STA304 Final Project: Causal Inference Analysis between the Factor Immunization Coverage with other Possible factors on Life Expectancy

Qihui Huang, 1004932905

12/7/2020

Github Repo

https://github.com/heyhaileyhhh/304finalreport_qh.git

Abstruct

Life expectancy has always been an interesting topic, and lots of statistical researches have examine this using different kinds of indexes. Since immunization has not been taken into consideration yet, we will look into this factor and find out the relation to life expectancy. This question-driven is about the causal inference and multiple linear relation between factors that influencing life expectancy. Propensity score matching, logistic regression and multiple linear regression model will be presented in this report, so that all analysis will be proceed based off these models mainly.

Key Words

Propensity Score, Causal Inference, Life expectancy, Immunization, Observational study, Health.

Introduction

Nowadays, the concept of causality has become more and more popular in the domain of Statistics studies. Causal inference examines the potential cause and effect relations, that is how the outcome will change when the cause varies. It has been widely used for analysis in health, social, behavioral and economic sciences. With the techniques like propensity score matching being introduced, a quasi-experimental method, it helps us create artificial control groups to estimate the relation of the possible cause and effect conveniently. Usually, when we do not randomly assign data into two groups, causality cannot be inferred there because of the non-randomization, but propensity score matching could minimize effect of selection bias therefore the influence of confounding factors will be reduced effectively. Moreover, observational data extracted from WHO is used in this report instead of complete experimental data. As observational data is much more detailed and also a reliable source containing related rich information, more insights and creative observations/analysis can be performed here.

The main goal of this report is to find out that whether a causal relation between life expectancy and immunization coverage is presented in countries around the world in a 15 years period. Other related

factors are also covered. Life expectancy has always been an intriguing topic, and there have been lots of studies conducted in the past on factors affecting life expectancy, like personal income, morality, alcohol consumption, GDP and health care insurance etc. During COVID-19 pandemic, topics like immunization and vaccination have become prominent, motivating us to explore the effect of the vaccines. As COVID vaccines are still in approval process, people all around the world are expecting the release of the vaccines that would lower down the number cases and benefit all people's personal health in long run. Therefore it is non-negligible when we consider the factors like immunization rate that may impose effects on life expectancy. This motivates us to explore the casual inference between the life expectancy and vaccination converge, and build a regression model based off multiple factors, considering data in a decade period all around the world. (Kumar Rajarshi 2017)

The data set from (Anthony Goldbloom 2010) will be used in this analysis. We aim to find causal relations between factors with life expectancy. Other factors like countries economic condition, mortality and average education level will also be taken into consideration. In Methodology section(section 2), data cleaning is performed and model is built here. Propensity score matching is used here to discern whether individuals get vaccinated already and whether individuals die before the average age of population. Results of propensity score matching will be presented in Result section(section 3). Interpretation of the results, graphs, weakness, conclusion and further steps will be included in the Discussion part(section 4).

Methodology

Data Description and Data Cleaning Process

Some definitions of variables in the data set:

- *life_expectancy*: Life Expectancy in age
- *status*: A categorical variable, country's Developed or Developing status
- *adult_mortality*: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- *Alcohol*: Alcohol, recording per capita (15+) consumption (in litres of pure alcohol)
- *healthcoverage_percent*: Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- *total_expenditure*: General government expenditure on health as a percentage of total government expenditure (%)
- *BMI*: Body Mass Index
- *Schooling*: Number of years of Schooling(years)
- *vaccine_total_over_ave*: A binary variable recording whether a country's vaccine coverage is over the world's average vaccine coverage or not.

The original data set is obtained from the World Health Organization (WHO) data repository. It contains detailed health related factors from 193 countries and corresponding economic status of each country from the United Nation website. It records people's health status more than a decade, from 2000 to 2015. The data set has now been updated and reduced to 22 variables with 2938 rows by the user on Kaggle website, which contains mainly 5 categories: mortality factors, immunization related factors, economical factors, social/educational factors and personal habit factors.

I have cleaned the data set sourced from the Kaggle website by performing some basic data cleaning steps, including mutating and choosing specific variables. There are in total 22 variables including both categorical and numerical, and these are all potential factors that may impose different levels of effect on life expectancy; 13 of them have been chosen and put into a new cleaned data set, like *life expectancy*, *country*, *BMI* and different types of vaccines etc.. These variables are more representative and clean compared to other factors, and will be easily manipulated in the later analysis.

