Analysis of Factors Influencing Life Expectancy Globally

Abstract: Our report details the best model we found to estimate life expectancy from data given on each country from 2000-2015 provided by the World Health Organization (WHO). We used a multiple linear regression model for the analysis. After determining that the HIV/AIDS variable needed to be log transformed, we employed backward selection to select the variables to best fit our model. We then studied each observation in our model through influential analysis to further investigate what countries and their corresponding years have an impact on our model. For further research we would like to study different groupings and some machine learning algorithms to find unique and possibly more accurate models to better predict average life expectancy of a country.

Introduction

The World Health Organization (WHO) established in 1948 is a specialized agency of the United Nations with the responsibility of advocating for universal healthcare and responding to health emergencies. WHO's data on various health and socioeconomic factors provides insight into issues that are affecting people globally. Through this project, we aim to find relationships between life expectancy and other health, social, and economic factors to understand if a country's average life expectancy is influenced by these factors. Creating a predictive model that best represents life expectancy can help aid physicians to understand the risks to their patients lives on a global scale. Moreover, the relationships found with this project can help WHO understand where and how to allocate their resources to increase life expectancy in different countries. We hope to bring light to factors such as schooling, preventable diseases, etc. so that we can show what's important in increasing a country's life expectancy.

Data Exploration

Our chosen dataset for this exploration and analysis was a Life Expectancy Dataset from the World Health Organization (WHO). This dataset lists 2,938 observations made up of 193 countries between the years 2000-2015. After data cleaning by removing missing values, there were 1,853 observations to base our model on. Our data includes different factors affecting life expectancy such as various diseases, economic factors, and social factors. Our response variable was life expectancy while our predictor variables were the factors:

- Alcohol: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- HealthExpenditure: General government expenditure on health as a percentage of total government expenditure (%)
- BMI: Average Body Mass Index of entire population
- HIV_AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years)
- Hepatitis B: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles: Number of reported cases for measles per 1000 population
- Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
- Schooling: Number of years of Schooling (years)
- GDP: Gross Domestic Product per capita (in USD)
- HDI_Income: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

Our initial data exploration showed that all life expectancies in the dataset had a left skewed normal distribution where the most common life expectancies were within the range 72-76. In addition, all our predictor variables appear roughly linear when plotted against life expectancy. (Appendix A)

Methods

We decided to use a multiple linear regression model to estimate the life expectancy of a country based on our data from the years 2000-2015. To model our data and determine which variables would give us the best model, we performed a backward selection methodology.

To do this we first fit a model using all our predictors. Then we looked at our summary output and determined which variables were significant using a significant level of 0.05 (Appendix B). The variables we found to be significant were then removed. In order to verify that our reduced model was a better fit, we performed an anova test to compare the full model to the reduced model. Next, we looked to see if our model had any variables with a high multicollinearity by looking at the variance inflation factors (VIF) of each variable and seeing if any were over 10.

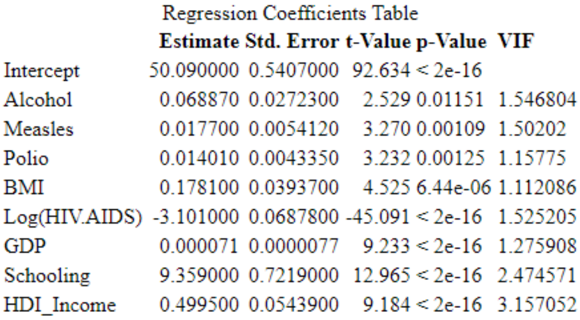
To check the adequacy of our model we then performed residual analysis. To determine this we first created a histogram of the studentized residuals of our model and a QQ plot of the studentized residuals vs t-Quantiles (Appendix C). By looking at both of these plots we could determine if our model was approximately normal and confirm our normality assumption. We then looked to see if our variance of errors was relatively equal by plotting the R-Student residuals versus our fitted values and our regressors (Appendix D). We then checked if the residuals were randomly distributed across the x-axis on each plot, and if not we determined what transformation on the regressor or the response variable was needed. If we performed a transformation on our data then we would follow the same steps above with the new data.

