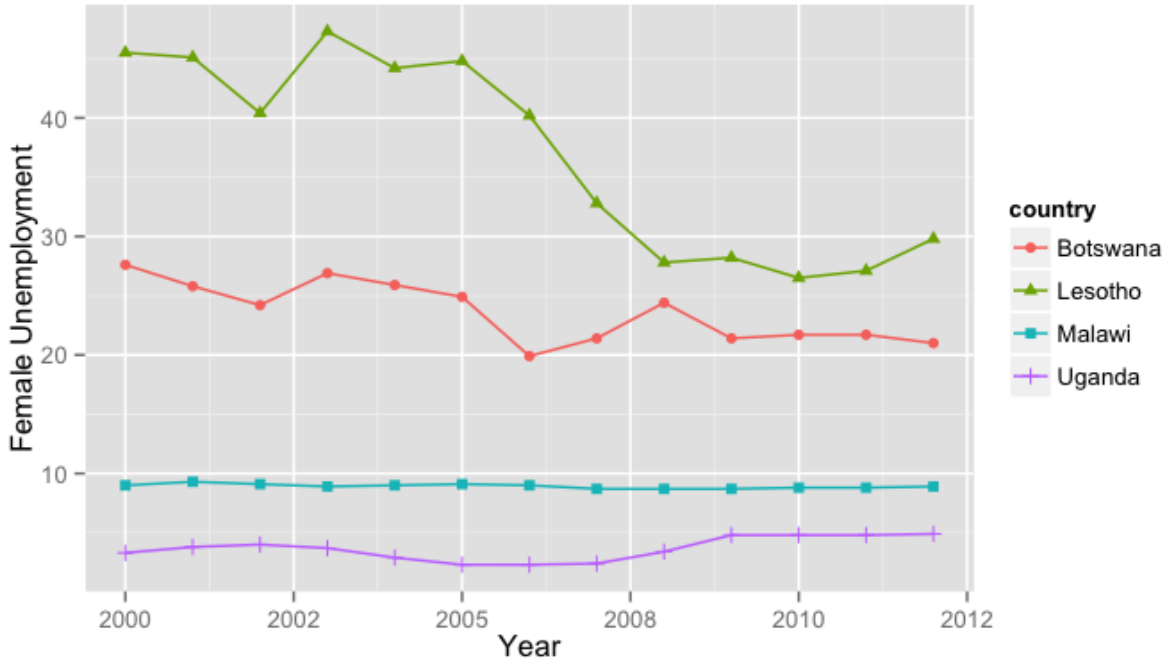


Figure 10 shows the school enrollment rates of females per country over time. The rate of females enrolled in schools showed a general decreasing tendency with some backlashes for Uganda. In Lesotho the female school enrollment rate also decreased over time but stagnated between 2009 and 2012. Botswana had a very low and stable female school enrollment rate between 2000 and 2009. Malawi had a high show a steep decrease in HIV incidence rates from 2000 to 2012. Malawi is even reaching to an HIV/AIDS incidence rate that is lower than 0.5 in 2012. The HIV/AIDS incidence rate of Uganda in contrast increases continuously between 2000 and 2012. The HIV/AIDS incidence rate of Lesotho is slowly decreasing over time, however there were to backlashes one in 2009 and one in 2011.

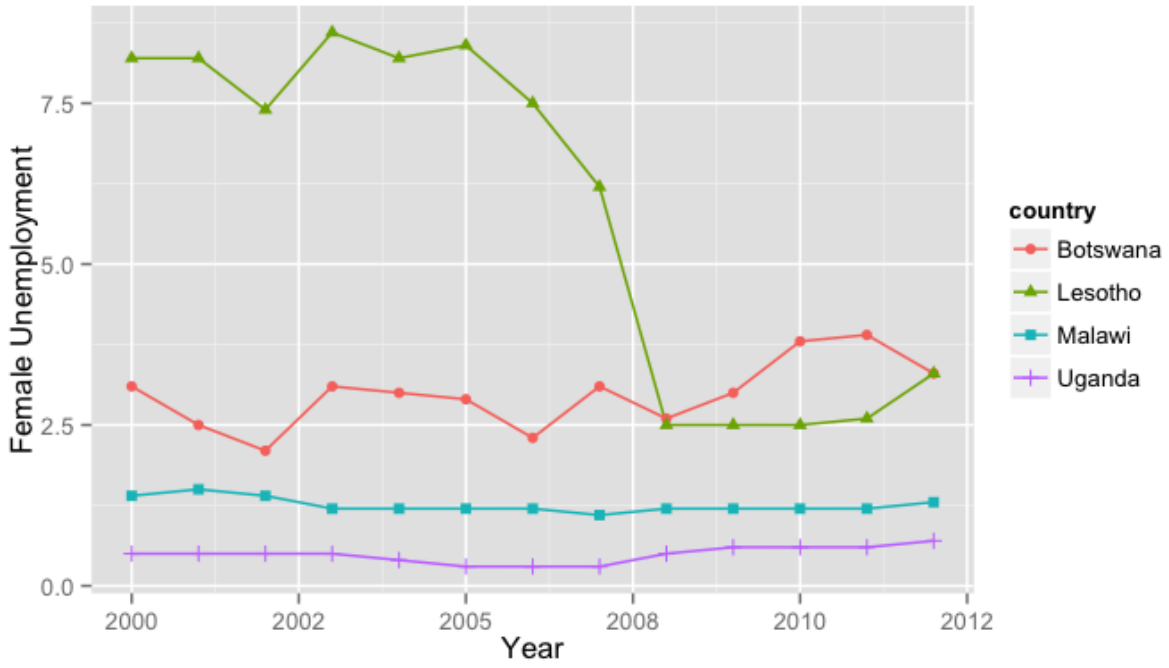
Female School Enrollment in Selected Countries



Female Unemployment in Selected Countries



Female Unemployment Share in Selected Countries



10 Inferential Statistics

10.1 Data Imputation

The data loaded from the Worldbank contained a lot of missing data because some countries do not report information on specific variables on a yearly basis. A high share of missing values creates several problems and limits the data analysis. The paper imputed the missing data by using the "Amelia" package. "Amelia" generates by default five multiple, complete datasets that contain estimations of missing data points.

To enable the data imputation highly collinear variables had to be dropped. To test for multicollineraity this paper used variance inflation factors (VIF). The VIF showed that in a simple OLS regression model integrating all independent variables, three variables were highly multicollinear and had a variance inflation higher than the threshold of 10. We tested the multicollinearity between the variables and found that there was high multicollinearity between GDP and GDP per capital and between Primary Education and Female School Enrollment. In a first step, we excluded one of these multicollinear variables for each group based on their explanatory strength for our research question, Primary Education and GDP. After having further problems of multicollinearity with GDPpc we also excluded Healthcare Expenditure and two variables that were created for the descriptive statistics part only, when showing the difference of HIV/AIDS Incidence over time (Incidence2 and IncidenceDif).

10.2 Model 1 - Comparison of countries with an HIV Incidence Rate below and above the Median

For Model 1 logistic regressions are used for predicting the likelihood that a country has a low HIV incidence rate (the dependent variable Y is equal to 1, rather than 0) given certain values of the explanatory variable. As explained in the descriptive statistics part, the dependent variable of Model 1 is a dummy variable, that takes the value of 1 if the HIV incidence rate is below the median and 0 for countries with an HIV incidence rate above the median.

