Socioeconomic factors on HIV/AIDS:

a cross-country comparative study

11812532 Luo Yiling

Abstract

People and countries are paying more and more attention to AIDS in recent years, thus an efficient and reliable method to evaluate the aid's rate with the countries' other factors is crucial to this situation.

To ensure the reliability of our model and simplify the modeling process, we detect the missing data at the very beginning. With a very short data visualization, it's clear to see some basic parameters and the distribution of the factors.

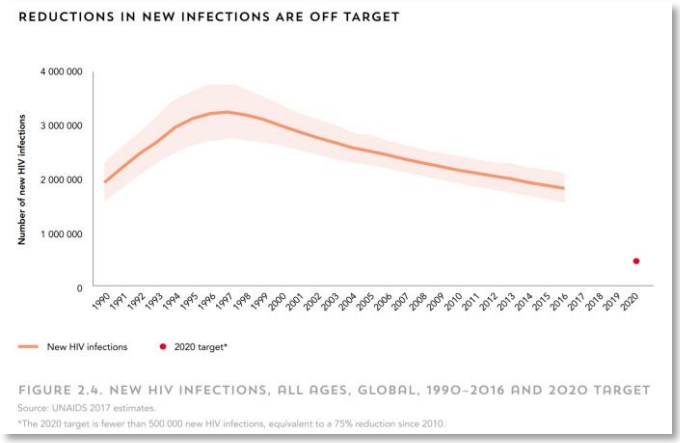
When building models, firstly, using full factors, by some tests and plots it's better to take LOG transformation, so we take LOG and fit the first full model. After detecting the multicollinearity and condition indexes, we delete some highly correlated factors. Then we use stepwise method to find the reduced model, and test its degree of significance and VIFs. Then test its residuals, linearity, homoscedasticity, independence and the outliers.

Based on our model and analysis, we can study some extension problems such as the connection between highly correlated factors, the common of the highest aid's rate countries and the relationship between the percentage the government spend on health and aids. Also, we can get from the model that the higher the agriculture is, the lower the fertility is and the higher the urbanPop is, the lower the rate is, which have some practical significance.

1 Introduction

1.1 Background

With the development of health consciousness, people caring more on some widespread diseases such as AIDS. In order to arise the public's awareness and help to reduce the infectious rate(the plot of new HIV infections number are shown above[1]), experts collect the data from different countries and make a dataset of the prevalence rate and many other factors among these countries (given in *data.xlsx*), willing to find the relation and correlation between these factors, using some model to explain the variables and coming out some useful advices to the people.



Graph 1: New HIV Infections[1]

1.2 Questions of interests

1. What's the best model to fit the data?
2. Which factors have significant contribution to the countries' aid rate?
3. What's the common of the country which has high aid rate?

2 Missing Data Detection

We first see if this data has missing points. After using function *is.na(data)*, we can see that data don't have missing point.

3 Data Visualization

3.1 summary of the data

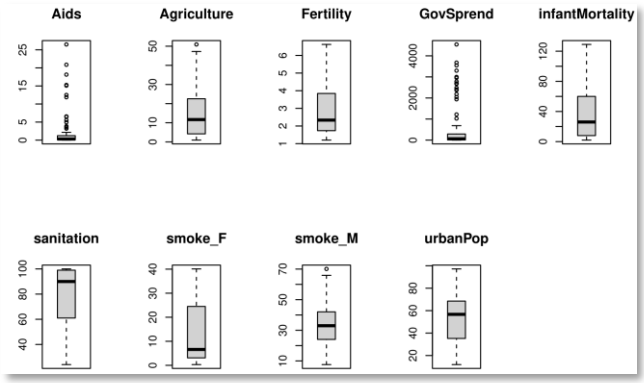
Country	Aids	agriculture	fertility	GovSpend	infantMortality
Length:95	Min. : 0.010	Min. : 0.9514	Min. :1.201	Min. : 1.0	Min. : 2.00
Class :character	1st Qu.: 0.090	1st Qu.: 4.2046	1st Qu.:1.736	1st Qu.: 11.0	1st Qu.: 8.00
Mode :character	Median : 0.310	Median :11.6700	Median :2.334	Median : 64.0	Median : 26.00
	Mean : 2.078	Mean :15.1005	Mean :2.912	Mean : 573.3	Mean : 36.94
	3rd Qu.: 1.235	3rd Qu.:22.5465	3rd Qu.:3.845	3rd Qu.: 291.0	3rd Qu.: 60.00
	Max. :26.490	Max. :50.9353	Max. :6.623	Max. :4542.0	Max. :129.00

