JSC370 Midterm

Shiyuan Zhou

2022/2/27

Introduction

How humans can live longer is one of the most debated topics in human history. In an era of rapid advances in medicine, education, and technology, human health has improved significantly but will humans necessarily live longer? The differences between countries are not only due to race and region but the imbalance on health care and medical technologies across countries is everywhere. However, in many cases, it is difficult to spread or teach advanced medical treatments to other countries. Does the absence of advanced medical treatments determines lower average life expectancy? Obviously, the answer is no. Medical treatment is not the only factor that determines the life expectancy of human beings but also the climate and environment, social factors, etc.

In fact, most of the influencing factors come from the government and the social organizations involved. We have a certain measure of whether the government is making a difference in humanistic care, which is the Human Development Index (HDI). The HDI is defined as a summary measure of average achievement in key dimensions of human development, like health. Also, social care as well as health care development cannot be achieved without the government expenditure. That is, at the theoretical level, both the HDI and government expenditure on health care may have an impact on the health of people, leading to an increase in their average life expectancy. However, governments' decisions heavily depends on the level of development of the country, i.e., whether a country is a developed country or not also affects the health policy and the standard of living of the population. These thoughts have led me to wonder whether life expectancy has a stronger relationship between human development index or government health care spending. Additionally, whether these relationships will be altered by the degree of development.

Github Repository: https://github.com/ZhouEEEEEE/JSC370-Midterm.git

Research Question

Is government health expenditure have higher impact on life expectancy than Human Development Index? Does it also depends on the development status of the country?

Methods

Used R packages

Here are the following R packages that I used for this portfolio.: data.table dtplyr dplyr ggplot2 mgcv zoo leaflet ggpubr

Data Source

The Data that I used to answer my research question is based on the WHO data and published on Kaggle by Kumar Rajarshi. This dataset includes values social factors of 193 countries from 2000 to 2015 and the life expectancy in age. In our research question, we are aim to compare the impact of government health expenditure and Human Development Index on life expectancy. These two predictors are represent by 'Total expenditure' and 'Income composition of resources' in our dataset. The target is life expectancy. Since we also stated that social factors may have a big difference between developed and developing countries. We sill also include the binary variable 'Status' that idicate the development status of a country. All of these variables will change across years. Here are the variable details.

life expectancy: Life Expectancy in age Total expenditure: General government expenditure on health as a percentage of total government expenditure (%) Income composition of resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1) Status: Developed or Developing status

Link of data: https://www.kaggle.com/kumarajarshi/life-expectancy-who

Data Checking

Before answering our research question, we need to do Exploratory Data Analysis first to find issues in our data, clean our data, and make summary statistics, plots, and graphs for our key variables.

Check number of missing values in each column

Table 1: Number of missing values for each variable

	num_na
Country	0
Year	0
Status	0
Life expectancy	10
Adult Mortality	10
infant deaths	0
Alcohol	193
percentage expenditure	0
Hepatitis B	553
Measles	0
BMI	34
under-five deaths	0
Polio	19
Total expenditure	225
Diphtheria	19
HIV/AIDS	0
GDP	448
Population	652
thinness 1-19 years	34
thinness 5-9 years	34
Income composition of resources	167
Schooling	163

The table I presented is the number of missing values in each columns. For example, there are 167 missing values in Income composition of resources. We will do the missing value imputation in the next section.

Check dimensions of our data

Table 2: Summery table of the dimensions of our data

axis	value
num_observations	2937
num_variables	22

We have 2937 number of observations and 22 number of variables in our dataset.

Check the summary statistics of required numeric variables

Table 3: Summary statistics of required variables

Life expectancy	Total expenditure	Income composition of resources
Min. :36.30	Min. : 0.370	Min. :0.0000
1st Qu.:63.10	1st Qu.: 4.260	1st Qu.:0.4930
Median :72.10	Median : 5.755	Median :0.6770
Mean :69.22	Mean : 5.938	Mean :0.6275
3rd Qu.:75.70	3rd Qu.: 7.492	3rd Qu.:0.7790
Max. :89.00	Max. :17.600	Max. :0.9480
NA's :10	NA's :225	NA's :167

The summary statistics of key variables help us to find the issues and reliability of our data. According to the summary table we get, variable 'Life expectancy' and 'Total expenditure' do not have big issues and in our estimated bound(life expectancy should be greater than 0 and less than 100, total expenditure should be greater than 0 and less than 100 since it represents proportion). However, the variable 'income composition of resources' has minimum values equals to 0. Since this variable indicate human development index, its impossible to have 0 values. According to the worldpopulationreview.com, the country with lowest HDI in 2019 is Niger with 0.394. Hence, 0 income composition should be removed from the data set in order to prevent wrong model fitting.