As Odds and Odds ratios are difficult to interpret and do not allow to make direct statements about predictors in our model, predicted probabilities were calculated after running the logistic regressions. The interpretation of the logistic regression results presented in table 1 will thus focus only on the significance of the variables.

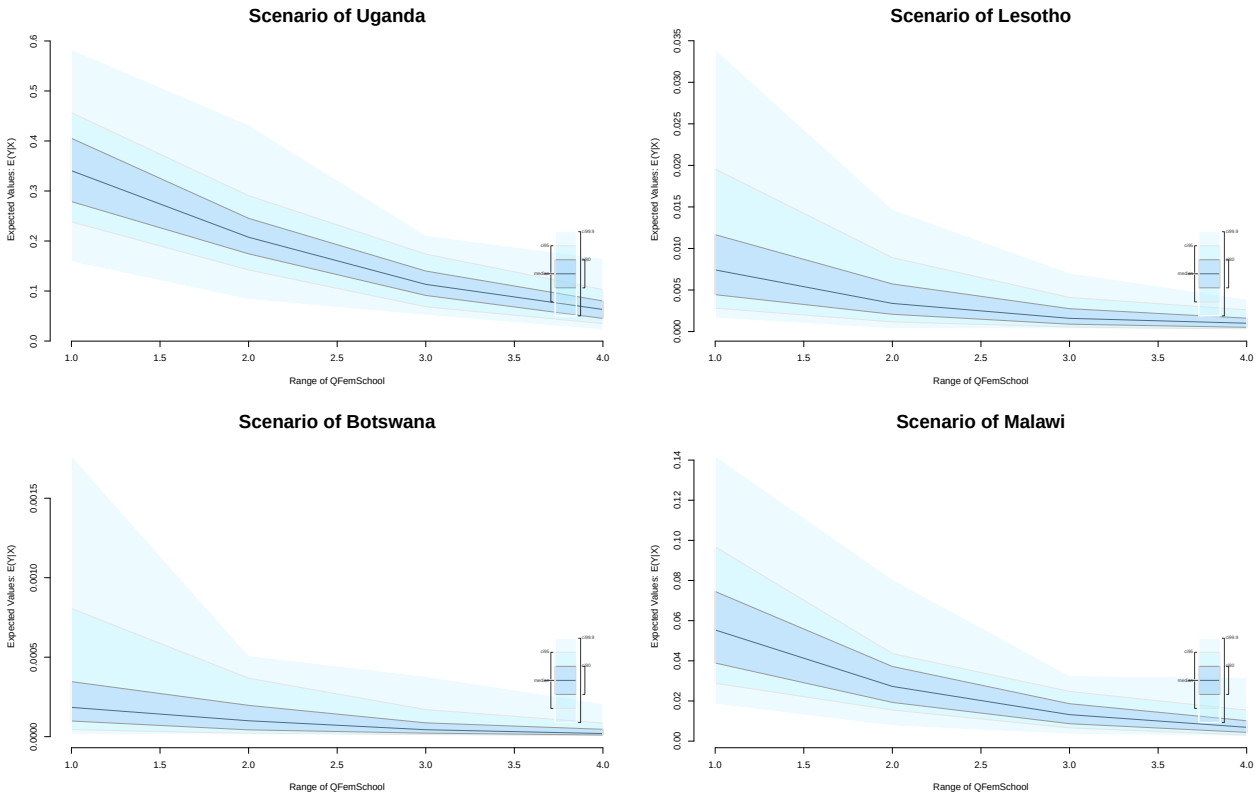
Table 1: Logistic Regression Results of Model 1

Variables	Coefficients	Std. Error	T-Statistic	P-Value
Constant	-102.92	9.04	-11.38	0.00
GDP per capita	-0.75	0.32	-2.34	0.02
Share of Rural Population	-0.97	0.42	-2.33	0.02
CO2 Emissions per capita	-1.02	0.19	-5.42	0.00
Healthcare Expenditure	0.37	0.36	1.05	0.30
Access to Water	0.57	0.78	0.74	0.46
Access to Sanitation	0.23	0.24	0.96	0.34
Life Expectancy	30.20	2.35	12.85	0.00
Immunisation against DPT	-1.50	1.36	-1.10	0.28
Immunisation against Measles	1.67	1.47	1.14	0.26
Female School Enrollment	-3.70	0.60	-6.19	0.00
Share of Female Unemployment	-0.03	0.03	-0.79	0.43

From the logistic regression results presented in Table 1 it can be seen that the variables GDP per capita, Share of Rural Population, CO2 Emissions per capita, Life Expectancy and Female School Enrollment are showing statistically significant results at the 95 percent confidence level. The variables Immunization against DPT, Immunization against Measles, Access to Water, Access to Sanitation and Healthcare Expenditure and Share of Female Unemployment are however not statistically significant.

A detailed analysis of the predicted probabilities will be given for the main variable of interest Female School Enrollment. To calculate the predicted probabilities, quantiles of Female School Enrollment are computed and applied to four different scenarios. Each scenario fixes the mean values of the explanatory variables (excluding Female School Enrollment) to those of the case studies that were introduced in the descriptive statistics part.

The predicted probabilities were calculated using the ZELIG-package as ZELIG has the option to handle multiply-imputed data frames. The predicted probabilities plots show that a higher quartile of Female School Enrollment triggers a lower HIV/AIDS incidence. At the fourth quartile, the HIV/AIDS incidence is lowest in all four graphs. The uncertainty around the estimates is reflected by the bandwidth of the plots. It can be concluded, that the effect of Female School Enrollment is going in the same direction for all four scenarios even though, the expected values for the predicted probabilities differ. For the Scenario of Uganda the strongest effect can be observed. **Predicted Probabilities for Female School Enrollment**



From the Residuals vs. Fitted plot (see Appendix) it can be assumed, that Model 1 faces problems of heteroscedasticity, as a smaller dispersion around the regression line can be observed for lower values of X compared to higher values of X.

10.3 Model 2 - Focusing on Countries with an HIV Incidence Rate above the Median

As explained in the previous section, the dependent variable for the second empirical model is only considering countries with an HIV/AIDS incidence rate higher than the median. Further, the dependent variable for this model was logged to better approach a normal distribution.

From the Residuals vs. Fitted plot (see Appendix) it can be assumed, that Model 2 faces problems of heteroscedasticity, as a smaller dispersion around the regression line can be observed (for lower and higher values of x). Given the presence of heteroscedasticity in Model 2, the regression will be estimated with robust standard errors to decrease the heteroscedasticity of the error term.

(It cannot be controlled for fixed effects of Model 2, because the degrees of freedom are too low.)

From the OLS regression output presented in Table 2 it can be seen that the variables Share of Rural Population, CO2 Emissions per capita, Life Expectancy and Female School Enrollment are still statistically significant at the 95 percent confidence level (like the regression outputs in Model 1). Furthermore, the variable share of Female Unemployment became significant in this model (it was insignificant in Model 1). GDP per capita is only significant at a 90 percent confidence level.

Surprisingly, the sign of the coefficient for Female School Enrollment does not match our theoretical assumptions on the direction of the effect and the finding is also opposite to the effect that was initially found for the predicted probabilities of Model 1. According to the regression output, a one percent increase in Female School Enrollment triggers a 1.38 percent increase in HIV/AIDS incidence.