sanitation	smoke_F	smoke_M	urbanPop
Min. : 24.00	Min. : 0.30	Min. : 7.60	Min. :11.92
1st Qu.: 61.00	1st Qu.: 3.10	1st Qu.:24.05	1st Qu.:35.22
Median : 90.00	Median : 6.60	Median :33.00	Median :56.68
Mean : 80.01	Mean :12.77	Mean :33.93	Mean :53.43
3rd Qu.: 99.00	3rd Qu.:24.50	3rd Qu.:42.05	3rd Qu.:68.51
Max. :100.00	Max. :40.10	Max. :70.10	Max. :97.18

Table 1: summary of the data

From the summary, we can see that there are 95 countries in total, the median and mean of the percentage of aids are 0.310 and 2.078, and other factors' median/mean/minimum/ maximum and two quantiles are shown in **Table 1**.

3.2 Boxplot of all factors



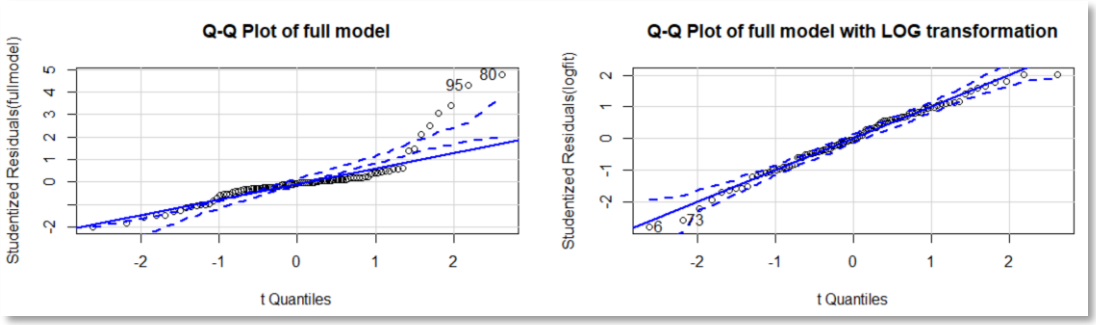
Graph 2: Boxplot of all factors

4 Linear Model Building Process

4.1 Full Model

4.1.1 Full Model Building & Testing Transforming Method

We use all factors to construct a linear model. First, we should find out if we should take some transform to the model. In this case, we try Q-Q plot for the normality, and try spreadLevelPlot() for the homoscedasticity, and it shows that **the suggested power transformation is -0.007421982**, in order to simplify the model, we take **LOG transform** to the full model.



Graph 3: Q-Q plot of full model and LOG transformation

4.1.2 LOG Full Model Building

```
Call:
lm(formula = log(Aids) ~ (agriculture + fertility + GovSpend +
    infantMortality + sanitation + smoke_F + smoke_M + urbanPop))

Residuals:
    Min      1Q  Median      3Q     Max
-3.5738 -0.8069 -0.0499  0.9358  2.7489

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.2915829  1.7297883  -0.169 0.866534
agriculture     -0.0777436  0.0193399  -4.020 0.000124 ***
fertility        0.5633724  0.2652386   2.124 0.036540 *
GovSpend        -0.0002359  0.0001940  -1.216 0.227237
infantMortality  0.0168222  0.0114114   1.474 0.144091
sanitation      -0.0066838  0.0132212  -0.506 0.614476
smoke_F          0.0236865  0.0190162   1.246 0.216295
smoke_M         -0.0188785  0.0155143  -1.217 0.226993
urbanPop        -0.0146703  0.0102646  -1.429 0.156566
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.403 on 86 degrees of freedom
Multiple R-squared:  0.5044,    Adjusted R-squared:  0.4583
F-statistic: 10.94 on 8 and 86 DF,  p-value: 1.62e-10
```

