Analysis on how social factors influence the annual increase of HIV/AIDS infection within country level from observational data

Apply multivariate linear regression models with Regression Discontinuity Design (RDD) to observe the causal effect of the social factors and test against models with a test set

Junming Zhang

Tuesday, December 22, 2020

Abstract

This report focused on observing how social factors within a country level affect the spread of HIV and AIDS by building multivariate linear regression models and compare the statistics summary and diagnosis of two models. It is found that it is not capable of concluding the social factors found have absolutely strong influence on the HIV/AIDS spread, however, some social factors have significant causality on the HIV/AIDS spread by testing the behavior of the linear regression model before and after the threshold, concluding from the RDD. And also, the multilevel model works better by comparing the diagnosis statistics of models. This is important because it shows for what social factors the government invest more to control in order to contain the AIDS infection, a disease cannot be cured currently, and also suggests what kind of regression model is preferable for the similar task (for instance, make policies based on observing what social factors matter in country/region/territory level) and also inspire other to how to build a statistics model when the observational study is inter-regional, and also how to select model with according to datasets the research community has.

key words: HIV/AIDS, social factors, multivariate linear regression, model selection, regression discontinuity design (RDD), observational study, causal inference

Please click **"here"** to access the GitHub repository for all work.

Introduction

Human immunodeficiency virus infection and acquired immunodeficiency syndrome (HIV/AIDS) is a disease caused by the infection of the human immunodeficiency virus (HIV) (Sepkowitz, 2001). If someone is infected, the disease has a long incubation period without any symptom (CDC, 2020), the average incubation period for transfusion cases is about 7.66 years (Anderson & Medley, 1988). As the infection develops, it interferes the functionality of the host's immune system, and increases the risk of being infected by opportunistic infections such as tuberculosis. The symptom after the development of the infection is named as acquired immunodeficiency syndrome (AIDS) (CDC, 2020). HIV spread is caused by unprotected (i.e., no condom) sex (even for oral and anal sex), infectant blood transfusions, hypodermic needle injection (such as taking some drugs), and mother-to-child transmission during pregnancy, delivery, or breastfeeding (Rom & Markowitz, 2015). The virus cannot spread through some body fluids, like sweat, tears and saliva (CDC, 2005). There is no cure or vaccine for this disease, however, the disease can be slowed down by the antiretroviral treatment and with this treatment, victims may have a near-normal life expectancy (CDC, 2020).

Currently HIV is not only a form of illness, but also has large impact on the society, such as a source of discrimination and many misconceptions, and also economic conditions (UNAIDS, 2006). And the prevention and remedy of AIDS also has conflicts with religions (McCullom, 2013). This means that AIDS may have some social features and attributes, and according to CATIE, Canada's source for HIV and hepatitis C

information, "Many factors in our society, including poverty, physical and sexual abuse, lack of education, homelessness, stigma, addiction, violence, untreated mental health problems, lack of employment opportunities, powerlessness, lack of choice, lack of legal resident status and lack of social support, play a role in HIV infection..." (original text cited from the source) (CATIE, 2018). It is necessary to learn how the society can influence the AIDS spread.

To make the observational study and obtain a tool to predict the AIDS spread, it is worth trying to build a linear regression model which indicates the relationship between the AIDS spread, which is defined as the annual increasing cases within a country level, and several social features as attributes. And in this report, two linear models will be built and tested. The reason to use the linear regression model is that all the social features are quantified and thus all the analysis is made with respect to specific numeric data, and the prediction is also numeric. Linear regression model works based on numeric analysis and makes predictions on continuous change, which is the case for AIDS cases increase per year, based on the change of the social attributes. Meanwhile, the diagnosis of linear regression model suggests if some variables can be rejected for the targeted prediction if they do not have strong effect on the response. And the reason to construct two models is to compare and contrast the advantages and disadvantages of two models and observe how the model selection influences the prediction, and then choose a superior model to work on the future prediction. The first model is built with the general linear regression construction tool and the second model is built with multilevel linear regression. To analyze how the social factors selected influence the HIV/AIDS infection, a common way is to make the causal inference between the social factors and the result, and the causal link is drawn to the population from the observational data. One method to make the observational study is Regression Discontinuity Design (RDD), which observes the causality when predictors are numeric, continuous and have a threshold (discontinuity) to determine if the treatment is implemented.

In the rest of the report, the datasets used to make the analysis and test (about the number of new AIDS infections each country per year and the quantified social features in each country per year such as HDI, HAQ), and also the model, how models are constructed (mathematical notation and variables) and why the models selected to test (actually, a linear regression model will be built) and show the diagnosis of two linear regression models, and how the RDD is implemented, in general, the introduction section about the data and the model is referred to as **Methodology**. In the **Results** section, all the figures and tables generated from the diagnosis about the accuracy and statistical summary of model will be displayed, the causality inference of each variable (observed from RDD) will be made, and also may include the analysis on the tables and figures. In the **Discussion**the causality observed from the model will be summarized, all the work done will be summarized, and the meaning of the results will be explained. Also, an analysis of the weakness of the model will be provided and the shortcomings and possible improvements (further steps can be done) will be listed, and the suggestion on model selection will be given. In the final section (**References**), all references used in this report are provided.

Methodology

In this section, the datasets used in the observational study and the model built are introduced.

Data

This subsection is about the dataset used to build the model. The dataset used to train in the model is `train_set.csv`, and the dataset used to test the model is `test_set.csv`. The `train_set.csv` and the `test_set.csv` are splitted from a large data table combined by 8 datasets (all datasets are observational data rather than experimental data), and the most processes of computing new data and combining datasets used to build the model is generated by `datagen.R`. To combine the 8 datasets, "Year" and "Entity" are two variables used as a key. In these 8 datasets, `new-cases-of-hiv-infection.csv` and `world-population-by-world-regions-post-1820.csv` provide the response variable, `new_cases_share_5_years_later`, computed from the variable `new_cases_5_years_later` in `new-cases-of-hiv-infection.csv` and the variable `population_size_5_years_later`

in `world-population-by-world-regions-post-1820.csv` ($new\_cases\_share\_5\_years\_later = \frac{new\_cases\_5\_years\_later}{population\_size\_5\_years\_later}$). Then, the predictors data are obtained from the datasets `human-development-index.csv`, `healthcare-access-and-quality-index.csv`, `average-real-gdp-per-capita-across-countries-and-regions.csv`, `recent-ILO-LFP.csv`, `share-of-population-urban.csv` and `share-with-alcohol-or-drug-use-disorders.csv`, and the predictor variables provided by the 6 datasets are `hdi`, `haq`, `gdp_per_capita`, `female_emply_rate`, `urb_rate`, and `drug_alcohol_disorder_share` respectively.