After we found the best model from our data we then performed Residual and Influential analysis. For our residual analysis we looked at three bar graphs of our Standardized residuals, Studentized residuals, and R-Student residuals to determine whether we had any observations that were potentially y-axis outliers and that we need to investigate further. For our influential measures we studied the Hat, Cooks, DFBETAS, DFFIT, COVRATIO measures to determine which countries and their corresponding years are potentially influential to our model and need further investigation. The Hat measure shows us which observations are leverage points. Both the Cooks and DFBETAS measures, determine each observation's influence on our regression coefficients. The DFFIT measures the amount of deletion influence on the predicted or fitted value. The COVRATIO measure determines each observation’s influence on the precision of estimation to our model.

Results

On our first fitting of the full model and then performing backward selection we found that Schooling, HIV_AIDS, Measles, Polio, BMI, GDP, and HDI_Income were significant when comparing their p-values to our significance level of 0.05 (a full summary output can be found in Appendix B). Taking our reduced model as our null hypothesis and the full model as the alternative hypothesis and running an anova test, we got a F-value of .5216 and a p-value of .6675 which is greater than a significance level of 0.05. Therefore we kept the null and accepted the reduced model as our new model. We then checked our normality assumption and concluded our model was approximately normal (Appendix C). Although, it was pointed out in the qq plot that a few observations were outliers, in particular the country Antigua and Barbuda in the years 2003 and 2004.

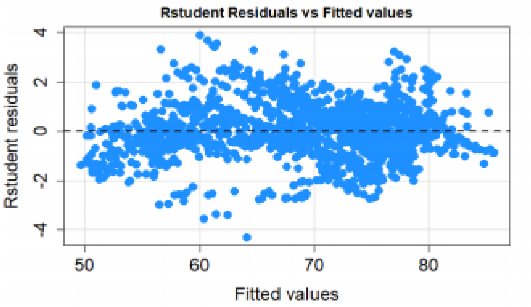
Looking at the plot of our R-Student residuals versus fitted values (Appendix D), we found that our model was right-skewed. We then looked at the R-Student residuals vs each regressor and found that HIV/AIDS formed an inverse exponential curve (Appendix F) instead of being randomly distributed. To remedy this, we applied a logarithmic transformation on the HIV/AIDS variable and refitted our model. The backward selection method found that Schooling, Log(HIV_AIDS), Measles, Polio, BMI, GDP, HDI_Income, and now Alcohol were significant. The details of our final model can be seen in the Regression Coefficients table to the right. Then, with the reduced model as the null and the full model as the alternative, we performed another ANOVA test and got an F-value of 0.66 and a significant p-value of 0.517. Thus, we accepted the reduced model as our new model.



Regression Coefficients Table

	Estimate	**Std. Error**	**t-Value**	**p-Value**	**VIF**
Intercept	50.090000	0.5407000	92.634	< 2e-16	
Alcohol	0.068870	0.0272300	2.529	0.01151	1.546804
Measles	0.017700	0.0054120	3.270	0.00109	1.50202
Polio	0.014010	0.0043350	3.232	0.00125	1.15775
BMI	0.178100	0.0393700	4.525	6.44e-06	1.112086
Log(HIV.AIDS)	-3.101000	0.0687800	-45.091	< 2e-16	1.525205
GDP	0.000071	0.0000077	9.233	< 2e-16	1.275908
Schooling	9.359000	0.7219000	12.965	< 2e-16	2.474571
HDI_Income	0.499500	0.0543900	9.184	< 2e-16	3.157052

Looking at our Regression Coefficients Table (above) we saw that we have no multicollinearity since all the VIF values are less than 10. After this we performed a residual analysis (Appendix E) and confirmed our normality assumption with no observations far outside the qq plot range of tolerance. Looking at the plot to the right, of RStudent Residuals vs Fitted Values, it can be seen that our model is randomly distributed across the x-axis. This tells us that the model is a good linear fit for our data. We can also see in the R-Student residual versus Regressors (Appendix F) that all are randomly distributed, confirming our models linearity. From this analysis, we determined the final model that best represents the WHO data set. The linear equation of the final model is:



Life Expectancy = 50.090000 + (0.068870)Alcohol + (0.017700)Measles + (0.014010)Polio + (0.178100)BMI - (3.101)log(HIV/AIDS) + (0.000071)GDP + (9.359)Schooling + (0.499500)HDI_Income

Lastly, we looked for potential y-axis outliers and found that according to the Standardized and Studentized residual graphs (Appendix G), there were 14 potentially influential outliers (Appendix H). However, from the R-Student residual graph (Appendix G), we only found Sierra Leone 2014 to be potentially influential. Next, looking at our influential measures, we first found 59 observations (Appendix I) to be potentially influential from the Hat measure. We then found that no observations were potentially influential on our model's regression coefficients according to both the Cooks and DFBETAS measure. From our DFFIT measure (Appendix J) there were 37 potentially influential observations on the amount of deletion influence they had on the fitted values. Our last influential measure, the COVRATIO measure (Appendix K), found 169 potentially influential observations on the overall precision of estimation for our model. For each of these potentially influential observations further investigation is needed.

