Math 3340

Dataset 10: Model selection of influencing factors of life expectancy based on Rstudio

By

Ian Shipley

Wenjian Zhang

Zengzi Han

Dec.12, 2020

***Abstract***

The end goal of this project is to discover which of these variables are the most important when it comes to increasing life expectancy and what are the impacts of not keeping a population healthy. Once the raw data is standardized, we can create a correlation coefficient matrix to see which variables have an impact on which variables. It'll be important to note where the large coefficients are as well as the smaller or 0 coefficients to know what has a large impact and what has little impact. Once this is done, we can use the data to logically conclude and relate real life situation and meaning to find a solution to our problem.

***Introduction***

Preserving human life is an extreme importance to everyone, and it is important to know how which precautions must be appropriated and what steps must be taken in order to solve this problem. Using the data provided by WHO (world health organization) and GHO (Global health organization) it is possible to find connections between the average life expectancy and other factors in a nation. This data set contains 29 different variables for the purpose of studying the health and wellbeing of other nations. There are three main questions we are looking to answer using this data. We will use this data to find what measures nations with low average life expectancy should take in order to increase their life expectancy as well as what all nations can focus on to keep their life expectancy high. Isolating the unimportant variables is also important to note so that we can show not only what is important but what is not important. The data and variables provided should give us the information provided to answer these questions. It is important to understand all of the variables given including what they represent.

***Data description analysis,***

The 29 variables are the country and the current year the data was taken. Their status of whether they are a developing or developed nation. The average life expectancy of citizens in that nation. The adult mortality rate which is the rate at which people between the ages of 15-65 die per 1000 people. Infant deaths which is the number of children ages 0-14 per 1000 children. Alcohol which is the consumption of alcohol per capita. Percentage expenditure which is expenditures based on health as a percentage of total GDP. Hepatitis B, the percentage of one-year-olds that are vaccinated. Measles, the reported number of cases per 1000 people. BMI, the average Body mass index of the people living in the nation. Under five deaths, the number of deaths of citizens under the age of 5 per 1000 people. Polio, the number of one-year-olds vaccinated for polio per 1000. Total expenditure, the amount the government spends on health as a percentage of all government spending. Diphtheria the number of one-year-olds vaccinated for Diphtheria. HIV/AIDS, the number of deaths per births. GDP, the gross domestic product per capita. Population, the population of the nation. Thinness 10-19 years, the rate of children between ages 10-19 are considered thin. Thinness 5-9 years, the rate of children between ages 5-9 are considered thin. Income composition of resources, how wealthy a nation is. Finally Schooling, the average number of years of schooling. With these variables we will conduct a finding to conclude which are important to the health of an individual along with how can we raise the life expectancy of struggling nations.

The Data originally collected by GHO (Global Health Organization) contains a file with a list of 193 countries detailing the life expectancy of its citizens among many othe. The data set contains multiple data results that of which include the life expectancy of a nation, the development status, the population the schooling available along with many others. In total, there were 2938 observations across 193 countries and 15 years (2000-2015) along with 29 different variables for each year for each country. In the R code, we’ve assigned the year, development status, the life expectancy and the country as factor variables. The Data must then be standardized so that we can calculate a more precise correlation coefficient. Once this is found we can create a correlation coefficient matrix which will show either a positive or negative relationship between two variables on a –1 to +1 scale. Correlation coefficients with a high magnitude correlation are values we’ve considered to be either 0.7 or above or –0.7 and below. Based on the results, we can observe that there is high correlation between life expectancy and schooling, income composition of resources and adult mortality. On the other hand, we can also see which variables have little or no correlation between them by finding the correlation values closest to 0. For instance, we can see that population has low a correlation with GPD, HIV, life expectancy, schooling among many other variables. Lastly, we can create scatterplots to help visualize the relationship between two variables. Starting in the top left and moving diagonally down to the bottom right, we have the graphs that represent A variables fluctuation over all the data points. This shows the distribution of the results graphicly. For example, looking at the GDP graph compared to the schooling graph we can see there is little change between the GDP per capita measured, meaning most countries had around the same level of GDP. While the schooling graph resembles more of a bell curve which shows that the data more is normally distributed as well as showing the data has a wide range of possible values. Looking at these graphs we can also see visually which two variables have a high magnitude correlation by looking at the variable vs number of observation graphs discussed earlier and comparing them to each other. We know that by looking at the correlation coefficients of schooling and Income composition of resources (0.78) they have high positive correlation. By creating a scatterplot of these two variables we can see a linear relationship between the two as well as noticing the similarities between the observation graphs. The scatterplots give us another way to visualize correlation.