Both response and predictor variables are all **numeric**, and that is one reason why the linear regression model is built. Then the average of each predictor variable is computed and named in the pattern "`predictor_avg`", those averages are used as the threshold in RDD. And then, in order to show the causality quantitatively, dummy variables are generated, which are binary variables with value 1 or 0, which is named in the pattern "`predictor_over_avg`", it is 1 if $predictor > predictor\_avg$, and 0 otherwise. The dummy variables are used to observe the causal link between the quantity of the predictors and the quantity of the responses, and if the coefficient of one dummy variable is high, the corresponding predictor has strong effect on the response.

In this paragraph, the features and the knowledge of each variable to construct the model is interpreted below:

- `new_cases_share_5_years_later`: numeric data, this is the ratio between the number of new AIDS/HIV infection cases of the country and the population of the country after 5 years of the predictors (for example, if the predictors are collected in 2005, the responses are collected in 2010, collect data in this way because the incubation period is about 7.66 years, and the victims do not have symptoms immediately), and this is the data the model predicts. The share value is used instead of `new_cases_5_years_later` because each country has tremendously different population size, and population is an important factor to consider when the predictors are unified indices or percentage. The population size is the variable `population_size_5_years_later`. And $new\_cases\_share\_5\_years\_later = \frac{new\_cases\_5\_years\_later}{population\_size\_5\_years\_later}$, which combines the `new_cases_5_years_later` and the `population_size_5_years_later` by ratio. Since this is the response variable, there is no threshold for this variable, and therefore no average and dummy variable for this variable.
- `hdi`: numeric index, human development index (HDI) is an composite statistic index of the life expectancy, education and per capita income indicators. If the HDI is high, then the life expectancy, the education level and the gross national income GNI (PPP) per capita are high as well. This variable is used because based on the citation in the introduction section, poverty and education might be factors to consider. The dataset about this index is chosen not only because it reflects multiple social factors, but also this dataset covers most countries in the world. The average of this variable in the dataset is `hdi_avg`, and the dummy variable is `hdi_over_avg`.
- `haq`: numeric data, the Healthcare Access and Quality (HAQ) index is a measurement on a scale from 0 (worst) to 100 (best) based on death rates from 32 causes of death that actually can be cured. This variable is used because it reflects the medical quality of a country and might be relative to HIV/AIDS infection since contaminated blood transfusion is one way HIV spreads. The dataset about this index covers most countries in the world. The average of this variable in the dataset is `haq_avg`, and the dummy variable is `haq_over_avg`.
- `gdp_per_capita`: numeric data, this is the GDP per capita for each country from 1870 to 2016 for the most countries and territories in the world. Gross domestic product (GDP) is the measurement the market value of all the final goods and services produced during a fixed time period, and "per capita" means "per person". This metric is one of the most important indices of the economic situation of a country, and this index may help discuss the relationship between the national economy and AIDS/HIV infection. All countries report this data and thus the relative data set covers nearly all countries/territories in the world. The average of this variable in the dataset is `gdp_per_capita_avg`, and the dummy variable is `gdp_per_capita_over_avg`.
- `female_emply_rate`: numeric number, the labor force participation rate of female is the proportion of the female population with age at least 15 that supply labor force for the production of goods and

services during a fixed period. The rate is used to describe the gender inequality, ideally, the higher the female employment rate, the less the gender inequality. However, this does not take into consideration of the employment diversity and income of women. Although there are some datasets about the gender inequality indices, but those datasets do not cover half of countries in the world, therefore this dataset is the most "ideal" one since it covers nearly all countries/territories although there are flaws listed above. The average of this variable in the dataset is `female_emply_rate_avg`, and the dummy variable is `female_emply_rate_over_avg`.

- `urb_rate`, numeric data, share of urban population in a country in a specified period. This rate is chosen because there are more communications between people in urban area, which also means AIDS/HIV may have higher possibility to spread, thus share of urban population is also considered. The average of this variable in the dataset is `urb_rate_avg`, and the dummy variable is `urb_rate_over_avg`.
- `drug_alcohol_disorder_share`, numeric data, this value is the percentage of people in a country has alcohol or drug disorder in a specified time, this proportion is taken into consideration because based on the introduction, addiction may contribute to AIDS/HIV infection. The average of this variable in the dataset is `drug_alcohol_disorder_share_avg`, and the dummy variable is `drug_alcohol_disorder_share_over_avg`.

To introduce each dataset, including source, population, approach to collect the data, the population, the frame, and the sample, a table is shown below. The datasets are the census data collected by governments, NGOs, academia and other kinds of organizations in most of the countries/territories in the world. The source column lists the data source labelled in the reference section. The dataset can be downloaded from the link given in the reference section. And since the datasets are all census data collected by the government, it is hard to give the detailed information for columns about how to collect the data, the information about finding respondents, how to find respondents and how to deal with non-responses.

Table 1: introduction on datasets

dataset	variable	data_source
new-cases-of-hiv-infection.csv	new_cases_5_years_later	reference 1
world-population-by-world-regions-post-1820.csv	population_size_5_years_later	reference 2
human-development-index.csv	hdi	reference 3
healthcare-access-and-quality-index.csv	haq	reference 4
average-real-gdp-per-capita-across-countries-and-regions.csv	gdp_per_capita	reference 5
recent-ILO-LFP.csv	female_emply_rate	reference 6
share-of-population-urban.csv	urb_rate	reference 7
share-with-alcohol-or-drug-use-disorders.csv	drug_alcohol_disorder_share	reference 7
new-cases-of-hiv-infection.csv	new_cases_5_years_later	reference 8

Table 2: introduction on datasets, continue

dataset	approach_to_collect_data
new-cases-of-hiv-infection.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered
world-population-by-world-regions-post-1820.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered
human-development-index.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered
healthcare-access-and-quality-index.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered
average-real-gdp-per-capita-across-countries-and-regions.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered

dataset	approach_to_collect_data
recent-ILO-LFP.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered
share-of-population-urban.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered
share-with-alcohol-or-drug-use-disorders.csv	all data collected by government, NGO or academia, and the data are census data, thus all information can be covered