The sign of the coefficient for the variable Female Unemployment compared to total unemployment is

supporting the intuition of this paper: a one percent decrease in Female Unemployment compared to total unemployment is *ceteris paribus* followed by a 0.13 percent decrease in the HIV/AIDS incidence rate.

Table 2: OLS Regression Results of Model 2 with robust standard errors

Variables	Coefficients	Std. Error	T-Statistic	P-Value
Constant	15.89	1.68	9.44	0.00
GDP per capita	0.15	0.08	1.91	0.06
Share of Rural Population	0.55	0.13	4.23	0.00
CO2 Emissions per capita	0.13	0.04	2.88	0.00
Healthcare Expenditure	-0.11	0.10	-1.06	0.29
Access to Water	0.25	0.20	1.23	0.22
Access to Sanitation	0.00	0.07	0.03	0.98
Life Expectancy	-7.06	0.29	-24.22	0.00
Immunisation against DPT	0.29	0.34	0.84	0.41
Immunisation against Measles	-0.16	0.36	-0.44	0.67
Female School Enrollment	1.40	0.15	9.45	0.00
Share of Female Unemployment	0.13	0.02	6.99	0.00

11 Findings

Provided the special interest of this paper on the social and community networks level of the rainbow model, this section will mostly focus on the findings regarding the impact of female school enrollment and the differente of female to total unemployment on HIV incidence rates. As it was mentioned above, the interest in this particular level of Dahlgren and Whitehead's model is explained by the insufficient amount of research on the significance of such factors in explaining disease specific determinants of health. Furthermore, female schoole enrollment and the difference between female and total unemploymentis used in this paper to operationalize the social and community network factors.

The estimation of Model 1 found that female school enrollment was highly significant in explaining whether countries have an HIV/AIDS incidence rate above or below the median and the predicted probabilities showed that the effect was going in the assumed direction, namely that an increase in female school enrollment increased the probabilities of having an HIV/AIDS incidence rate lower than the median. In Model 2 however, the sign of this coefficient surprisingly changed. In this case - unlike in model 1 - an increase in female school enrollment leads, *ceteris paribus*, to an increase in HIV/AIDS incidence rates.

A potential explanation for the difference in the significance of female school enrollment relates to the sample that each model is analysing. While model 1 looks at all countries for which data was available, model 2 only looks at countries where HIV incidence rates were above the median. Looking closer into these countries, it was noticed that most of these countries are low-income African countries. As Dupas has observed, HIV prevention campaigns in some of these countries (the author looks mostly at West African countries) focus exclusively on abstinence (Dupas 2009). Because abstinence campaigns often fail to increase the use of condoms, by focusing on ineffective prevention strategies, these countries do not benefit from the potential reduction of HIV/AIDS incidence rates that female school enrollment could bring.

With regards to the difference between female and total unemployment, model 1 found no statistical significance. In Model 2 however the effect of this variable became significant and the direction of the effect was in line with the original assumptions for this variable, i.e. that a higher difference in female to total unemployment would lead, *ceteris paribus*, to higher HIV/AIDS incidence rates. An explanation for the insignificant effects of relative female unemployment in Model 1 and the statistically significant effect in Model 2 is assumably lying in the problems of the coding of the variable. This paper the difference between female and total unemployment as a proxy for female labour market participation because data on female

labour market participation was not available. The proxy might however come with inaccuracies in the measurement. In addition, in low- and middle-income countries with a high share of agricultural production, female unemployment might be overestimated given that these countries tend to neglect the value created by female domestic employment.

(Sachs and Malaney 2002)

12 Conclusions and Policy Recommendations

The first that is worth evaluating is whether the findings of this research are in line with the results of existing evidence. As it was already mentioned, the link between most levels of Dahlgren and Whitehead's model (all except social and community networks level) has already been explored and they have been found to be significant determinants of HIV/AIDS incidence rates. This research however found that only a limited number of variables within each level were significant: while the share of rural population, CO2 emissions per capita and life expectancy were found to be significant at the 95 percent confidence level in both regression models, all the other variables included in this research (GDP per capita, healthcare expenditure, access to water, access to sanitation, immunisation against DPT and immunisation against measle) were not statistically significant across both model specifications. These findings cast doubts on the general applicability of Dahlgrehn and Whitehead (**???**) model to specific diseases like HIV/AIDS.

Further, the general importance of social and community networks for determining HIV/AIDS incidence can not be fully confirmed by the findings of this paper. As mentioned before, Share of Female Unemployment is only statistically significant in Model 2 and the direction of the effect of female school enrollment changes across the different model specifications. Nevertheless, Female School Enrollment showed a strong statistical significance across both models and can therefore be assumed to be important even if potentially ambiguous determinant for HIV/AIDS incidence.

On the basis of these conclusions, the following policy recommendations can be given:

INVEST IN GIRLS EDUCATION: If female schooling is a significant determinant of HIV, investing in female education is not only a good investment because it will improve the skills of women and increase their opportunities to get a job but also because it will make the population healthier thus reducing the costs of the health system.

FURTHER EXPLORE THE LINK BETWEEN FEMALE LABOUR PARTICIPATION AND HIV INCIDENCE: When looking at the difference between female and total unemployment, we realised that this indicator is not exactly portraying the variable that we wanted to analyze, namely female labour participation. It would be therefore interesting to test whether this variable is a good determinant of HIV before reaching any conclusion.

LOOK AT MORE DETERMINANTS OF SPECIFIC DISEASES: We realised that there is a gap in the literature exploring disease specific determinants of health regarding the impact of social and community level factors. It would be interesting to evaluate the determinants of other diseases and see whether these coincide with the determinants of HIV.

13 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. Especially, a better operationalization for the social and community networks would have been beneficial for this paper. Famale

labour market participation would have been a better measure instead of taking the comparison of the female unemployment rate to the total unemployment rate as a proxy. Further interesting variables for the operationalization of this level could for example be measures on social cohesion, civil society engagement and trust in neighbours. Data on these variables were however neither available for most of the countries in our sample nor not available on a yearly basis. To deal with this limitation, the research could only use the theoretical framework as an instrument to guide the selection of variables but cannot utilise the findings to test the validity of the model.

A limitation specific to Model 1 is that the dummy variable used in this paper (differentiating between countries with an HIV/AIDS incidence rates above and below the median) is not the most accurate variable to measure progress in the achievement of Target 6.A of the MDGs ("have halted by 2015 and begun to reverse the spread of HIV/AIDS") (United Nations 2014). A plain dummy variable on whether MDG 6.a was achieved or not is however not suitable for identifying the specific determinants of HIV/AIDS at the country level as it is not differentiating between the relative size of the effect (an increase from 0,001 to 0,002 would be equally bad as a change from 0,1 to 0,3). Future research could however try to solve this problem for example by adding weights or ratios, which would facilitate a clearer explanation of why some countries will fail to achieve MDG 6.a.

A further limitation of Model 2 is that the dependent variable (logged HIV/AIDS incidence rate) for the sample of countries with a higher HIV/AIDS incidence rate is still not perfectly normally distributed after the log transformation. According to Cribari-Neto and Zeileis (Cribari-Neto and Zeileis 2010) a log transformation faces some disadvantages compared to a beta regression, which is an alternative method for dealing with non-normal frequency distributions of the dependent variable. According to the authors, a beta regression model is especially useful, for continuous dependent variables that are restricted to the standard unit interval (0, 1). A beta regression would therefore constitute an interesting additional approach to the OLS regression in Model 2.