Table 2: summary of LOG full model

From **Table 2**, as **H_0 means all the coefficients of the factors (except the intercept) are 0**, and **H_1 means at least one factor has its coefficient not 0**. we discover that p-value is 1.62×10^{-10}, so we **reject null hypothesis** and the model is meaningful. Then we see the P-value of each factor, we can find *agriculture and fertility* has small P-value(<0.05) thus is significant. But as we should test the multicollinearity, so these factors may not be the final one.

4.1.3 LOG Full Model Testing

As some of the factors are not very significant, so we first turn to see the pairwise correlations of all factors.

	agriculture	fertility	GovSpend	infantMortality	sanitation	smoke_F	smoke_M	urbanPop
agriculture	1.0000	0.7105	-0.5153	0.7047	-0.6834	-0.5077	-0.2398	-0.6906
fertility	0.7105	1.0000	-0.4058	0.9077	-0.8218	-0.5320	-0.6233	-0.6142
GovSpend	-0.5153	-0.4058	1.0000	-0.4856	0.4659	0.6136	0.0142	0.5367
infantMortality	0.7047	0.9077	-0.4856	1.0000	-0.8459	-0.5781	-0.5267	-0.6585
sanitation	-0.6834	-0.8218	0.4659	-0.8459	1.0000	0.4982	0.4531	0.6376
smoke_F	-0.5077	-0.5320	0.6136	-0.5781	0.4982	1.0000	0.3494	0.6135
smoke_M	-0.2398	-0.6233	0.0142	-0.5267	0.4531	0.3494	1.0000	0.2266
urbanPop	-0.6906	-0.6142	0.5367	-0.6585	0.6376	0.6135	0.2266	1.0000

Table 3: pairwise correlation

From the **Table 3**, we can see that some factor such as **infantMortality&fertility** and **fertility&sanitation** has very large pairwise correlation, so we define that this model has multicollinearity, also we have very large condition numbers. So, we need to drop several highly

correlated independent variables to fit the model better. In this case, we choose the correlation which are larger than 0.8, delete infantMortality and sanitation.

```
> e$val
[1] 1.399393e+08 8.584923e+05 1.337407e+05 2.126176e+04 1.208868e+04 5.641298e+03 4.781632e+03 3.664983e+01
> sqrt(e$val[1] / e$val)
[1]    1.00000   12.76738   32.34730   81.12790  107.59211  157.49992  171.07311 1954.04171
```

Table 4: condition indexes

4.2 Reduced Model

4.2.1 Reduced Model Building (Stepwise)

We use stepwise method to find the model with the smallest AIC. From the output of stepwise method, we can see that the factors remain in the reduced model are agriculture, fertility and urbanPop.

```
Step:  AIC=69.7
log(Aids) ~ agriculture + fertility + urbanPop

             Df Sum of Sq    RSS     AIC
<none>                    181.89  69.703
+ GovSpend    1     1.754 180.13  70.782
- urbanPop    1     6.556 188.44  71.067
+ smoke_M     1     0.951 180.94  71.205
+ smoke_F     1     0.274 181.61  71.560
- agriculture 1    34.917 216.80  84.386
- fertility   1   113.872 295.76 113.888

Call:
lm(formula = log(Aids) ~ agriculture + fertility + urbanPop)

Coefficients:
(Intercept)  agriculture    fertility     urbanPop
   -1.95120     -0.07666      1.03209     -0.01670
```