Table 3: introduction on datasets, continue

dataset	population	frame	sample
new-cases-of-hiv-infection.csv	global: all countries/territories in the world, national: all AIDS patients in the country	global: list of countries/territories in the world; national: list of patients in the government and relative organizations	global: the countries/territories have and can provide the relative data in the specified time, national: all AIDS patients in the country and reported
world-population-by-world-regions-post-1820.csv	global: all countries/territories in the world, national: all people in the country	global: list of countries/territories in the world; national: list of population data kept by the government	global: all countries/territories in the world, national: all people in the country that been connected
human-development-index.csv	global: all countries/territories in the world, national: all people in the country	global: list of countries/territories in the world; national: list of population, education and economic data of government and academia	global: all countries/territories in the world, national: all people in the country and been visited and reported
healthcare-access-and-quality-index.csv	global: all countries/territories in the world, national: all people who died because of the 32 diseases	global: list of countries/territories in the world; national: list of people died because of 32 diseases kept in the medical organizations	global: all countries/territories in the world, national: all people who died because of the 32 diseases and been reported
average-real-gdp-per-capita-across-countries-and-regions.csv	global: all countries/territories in the world, national: all the market value produced	global: list of countries/territories in the world; national: list of economy data in the government and academia	global: all countries/territories in the world, national: all the market value produced and can be found and reported
recent-ILO-LFP.csv	global: all countries/territories in the world, national: all women who are employed	global: list of countries/territories in the world; national: list of economy and population data in the government and academia	global: all countries/territories in the world, national: all women who are employed and can be found and reported
share-of-population-urban.csv	global: all countries/territories in the world, national: all people living in urban area	global: list of countries/territories in the world; national: list of people living in urban areas kept in the government	global: all countries/territories in the world, national: all people living in urban area and can be found and reported

dataset	population	frame	sample
share-with-alcohol-or-drug-use-disorders.csv	global: all countries/territories in the world, national: all people who have the problem of durg and alcohol disorder	global: list of countries/territories in the world; national: list of people have disorders from drug or alcohol abuse	global: all countries/territories in the world, national: all people who have the problem of durg and alcohol disorder and can be found and reported

Table 4: introduction on datasets, continue

dataset	info_find_respondent	deal_with_nonresponse
new-cases-of-hiv-infection.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
world-population-by-world-regions-post-1820.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
human-development-index.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
healthcare-access-and-quality-index.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
average-real-gdp-per-capita-across-countries-and-regions.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
recent-ILO-LFP.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
share-of-population-urban.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered
share-with-alcohol-or-drug-use-disorders.csv	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered	NA, research supported by the local organizations including government and academia, and all datasets are census data, thus all information can be covered

These datasets are all census data from nearly all countries and territories in the world, therefore the datasets can completely and objectively reflect the situation in each country and there are abundant amount of data to build the model even if some data have to be separated as the test data. However, the problems with the datasets are that the data points distribute extremely non-uniformly and the scales for each variable differ largely. For the first problem, a linear regression model with RDD is designed, so there are two regression lines which can fit the data points better, the average for each data point is used as the threshold, and here is

the distribution diagram:

figure 1: data distribtuion and simple MLR with RDD for train set

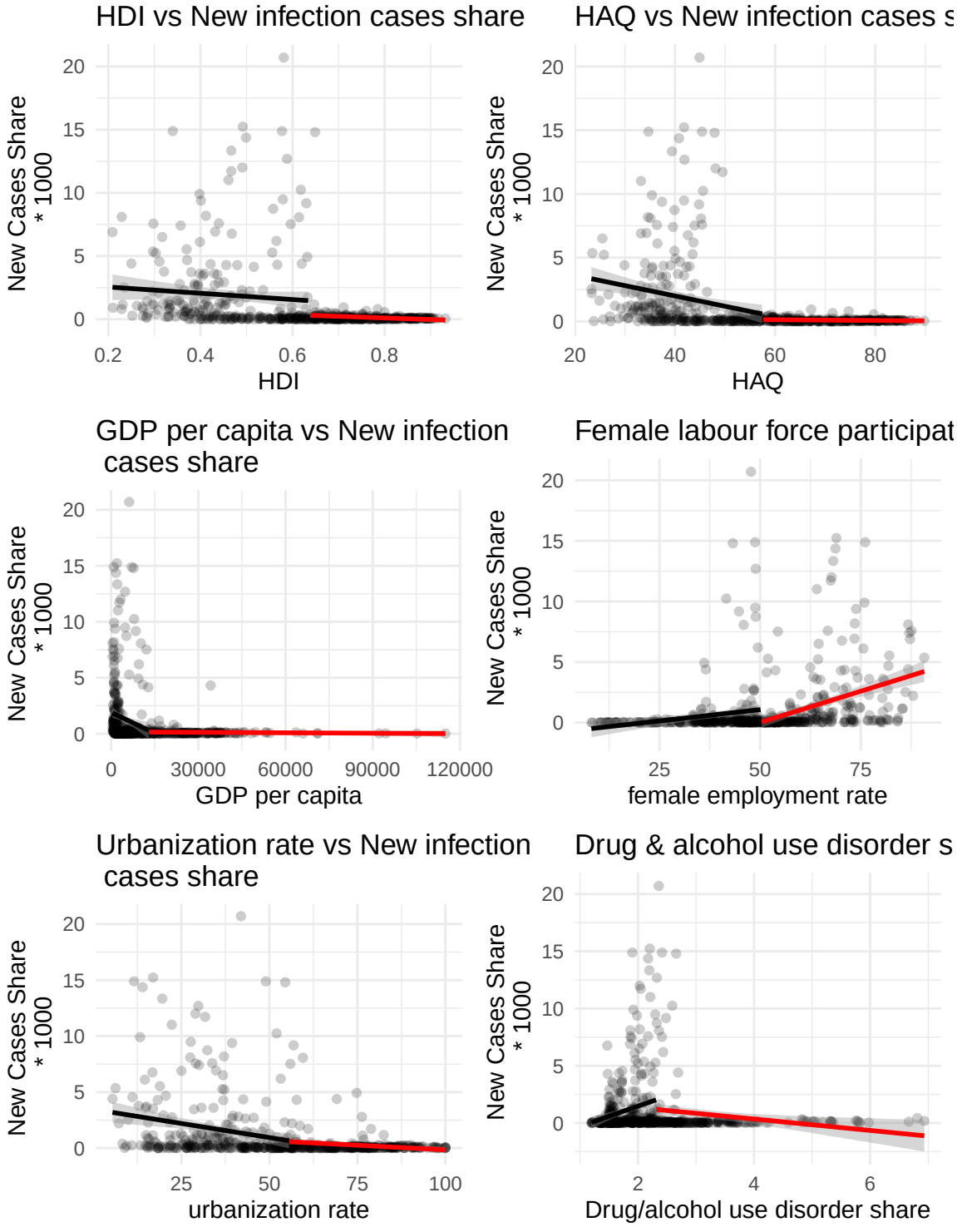
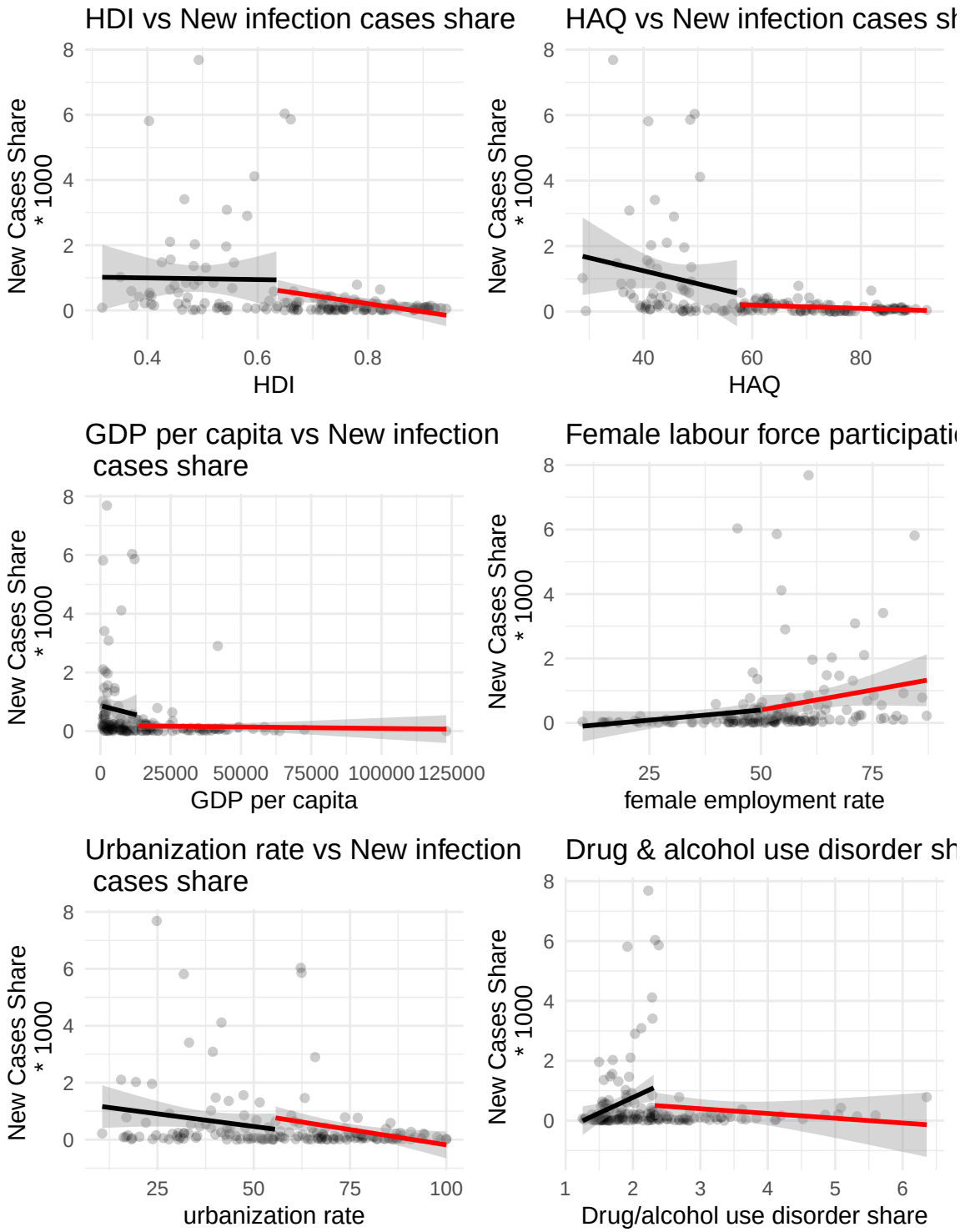