14 Appendix

Descriptive Statistics

Table 3: 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
GDPpc	945	5,296.0	5,153.0	441.2	30,875.0
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
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
lGDPpc	945	8.1	1.0	6.1	10.3
lRural	988	4.0	0.4	2.2	4.5
lCO2	829	−0.6	1.6	−5.6	3.6
lHCexpend	962	1.7	0.4	0.6	2.9
lWater	982	4.3	0.3	3.2	4.6
lSanitation	983	3.7	0.8	1.9	4.6
lUnemploym	988	1.9	0.7	−0.5	3.7
lHCexpendpc	962	4.1	1.2	0.9	7.0
lFemSchool	828	4.6	0.3	3.0	5.1
lLifeExpect	988	4.1	0.2	3.6	4.4
lDPT	988	4.3	0.3	2.9	4.6
lMeasles	988	4.3	0.3	2.8	4.6
lIncidence	988	−2.4	1.6	−4.7	1.5
QFemSchool	828	2.5	1.1	1	4
ShFemUnempl	988	1.5	2.9	−3.4	23.9
Dummy	988	0.5	0.5	0	1

Figure 3:HIV Incidence Rates per Country over Time

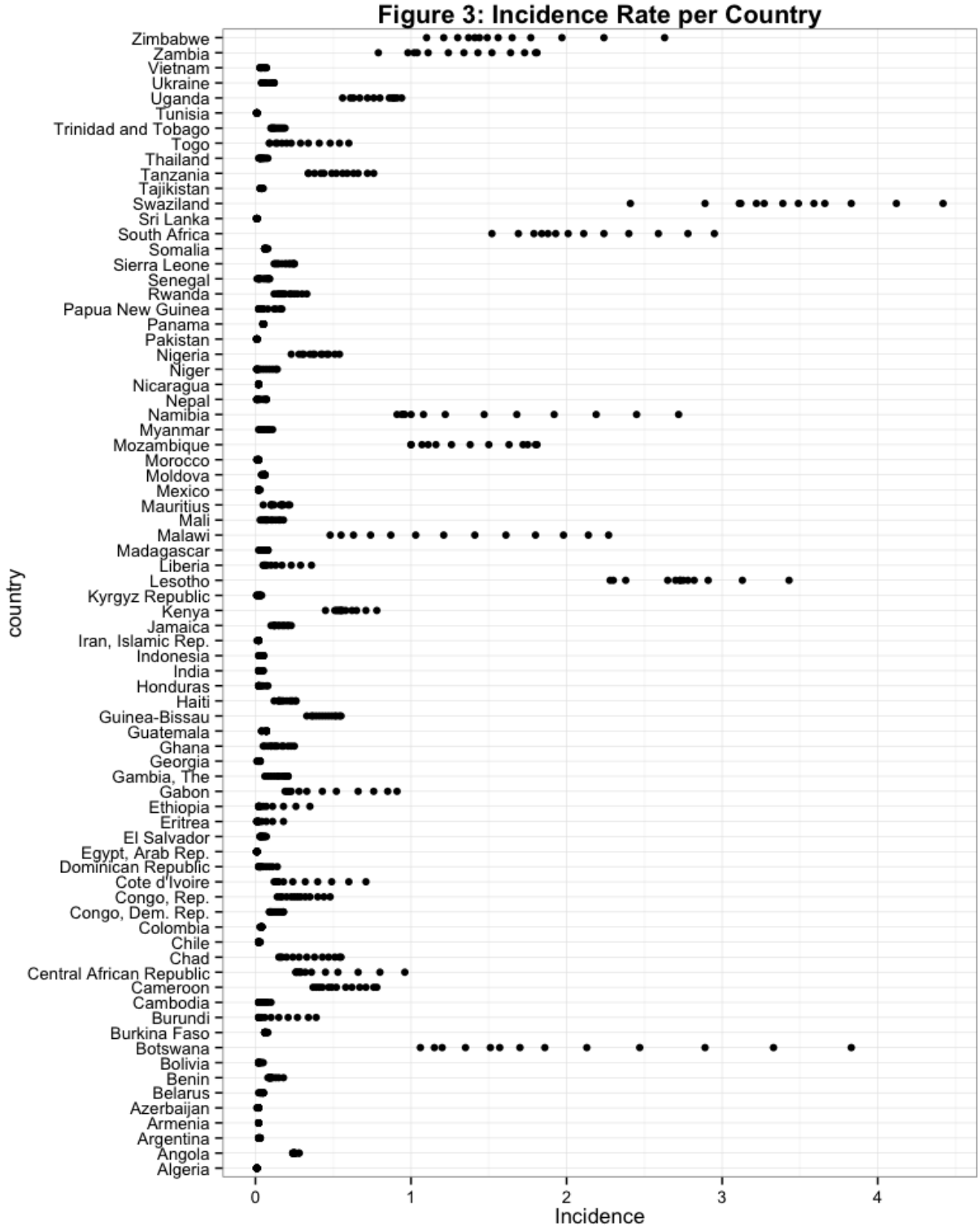


Figure 4:Change in HIV Incidence Rate compared to Previous Years per Country

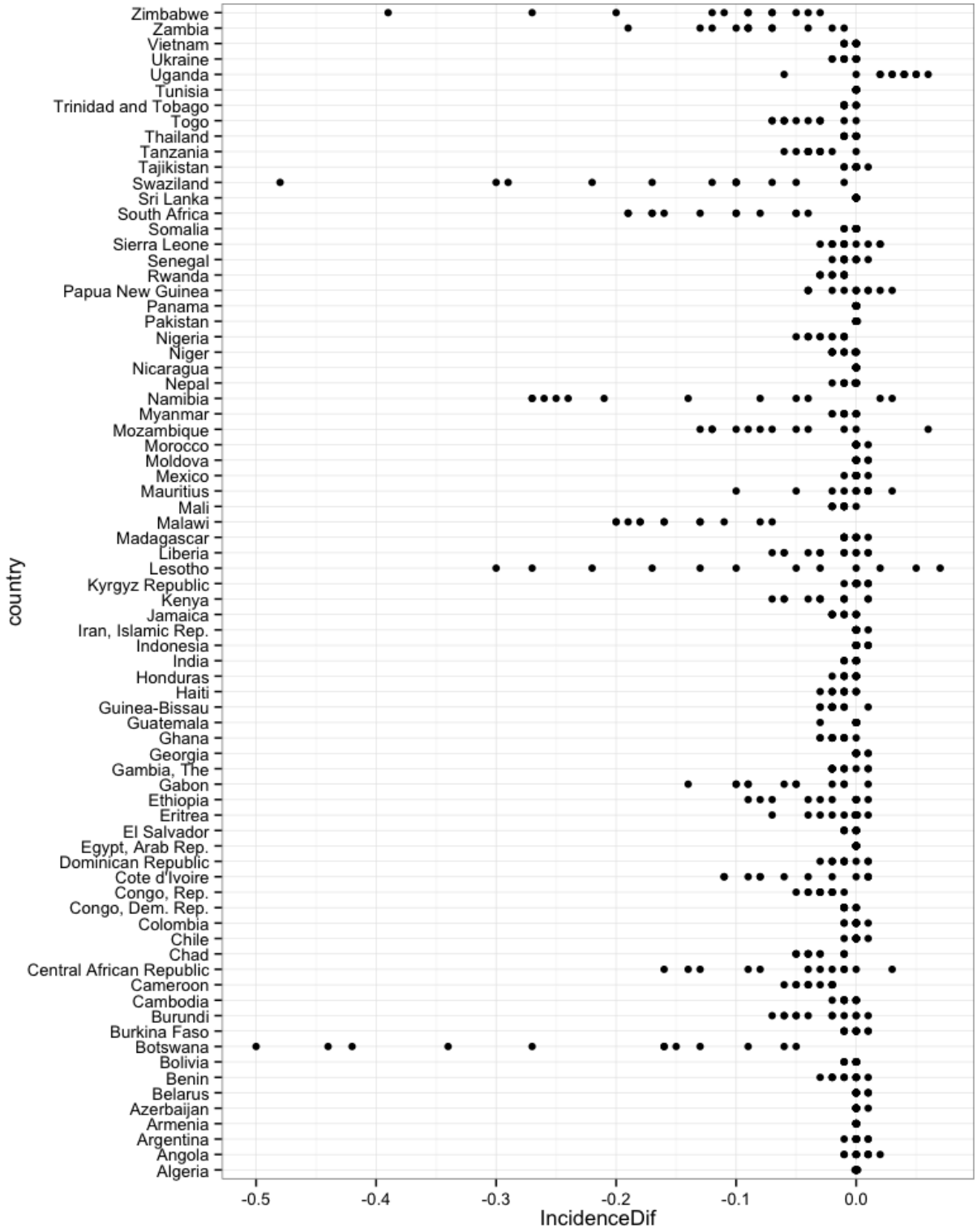


Figure 5:Scatterplot of variables for socio-economic, cultural and environmental conditions