Conclusion and Discussion

Our regression analysis finds that a country's life expectancy is best predicted by its alcohol consumption per capita, measles case rate, polio immunization rate, average BMI, GDP per capita, average education level, Human Development Index in terms of its income composition of resources, and the natural log of its death rate due to HIV/AIDS. Specifically, if a country is looking to increase its life expectancy, it should focus on growing its economy, educating its citizens, using its resources more productively, immunizing its population against polio, and encouraging its citizens to drink less alcohol.

Our influential analysis found several countries with multiple years of data that were potentially influential on our model, and further study is required to ascertain why exactly that is the case for each observation. For example, Afghanistan remained potentially influential from 2000-2014 to the overall precision of estimation for our model, which stands out from other countries. This could be explained by the decades-long war it has been in, which must have had a detrimental effect on its life expectancy. However, this is simply speculation, and further investigation is needed to understand the reason for this anomaly in our data. One path we could take is comparing Afghanistan's life expectancy data from before and after the war and seeing if there was a significant difference in life expectancy.

This is not the only area where further study could be undertaken. We also would like to study more data from before and after 2000-2015 to possibly create a better model to calculate a country's average life expectancy. Additionally, we could explore some ways to create an even better-fitting model than the one we have, by grouping different countries together (by region, developing vs. developed status, etc) and seeing which grouping results in the best model. Another method such as k-means clustering or something similar could be performed and observe if that results in new and interesting ways to better organize our data.

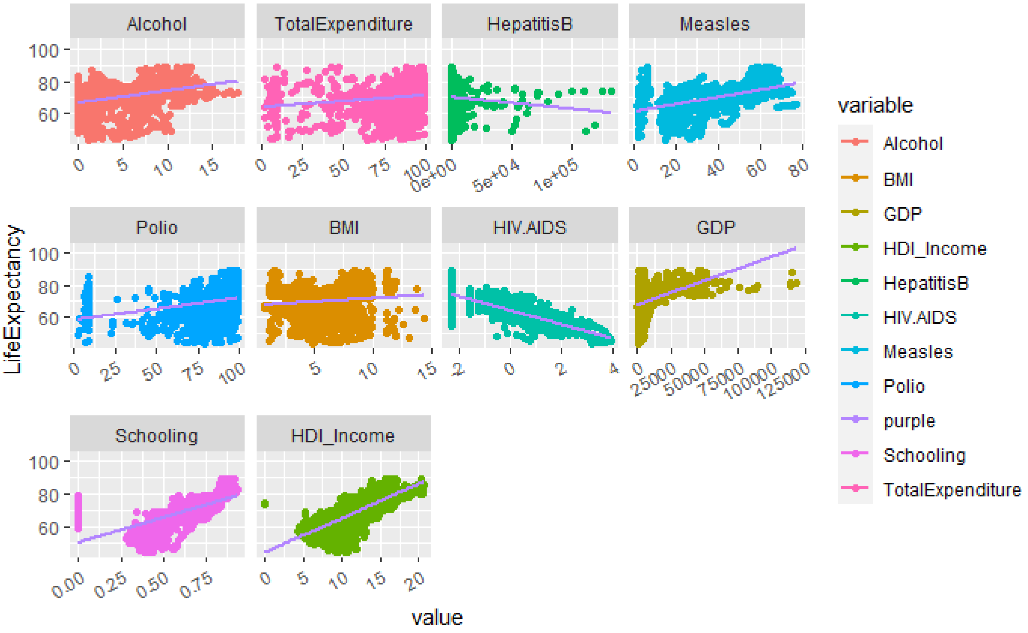
References

Rajarshi, Kumar. “Life Expectancy (WHO).” Kaggle, 10 Feb. 2018, www.kaggle.com/kumarajarshi/life-expectancy-who.