For convenience's sake, especially in matching propensity scores, a new binary variable is created, measuring whether a country's immunization coverage exceeds the world's average or not. This is simply done by adding all general vaccines coverage rate together for each country, then compared each of them with the mean of the world vaccine coverage. In this case, 1 represents the country has met the world's average immunization coverage, and 0 represents the country has not met the criteria here yet. Moreover, renaming variables and removing rows with NA or invalid numbers are done in this part as well.

Note: Packages used for cleaning and modeling are: (Robinson, Hayes, and Couch 2020),(Wickham et al. 2019),(Wickham and Hester 2020),(Gelman and Su 2020),(Wickham et al. 2020)

A Glimpse of Cleaned Data Set: Note that only part of the dataset is shown here

status	life_expectancy	adult_mortality	Alcohol	healthcoverage_percent	Hepatitis_B
Developing	65.0	263	0.01	71.279624	65
Developing	59.9	271	0.01	73.523582	62
Developing	59.9	268	0.01	73.219243	64
Developing	59.5	272	0.01	78.184215	67
Developing	59.2	275	0.01	7.097109	68
Developing	58.8	279	0.01	79.679367	66
Developing	58.6	281	0.01	56.762217	63
Developing	58.1	287	0.03	25.873925	64
Developing	57.5	295	0.02	10.910156	63
Developing	57.3	295	0.03	17.171518	64
Developing	57.3	291	0.02	1.388648	66
Developing	57.0	293	0.02	15.296066	67
Developing	56.7	295	0.01	11.089053	65
Developing	56.2	3	0.01	16.887351	64
Developing	55.3	316	0.01	10.574728	63

Model

The propensity score (probability of being treated) will be estimated by constructing a logistic regression model at first. Our treatment is immunization coverage in a country, and life expectancy is our outcome of interest, so that propensity score matching will be for the immunization coverage propensity. Nearest neighbor matching will be used next to find the closest match in the untreated group, keeping only the rows that are matched. Lastly, a matched sample will be run for common multiple regression analysis. Details of each steps and fitted model will be elaborated below.

Propensity Score Matching Process Firstly, logistic regression model is chosen because the treatment is a binary variable, and we want to predict the probability of it as an outcome based off other factors in the dataset. Here, the treatment immunization coverage is now presented as outcome based on different co-variates in our dataset. In this step, we think that predictors like countries' status, alcohol consumption, adult morality, GDP etc. are able to explain whether a country is treated or not. The model gives out the probability of the treatment. Propensity scores for different countries are obtained here and added to the data set, which is exactly the probability of a country's vaccine coverage over world's average.

Logistic regression model on treatment is defined as

$$log(\frac{\hat{p_{vaccine}}}{1-\hat{p_{vaccine}}}) = \hat{\beta}_0 + \hat{\beta}_1 * x_{statusDeveloping} + \hat{\beta}_2 * x_{adult\_mortality} + \hat{\beta}_3 * x_{Alcohol} + \hat{\beta}_4 * x_{total_expenditure}$$

$$+\hat{\beta}_5 * x_{GDP} + \hat{\beta}_6 * x_{BMI} + \hat{\beta}_6 * x_{Schooling} + \hat{\beta}_7 * x_{health\_coverage\_percent}$$

Table 1: Summary coefficient table of fitted logistic regression model between treatment vaccine coverage and other factors

Intercept/Predictors	Estimates	P-value
intercept	-2.0880	4.81e-06
statusDeveloping	-0.6484	0.0117
adult_mortality	-0.0013	0.0049
Alcohol	-0.0189	0.3301
total_expenditure	0.0756	0.0023
GDP	9.671e-05	9.60e-07
BMI	0.0016	0.6362
Schooling	0.7556	$<$ 2e-16
healthcoverage_percent	0.0007	4.84e-09

In **Table 1**, if we interpret each coefficients in the model, each $\hat{\beta}$ represents the relative change in log odds as the value for the predictor variable increases or decrease by one unit. For example, holding all other predictors constant, the one unit change in number of years attending school(0.7556) has stronger effect on log odds, comparing with the unit change in adult mortality(-0.0013), because of a larger absolute value of coefficients. The probability of vaccine coverage $\hat{p}$ for each specific cases(rows) can be calculated based on variables value using the same procedure. Notice that the probability might be against the actual vaccine coverage result, which means the propensity score/our forecast does not imply whether treated or untreated.