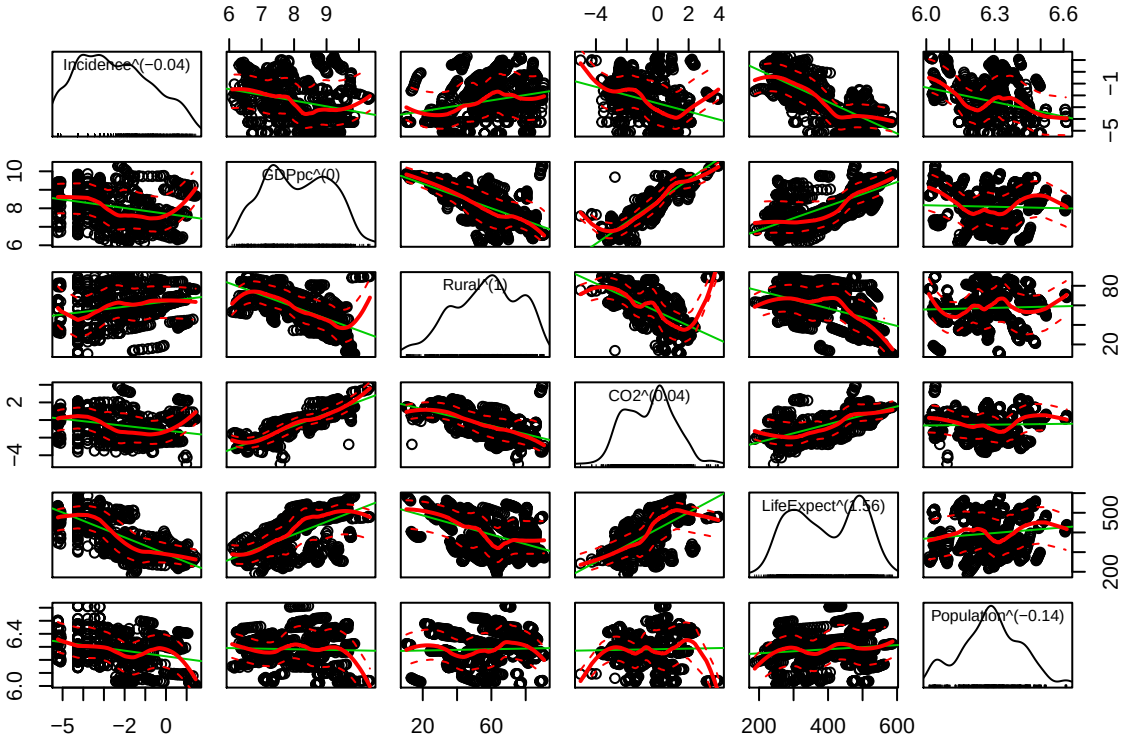


Figure 6:Scatterplot of variables for living and working conditions

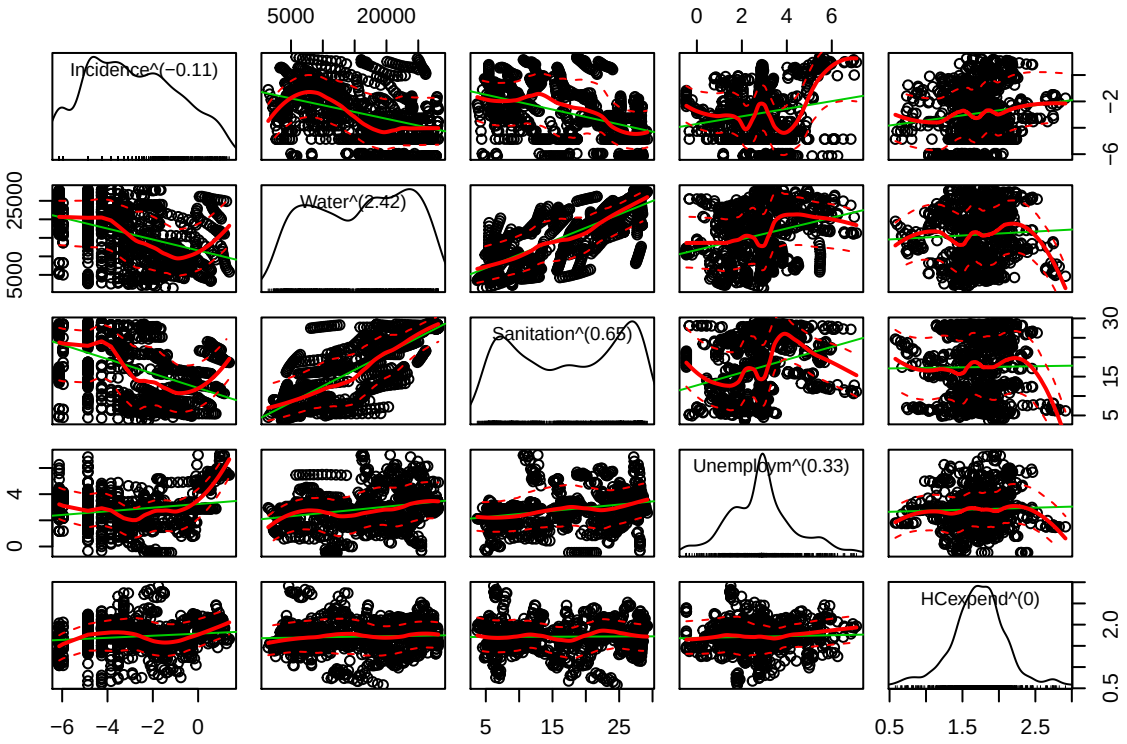
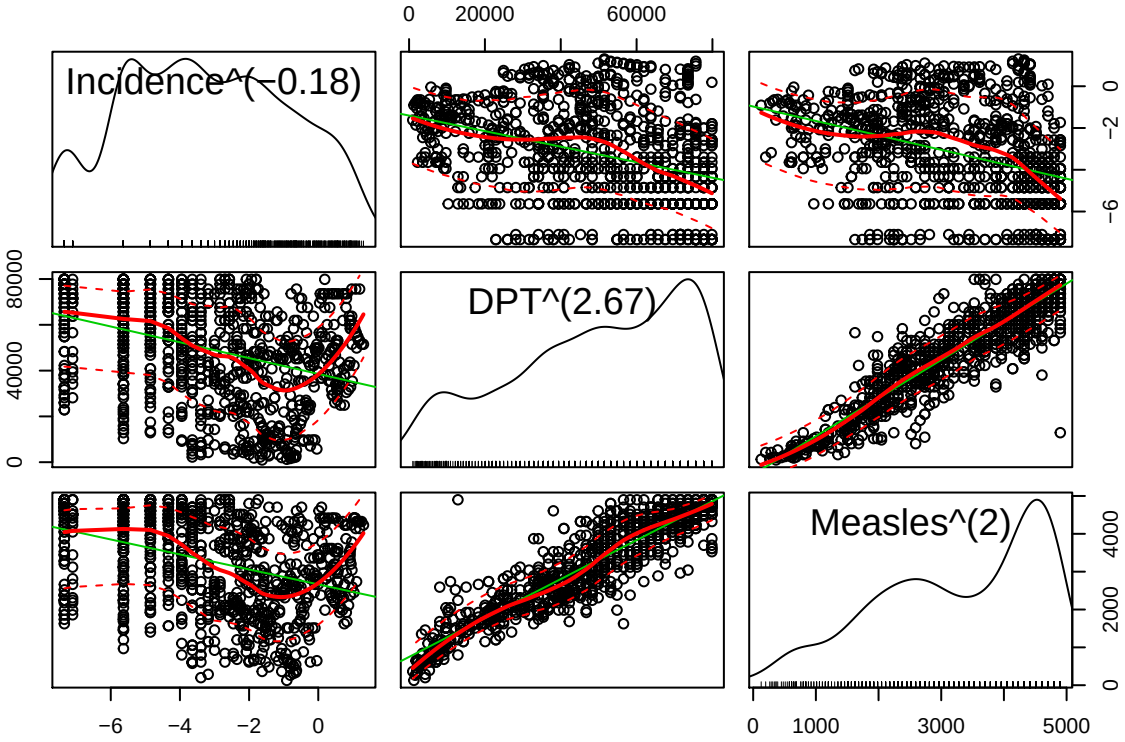