***Methods analysis***

It is very complicated to find the relationship between independent variables and dependent variables in a large number of data. In order to find the relationship between many variables and the correlation between independent variables and dependent variables, the model should be optimized. Backward model selection is the method used this time, and the model modification method is also attached. When using this method, we first need to set a alpha-out value according to the actual situation and use it as a benchmark to evaluate which variable will be removed, here the value of alpha-out is 0.05. The first step is to do a linear regression of all the variables. Compare their p-value, where the largest variable greater than alpha-out will be removed from the preselected model. However, it should be noted that when using this method, only one variable can be removed at a time. The next step is to do a linear regression for the remaining variables, also repeat the previous step to filter one by one. When the step runs until the P-value of all variables after linear regression is less than alpha-out, the filtering is stopped. At this point, the model is the final model. After the model is obtained, the model screening also needs to conform to Gauss-Markov Assumption, when E(ei)=0, var(ei)=theta2, e~Normal(0, theta2). This also needs to be combined with graphic information to modify the model, there are four common graphics Normal Q-Q, Residuals vs Fitted, Scale-location and Residual vs Leverage. Usually, we can detect whether residual conforms to normal distribution, whether the variance of R is equal, and whether it involves outliers.

***Results***

Going through the variables one by one, starting with life expectancy, we see that it has high correlation with schooling, Income composition of resources and adult mortality and has little correlation with Hep B, measles, under five deaths, total expenditure and population. Secondly, adult mortality is highly correlated with only life expectancy but since it has already been stated there's no need to repeat correlations since if A is correlated with B, then B is correlated with A. Adult mortality has little correlation with infant deaths, alcohol, Hep B, Measles, under five deaths, polio, total expenditure, Diphtheria and population. Infant deaths are highly corelated with only under five deaths and have low correlation with alcohol, percentage expenditure, polio, total expenditure, diphtheria, HIV, GDP and Income composition of resources. Alcohol’s only notable correlation are that it has low correlation with hep B, Measles, under five deaths, HIV and population. Percentage expenditure Is only highly correlated with GDP and had low correlation with Hep B, Measles, under five deaths, total expenditure and population. Hep B is not highly correlated with anything that hasn’t been stated yet and has low correlation with Measles, BMI, Total expenditure, HIV, GDP, Population, thinness 10-19, thinness 5-9 and income. Measles is not highly correlated with anything that hasn’t been stated yet and has low correlation with BMI, polio, total expenditure, diphtheria, HIV, GDP, Both thinness values and schooling. BMI is highly correlated with anything that hasn’t been stated yet and has low correlation with polio, Total expenditure, Diphtheria and population. Under five deaths only has low correlation with polio, total expenditure, diphtheria, HIV, GDP and Income composition of resources. Polio only has low correlation with total expenditure HIV, GDP, population, and both thinness results. Total expenditure only has low correlation with Diphtheria, HIV, GDP, population and Income composition of resources. Diphtheria only has low correlation with GDP, HIV, population and both thinness results. HIV only has low correlation with GDP, population, and both thinness results. GDP only has low correlation with population. Population has low correlation with Income composition of resources and schooling. Thinness from the age of 10-19 has high correlation with thinness from 5-9. Thinness from 5-9 has no new correlations. Lastly Income composition of resources is highly correlated with schooling.