Table 5: Stepwise method

```
Call:
lm(formula = log(Aids) ~ agriculture + fertility + urbanPop)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8345 -0.8763 -0.0973  1.0877  3.2653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.95120    0.77995  -2.502   0.0141 *
agriculture -0.07666    0.01834  -4.180 6.70e-05 ***
fertility    1.03209    0.13674   7.548 3.27e-11 ***
urbanPop    -0.01670    0.00922  -1.811   0.0734 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.414 on 91 degrees of freedom
Multiple R-squared:  0.4673,    Adjusted R-squared:  0.4497
F-statistic: 26.61 on 3 and 91 DF,  p-value: 1.91e-12
```

Table 6: summary of final model

4.2.2 Reduced Model Testing

We then use summay() to see that p-value of the model is 1.91×10^{-12}, so the model is meaningful, and three factors in the model is significant(6.70×10^{-5}, 3.27×10^{-11}, 0.0734 respectively) And we can test the VIF of three factors, which the result shows that all VIF is smaller than 10 (2.554701, 2.145796, 2.031220 respectively), so they can be explained individually. And the R^2 is 0.4673 which means about 46.73% of variation of the aids has been explained by the model, which have no extreme difference compared with the full one(50.44%), so the model is well.

The final model is : $\boldsymbol{E(y) = exp(-1.95120 - 0.07666x_1 + 1.03209x_2 - 0.01670x_3)}$,

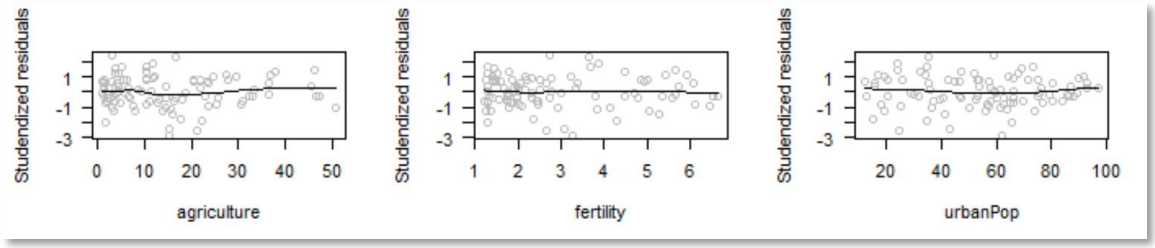
where x_1 is agriculture, x_2 is fertility, x_3 is urbanPop.

5 Model Diagnostic

If the model is well-defined, it should suit some properties such as linearity, homoscedasticity, independence, distribution and lack of outliers.

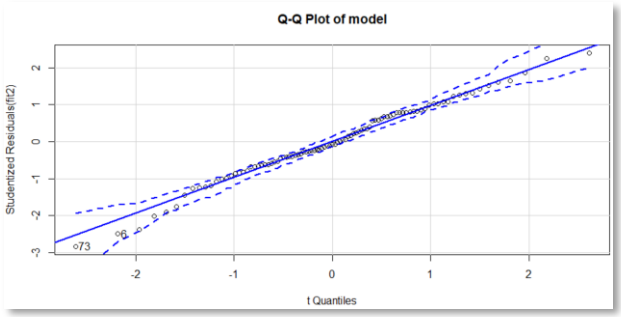
5.1 Testing Residuals

The graphs show the residual plots against each of the predictor variables, from the plots, we can see that the residuals are all around 0 and the fitted line nearly 0 and nearly straight, so the model is suitable.



Graph 4: Residual Plots against each of the predictor variables

5.2 Normality



Graph 5: Q-Q Plots of reduced model

If the residual values are met the normality assumption, the points on the Q-Q Plot should fall on the straight 45-degree line. The plot shows nearly all residuals fell in the 95%CI interval, also suit the normality.

5.3 Homoscedasticity

We use non-constant error variance test, the result shows that is nonsignificant (p-value is 0.28225), suggesting that we've met the constant variance assumption.

```
> ncvTest(fit2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.156242, Df = 1, p = 0.28225
```

Table 7: non-constant error variance test

```
> durbinWatsonTest(fit2)
 lag Autocorrelation D-W Statistic p-value
   1      -0.0858465      2.111686   0.598
 Alternative hypothesis: rho != 0
```

Table 8: Durbin-Watson test

5.4 Independence

We apply Durbin-Watson test, the nonsignificant p-value (0.598) suggests a lack of autocorrelation, and conversely an independence of errors.