figure 2: data distribtuion and simple MLR with RDD for test set



From the diagram, it can be seen that although RDD is used in the linear regression, there are still many outliers. Therefore, a multilevel linear regression model is considered and "Entity" (categorical data) is used to divide data into different cells (and "Entity" is not used to build the normal linear regression model). For the second problem, a "standardize" function is used to standardize each variable by:

$$standardize\_x_i = \frac{x_i - \mu}{\sigma}$$

however, this may make the model parameters (coefficient, β term in the model) far from expectation, although those parameters are dramatically large and hard to use for evaluation.

Model

In this subsection, the two linear regression models with built for the observational study will be discussed. Both of the two models are fitted with the train set split by `datagen.R`. Since the data points do not distribute uniformly as described in the **data** subsection, RDD is used to build the piecewise model to fit data points more flexibly, and the average of each parameters is used as the threshold in RDD.

Also, a dummy variable for each normal variable is used in the linear regression model to estimate the effect of the causality between that variable and the response. The larger the average of the dummy variable, the stronger the causality effect. The dummy variable has the format $\beta_{predictor\_over\_average} I(predictor > predictor\_avergae)$, where $\beta_{predictor\_over\_average}$ is the quantity of the causality effect explained above, which can be seen as the coefficient of the indicator, and $I(predictor > predictor\_avergae)$ is an indicator with binary value 1 and 0. If $I(predictor > predictor\_avergae) = 0$, the predictor is less than or equal to the threshold (so at the left side of the graph) and in the model, the causality effect is cancelled since the indicator is 0 and the causality effect (coefficient term of that indicator, $\beta_{predictor\_over\_avergae}$, mathematically speaking) is not added to the response. If the predictor is beyond the threshold ($I(predictor > predictor\_avergae) = 1$), the causality effect acts in the model (the response value os the addition of the value of variable and the causality effect of the variable since the indicator is 1 and the causality effect is not cancelled). Each variable and corresponding dummy variable has been explained in the **data** subsection.

In order to build the multivariate linear regression model to predict the share of new AIDS/HIV, two models are built. The first model is a norm linear regression model, built by the R function **lm** and has the mathematical expression:

$$
\begin{aligned}
y_{new\_cases\_share\_5\_years\_later} =& \beta_{hdi} x_{standardized\_hdi} + \beta_{haq} x_{standardized\_haq} + \\
& \beta_{gdp\_per\_capita} x_{standardized\_gdp\_per\_capita} + \\
& \beta_{female\_emply\_rate} x_{standardized\_female\_emply\_rate} + \\
& \beta_{urb\_rate} x_{standardized\_urb\_rate} + \\
& \beta_{drug\_alcohol\_disorder\_share} x_{standardized\_drug_alcohol_disorder_share} + \\
& \beta_{hdi\_over\_avg} I(hdi > hdi\_average) + \beta_{haq\_over\_avg} I(haq > haq\_average) + \\
& \beta_{gdp\_per\_capita\_over\_avg} I(gdp\_per\_capita > gdp\_per\_capita\_average) + \\
& \beta_{female\_emply\_rate\_over\_avg} I(female\_emply\_rate > \\
& female\_emply\_rate\_average) + \\
& \beta_{urb\_rate\_over\_avg} I(urb\_rate > urb\_rate\_average) + \\
& \beta_{drug\_alcohol\_disorder\_share\_over\_avg} I(drug\_alcohol\_disorder\_share > \\
& drug\_alcohol\_disorder\_share\_average) + \beta_0
\end{aligned}
$$