Furthermore, most of the P-values of each predictor are smaller than the significance level 0.05, which means they all have contributions in logistic regression that predict the probability of vaccine coverage over average in one country. A few of them are greater than 0.05, but we keep this for now since it does not exert large effect on our result by now.

Next, for every country which was actually treated(over world's average immunization coverage), we match the untreated group(under world's average immunization coverage) based on the similar propensity score.Basically, the propensity scores will be compared and matched from treated and untreated groups. After matching, the dataset has been reduced to 1226 observations with 17 columns, and there are exactly 613 pairs of treated and untreated observations. The dataset only keeps the observations that with matched propensity scores in treated and untreated groups.

total_expenditure	Diphtheria	GDP	Schooling	vaccine_total	vaccine_coverage_over_ave	fitted
4.20	9	21.76880	4.4	187	0	0.1520130
3.50	93	232.79455	4.7	279	1	0.1581297
3.14	98	287.42222	5.0	280	1	0.1654857
3.30	92	43.75445	5.2	276	1	0.1785284
3.30	94	297.82859	5.3	282	1	0.1788853
3.29	91	317.32943	5.3	273	1	0.1797005
3.24	9	482.14994	5.1	27	0	0.1812576
3.69	94	326.82564	5.2	282	1	0.1839738
3.60	96	582.77553	5.0	288	1	0.1849504
7.10	83	127.42966	5.2	242	0	0.1936776

Actual Causal Inference Analysis with Multiple Linear Regresion Model The multiple linear regression is chosen since the response variable here, life expectancy, is a numeric. The rest of the variables are not showing any outstanding evidence to use other models, so multiple linear regression model might be the best choice. Diagnostic of the fitted model will be checked in the Weakness section.

The final fitted multiple linear regression model is defined as:

$$\widehat{life\_expectancy} = \hat{\beta}_0 + \hat{\beta}_1 * x_{statusDeveloping} + \hat{\beta}_2 * x_{adult\_mortality} + \hat{\beta}_3 * x_{Alcohol} + \hat{\beta}_4 * x_{total\_expenditure} +$$

$$+\hat{\beta}_5 * x_{GDP} + \hat{\beta}_6 * x_{BMI} + \hat{\beta}_6 * x_{Schooling} + \hat{\beta}_7 * x_{health\_coverage\_percent} + \hat{\beta}_8 * x_{vaccine\_coverage\_over\_ave}$$

Table 2: Coefficients are extracted from the summary table of the model

Intercept/Predictors	Estimates	P-value
intercept	54.4500	< 2e-16
statusDeveloping	-0.1820	0.8380
adult_mortality	-0.0306	< 2e-16
Alcohol	-0.2710	2.19e-07
total_expenditure	-0.1446	0.0160
GDP	8.045e-05	0.147
BMI	0.0739	< 2e-16
Schooling	1.487	< 2e-16
healthcoverage_percent	0.0001	0.652
vaccine_coverage_over_ave	1.6980	3.52e-10

The final multiple regression model is produced after propensity score matching.The vaccine coverage predictor with other factors are added to predict the relationship between life expectancy and themselves. The intercept $\hat{\beta}_0$ represents the average of the life expectancy when other predictors have value 0. The rest of $\hat{\beta}$s before each predictors are the expected change in response variable life expectancy resulted from one unit change in this variable, holding other variables constant. If it is a categorical/binary variable, so then it measures the average difference in response variable based off different levels in this kind of variables.