“What We Do.” *World Health Organization*, World Health Organization, www.who.int/about/what-we-do.

Appendix

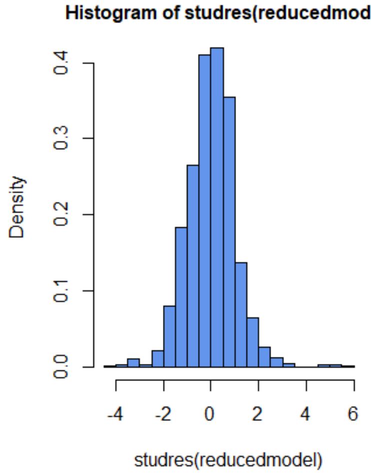
Appendix A: Graphs of Life Expectancy versus each predictor after transforming HIV/AIDS by a natural log.

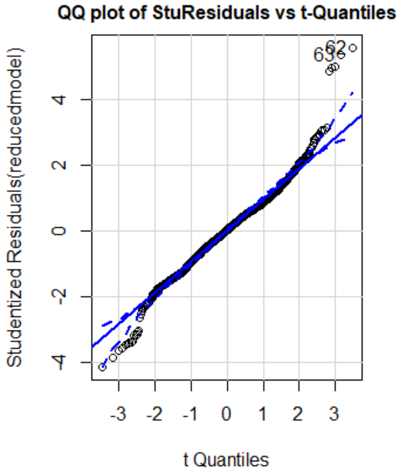


Appendix B: Summary output of the full model Pre-Transformation. From this we can see that Total Expenditure, Alcohol, and Hepatitis B were not significant and can exclude them from our model for now to create a reduced model which we compare to the full model here using an anova test.

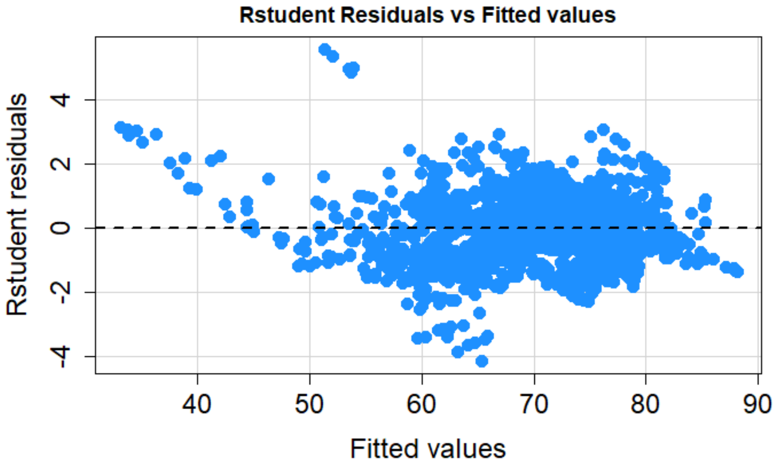
```
Call:
lm(formula = y ~ xTotalExpenditure + xSchooling + xHIV_AIDS +
    xAlcohol + xHepatitisB + xMeasles + xPolio + xBMI + xGDP +
    xHDI_Income, data = life)

Residuals:
     Min       1Q   Median       3Q      Max
-17.3650  -2.5226   0.1088   2.5485  23.4122

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.728e+01  6.098e-01  77.522  < 2e-16 ***
xTotalExpenditure  3.067e-03  4.477e-03   0.685  0.49338
xSchooling         1.020e+01  8.035e-01  12.692  < 2e-16 ***
xHIV_AIDS         -6.407e-01  1.791e-02 -35.774  < 2e-16 ***
xAlcohol          -2.902e-02  3.011e-02  -0.964  0.33527
xHepatitisB        4.096e-06  1.046e-05   0.392  0.69534
xMeasles           4.782e-02  5.956e-03   8.029 1.74e-15 ***
xPolio             2.932e-02  5.326e-03   5.504 4.23e-08 ***
xBMI               1.209e-01  4.384e-02   2.758  0.00588 **
xGDP               7.997e-05  8.563e-06   9.339  < 2e-16 ***
xHDI_Income        9.426e-01  5.910e-02  15.948  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.189 on 1842 degrees of freedom
Multiple R-squared:  0.7647,    Adjusted R-squared:  0.7634
F-statistic: 598.5 on 10 and 1842 DF,  p-value: < 2.2e-16
```

Appendix C: Both the histogram and QQ plot support the normality assumption, but Antigua and Barbuda in 2003 and 2004 were found to be outside the norm.