The model has the reponse, the share of the new cases infection after 5 years (the ratio between the new cases after 5 years and the population after 5 years), which is predicted from the HDI, HAQ, GDP per capita, the proportion of women employed, the rate of population in urban areas and the ratio of people with drug and alcohol disorder (all values are standardized), and each predictor variables are assigned with a coefficient ($\beta_{predictor}$) during the model construction. Except for those terms, the dummy variable for each predictor is also added to observe the causality effect, those dummy variables are indicators, which is 1 if the value of the corresponding predictor is beyond the threshold (average of the predictor values) and 0 otherwise, and the coefficient term ($\beta\_predictor\_over\_average$) for the dummy variables are the quantity of the causality effect. The detailed information of the dummy variable has been introduced above. The reason to standardize the variable has been explained **data** subsection, and since each linear regression has error, the term β_0 is

labelled in the mathematical expression as the residual of the regression, and the intercept in the model expression.

The plots of data and the simple linear regression line built by "lm" as above has been attached in the **data** section. Seen from the graph, there are many outlier data points. To make the model fit the data points better, considering the different social, medical and economic conditions in each country (sometimes the economic conditions in a country does not necessarily imply the medical conditions in that country, because policy is also important, Cuba and some other socialist countries are good examples), the "Entity" is used as a level variable and a random intercept ($(1|Entity)$), and it has the following mathematical expression ("Entity" is of categorical data type because it is qualitative and nominal):

$$
\begin{aligned}
y_{new\_cases\_share\_5\_years\_later} =& \beta_{hdi} x_{standardized\_hdi} + \beta_{haq} x_{standardized\_haq} + \\
& \beta_{gdp\_per\_capita} x_{standardized\_gdp\_per\_capita} + \\
& \beta_{female\_emply\_rate} x_{standardized\_female\_emply\_rate} + \\
& \beta_{urb\_rate} x_{standardized\_urb\_rate} + \\
& \beta_{drug\_alcohol\_disorder\_share} x_{standardized\_drug_alcohol_disorder_share} + \\
& \beta_{hdi\_over\_avg} I(hdi > hdi\_average) + \beta_{haq\_over\_avg} I(haq > haq\_average) + \\
& \beta_{gdp\_per\_capita\_over\_avg} I(gdp\_per\_capita > gdp\_per\_capita\_average) + \\
& \beta_{female\_emply\_rate\_over\_avg} I(female\_emply\_rate > \\
& female\_emply\_rate\_average) + \\
& \beta_{urb\_rate\_over\_avg} I(urb\_rate > urb\_rate\_average) + \\
& \beta_{drug\_alcohol\_disorder\_share\_over\_avg} I(drug\_alcohol\_disorder\_share > \\
& drug\_alcohol\_disorder\_share\_average) + (\alpha + a_j) + \beta_0
\end{aligned}
$$

In the above expression for multilevel multilvariate linear regression, every fixed term is the same as the first model except for the random effect term $\alpha + a_j$, which is decided by the random level "Entity". *alpha* is constant and a_j follows the normal distribution $\mathcal{N}(0, \sigma_a^2)$. Therefore, the behavior of the model varies based on the country and the RDD is also implemented in this model. The model is built by the tool **lmer** in the package **lme4**.

In order to compare and contrast the strength and weakness of two models, and also test the influence and the causality effect of each variable of social factors, some diagnosis has been done for two models, including test the R-squared, p-value and AIC/BIC value, and the table of these values are listed below.

Table 5: ANOVA table of simple MLR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
std_hdi	1	26.8697498	26.8697498	28.9652060	0.0000001
std_haq	1	7.4098652	7.4098652	7.9877287	0.0048789
std_gdp_per_capita	1	0.6961081	0.6961081	0.7503946	0.3867250
std_female_emply_rate	1	7.9569591	7.9569591	8.5774881	0.0035430
std_urb_rate	1	1.1090614	1.1090614	1.1955523	0.2746859
std_drug_alcohol_disorder_share	1	0.0925377	0.0925377	0.0997543	0.7522432
hdi_over_avg	1	0.7806471	0.7806471	0.8415264	0.3593583
haq_over_avg	1	0.7320976	0.7320976	0.7891908	0.3747291
gdp_per_capita_over_avg	1	0.0078329	0.0078329	0.0084437	0.9268187
female_emply_rate_over_avg	1	6.2544921	6.2544921	6.7422530	0.0096647
urb_rate_over_avg	1	0.0056990	0.0056990	0.0061435	0.9375537
drug_alcohol_disorder_share_over_avg	1	0.3081625	0.3081625	0.3321947	0.5646028
Residuals	556	515.7767875	0.9276561	NA	NA

From the ANOVA table above, for the simple MLR, generally the null hypothesis: the variable does not have significant influence on the response, cannot be rejected since the most p-values are greater than 0.05, except for the variable HDI, HAQ and female employment rate.

Table 6: AIC and BIC measurement of simple MLR

aic	bic	r_squared
1586.873	1647.687	0.0919423

AIC and BIC are a pair of information criteria methods to assess the quality of model fitting and penalizing the number of parameters, the more the parameters to fit the model, the higher the AIC and BIC. The higher the AIC and BIC, the worse the fitting quality is. The difference between AIC and BIC is that the BIC penalizes more on the number of parameters compared to AIC. And the AIC and BIC have been listed in the table above (1586.873 and 1647.687).

Another diagnosis statistic, R^2, describes the amount of the variance in the dependent variable which is predictable from the independent variable(s). The higher the R^2, the more variance can be described by the model, and the better the model. R^2 is also shown in the table, which is quite low (only 9%).