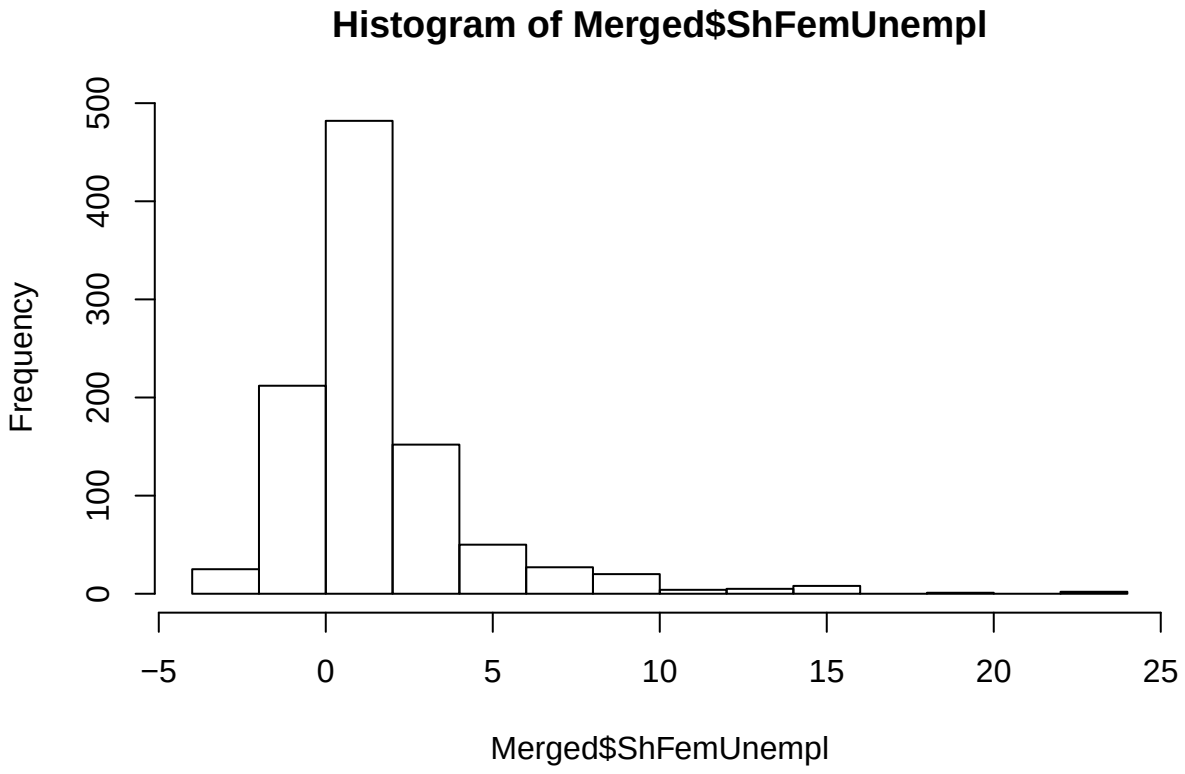


Figure 7:Scatterplot of variables for individual lifestyle factors



Histogram of Female Unemployment compared to Total Unemployment (not logged)