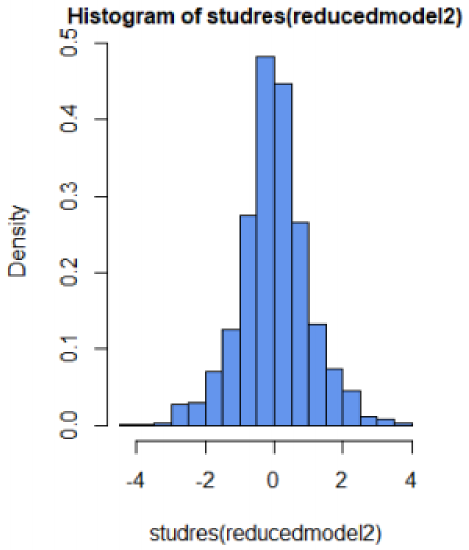


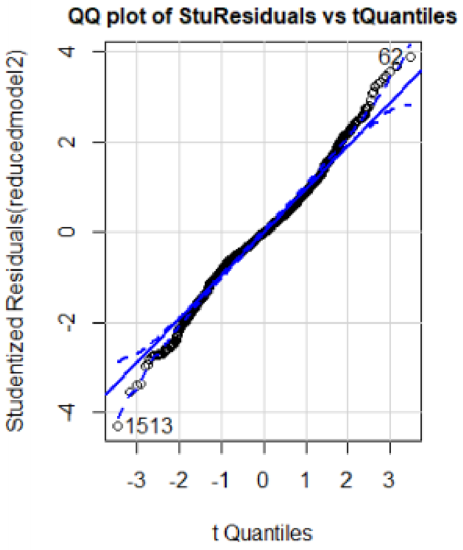


Appendix D: The plot of our model before our transformation is bunched a little closer to the right but other than that it looks randomly distributed.

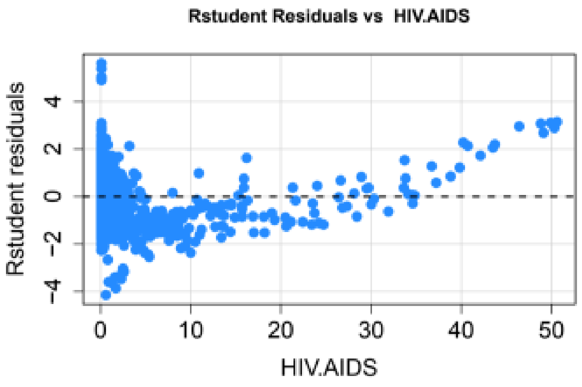


Appendix E: The Histogram supports our normality assumption, and the QQ plot is approximately normal also further supporting our normality assumption.

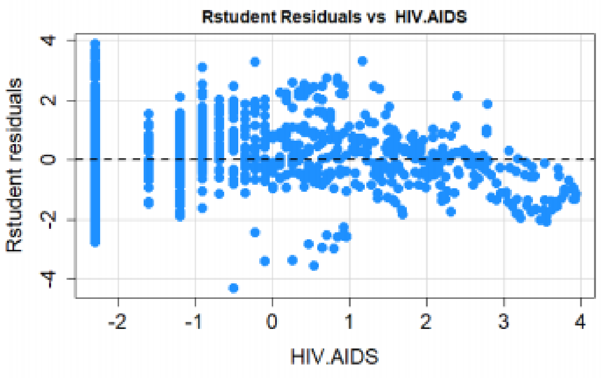




Appendix F: Plots of the R-Student Residual vs the regressor HIV/AIDS before and after transforming HIV/AIDS by a natural log and refitting a new model with the transformed variable.