Table 7: ANOVA table of multilevel MLR

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
std_hdi	0.0035551	0.0035551	1	567.0978	0.0396213	0.8422942
std_haq	0.0694418	0.0694418	1	568.1468	0.7739145	0.3793807
std_gdp_per_capita	0.0020302	0.0020302	1	502.9146	0.0226259	0.8804944
std_female_emply_rate	0.1920086	0.1920086	1	476.7081	2.1398964	0.1441706
std_urb_rate	0.1939529	0.1939529	1	445.3769	2.1615655	0.1422065
std_drug_alcohol_disorder_share	0.0004144	0.0004144	1	404.6826	0.0046182	0.9458532
hdi_over_avg	0.8363177	0.8363177	1	434.7646	9.3205906	0.0024050
haq_over_avg	0.0207039	0.0207039	1	431.1352	0.2307412	0.6312184
gdp_per_capita_over_avg	0.0053113	0.0053113	1	462.6052	0.0591936	0.8078829
female_emply_rate_over_avg	0.0095125	0.0095125	1	475.4289	0.1060145	0.7448713
urb_rate_over_avg	0.1954172	0.1954172	1	472.6852	2.1778848	0.1406726
drug_alcohol_disorder_share_over_avg	0.0000363	0.0000363	1	537.2276	0.0004051	0.9839494

From the ANOVA table above, all p-values of this model are greater than 0.05, which means the null hypothesis all variables cannot be rejected, and thus no variables are absolutely significant on the response. Thus, more selected variables strongly influence the response of the simple MLR than the multilevel MLR by comparing p-values.

Table 8: R squared table of multilevel MLR

R2m	R2c
0.0595417	0.9147372

Generally, the R squared table above shows two R squared values: the R^2 marginal and R^2 conditional.

The R^2 marginal, which includes the fixed variable only (the cell variable, or more formally, the random variable, is not counted). And it has the equation:

$$\frac{var(f)}{var(f) + var(r) + var(\epsilon)}$$

where $var(f)$ is the variance of fixed effects f, $var(r)$ is the variance of random effects r, and $var(\epsilon)$ is the variance of the model residuals ϵ. This describes the proportion of the variability explained by the fixed effects (Love, 2020).

The R^2 condtional, which includes both fixed and random variables (including the cell variable, or more formally, the random variable). And it has the equation:

$$\frac{var(f) + var(r)}{var(f) + var(r) + var(\epsilon)}$$

This describes the proportion of the variability explained by both random and fixed effects. Seen from the table, $R^2_{marginal} = 0.0595417$ and $R^2_{conditional} = 0.9147372$ (Love, 2020). If the $R^2_{conditional}$ is considered (random effect is counted), the multilevel MLR can describe much more variability than simple MLR. However, R^2 is not completely agreed by now (Love, 2020).

Table 9: AIC and BIC measurement table of multilevel MLR

aic	bic
829.9389	895.0971

AIC and BIC of the multilevel MLR are 829.9389 and 895.0971 respectively, which are much less than the AIC and BIC of the simple MLR, 1586.873 and 1647.687. Thus, the multilevel MLR are fitted better with the same train sets.

Thus, to conclude, 3 variables have strong effect on the response for the simple MLR while no on the response for the multilevel MLR. But the multilevel MLR explains more variability than the simple MLR by comparing conditional R^2 (although not widely recognized in practice) of the multilevel MLR and R^2 of the simple MLR. And the multilevel MLR is fitted better than the simple MLR with the same train set since the AIC and BIC of multilevel MLR are both much less than those of simple MLR.

Here is one important fact about the sampling and survey aspect to mention, Since the data are census data in each country/territory and cover the most countries in the world, all the datasets cover nearly all the cases in each country/territory in the world, the model can be fed with the adequate data. Also, based on the central limit theorem and law of large numbers, since there are many cases, the distribution of the average of data points converges to the normal distribution, and the average of the data points is closer to their expected value, and since the scales of each variable have huge gaps between each other, then to make the model summary statistics easier to read and evaluate (so not too large or high), each variable is scaled as:

$$standardize\_x_i = \frac{x_i - \mu}{\sigma}$$

based on the facts above.

Results

In this section, the result of the performance for two models running on the test set are shown below. This section mainly shows the correlation accuracy and the min max accuracy and mean absolute percentage error (MAPE), and also the causality effect reflected by the coefficient $\beta_{predictor\_over\_average}$ of the dummy variable $I(predictor > average)$.

The correlation between the actuals and predictions can be used as a form of accuracy measure. If the correlation accuracy is high, then the actuals and predicted values have similar directional movement, which means when the actuals values increase the predicteds also increase and vice-versa (Prabhakaran, 2017). Min max accuracy and MAPE are both measures of the prediction accuracy of statistical forecasting methods,

while

$$MinMaxAccuracy = \frac{min(actuals, predicteds)}{max(actuals, predicteds)}$$

and

$$MeanAbsolutePercentageError = mean(\frac{predicteds - actuals}{actuals})$$

If the min max accuracy is higher, the model predicts more accurately. And lower MAPE implies the model is more accurate.

Table 10: correlation accuracy of the simple MLR

	actuals	predicteds
actuals	1.0000000	0.2727435
predicteds	0.2727435	1.0000000

The table 10 above shows the correlation accuracy of the simple MLR to be 27.27435% ($\sigma_{pred|actual}$) by testing against the test set. This correlation accuracy uses the correlation between each pair of prediction made by the simple MLR with all the variables and actual in the test set, the higher the correlation accuracy, the better the model (the prediction develops in the same direction as the actual).

Table 11: min max accuracy and mean absolute percentage error of the simple MLR

min_max_accuracy	mape
0.5017672	0.2439018

The table 11 above shows the min max accuracy and MAPE of the simple MLR to be 50.17672% and 24.39018% respectively by testing against the test set. This min max accuracy and MAPE are computed with the formula above, with each pair of prediction made by the simple MLR with all the variables and actual in the test set. And the higher the min max accuracy and the lower the MAPE, the better the model (predictions closer to the actuals).

Table 12: correlation accuracy of the multilevel MLR

	actuals	predicteds
actuals	1.0000000	0.9454366
predicteds	0.9454366	1.0000000

The table 12 above shows the correlation accuracy of the multilevel MLR to be 94.54366% ($\sigma_{pred|actual}$) by testing against the test set. This correlation accuracy uses the correlation between each pair of prediction made by the multilevel MLR with all the variables and actual in the test set, the higher the correlation accuracy, the better the model (the prediction develops in the same direction as the actual).

Table 13: min max accuracy and mean absolute percentage error of the multilevel MLR

min_max_accuracy	mape
0.7335458	0.2706228

The table 13 above shows the min max accuracy and MAPE of the multilevel MLR to be 73.35458% and 27.06228% respectively by testing against the test set. This min max accuracy and MAPE are computed with the formula above, with each pair of prediction made by the multilevel MLR with all the variables and actual in the test set. And the higher the min max accuracy and the lower the MAPE, the better the model (predictions closer to the actuals).

By above, the multilevel MLR has much higher min max accuracy and correlation accuracy higher than the simple MLR with the test set. However, the multilevel MLR has slightly higher MAPE than the simple MLR. Therefore, generally the multilevel MLR has better performance than the simple MLR with regard to the accuracy in prediction.

