Socioeconomic factors on HIV/AIDS: a cross-country comparative study

Chenkai Wang [*]

Department of Statistics and Data Science, Southern University of Science and Technology

December, 2020

Abstract

This project mainly consists of a background investigation, data preprocessing, model building, selection, analysis, and outlier inspection. First, we conduct background investigation and preliminary analysis of possible factors affecting the prevalence rate of AIDS. Then we try different methods to build and determine the final model. Next, we perform linear regression testing and multicollinearity testing for the model. Finally, we pick out the outliers and give some possible explanations from a social custom perspective.

[*] email address: `11710619@mail.sustech.edu.cn`

1 Background information

This section consists of two parts: the first is a brief introduction to HIV and AIDS, and the second is a preliminary analysis of the factors influencing the prevalence rate of AIDS.

1.1 What are HIV and AIDS?

HIV is a virus that attacks cells in the immune system. The virus destroys a type of white blood cell in the immune system called a T-helper cell – also referred to as a CD4 cell – and uses them to make copies of itself. If HIV is left untreated, it may take up to 10 or 15 years for the immune system to be so severely damaged that it can no longer defend itself. However, the rate at which HIV progresses varies depending on age, general health, and background. With treatment, people living with HIV can enjoy a long and healthy life.

AIDS is a set of syndromes caused by HIV. A person is said to have AIDS when their immune system is too weak to fight off infection, and they develop certain symptoms and illnesses. This is the last stage of HIV when the infection is very advanced, and if left untreated will lead to death.

1.2 Factors influencing the prevalence rate of AIDS

The central idea is that HIV is not spread randomly, as tends to be the case with the bacteria that cause tuberculosis or the virus that cause the common cold. Instead, HIV is most often transmitted as a consequence of purposeful behavior that often has a strong economic foundation [1].

It is found that instability in socio-economic and political aspects in many sub-Saharan African countries was responsible for creating a suitable environment for spreading HIV/AIDS infection [2]. Also, in resource-poor countries, initiation and maintenance of highly active antiretroviral therapy have been associated with many challenges and problems such as the poor infrastructural base for the control programs; irregular or non-availability of drugs; poor drug adherence; laboratory monitoring of viral load; CD4 cell counts; full blood counts; electrolytes, kidney and liver functions [3]. It is known that antiretroviral therapy is one of the most effective ways to treat AIDS.

Therefore, it is very reasonable to link AIDS and economic factors. However, there are many variables in the given data that collectively reflect the economic level of a country, such as agriculture, sanitation, and urban population. Intuitively, these variables that can reflect the economic development level are not independent of each other. On the contrary, they are related to each other. For example, if the contribution of agriculture to GDP is relatively high, we have reason to believe that the urban population rate will be relatively low. Therefore, when we choose variables related to economic factors, we must choose a representative, namely, more significant variables.

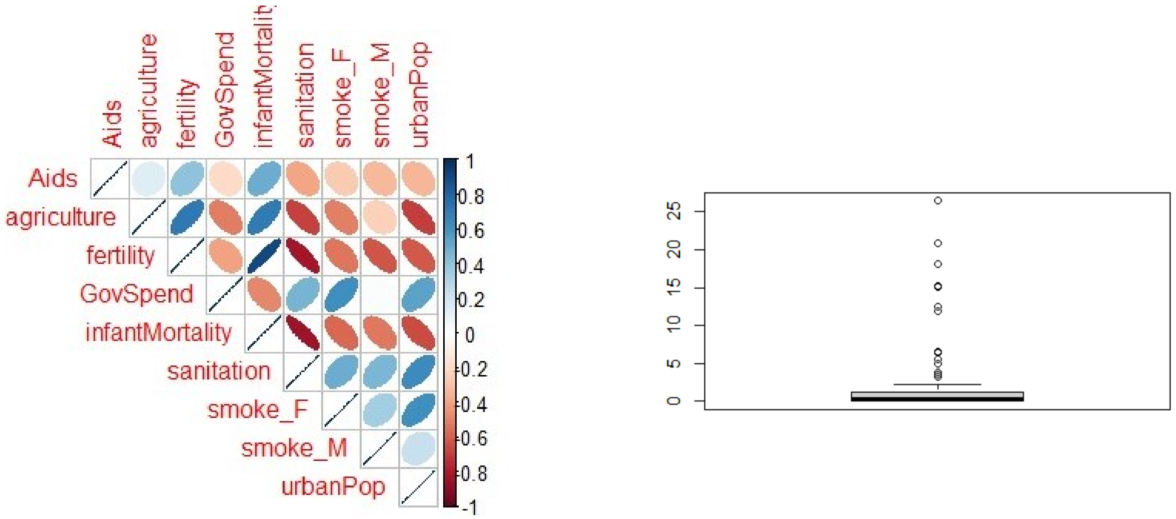
Meanwhile, it is shown that South Africa's health protocols for mother-to-child transmission of HIV are not comprehensive, including the absence of protocols for prenatal testing and pregnancy treatment. [4]