Testing for Multicollinearity of the Variables

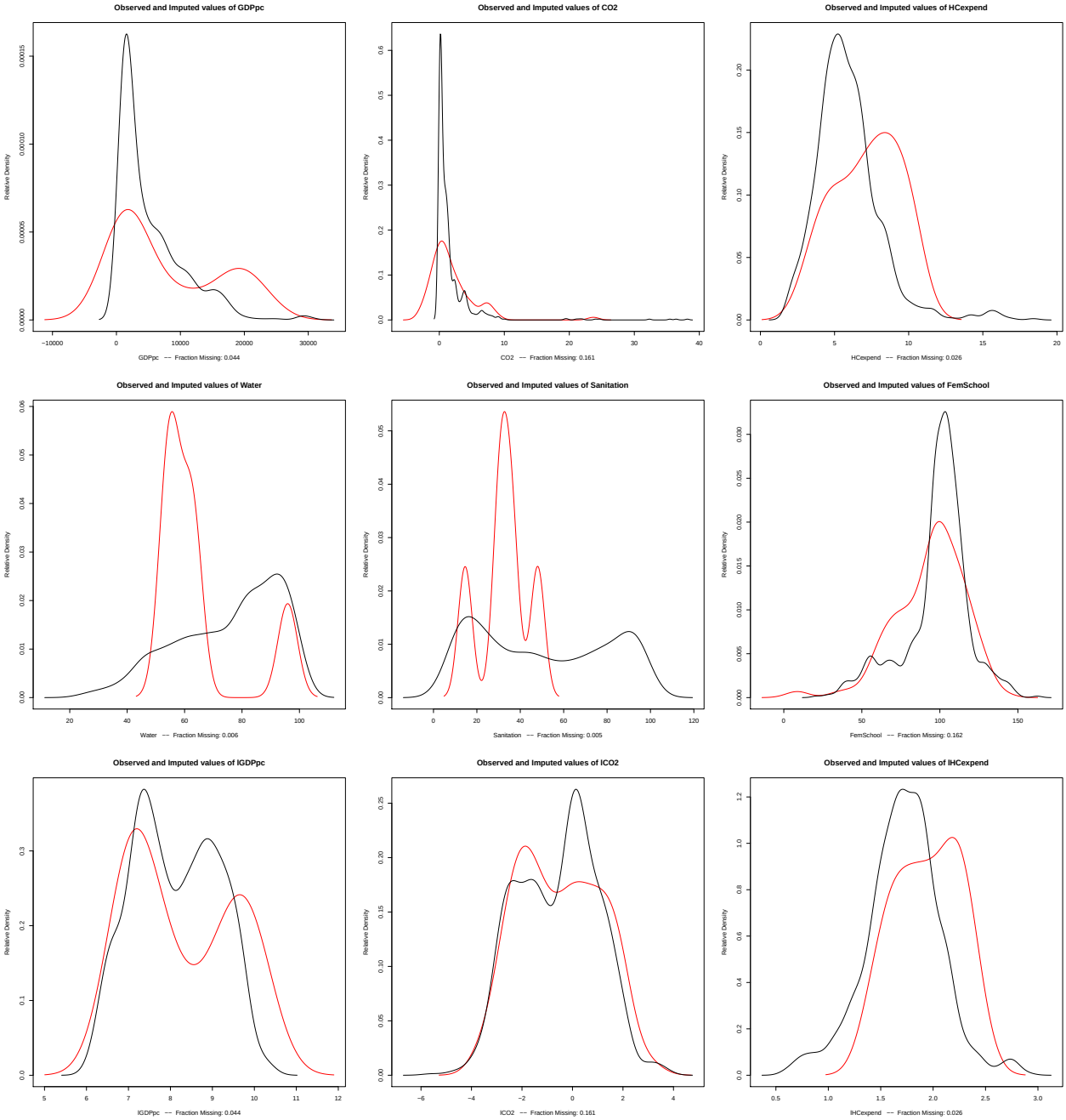
Table 4: Variance Inflation Factors - Table 1

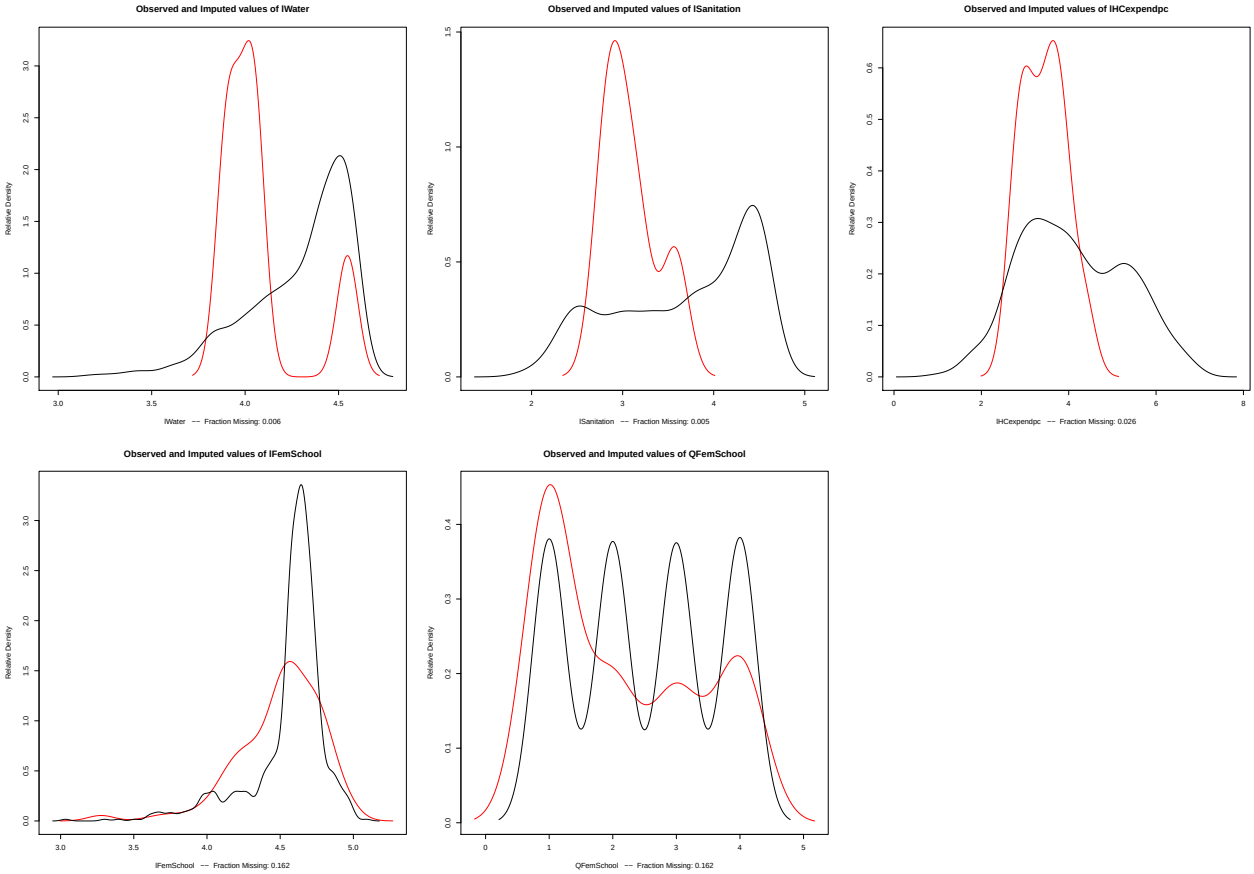
	vif
GDP	3.22
GDPpc	12.78
Rural	2.71
CO2	3.48
HCexpend	1.68
Primary	51.11
Water	3.94
Sanitation	4.57
Unemploym	2.80
HCexpendpc	7.25
ShFemUnempl	1.80
FemSchool	56.88
LifeExpect	5.55
DPT	8.74
Measles	9.42
Population	3.03
Incidence2	4.88
IncidenceDif	2.73