From **Table 2**:

- $\hat{\beta}_0$: Intercept: the average of the life expectancy is 54.45 years when other predictors have value 0.
- $\hat{\beta}_1$: The average difference between developing country and the baseline variable developed country is 0.182 years on life expectancy, holding other variables constant.
- $\hat{\beta}_2$: The expected change in life expectancy will decrease by 0.0306 years if adult mortality rate increases by one unit, holding other variables constant.
- $\hat{\beta}_3$:The expected change in life expectancy will decrease by 0.2710 years if alcohol consumption increases by one unit, holding other variables constant.
- $\hat{\beta}_4$:The expected change in life expectancy will decrease by 0.1446 years if total expenditure increases by one unit, holding other variables constant.
- $\hat{\beta}_5$:The expected change in life expectancy will increase by 8.045e-05 years if GDP increases by one unit, holding other variables constant.
- $\hat{\beta}_6$:The expected change in life expectancy will increase by 0.0739 years if BMI increases by one unit, holding other variables constant.
- $\hat{\beta}_7$:The expected change in life expectancy will increase by 1.487 years if years attending school increases by one unit, holding other variables constant.
- $\hat{\beta}_8$:The expected change in life expectancy will increase by 0.0001 years if health coverage percentage increases by one unit, holding other variables constant.

- $\hat{\beta}_8$:The expected change in life expectancy will increase by 1.698 years if vaccine coverage increases by one unit, holding other variables constant.

Table 3: Other statistics and criteria of the model

Statistics	Value
R-squared	0.7127
Adjusted R-squared	0.7105
Residual standard error	4.6420

Result

Interpretation of result in Model Section above

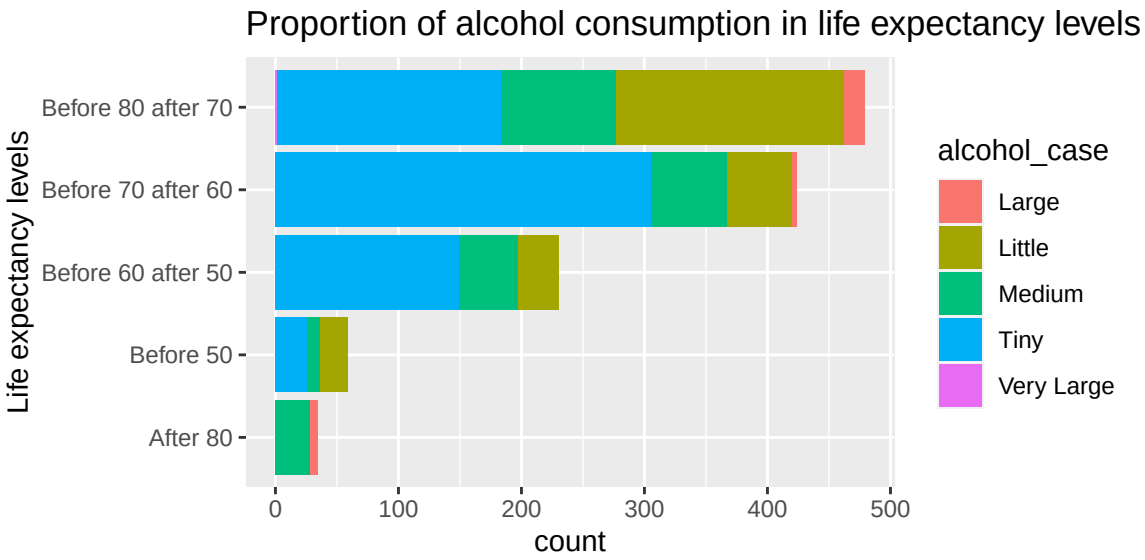
The **Table 2** above shows the relationship between each predictors and the response variable life expectancy. Comparing their absolute value of the coefficients, number of years attending school and countries with vaccine coverage over world's average seems like imposing more positive effects on estimated life expectancy, which means these two factors correlates with the response variable and increase the people's life expectancy over 1 year, 1.487 years and 1.698 years respectively. Causality can be confirmed here, between life expectancy and immunization coverage, and its p-value is also significant.

Other factors like the status of a country(developed or developing country), GDP, BMI(coefficient 0.0739), alcohol consumption(coefficient -0.2710) and expenditure on health care(coefficient 0.0001) are less likely on extending life expectancy since they are either not strongly correlated with the outcome or they are bringing negative effect on the result. #For instance, one unit increase in alcohol consumption brings down about 0.3 year of life expectancy, which imposes the most negative effect on our outcome; the variable GDP has tiny effect on life expectancy(coefficient 8.045e-05), the rise in life expectancy is so small so that can be ignored.