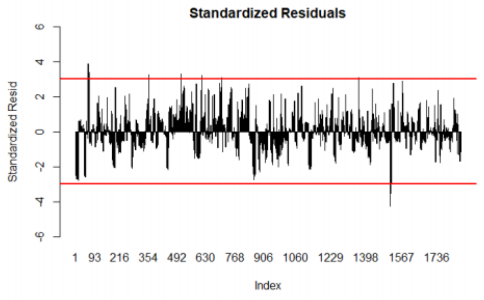


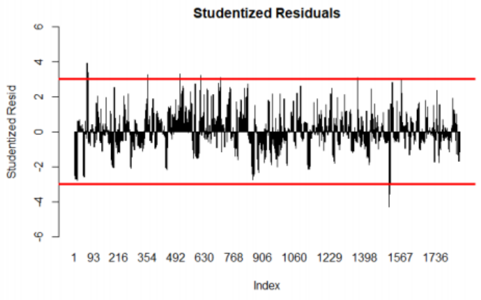
Residuals vs. HIV/AIDS **Pre-Transformation**

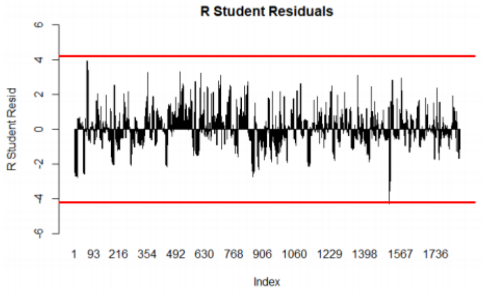


Residuals vs. HIV/AIDS **Post-Transformation**

Appendix G: The potential y-axis outliers found in the Standardized and Studentized Residuals can be identified in Appendix H. For the R-Student Residual plot the only y axis outlier was Sierra Leone 2014 that barely crosses the cutoff.







Appendix H: The table below contains the 14 observations found by the Standardized and Studentized residual plots to be potential y-axis outliers.

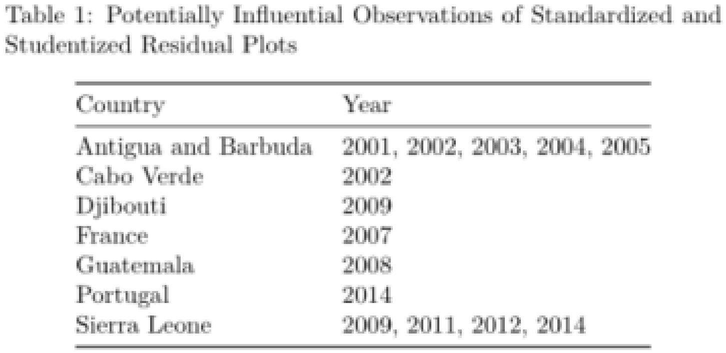


Table 1: Potentially Influential Observations of Standardized and Studentized Residual Plots

Country	Year
Antigua and Barbuda	2001, 2002, 2003, 2004, 2005
Cabo Verde	2002
Djibouti	2009
France	2007
Guatemala	2008
Portugal	2014
Sierra Leone	2009, 2011, 2012, 2014

Appendix I: Hat measures of all 37 potentially influential observations to investigate further.

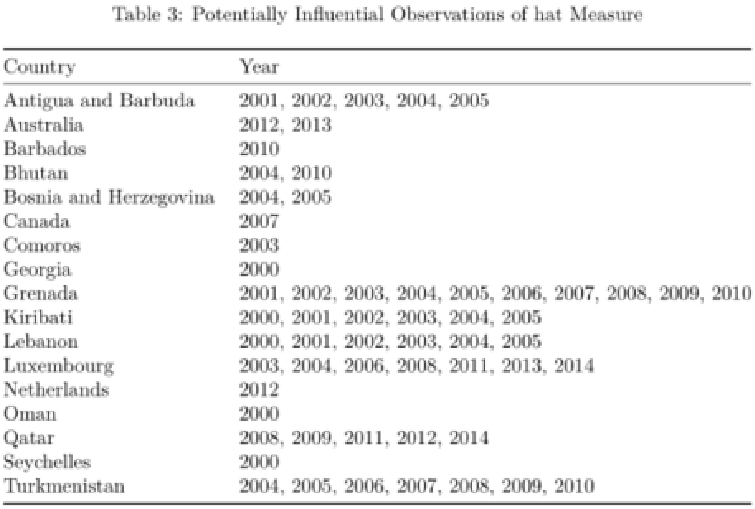


Table 3: Potentially Influential Observations of hat Measure

Country	Year
Antigua and Barbuda	2001, 2002, 2003, 2004, 2005
Australia	2012, 2013
Barbados	2010
Bhutan	2004, 2010
Bosnia and Herzegovina	2004, 2005
Canada	2007
Comoros	2003
Georgia	2000
Grenada	2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010
Kiribati	2000, 2001, 2002, 2003, 2004, 2005
Lebanon	2000, 2001, 2002, 2003, 2004, 2005
Luxembourg	2003, 2004, 2006, 2008, 2011, 2013, 2014
Netherlands	2012
Oman	2000
Qatar	2008, 2009, 2011, 2012, 2014
Seychelles	2000
Turkmenistan	2004, 2005, 2006, 2007, 2008, 2009, 2010

Appendix J: Table of the 37 potentially influential observations found by the DFFIT Measure to investigate further.