Hence we can conclude that the AIDS prevalence rate is also closely related to a country's medical and health development level. The higher the medical development level, the stricter the prevention and control of HIV/AIDS will be, the more rigorous the treatment plan will be, the more significant effect will be. Simultaneously, extensive publicity measures can further increase people's attention and awareness, which can prevent the further spread of HIV/AIDS among the population. Like economic factors, there are many variables in the given data that collectively reflect the level of medical and health development of a country. For example, in general, the higher the GovSpend is, the lower the infantMortality will be. When we choose this part of the variables, we should also choose representative variables.

2 Data preprocessing

This section mainly includes data importing and visualizing, which are as follows.

First, we import the data using packages in R and check for missing values. After that, we rearrange it according to the response variable, that is, the value of the AIDS prevalence rate. We can see that The country with the lowest prevalence rate is Bangladesh and the highest is Swaziland. Meanwhile, the average prevalence rate is 2.078.



(a) Correlation between variables

(b) Box figure based on the prevalence rate

Figure 1: Data visualization

Figure 1(a) confirms the conjecture in section1.2 that there is a correlation between variables. Therefore, it is crucial to choose significant predictable variables in model building. From figure1(b), it shows that the HIV prevalence rate in most countries is below 5, but some countries have very high infection rates. For these countries with significantly higher infection rates, we will analyze them separately.

3 Model building

In this section, we will build linear models. The idea of model building is as follows. The first is to build a full model and find significant predictive variables. Next, we only consider the significant predictive variables and use them to build a new model. Ultimately, the stepwise regression method is used to simplify the full model and get another new model. After that, we will select and determine the final model according to certain criteria.

3.1 Full model

First, we try to take all variables into consideration, and the results are in table 1.

From the summary of the full model, only agriculture and infantMortality are significant. These two variables can also generally reflect the country's economic development level and health development level. Therefore, it is appropriate to include these two variables in the model for analysis from a practical perspective. In order to further simplify and optimize our model and make it more convincing, we will delete other non-significant variables to get the reduced model.

Table 1: Full model with all predictive variables

	Dependent variable:
	AIDS
agriculture	−0.184***
	(0.054)
fertility	−0.320
	(0.741)
GovSpend	−0.0001
	(0.001)
infantMortality	0.104***
	(0.032)
sanitation	−0.0003
	(0.037)
smoke_F	0.030
	(0.053)
smoke_M	−0.023
	(0.043)
urbanPop	−0.053*
	(0.029)
Constant	5.284
	(4.832)
Observations	95
R^2	0.364
Adjusted R^2	0.304
Residual Std. Error	3.918 (df = 86)
F Statistic	6.140*** (df = 8; 86)
Note:	*p<0.1; **p<0.05; ***p<0.01

3.2 Reduced model 1

According to the results in section 2.1, we only take agriculture and infantMortality as predictive variables this time and name it reduced model 1. The specific results are shown below.

Table 2: Reduced model 1 with agriculture and infantMortality

	Dependent variable:
	AIDS
agriculture	-0.156^{***}
	(0.044)
infantMortality	0.107^{***}
	(0.016)
Constant	0.465
	(0.632)
Observations	95
R^2	0.336
Adjusted R^2	0.321
Residual Std. Error	3.871 (df = 92)
F Statistic	23.227^{***} (df = 2; 92)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that in reduced model 1, all variables are significant.

3.3 Reduced model 2

To avoid deleting some important variables, we use stepwise regression to get another model named reduce model 2.

Table 3: Reduced model 2 via stepwise regression

	Dependent variable:
	AIDS
agriculture	-0.193^{***}
	(0.048)
infantMortality	0.097^{***}
	(0.017)
urbanPop	-0.047^{*}
	(0.026)
Constant	3.954^{**}
	(1.990)
Observations	95
R^2	0.360
Adjusted R^2	0.338
Residual Std. Error	3.821 (df = 91)
F Statistic	17.026^{***} (df = 3; 91)
Note:	*p<0.1; **p<0.05; ***p<0.01

The step function in the R language can help us realize it. Due to space limitations, only the final results are shown here.

Comparing with the full model and reduced model 1, reduced model 2 has the biggest adjusted r square. Simultaneously, the urban population rate is also an important indicator in measuring the economic level. In general, the higher the urban population, the better the economic performance of a country [5], so we decide to include urbanPop as a predictive variable and use reduced model 2 as our final model. The least square regression equation is

$$Y = -0.19340X_1 + 0.09678X_2 - 0.0437X_3 + 3.9549$$

where X_1, X_2 and X_3 represents agriculture, infantMortality and urbanPop respectively.