5.5 Outliers

As outliers are different than other observations in some way or they exert disproportionate amount of influence on the results, so we need to use outlierTest() to test them.

The result shows that *No Studentized residuals with Bonferroni p < 0.05,* as the null hypothesis of the test is it's an outlier, as all point refuse the null hypothesis, so no outlier. (Also the largest studentized residual is -2.84 < |3|, also means no outlier.)

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferroni p <dbl>
73	-2.837661	0.0056152	0.53345

Table 9: Test of outliers

Above all, the model satisfies the properties.

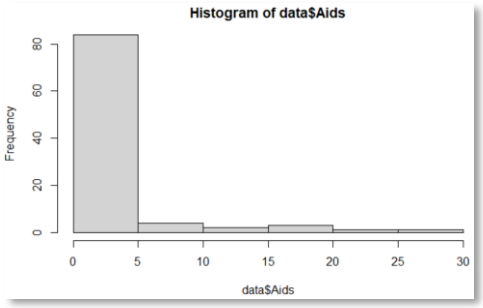
As the final model is $\boldsymbol{E(y) = exp(-1.95120 - 0.07666x_1 + 1.03209x_2 - 0.01670x_3)}$, so we know that the higher the agriculture is, the lower the fertility is and the higher the urbanPop is, the lower the rate is, which is also fit in with our ideas.

After this basic conclusion, we furthermore come to see some extension questions.

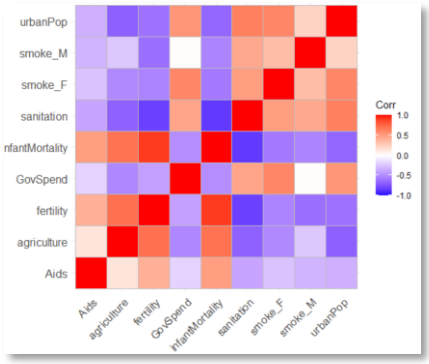
6 Extension Questions

6.1 Discussion of taking LOG transform

It's interesting that it should take log transform. But as it's shown in **Graph 6**, we can see that although much of the countries' aids rate are lied in a small interval, some countries have very high level of rate thus turn to a very big deviation.



Graph 6: histogram of aids



Graph 7: correlation plot of all factors

6.2 Correlation between factors

Let's see **Table 3 and Graph 7** the pairwise correlation of all factors. There are some interesting facts in the correlation table, thus we discuss two of them.

As we mentioned above, fertility, infantMortality and sanitation have very high pairwise correlation. These three factors are about child and infrastructure, which relate to the basic life. The matrix shows that per woman has more child, then number of children who dies before one year old is higher, which then indicate the woman give birth to even more children. And also, higher death rate and the lower adequate access to excreta disposal facilities are highly correlated, which shows the degree of their standard of living. So, for developing countries compared to developed countries are more likely to give birth to child and have lower living standard, so they are highly correlated and increase the aids' rate, which worsen the child's living condition.[2]

Also, an interesting fact is that the growing general government expenditure on health as percentage of total government expenditure will increase the women's smoking percentage to some extent (the correlation between two factors is 0.6136), but on the other hand, it has no influence with the men's smoking percentage (0,0142).