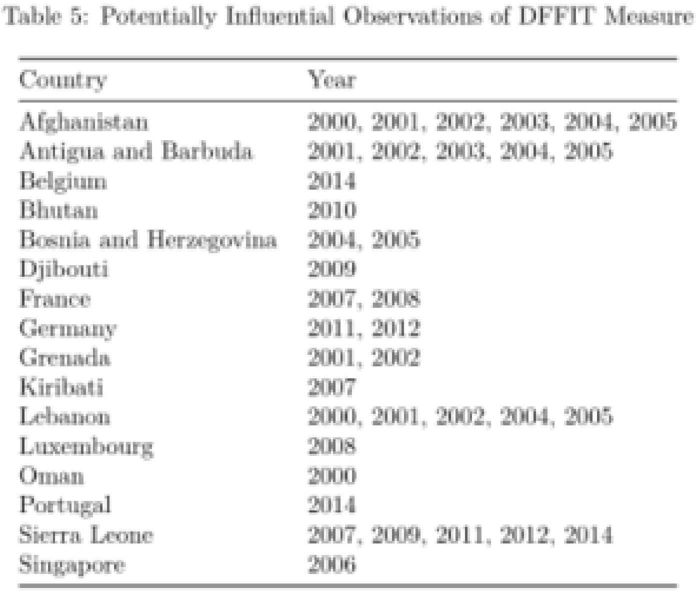


Table 5: Potentially Influential Observations of DFFIT Measure

Country	Year
Afghanistan	2000, 2001, 2002, 2003, 2004, 2005
Antigua and Barbuda	2001, 2002, 2003, 2004, 2005
Belgium	2014
Bhutan	2010
Bosnia and Herzegovina	2004, 2005
Djibouti	2009
France	2007, 2008
Germany	2011, 2012
Grenada	2001, 2002
Kiribati	2007
Lebanon	2000, 2001, 2002, 2004, 2005
Luxembourg	2008
Oman	2000
Portugal	2014
Sierra Leone	2007, 2009, 2011, 2012, 2014
Singapore	2006

Appendix K: The table for COVRATIO Measure of which 169 observations were found to be potentially influential and need further investigation.