Table 5: Variance Inflation Factors - Table 2

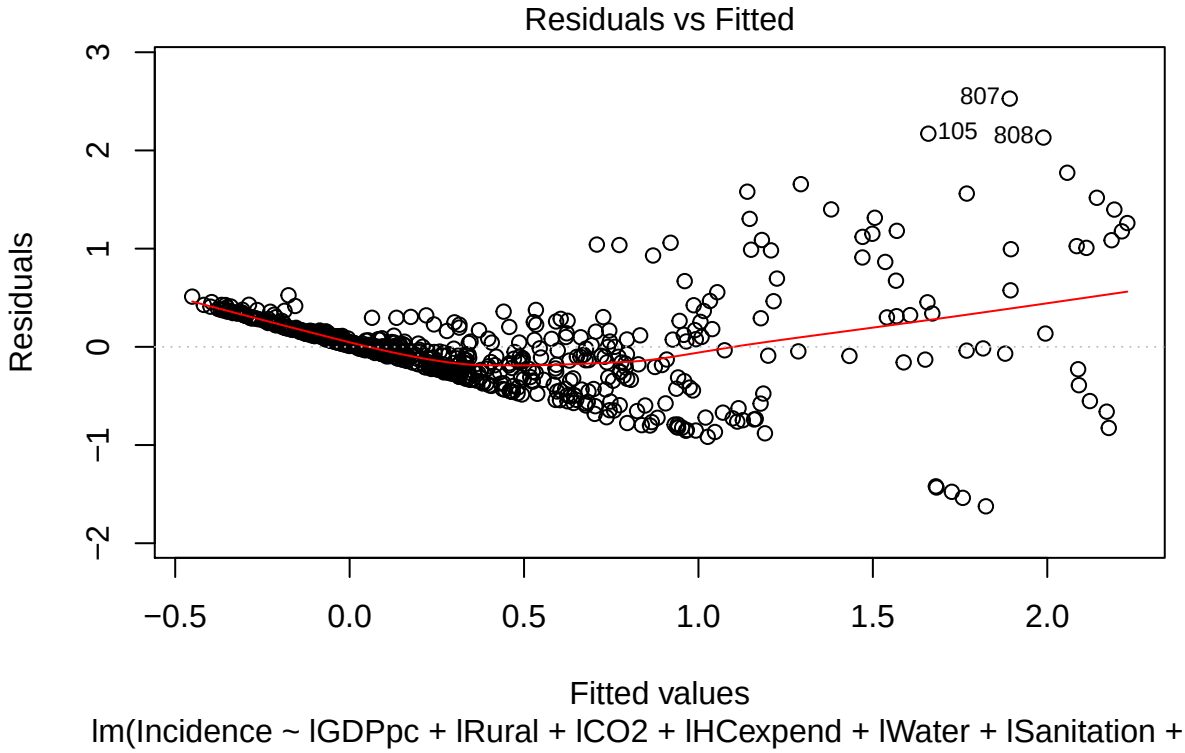
	vif
GDPpc	5.00
Rural	2.33
CO2	3.14
Water	3.76
Sanitation	4.24
HCexpend	1.20
ShFemUnempl	1.58
Unemploym	1.75
FemSchool	1.35
LifeExpect	3.34
DPT	8.55
Measles	8.97
Population	1.19

Data Imputation Matrix

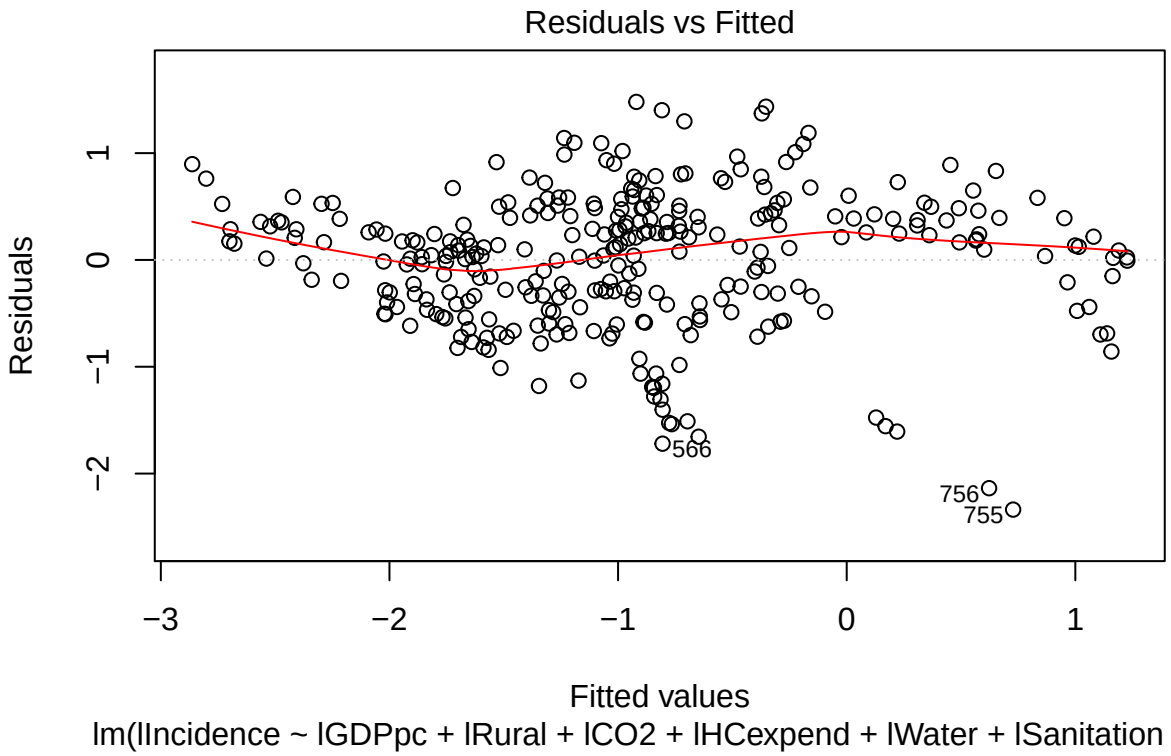



Observed and Imputed values of lWater
Relative Density
lWater --- Fraction Missing: 0.006
Observed and Imputed values of lSanitation
Relative Density
lSanitation --- Fraction Missing: 0.005
Observed and Imputed values of lHCexpendpc
Relative Density
lHCexpendpc --- Fraction Missing: 0.026
Observed and Imputed values of lFemSchool
Relative Density
lFemSchool --- Fraction Missing: 0.162
Observed and Imputed values of QFemSchool
Relative Density
QFemSchool --- Fraction Missing: 0.162

Testing for Heteroscedasticity - Model 1



Testing for Heteroscedasticity - Model 2



Cribari-Neto, Francisco, and Achim Zeileis. 2010. "Beta Regression in R." *J. Stat. Softw.*, 1–24.

Dahlgren, Göran, and Margaret Whitehead. 1991. "Policies and Strategies to Promote Social Equity in Health." *Institute for Futures Studies*.

Dupas, Pascaline. 2009. *Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya*. National Bureau of Economic Research.

Gerring, John. 2008. "Case Selection for Case-Study Analysis: Qualitative and Quantitative Techniques." *J. Box-Steffensmeier, HE Brady, & D. Collier (Reds.), Oxford Handbook of Political Methodology, Oxford University Press, Oxford*, 645–84.

Haacker, Markus. 2002. "The Economic Consequences of HIV/AIDS in Southern Africa." *IMF Working Paper* 02 (38).

Hurrelmann, Klaus. 1989. "Human Development and Health." *Springer-Verlag*.

Kalemli-Ozcan, Sebnem. 2011. "AIDS, "Reversal" of the Demographic Transition and Economic Development: Evidence from Africa." *Springer-Verlag*.

Sachs, Jeffrey, and Pia Malaney. 2002. "The Economic and Social Burden of Malaria." *Nature* 415 (6872): 680–85.

Talbott, John R. 2007. "Size Matters: The Number of Prostitutes and the Global HIV/AIDS Pandemic." *PLoS ONE*.

United Nations. 2014. "Millennium Development Goals." http://www.un.org/millenniumgoals/aids.shtml.