Judging from the P-value and other statistics in **Table 2** and **Table 3**, variables like GDP, health care coverage percentage and status of a country have significance level larger than 0.05, which means these variables may not have a linear relation with the response variable life expectancy, so that they almost contribute nothing on predicting life expectancy. The adjusted R^2 (0.7105) showed in **Table 3** is fairly close to 1, the largest value that R^2 can get. It measures how close are the data to fitted line. In this case, the model is valid since most of them are close to what we predict.

Interpretation of result from other plots and tables

Plot 1: Relationship between Alcohol Consumption and Life Expectancy



The barplot exhibits the proportion of alcohol consumption per person in different age groups over 15 years. According to the legend shown in the graph, it is obvious that people who consume a tiny bit of alcohol take up the largest part in each age group. By tiny bit, we mean under 3.0 liters per capita. Medium and little amount of consumption is also common in this case. Compared with the people who tend to have longer life expectancy, ones who are short-lived(especially people who died before 50) tend to drink on "Little" level as it takes up almost half of the population.

Plot 2: Relationship between vaccine coverage and Life Expectancy

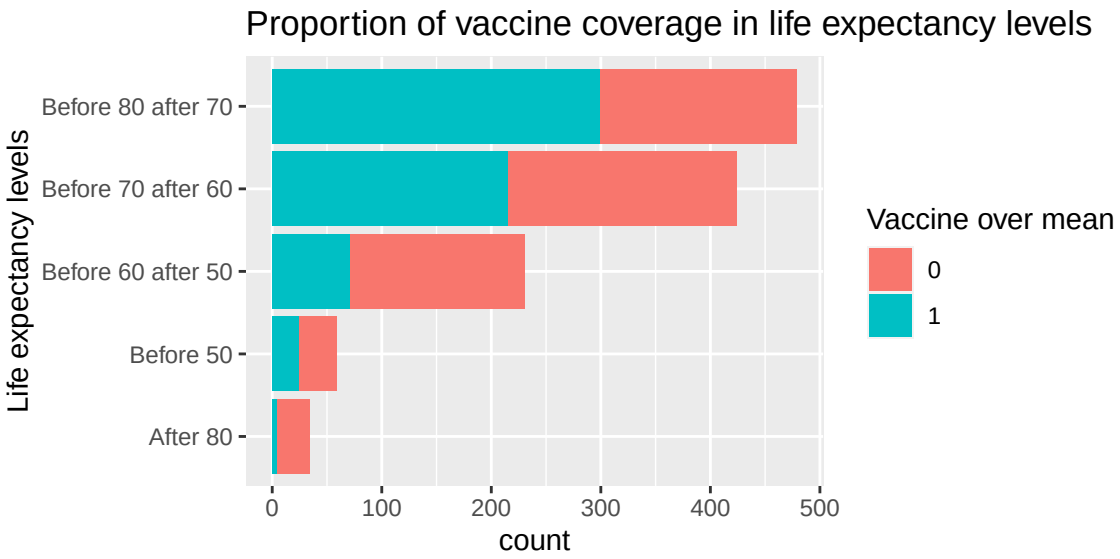
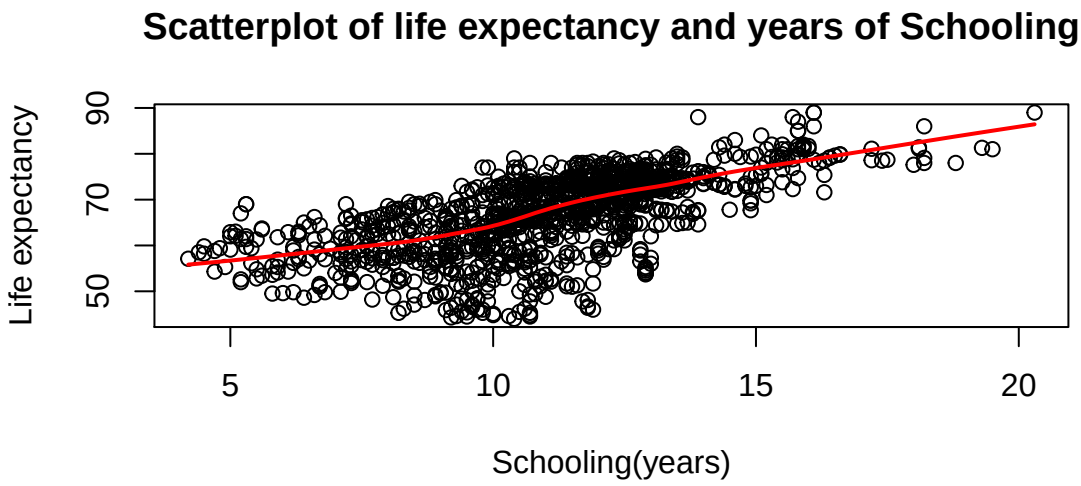