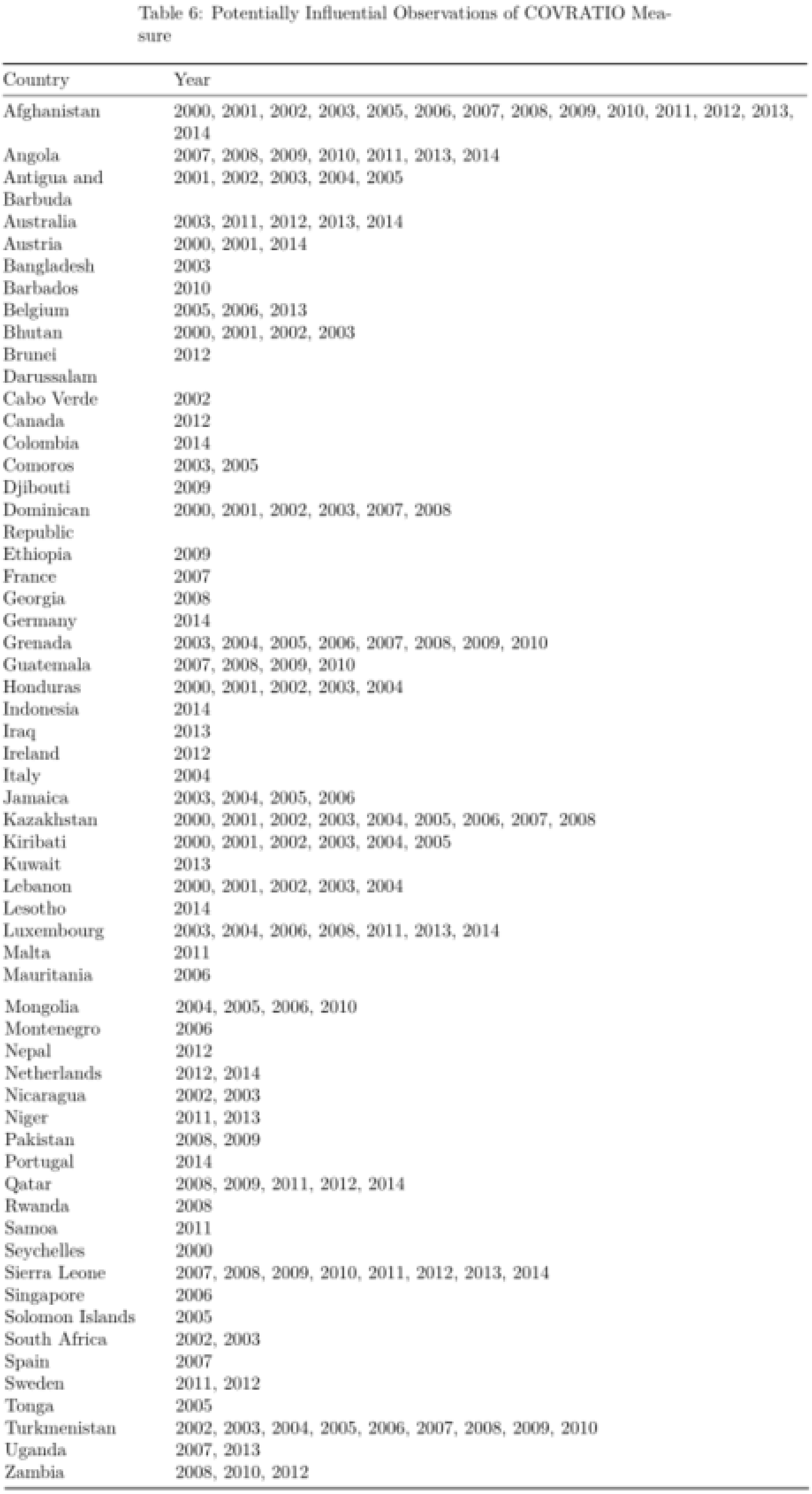


Table 6: Potentially Influential Observations of COVRATIO Measure

Country	Year
Afghanistan	2000, 2001, 2002, 2003, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014
Angola	2007, 2008, 2009, 2010, 2011, 2013, 2014
Antigua and Barbuda	2001, 2002, 2003, 2004, 2005
Australia	2003, 2011, 2012, 2013, 2014
Austria	2000, 2001, 2014
Bangladesh	2003
Barbados	2010
Belgium	2005, 2006, 2013
Bhutan	2000, 2001, 2002, 2003
Brunei Darussalam	2012
Cabo Verde	2002
Canada	2012
Colombia	2014
Comoros	2003, 2005
Djibouti	2009
Dominican Republic	2000, 2001, 2002, 2003, 2007, 2008
Ethiopia	2009
France	2007
Georgia	2008
Germany	2014
Grenada	2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010
Guatemala	2007, 2008, 2009, 2010
Honduras	2000, 2001, 2002, 2003, 2004
Indonesia	2014
Iraq	2013
Ireland	2012
Italy	2004
Jamaica	2003, 2004, 2005, 2006
Kazakhstan	2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008
Kiribati	2000, 2001, 2002, 2003, 2004, 2005
Kuwait	2013
Lebanon	2000, 2001, 2002, 2003, 2004
Lesotho	2014
Luxembourg	2003, 2004, 2006, 2008, 2011, 2013, 2014
Malta	2011
Mauritania	2006
Mongolia	2004, 2005, 2006, 2010
Montenegro	2006
Nepal	2012
Netherlands	2012, 2014
Nicaragua	2002, 2003
Niger	2011, 2013
Pakistan	2008, 2009
Portugal	2014
Qatar	2008, 2009, 2011, 2012, 2014
Rwanda	2008
Samoa	2011
Seychelles	2000
Sierra Leone	2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014
Singapore	2006
Solomon Islands	2005
South Africa	2002, 2003
Spain	2007
Sweden	2011, 2012
Tonga	2005
Turkmenistan	2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010
Uganda	2007, 2013
Zambia	2008, 2010, 2012