Collaborative Research Project - Assignment 3

24 October 2014

Contents

1. Descriptive Statistics 1

Create cleaner covariate labels 6

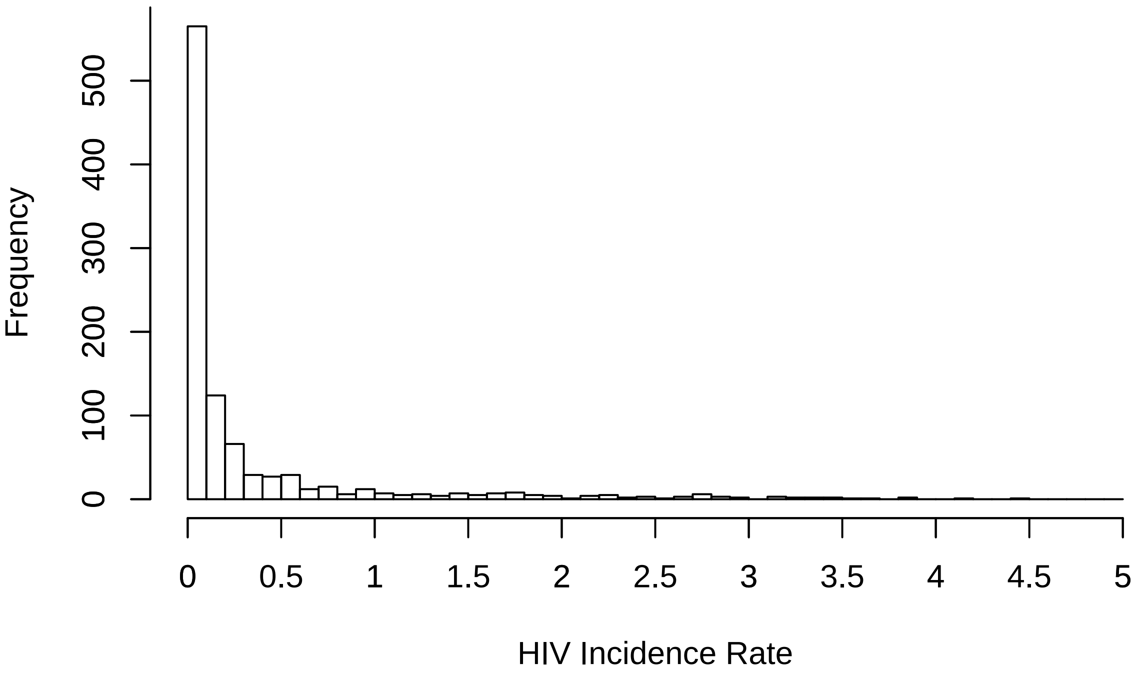
To automate the bibtex packages file: repmis::LoadandCite(pkgs, file = 'RpackageCitations.bi Note: Use a file name that is different from your literature BibTeX file! b')

1. Descriptive Statistics

For the discriptive statistics scatterplots and histograms will be shown in order to understand the distribution of the variables.

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



In most regions of the world HIV incidences decreased between the period of 2000 to 2015 (see Figure 2).

[Figure 2: Incidence Rate over Time] (/fig2.png)

When plotting the incidence rates per country (Figure 3) the countries with high incidence rates can be identified and the ranges of the observations per country are shown.