Table 4: Proportion of vaccine coverage in Life Expectancy

life_expectancy_case	proportion
After 80	0.1176471
Before 50	0.4067797
Before 60 after 50	0.3086957
Before 70 after 60	0.5070755
Before 80 after 70	0.6242171

Examining **Table4** and **Plot 2**, it shows the relationship between vaccine coverage and life expectancy. We notice that longer life expectancy cases generally have wider coverage of vaccination rate, while the shorter expectancy cases have lower coverage but not so much. In the age group from 60 to 70 and 70 to 80, the immunization coverage in these countries are all over the half of population in this category, around 51% got vaccinated in the 60-70 group and 12% more in 70- 80 age group. Other than these two groups, populations in countries with shorter life expectancy, like 50-60 and younger than 50 groups have limited coverage in vaccination, only about 30%-40% people have received vaccinations. As people with life expectancy over 80 have fewer cases, 11% immunization coverage is not significant here.

Plot 3: Scatterplot of Schooling and Life Expectancy



As the number of years attending school is significant and contributes the most in prolonging life expectancy other than the vaccination coverage, a model is fitted between only life expectancy and years of schooling. It is clear that there is a relatively strong and positive relationship between the explanatory and the response variable. Life expectancy grows as the schooling years grows, which means these two are linearly correlated and the longer ones attended the school, ones are more likely to live longer than others,

Discussion

Summary

So far, we conduct analysis on causal inference between life expectancy and potential factors that may have the ability to influence life expectancy, mainly on the cause-effect relation of factor vaccination coverage and life expectancy. Other factors like number of years in school, people's fitness measured in BMI, personal habits like alcohol intake and spending in healthcare are also taken into consideration. To accomplish our goals, propensity score matching is used here to reduce the bias therefore we can infer whether the casualty is presented. Furthermore, the multiple linear regression model is built for assessing the effects of all factors together to the life expectancy. Factors that are significant or strongly correlated with the response are picked out and investigated deeply. Details and interpretation of results will be discussed below.

Conclusion

In conclusion, there is a causality relation between vaccination coverage and life expectancy, based on 15-year period data of 193 countries around the world, meaning that wider coverage of immunization results

in longer life expectancy. Countries with vaccination(Hepatitis B, Polio, Diphtheria tetanus toxoid vaccine) coverage over average world's average coverage tends to result in a longer life expectancy. The propensity score matching technique provides us with a matched dataset with minimized selection bias; after fitting the final linear regression model, the p-value of vaccination coverage over average variable is significant, which shows the cause-effect relation is actually presented here. To further investigate the relation between vaccination coverage and life expectancy, plot and proportion tables are drawn to visualize the relationship. Reading from the proportion table **Table 4**, it is clear that people with longer life expectancy have been vaccinated when they are young. People who do not get general types of vaccines are more likely to be short-lived, died before 60 years old. After all, the vaccination is important for ourselves and world's health system, and it will necessarily extend people's life, control the diseases that we have already known.