Taking a closer look at the results of the data, it is important to understand where in the real-world can we apply this information. A lot of the results seem to be obvious in terms of why two variables are correlated with each other. For starters we can look at both thinness results of children between the ages of 5-9 and 10-19. It makes logical sense that if there are thin children in one age group, that there would also be thin children in the other age group and vice versa. These kinds of results are not extremely interesting nor insightful, so it is not of extreme importance to totally dissect all the information provided. We can break the different variables into different subcategories, for example, in the data set, there are a multitude of different diseases considered. Those diseases can be divided into two different groups. Diseases measured by how many deaths per 1000 births and Diseases measured by the population of citizens over the age of one that are vaccinated. The three diseases measured by vaccination, Polio, Diphtheria and Hep B, are all primarily correlated with each other. Their highest magnitude correlation coefficients are with each other, detailing the struggle with vaccines as a whole and not just a certain disease or outbreak of a disease across the world. Looking at the other types of diseases measured, measles and HIV, it is interesting to see that they share a correlation coefficient of exactly zero. While there is a massive connection between the 3 vaccine diseases, there is absolutely zero evidence in this research that there is a connection between HIV and measles. Taking a closer look at this result shows that measles is connected to infant death and HIV is connected to adult mortality which would give some explanation as to why the two are not related, but it is still interesting that their correlation coefficient is exactly zero since you'd believe that nations with bad healthcare wouldn’t be able to treat either and they would be more prominent in similar areas. There is another group of similar data sets we can create by putting together all economic factors in a set, GDP, Income composition of resources, total expenditure and percentage expenditure. GDP and percentage expenditure being so closely related isn't a surprise, since one is a measure of the other and their respective correlation coefficients with other variables are very close in value as well. Total expenditure is unusual since it isn't closely related to anything making it seem obsolete in the data set with it’s highest correlation coefficient being schooling at 0.24. Which leaves Income composition of resources which is an index measurement of a relative income source, or how much income is in a nation, or generally how wealthy the nation is. According to the results, Nations with high income composition of resources tend to have better schooling and a much high life expectancy. Relating this to the real world makes sense, since nations with good schooling and high life expectancy tend to mostly be qualities of well-developed nations which would result in them being wealthier. Alcohol, population and schooling are three somewhat miscellaneous data sets that don’t really fit in well to any category, nether the less, they are still important to discuss, especially schooling which out of all factors has the biggest impact on life expectancy. The data shows that the most important factor a nation can invest in is education. In regards to population, the only major factor involved is child infant deaths and under five deaths which is interesting to see. It leads up to believe that larger populated nations, still do not have the required medical equipment to keep children form dying. The second highest correlation value with population is measles, a disease normally found in children. In terms of alcohol, if it were to cause more deaths, it would have a high adult mortality rate correlation coefficient, since infants are not dying of alcohol related issues. Strangely enough however, not only does alcohol seem to not have a strong relationship with adult mortality, but the relationship is also slightly negative, meaning the higher the alcohol consumption, the lower the adult mortality rate is. Lasty we can take the remaining data sets and categorize them under health and wellbeing. These variables would include life expectancy, adult mortality, infant deaths, under five deaths, BMI and both thinness results. Here we have two sets of data that can be considered the same being infant deaths and under five deaths along with the two thinness data sets. BMI is the most overarching of these sets having somewhat large correlations with a multitude of variables which isn't surprising since BMI in itself is an overarching health measurement. A BMI also considers how thin you are in its measurements, resulting in BMI and thinness being very similar as well. In terms of adult mortality, and infant deaths there is not anything surprising that has not been discussed earlier in the conclusion, which leaves only life expectancy. Life expectancy has multiple large factors. Adult mortality clearly plays a huge factor since the majority of a nation's population are adults. Between the 5 diseases listed in the data set, it is HIV which has the greatest impact on life expectancy which is important to note for countries with low life expectancy. The two other major impacts are schooling and Income composition of resources which are closely related since, according to the data, it is difficult to have one and not the other.

***Conclusion***

After taking a closer look and analyzing our results and giving them real world applications there are multiple important takeaways in relation to life expectancy. The most critical result is the importance of education and its connection with life expectancy being the highest out of all variables. If all nations could heighten their average years in school, life expectancy would surely follow. Heightening a nations income composition of resources or its GDP is a good stepping stone for increasing education levels. Vaccinations and HIV are also detrimental to average life expectancy and would need to be addressed as well in order to solve this global issue. What’s important for every nation however is to be aware of problems and challenges that could arise with an increasing population such as infant deaths, and measles. Important as well to note that alcohol as well as total expenditure are not as important to average life expectancy as some might believe.

***Appendix***

[Daniel Murrell, M.D.](https://www.healthline.com/medical-team) (2020, April 24). *Facts About HIV: Life Expectancy and Long-Term**Outlook.*Retrieved December 11, 2020 from[**https://www.healthline.com/health/hiv-aids/life-expectancy**](https://www.healthline.com/health/hiv-aids/life-expectancy)

KumarRajarshi. (2017). *Life Expectancy (WHO)**Statistical Analysis on factors influencing Life Expectancy.* Retrieved December 11, 2020 from [**https://www.kaggle.com/kumarajarshi/life-expectancy-who**](https://www.kaggle.com/kumarajarshi/life-expectancy-who)

[Max Roser, Esteban Ortiz-Ospina & Hannah Ritchie](https://ourworldindata.org/team). (2013). *Life Expectancy.* Retrieved December 11, 2020 from <https://ourworldindata.org/life-expectancy>

[National Center for Immunization and Respiratory Diseases](https://www.cdc.gov/ncird/)& [Division of Viral Diseases](https://www.cdc.gov/ncird/DVD.html). (2020, November 5). *Complications of Measles.* Retrieved December 11, 2020 from[**https://www.cdc.gov/measles/symptoms/complications.html - :~:text=Subacute sclerosing panencephalitis (SSPE) is,fully recovered from the illness.**](https://www.cdc.gov/measles/symptoms/complications.html#:~:text=Subacute%20sclerosing%20panencephalitis%20(SSPE)%20is,fully%20recovered%20from%20the%20illness.)

Statistics Canada. (2009, November 20). *Composition of income.* Retrieved December 11, 2020 from <http://www.census2011.net/census-recensement/2006/ref/dict/pop020a-eng.cfm>