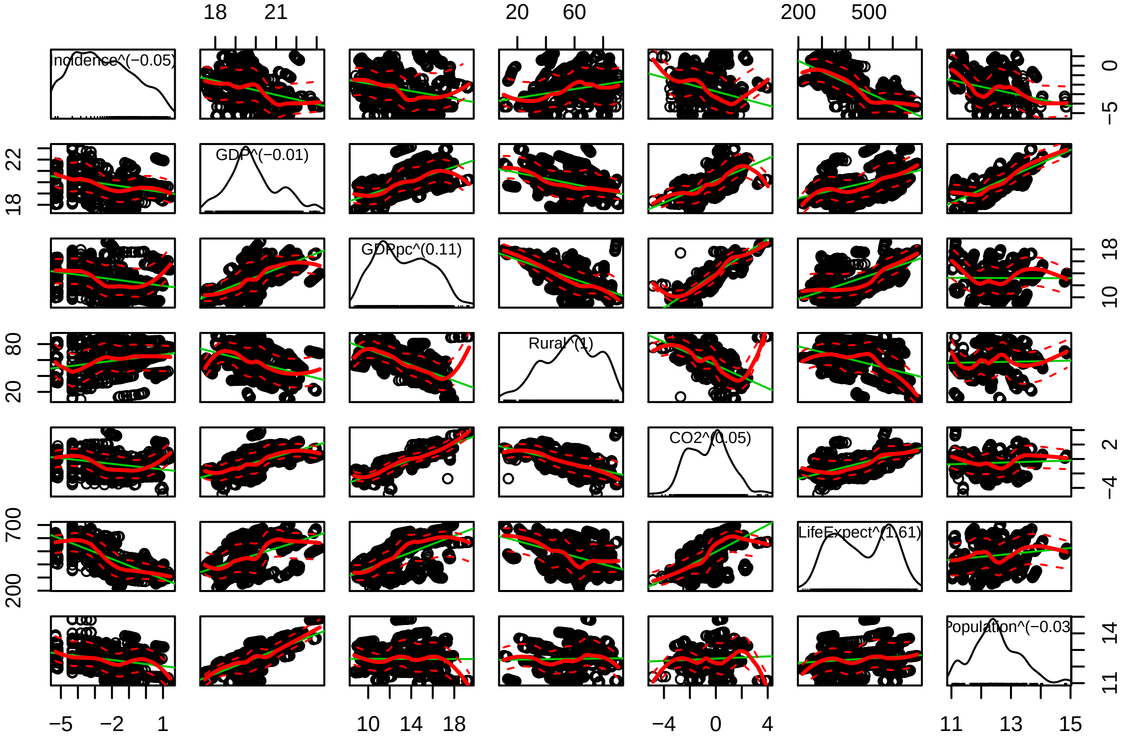
[Figure 3: Incidence Rate per Country] (/fig3.png)

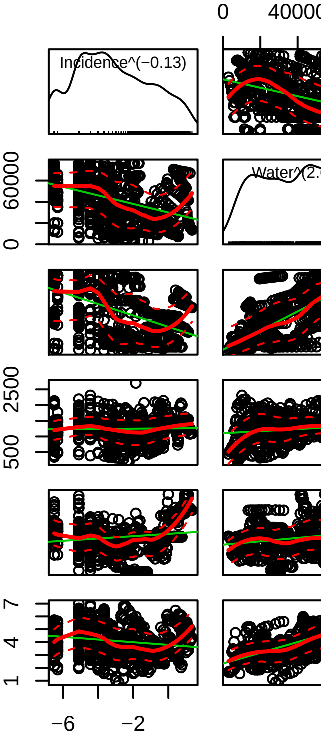
As our research question is investigating why MDG 6.A is not being reached by some countries, we are not only interested in the general HIV incidence rate but also in the decrease in the incidence rate from 2000 to 2015. As stated in the introduction Target 6.A of the MDGs specifies that countries should “have halted by 2015 and begun to reverse the spread of HIV/AIDS”" (United Nations (2014)).