The last two tables display the causality effect by showing the coefficient on the dummy variables introduced in the **data** and **model** subsection of **Methodology**. The larger the absolute value of the coefficient of the dummy variable, the higher the impact of the predictor to the response.

Table 14: the estimation of coefficients of variables (model parameters) for simple MLR

term	estimate
(Intercept)	0.2357973
std_hdi	0.1503039
std_haq	-0.2544020
std_gdp_per_capita	0.0487247
std_female_emply_rate	0.2295335
std_urb_rate	-0.0810418
std_drug_alcohol_disorder_share	0.0197957
hdi_over_avg	-0.0724593
haq_over_avg	-0.1391804
gdp_per_capita_over_avg	-0.0062665
female_emply_rate_over_avg	-0.3342499
urb_rate_over_avg	0.0096334
drug_alcohol_disorder_share_over_avg	0.0763930

Table 15: the estimation of coefficients of variables (model parameters) for multilevel MLR

term	effect	group	estimate
(Intercept)	fixed	NA	-0.0431524
std_hdi	fixed	NA	0.0258655
std_haq	fixed	NA	-0.1113570
std_gdp_per_capita	fixed	NA	0.0064600
std_female_emply_rate	fixed	NA	0.1027655
std_urb_rate	fixed	NA	-0.1540221
std_drug_alcohol_disorder_share	fixed	NA	0.0053343
hdi_over_avg	fixed	NA	0.2431783
haq_over_avg	fixed	NA	-0.0378202
gdp_per_capita_over_avg	fixed	NA	0.0192872
female_emply_rate_over_avg	fixed	NA	-0.0243872
urb_rate_over_avg	fixed	NA	-0.1228652
drug_alcohol_disorder_share_over_avg	fixed	NA	0.0021096
sd___(Intercept)	ran_pars	Entity	0.9486737
sd___Observation	ran_pars	Residual	0.2995463

The table 14 above shows that the variable HDI, HAQ, the female labour participation rate and the share of people have drug or alcohol disorder have significantly stronger causality effect with the response than all other parameters for the simple MLR, with -0.0724593, -0.1391804, -0.3343499, and 0.0763930 respectively, which means the response has the quantity of the dummy variable coefficient more or less value if the corresponding predictor of the dummy variable is beyond the threshold (average in this model) due to the RDD design. And from the table 15, it can be seen that the coefficients of the random terms are much larger than the coefficients of the fixed terms in terms of the absolute value. Then, the coefficients of dummy variables show that HDI (0.2431783) and urbanization rate (-0.1228652) have the strongest causality effect on the response of the multilevel MLR and others are nearly the same. And it is worth noting that if the country/territory is considered in addition for each sample while building the model, the causality strength of variables may be different, and the order of the causality effect may change due to various model selections.

Discussion

In this project, an observational study about the influence of selected social factors on the annual AIDS/HIV infection cases are made, and the impact is tested by building the linear regression model from two aspects: first, whether the social factor is a significant factor than influences the AIDS/HIV infection, second, how strong is the causality effect asserted by each social factor on the AIDS/HIV infection. To build such a model, the observational data collected from each country from 1990 to 2015 are used to feed the model. Before building the model, all the census datasets are combined into a huge one and then split to a train set and test set, the work is done by the R code `datagen.R`, the distribution of the data points have been discussed and plotted in the subsection **data** of the section **Methodology**. To compare and contrast the performance and select a better linear regression model, two functions are used to build the model, the first one is **lm**, which builds a simple multivariate linear regression (MLR); and the second one is **lmer**, from the package **lme4**, which can build a multilevel MLR. All the two models are built with the numerical variables listed in the **Methodolgy** section, the introduction and the source of all the variables to build two models, have been listed in the **Methodolgy** section as well (table 1 to 4 and also items in the **data** subsection of the section **Methodology**). They generally share the same variables except that the multilevel MLR requires a level factor, the categorical variable "Entity". And also, the Regression Discontinuity Design (RDD) is used to build two models, which helps to observe the causality link with the observational data, because seen from the datasets in figure 1 and 2, the data points distribute unevenly, all datasets have data points left skewed or right skewed. Thus, it is reasonable to build a piecewise regression and then RDD is used. Before building the model and testing the model, the predictor variables required are scaled by the customized function `standardize` because the scales for each variable have huge gaps between each other, and if the data are not scaled, the model parameters will be extremely large or small. Note that the multilevel MLR is built with the level variable "Entity", the country/territory each observation obtained from. After two models are built, the model statistics for diagnosis are collected, including p-values, R^2, AIC and BIC (please read the table 5 to 9). Finally, two models are tested by the test set and the accuracy and causal effect are recorded as the result.

The p-values show that for the simple linear regression, HDI, HAQ and female employment rate are significant to the response because the null hypothesis, the variables do not have significant impact on the variable, for them can be rejected, while all variables for the multilevel MLR have the possibility to be dropped. Also, based on the AIC and BIC value comparison, the multilevel MLR is fitted with the train set better than the simple MLR (AIC and BIC of multilevel MLR are all less than the simple MLR). The conditional R^2 (both fixed and random effects are included, which is special for the multilevel model) of the multilevel MLR is much higher than the R^2 of the simple MLR, which means the multilevel MLR can describe more variability than the simple MLR, nevertheless, the R^2 for the multilevel MLR has not been commonly recognized currently. Then to summarize the result on the performance of the prediction-making from the two models, the correlation accuracy, min max accuracy and MAPE are computed and shown in the table 10 to 13. The correlation accuracy of multilevel MLR is much higher than the simple MLR, it means that the predictions made by the multilevel MLR have much higher chance to change in the same trend with the actuals than the simple MLR, simply speaking, if the actuals increase/decrease, the predictions made by multilevel MLR is more likely to increase/decrease than the simple MLR. Also, the simple MLR has much lower min max

accuracy and slightly lower MAPE than the multilevel MLR, which means in general, multilevel MLR tends to make predictions closer to the actuals than the simple MLR.

To check the causality effect on of each variable on each model, each model has the dummy variable for each predictor variable. The larger the coefficient of the dummy variable, the stronger the causality link between the response and the predictor, and this is clear when the predictor is beyond the threshold (average of the data for each variable). Summary of the model parameters of two models are shown in the table 14 and 15. The table shows that for the simple MLR, the variable HDI, HAQ, the female labour participation rate and the share of people have drug or alcohol disorder have the strongest causality link. However, for the multilevel MLR, the HDI and the share of population in urbanization have the strongest causality link. Therefore, if the extra factors are counted, for instance, in this case, the country/territory is considered as the level variable, the causality link between the predictor and the response differs, and the order also differs.