Herd immunity is introduced these days during pandemic, and the ideal way to achieve this is to get vaccination. It means that when most people in a community are immune to the diseases, the rest of immune-compromised people are less unlikely to get the disease. So vaccines provide an immune response not only for healthy, normal people, but also protect the elderly, children, new born babies and people who are too weak to get vaccinated. Vaccines like Hepatitis B, Polio and Smallpox effectively controlled the spread of highly contagious diseases. Overall vaccination is important that could save thousands of people's lives as well.(Clinic 2019)

Other factors that are significant in p-value are also noteworthy. Alcohol consumption is considered to be a major predictor in estimating life expectancy. From **Plot 1**, populations are not over-drinking in most age groups. Most of them intake alcohol in a moderate quantity. At this consumption, the age group of 60-70 years has most of it. This shows that typical alcohol consumption at this amount is more likely to have a life expectancy falling in the 60-70 section, implying that small amount of intake (less than 3.0 liters per capita) will not affect our health system in a negative way; in contrast, drink moderate level of alcohol drinks occasionally may even beneficial to our bodies. But we cannot simply conclude from this graph, as there are some large amounts of alcohol intake in long-lived groups. This usually varies in different individuals, but keeping a good drinking habit will definitely prolong people's life.(Miller 2010)

Factor years attending school is also significant, and is strongly correlated with life expectancy. Even though we cannot directly conclude a causality between schooling and one's life duration, we could say with increasing number of years studying, people are inclined to obtain a longer lifetime. People study for a longer time to gain more knowledge about diseases, health care, fitness etc., so that they are able to and intend to acquire advanced medical resources. The dataset shows that the population with its citizens having more time at school spent more money on health care, also alcohol consumption is moderate, and BMI(body mass index) that measures one's fitness is in good standard. This raises our awareness of receiving education, even if the importance may not be reflected immediately in our life but will benefit people in the long run.

Weakness/Drawback & Next Steps

For the a dataset I am working on, since the data was collected from 2000-2015, it is kind of outdated as the latest data is five years ago. Data and information changes rapidly when countries around the world are developing fast, so that the dataset might not be fitted into our current situation. But the data are detailed and authentic from the World Health Organization, so it is still meaningful since the general trend is not going to be reversed suddenly in a few years. Also, the technique like propensity score matching may also have some potential drawbacks. This method is trying to mimic complete randomization of data, but then turns out imbalance and bias the data are even more significant, because covariates value might be away from the treated value. We could use other method like caliper matching, radius matching or Mahalanobis distance, but these method are often complicated than the propensity score matching.

For the model that was built previously, there are a few variables that are not significant which they probably do not follow a linear model, and it has not been taken out. So using these less significant variables might bias the result. The way to find a much more fitted model is to use backward elimination with BIC procedure. It will automatically select significant variables step by step, by removing the largest P-value(less significant)

one at a time. The final model is stated as $\widehat{life\_expectancy} = \hat{\beta}_0 + \hat{\beta}_1 adult\_mortality + \hat{\beta}_2 Alcohol + \hat{\beta}_3 GDP + \hat{\beta}_4 BMI + \hat{\beta}_5 Schooling + \hat{\beta}_6 vaccine\_coverage\_over\_ave$

Concluding Remarks

The intention of this paper is to show statistical and scientific observations of factors that affecting the life expectancy, based off real data collected from World Health Organization(WHO). Even though the dataset is collected five years ago, we can still use it to estimate the general trend of it. As there is actually a causal relationship between vaccination coverage and life expectancy, getting vaccinated is crucial for not only protecting ourselves but also for our community, especially during COVID-19 pandemic. Moreover, developing healthy habits are also important in prolonging life, like limiting alcohol intake, exercising regularly, doing physical examination etc..

References

Anthony Goldbloom. 2010. *Kaggle Data Science Company*. Google. www.kaggle.com.

Clinic, Mayo. 2019. *Herd Immunity and Covid-19 (Coronavirus): What You Need to Know*. https://www.mayoclinic.org/herd-immunity-and-coronavirus/art-20486808.

Gelman, Andrew, and Yu-Sung Su. 2020. *Arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. https://CRAN.R-project.org/package=arm.

Kumar Rajarshi. 2017. *Life Expectancy (Who): Statistical Analysis on Factors Influencing Life Expectancy*. World Health Organization. https://www.kaggle.com/kumarajarshi/life-expectancy-who.

Miller, Kelli. 2010. *For Some, Moderate Drinking May Prolong Life*. WebMD. https://www.webmd.com/food-recipes/news/20100824/moderate-drinking-may-prolong-life#:~:text=Aug.,three%20alcoholic%20drinks%20per%20day.

Robinson, David, Alex Hayes, and Simon Couch. 2020. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. *Welcome to the tidyverse*. *Journal of Open Source Software*. Vol. 4. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2020. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.