4 Model diagnostics

In this section, we will test and analyze the model. Testing includes linear regression model hypothesis testing and collinearity testing. Meanwhile, we will find outliers in the model and conduct further analysis.

4.1 Model testing

This section including five major test in linear regression. Results are as follows.

4.1.1 Test for normality

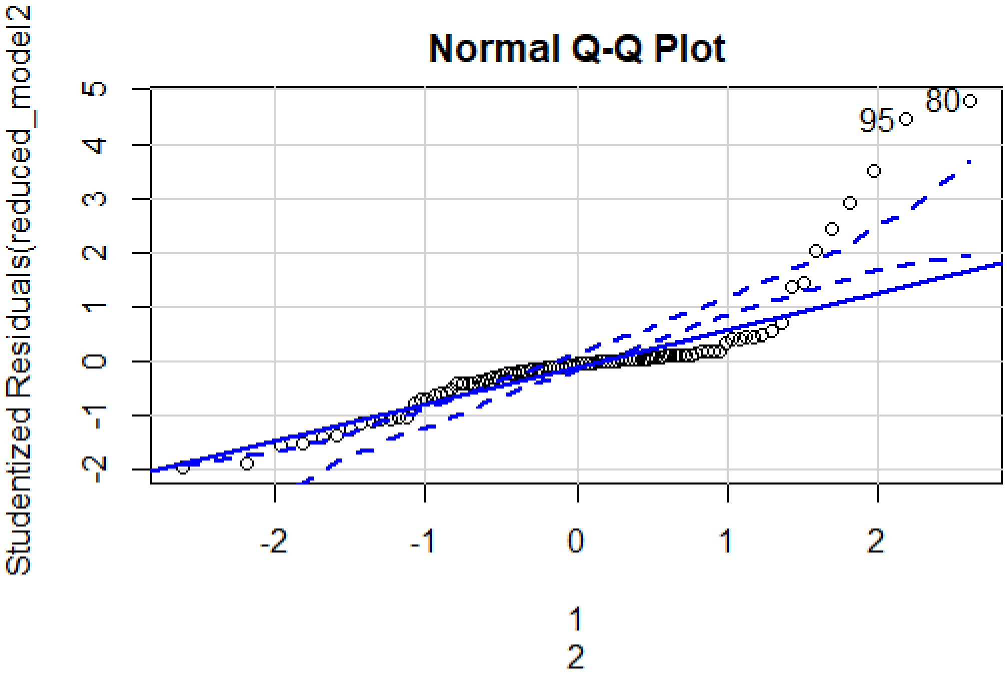


Figure 2: Normal Q-Q plot

From the Q-Q plot, we will reject the normality.

4.1.2 Test for the independence of error test

```
lag Autocorrelation D-W Statistic p-value
1       0.09744602        1.626292    0.074
Alternative hypothesis: rho != 0
```

The non-significant p-value and small D-W statistic(< 2) suggests a lack of autocorrelation, and conversely, independence of errors.