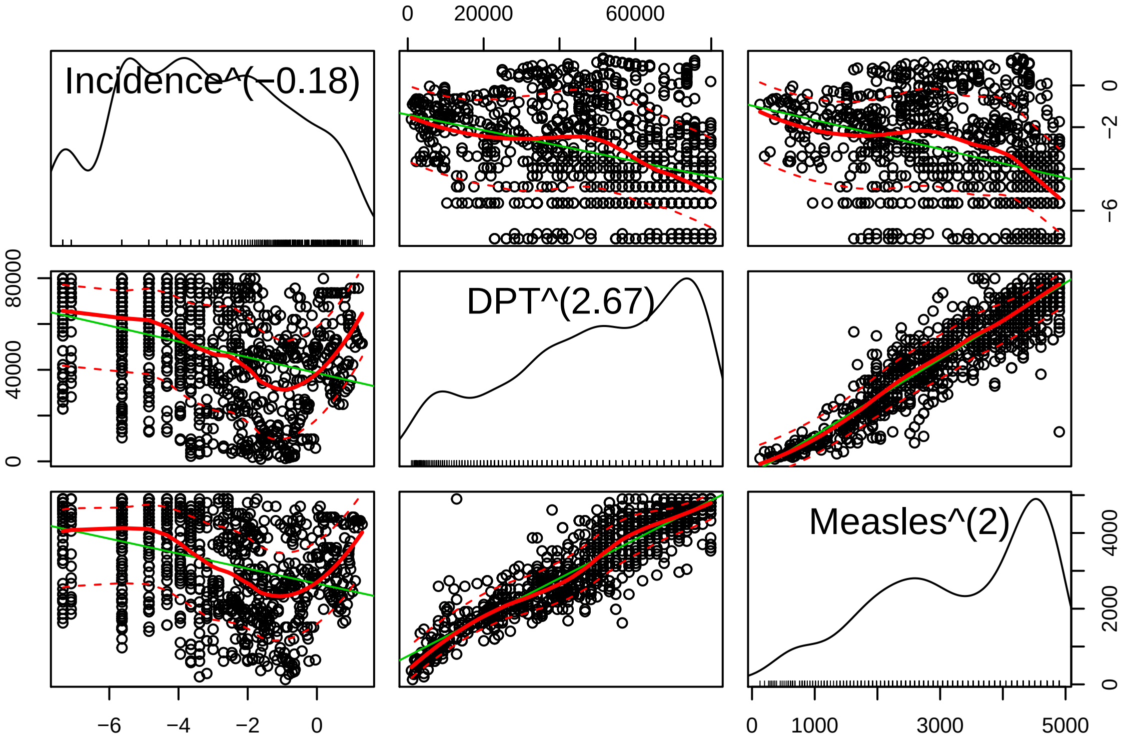
For this purpose, the dependent variable was lagged by one period and the difference between the lag and the current year was calculated. Further, a dummy variable was created assigning a value of zero for those observations where the incidence rate decreased compared to the previous year or stayed the 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 4: Incidence Rate per Country] (/fig4.png)

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







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 inferential statistics part.

As the dependent variable is coded as a dummy, being 0 if MDG 6.A was fulfilled and 1 for countries that don't fulfill MDG 6.A logistic regressions are used for the inferential statistics part. As Odds and Odds ratios are difficult to present to an audience that is not only consisting of statistical experts predicted probabilities are calculated following the logistic regressions. The interpretation of the results will only focus on the predicted probabilities.

```
##        lGDP      lGDPpc      lRural        lCO2    lHCexpend      lWater
##           2          24           2           7            3           3
## lSanitation  lUnemploym    lPrimary lHCexpendpc   lFemUnempl  lFemSchool
##           4          26          44          13           25          48
## lLifeExpect        lDPT    lMeasles
##           3           8           8
```

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.

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
##           2           2           3           1           3           3
##  lFemUnempl  lFemSchool lLifeExpect        lDPT    lMeasles
##           1           1           3           8           9
```

```
##
## Call:
## glm(formula = DDif ~ lGDPpc + lRural + lCO2 + lHCexpend + lWater +
##     lSanitation + lFemUnempl + lFemSchool + lLifeExpect + lDPT +
##     lMeasles, family = "binomial", data = Merged)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.829  -0.459  -0.397  -0.306   2.739
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8168     7.6724   -0.11   0.9152
## lGDPpc       -0.3365     0.3920   -0.86   0.3906
## lRural        0.3892     0.5344    0.73   0.4664
## lCO2          0.0211     0.2284    0.09   0.9265
## lHCexpend     0.1214     0.5185    0.23   0.8149
## lWater       -1.0915     1.0351   -1.05   0.2917
## lSanitation   1.0683     0.3880    2.75   0.0059 **
## lFemUnempl    0.0616     0.1899    0.32   0.7457
## lFemSchool    0.9319     0.8421    1.11   0.2685
## lLifeExpect  -0.8449     1.3825   -0.61   0.5411
## lDPT         -2.2652     1.5667   -1.45   0.1482
## lMeasles      2.0808     1.7479    1.19   0.2339
## ---
## Signif. codes:  0 '***' 0 '**' 0 '*' 0 '.' 0 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 365.47  on 612  degrees of freedom
## Residual deviance: 348.84  on 601  degrees of freedom
##   (375 observations deleted due to missingness)
## AIC: 372.8
##
## Number of Fisher Scoring iterations: 5
```