6.3 Discussion of high-AIDS rate countries

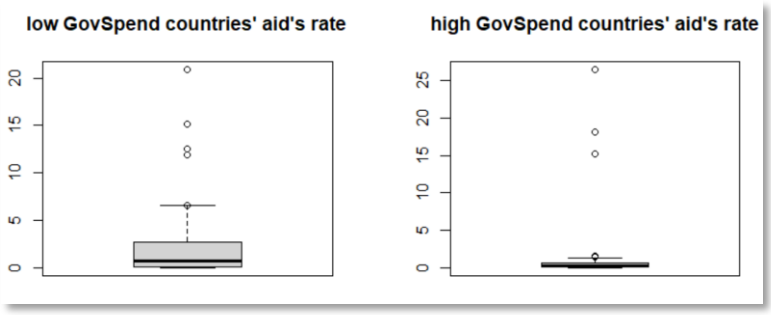
Country <chr>	Aids <dbl>	agriculture <dbl>	fertility <dbl>	GovSpend <dbl>	infantMortality <dbl>	sanitation <dbl>	smoke_F <dbl>	smoke_M <dbl>	urbanPop <dbl>
Swaziland	26.49000000	11.6606836	3.8680	97	110	59	3.2	14.6	23.94
Zimbabwe	20.87000000	16.7421875	3.6570	12	68	63	4.4	25.5	35.48
South Africa	18.14000000	3.1448629	2.7200	156	55	79	9.1	27.5	58.82
Namibia	15.28000000	10.3863325	3.6750	140	46	50	10.9	38.6	34.56
Zambia	15.09000000	23.4172478	6.0610	18	102	59	5.0	21.7	34.96

Table 10: high-AIDS rate countries

From the top-five rank list, we can see that all countries are from Africa, which indicate that compared to other continents, Africa has more serious problem with AIDS. “Current suggests that 60 percent of those infected with HIV may live in Africa… in some hospitals it may be the most common cause of death among adults in medical wards.” [3] And compared other factors in **Table 1**, we can see that except for South Africa, other four countries’ each level are fall behind the 1st quantile, falling behind the world average level.

6.4 The relationship between GovSpend and Aids

As we are curing about if the government spend on health is related to aid’s rate. Use these results and separate our data into two groups (one is lower than median 64.0 and the other is greater than 64.0), renamed as **low GovSpend** and **high GovSpend**, in order to see if aids’ rate have difference among these two groups (As GovSpend is not a significant factor in the final model, so it’s meaningful to do this split).



Graph 8: boxplot of low and high GovSpend countries’ aid’s rate

From the scale and boxplot, it seems that low GovSpend countries’ aid’s rate has much wide deviation compared to high GovSpend countries. Then we want to see more clearly about the mean,

as these two groups' distribution is not strictly as normal distribution and the observations is not big, so we use bootstrap ANOVA to see it these two groups have same mean (346 iterations were required to reach the stopping rule). The result shows that p-value is 0.2254, and thus we cannot reject the null that they have same mean, so it means there are no difference between different degrees of government spend on health.

Component 1 :

	Df	R Sum Sq	R Mean Sq	Iter	Pr(Prob)
GovSpendDifference	1	20.28	20.285	346	0.2254
Residuals	93	2054.25	22.089		

Table 11: Bootstrap ANOVA

7 Conclusion

The final model is: $\boldsymbol{E(y) = exp(-1.95120 - 0.07666x_1 + 1.03209x_2 - 0.01670x_3)}$, where $\boldsymbol{x_1}$ is agriculture, $\boldsymbol{x_2}$ is fertility, $\boldsymbol{x_3}$ is urbanPop. This model satisfies some assumptions discussed above which means it's meaningful and has some practical significance. These three factors occupy an important significance and showing that the higher the agriculture is, the lower the fertility is and the higher the urbanPop is, the lower the rate is.

Reference

[1] Ending AIDS progress towards the 90-90-90 targets. Global AIDS update, 2017. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS; 2017. Retrieved from http://www.unaids.org/sites/default/files/media_asset/Global_AIDS_update_2017_en.pdf

[2] https://www.humanium.org/en/children-world/

[3] Wilkins, H. A. (1992). HIV infection in africa -- AIDS in africa edited by peter piot, bila kapita and J. B. O. were / AIDS in africa: Its present and future impact by tony barnett and piers blaikie. Nature, 356(6368), 393. Retrieved from https://search.proquest.com/scholarly-journals/hiv-infection-africa-aids-edited-peter-piot-bila/docview/204418381/se-2?accountid=162699

[4] Leeson, G. W. (2018). The growth, ageing and urbanisation of our world. Journal of Population Ageing, 11(2), 107-115. Retrieved from http://dx.doi.org/10.1007/s12062-018-9225-7