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

Friday, December 18, 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 statistics model when the observational study is inter-regional.

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

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 analysis 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 numerical data, and the prediction is also numerical. Linear regression model works based on numerical 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. 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. 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, in general, the introduction section about the data and the model is referred to as which implemented by RDDtextbf{Methodology}. In the **Results** section, all the figures and tables generated from the predictions made by the model or the some diagnosis about the accuracy and statistical summary of model will be displayed, and also may include the analysis on the tables and figures. In the **Discussion**the causality observed from the model will be summarized, section, 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 possible improvements (further steps can be done) will be listed. In the final section (**References**), all references used in this report are provided.

Results

I got marks deducted form general format in PS3, please provide me some ways to generate clean table with the diagnosis of the model and the dataset from tidyverse and how to number each figure.

References

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

2. 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).

3. 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.

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

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

6. 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
7. Kamil Bartoń (2020). MuMIn: Multi-Model Inference. R package version 1.43.17. https://CRAN.R-project.org/package=MuMIn
8. 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
9. 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
10. 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
11. CDC. (2020, November 03). About HIV/AIDS. Retrieved December 09, 2020, from https://www.cdc.gov/hiv/basics/whatishiv.html
12. Anderson, R. M., Medley, G. F. (1988). Epidemiology of HIV infection and AIDS. Aids, 2. doi:10.1097/00002030-198800001-00009
13. Rom, W. N., Markowitz, S. B. (2015). Environmental and Occupational Medicine. Philadelphia, US: Lippincott, Williams, Wilkins.
14. 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
15. U. (2006). 2006 report on the global AIDS epidemic. Geneva: UNAIDS.
16. 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/
17. 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