```
##      lGDPpc      lRural        lCO2   lHCexpend      lWater lSanitation
##           7           2           6           1           3           3
##  lFemUnempl  lFemSchool lLifeExpect        lDPT    lMeasles
##           1           1           2           8           9
```

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

```
##             2.5 % 97.5 %
## (Intercept)   -16     14
## lGDPpc         -1      0
## lRural         -1      1
## lCO2           -0      0
## lHCexpend      -1      1
## lWater         -3      1
## lSanitation     0      2
## lFemUnempl     -0      0
## lFemSchool     -1      3
## lLifeExpect    -3      2
## lDPT           -5      1
## lMeasles       -1      6
```

Exporting the Predicted Probabilities with Stargazer

```
##
## \begin{table}[!htbp] \centering
##   \caption{Logistic Regression Estimates of HIV Incidence according to
##                 Female Schooling}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & DDif \\
## \hline \\[-1.8ex]
##  (Intercept) & $-$0.00 \\
##   & (0.00) \\
##   & \\
##  lGDPpc & 0.00 \\
##   & (1.00) \\
##   & \\
##  lRural & 0.00 \\
##   & (0.00) \\
##   & \\
##  lCO2 & 0.00 \\
##   & (1.00) \\
##   & \\
##  lHCexpend & $-$1.00 \\
##   & (1.00) \\
##   & \\
##  lLifeExpect & 1.00$^{***}$ \\
##   & (0.00) \\
##   & \\
##  lLifeExpect & 0.00 \\
##   & (0.00) \\
##   & \\
##  lWater & $-$1.00 \\
##   & (1.00) \\
##   & \\
##  lSanitation & $-$2.00 \\
##   & (2.00) \\
##   & \\
##  lDPT & 2.00 \\
##   & (2.00) \\
##   & \\
##  lMeasles & 0.00 \\
##   & (1.00) \\
##   & \\
##  lFemUnempl & $-$0.00 \\
##   & (1.00) \\
##   & \\
##  1st Quartile FemSchool & 1.00 \\
##   & (0.00) \\
##   & \\
##  2nd Quartile FemSchool & 4.00 \\
##   & (8.00) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 613 \\
## Log Likelihood & $-$171.00 \\
## Akaike Inf. Crit. & 370.00 \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

If you want to automatically generate tables from regression model output objects, texreg is a good package to turn to. First estimate your models:

L1 <- glm(admit ~ gre, data = Admission, family = 'binomial')

L2 <- glm(admit ~ gre + gpa, data = Admission, family = 'binomial')

L3 <- glm(admit ~ gre + gpa + as.factor(rank), data = Admission, family = 'binomial') Then use the stargazer function to create a results table. For PDFs set type = 'latex'. There are many stylistic modifications you can make with this function.

Create cleaner covariate labels

labels <- c('(Intercept)', 'GRE Score', 'GPA Score', '2nd Ranked School', '3rd Ranked School', '4th Ranked School')

stargazer::stargazer(L1, L2, L3, covariate.labels = la title = 'Logistic Regression Estimates of Grad Sch digits = 2, type = 'latex', header = FALSE)

LIMITATIONS

The paper had to make some compromises regarding its original aims that were outlined in the first research proposal.

1. Due to huge amounts of missing data and some problems with multicollinearity lots of variables had to dropped and could ultimately not be integrated in the logistic regression models. Therefore, the categories that were supposed to be tested

are facing some limitations A simpler and more common model on the main determinants of health is the ???rainbow model???, developed by Dahlgren and Whitehead (Dahlgren and Whitehead (1991), p.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.