Thus, by the diagnosis and the running result of two models, multilevel MLR is preferred with respect to accuracy and data fitting, and HDI (an index from education, personal income and life expectancy) might be the most important social factor to consider for containing AIDS/HIV infection for each country, since it cannot be rejected in the simple MLR, and has relatively strong causality effect with the AIDS/HIV infection for both models. And also, model selection might be important not only because of improving accuracy and data fitting, but also weighting the variables when building the model, for example, some social factors cannot be rejected for the simple MLR but all might be rejected for the multilevel MLR, and also each variable has different causality effect for the simple MLR and multilevel MLR. Therefore, model selection for predicting AIDS/HIV spread and other infectious diseases by social factors and government policies may vary based on the datasets researchers have.

To summarize, there are some drawbacks in the research design. First, since the data distribute extremely unevenly, seen from the data plot, it might not be even not adequate to make piecewise linear regression model, the figures of data actually shown that there are many outliers for the simple MLR, and also the min max accuracy and MAPE for both models are not ideal. Second, the scales data in each dataset are obviously far away from each other, and the scaling must be done in order to make the causality inference and other statistical observation. However, the scaling may lead to extra error in model construction. And finally, when making causal inference with RDD, it is important to find a precise threshold, but actually, the threshold in this experiment is not clear, and the average is just a pre-assumed threshold. The causality inference can be more advanced if the precise threshold can be found even if such a threshold is not pre-set.

To improve the model performance based on the problems listed above, here are some possible approaches, or future steps can be done. For the first problem, it is possible to test polynomial regressions with different highest orders, and also both single level and multilevel, although there might be problems of overfitting or underfitting. For the second problem, it is possible to consider computing indices, like HDI (human development indices), with a scale in an explicit range for all variables (both responses and predictors), and then the model parameters will not be that extreme, and the data scaling can be avoided. For the last problem, the FUZZY technique can be applied to find thresholds while building models with RDD if the threshold is unclear. A clearer cut means the data beyond the threshold are more likely under the treatment, and the causality effect can be tested and found more definitely.

References

1. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018. (from http://ghdx.healthdata.org/gbd-results-tool, download from https://ourworldindata.org/hiv-aids, Number of new cases of HIV, 1990 to 2017)
2. Gapminder, HYDE (2016) and United Nations Population Division (2019) (from http://www.gapminder.org/ and https://www.gapminder.org/data/documentation/gd003/, download from https://ourworldindata.org/world-population-growth, World population by region)

3. United Nations Development Programme (UNDP), UNDP (from http://hdr.undp.org/en/indicators/137506#, download from https://ourworldindata.org/search?q=human+development+index, Human Development Index, 1980 to 2017)

4. Global Burden of Disease Study 2015. Global Burden of Disease Study 2015 (GBD 2015) The Lancet in May 2017 in "Healthcare Access and Quality Index based on mortality from causes amenable to personal healthcare in 195 countries and territories, 1990–2015: a novel analysis from the Global Burden of Disease Study 2015." (from http://ghdx.healthdata.org/record/global-burden-disease-study-2015-gbd-2015-healthcare-access-and-quality-index-based-amenable, download from https://ourworldindata.org/search?q=healthcare+access+and+quality+index, Healthcare Access and Quality Index, 1990 to 2015)

5. Maddison Project Database, version 2018. Bolt, Jutta, Robert Inklaar, Herman de Jong and Jan Luiten van Zanden (2018), "Rebasing 'Maddison': new income comparisons and the shape of long-run economic development", Maddison Project Working paper 10 Historical data and reconstructions from a large number of different sources. See the accompanying paper for details (from https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2018 and http://www.ggdc.net/maddison/oriindex.htm,
download from https://ourworldindata.org/search?q=gdp+per+capita, GDP per capita, 1870 to 2016)

6. World Bank – World Development Indicators, International Labour Organization, ILOSTAT database. Data retrieved in September 2018. (from http://data.worldbank.org/data-catalog/world-development-indicators, download from https://ourworldindata.org/female-labor-supply, Female participation in labor markets, country by country)

7. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2016 (GBD 2016) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2017. (from http://data.worldbank.org/data-catalog/world-development-indicators, download from https://ourworldindata.org/urbanization, Share of people living in urban areasUN (1960 to 2017))

8. Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2016 (GBD 2016) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2017. (from http://ghdx.healthdata.org/gbd-results-tool, download from http://ghdx.healthdata.org/gbd-results-tool, Share of the population with alcohol or drug use disorders, 1990 to 2016)

9. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software,4(43), 1686, https://doi.org/10.21105/joss.01686

10. Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, 82(13), 1-26. doi: 10.18637/jss.v082.i13 (URL: https://doi.org/10.18637/jss.v082.i13).

11. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

12. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

13. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

14. Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

15. Kamil Bartoń (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn

16. David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.1. https://CRAN.R-project.org/package=broom

17. Ben Bolker and David Robinson (2020). broom.mixed: Tidying Methods for Mixed Models. R package version 0.2.6. https://CRAN.R-project.org/package=broom.mixed

18. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

19. Sepkowitz, K. A. (2001, June 07). AIDS - The First 20 Years: NEJM. Retrieved December 09, 2020, from https://www.nejm.org/doi/full/10.1056/NEJM200106073442306

20. CDC. (2020, November 03). About HIV/AIDS. Retrieved December 09, 2020, from https://www.cdc.gov/hiv/basics/whatishiv.html

21. Anderson, R. M., Medley, G. F. (1988). Epidemiology of HIV infection and AIDS. Aids, 2. doi:10.1097/00002030-198800001-00009

22. Rom, W. N., Markowitz, S. B. (2015). Environmental and Occupational Medicine. Philadelphia, US: Lippincott, Williams, Wilkins.

23. CDC. (2005, February 4). HIV and Its Transmission. Retrieved December 09, 2020, from https://web.archive.org/web/20050204141148/http://www.cdc.gov/HIV/pubs/facts/transmission.htm

24. U. (2006). 2006 report on the global AIDS epidemic. Geneva: UNAIDS.

25. McCullom, R. (2013, February 26). An African Pope Won't Change the Vatican's Views on Condoms and AIDS. Retrieved December 09, 2020, from https://www.theatlantic.com/sexes/archive/2013/02/an-african-pope-wont-change-the-vaticans-views-on-condoms-and-aids/273535/

26. CATIE. (2018). The Social Determinants of Health and Structural Interventions. Retrieved December 09, 2020, from https://www.catie.ca/en/hiv-canada/8/8-1

27. Love, K. (2020, October 5). R-Squared for Mixed Effects Models. The Analysis Factor. https://www.theanalysisfactor.com/r-squared-for-mixed-effects-models/

28. Prabhakaran. (2017). Linear Regression With R. R-Statistics.Co. http://r-statistics.co/Linear-Regression.html