4.1.3 Test for linearity

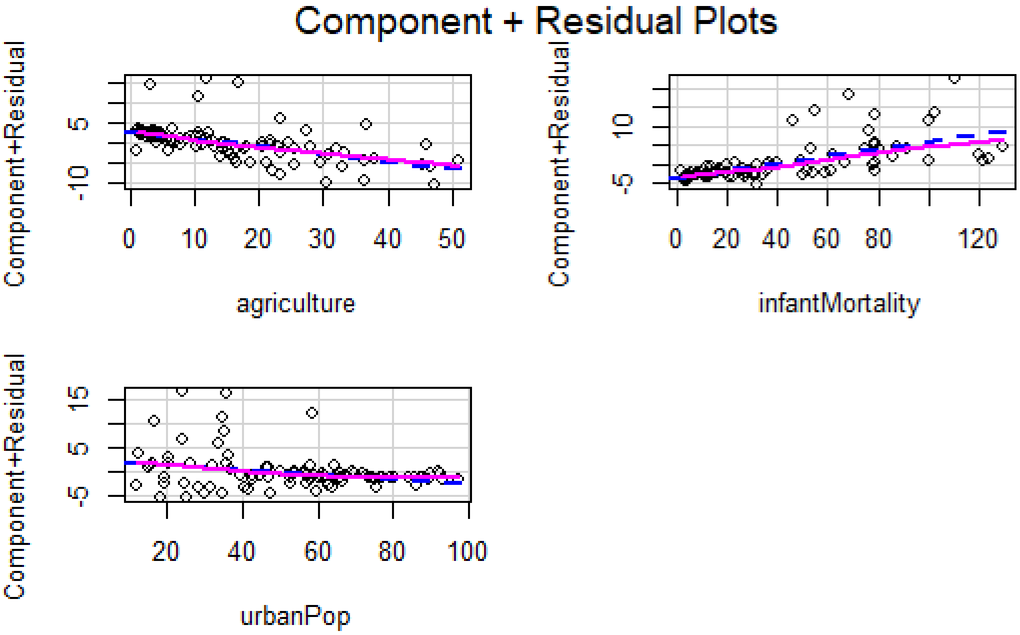


Figure 3: Partial-residual plots for the linear model

From the picture above, we can conclude that the model satisfies linear relation.

4.1.4 Test for homogeneity of variance

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 119.2528, Df = 1, p = < 2.22e-16
```

Since the p value is less than 0.05, we can conclude that the model does not meet the assumption of homogeneity of variance.

4.1.5 Test for multicollinearity

Here we use the vif function in R to test, and the results are as follows.

```
agriculture     infantMortality     urbanPop
2.422470        2.237460            2.153083
```

The result shows no significant collinearity in the final model since the absolute value of all the results is lower than 4.

4.2 Test for outliers

```
        rstudent      unadjusted p-value      Bonferroni p
80      4.786948      6.6088e-06              0.00062784
95      4.448903      2.4627e-05              0.00233950
```

As shown in the result of the outlier test, Swaziland(80) and Zimbabwe(95) are considered as outliers. Furthermore, we will compare the outliers, and the results are as follows.

Table 4: Values of average and outliers

	average	Zimbabwe	Swaziland
AIDS	2.078	20.870	26.490
agriculture	15.100	16.742	11.661
infantMortality	36.937	68.000	110.000
urbanPop	53.425	35.480	23.940

4.3 Outlier analysis

Reasons in Zimbabwe:

1. The proportion of women using condoms is extremely low, which only reaches 1% at present [6]. Moreover, the prevalent mode of transmission is unprotected heterosexual sex[7]. None-barrier methods are the most predominant contraceptive methods of choice among Zimbabwean women, with the contraceptive pill being the most popular [7].

2. The illegal nature of sex work and homosexuality in Zimbabwe presents huge barriers for sex workers and men who have sex with men from accessing HIV services [8].

Reasons in Swaziland:

1. The spread of the epidemic in Swaziland is fueled by behavioral, structural, and biological drivers, which include multiple and concurrent sexual partners, low and inconsistent use of condoms, inter-generational sex, mobility and migration, commercial sex, early sexual debut, gender inequality, and sexual violence, low levels of male circumcision [9].

2. Due to income inequality, many young Swazi girls engaging in transactional sex or sex for favors and in the process risking HIV infection [9].

5 Summary

In this project, our main work includes problem background investigation, data preprocessing, model building and analysis, and outlier test. Our final model can make a reasonable explanation for most of the given data. As for outliers, we give some possible explanations based on the actual situation. Maybe it is not a perfect linear regression model and it fails to meet all assumptions. Nevertheless, there is no perfect model for actual problems. A model that can solve the problem is good, right?

References

[1] Plamen Nikolov. The aids epidemic and its economic roots. *Harvard Health Policy Review*, 2009.

[2] Ramaswamy Premkumar and Achilles Tebandeke. Political and socio-economic instability: does it have a role in the hiv/aids epidemic in sub-saharan africa? *SAHARA-J: Journal of Social Aspects of HIV/AIDS*, 8(2):65–73, 2011.

[3] OR Obiako and HM Muktar. Challenges if hiv treatment in resource-poor countries: A review. *Nigerian journal of medicine*, 19(4):361–368, 2010.

[4] Memoona Hasnain. Antenatal hiv screening and treatment in south africa: Social norms and policy options. *African journal of reproductive health*, pages 77–85, 2004.

[5] `https://stats.unctad.org/handbook/Population/Total.html`.

[6] D Wilson and A Mehryar. The role of aids knowledge, attitudes, beliefs and practices research in sub-saharan africa. *AIDS (London, England)*, 5:S177, 1991.

[7] Christopher Mafuva and Hilda T Marima-Matarira. Hormonal contraception and hiv/aids transmission: challenges for zimbabwe's reproductive health service providers in promoting informed contraception choices. *Journal of public health in Africa*, 4(2), 2013.

[8] `https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/zimbabwe`.

[9] Johanes A Belle and Nokuthula N Gamedze. Behavioral factors contributing to the transmission of hiv and aids amongst young women of mbabane in swaziland. *African Health Sciences*, 19(3):2302–2311, 2019.