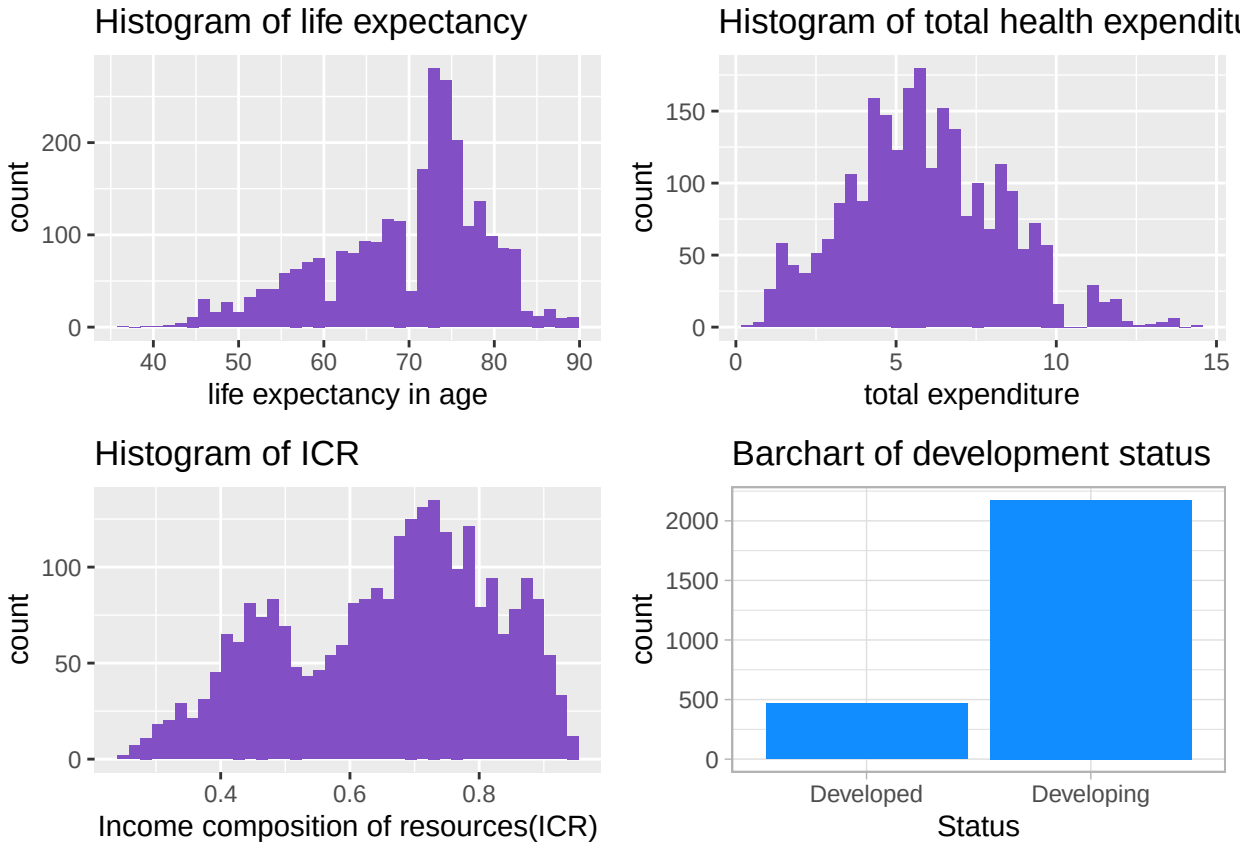
We removed the observations with 0 income composition and the new summary statistics is as followed.

Table 4: New summary statistics of required variables

Life expectancy	Total expenditure	Income composition of resources
Min. :36.30	Min. : 0.370	Min. :0.2530
1st Qu.:63.80	1st Qu.: 4.270	1st Qu.:0.5230
Median :72.30	Median : 5.730	Median :0.6870
Mean :69.59	Mean : 5.889	Mean :0.6584
3rd Qu.:75.70	3rd Qu.: 7.470	3rd Qu.:0.7840
Max. :89.00	Max. :14.390	Max. :0.9480
NA's :3	NA's :174	NA

Check Distribution of required variables

We need to check the distribution of our variables. This helps to determine outliers, skewness, and whether it is appropriate to fit the model.



According to the histograms we have for those three numeric variables, their distribution is almost normal, indicating linear model may be better options. However, there are also several issues. Firstly, the distribution of ICR is bimodal, the distribution of life expectancy is left skewed, and the distribution of total health expenditure is right-skewed. Secondly, the peak value of life expectancy and health expenditure have a very high count, which may pull our model become more centralized. Thirdly, according to the bar chart, the number of developing county is much more that that of developed country, which means, if we add status variable to our model, the data of developing country may pull our model and become biased.

Data Wranggling

Missing values imputation

Firstly, we need to handle the missing values by imputation. We use mean value of current column to impute by for looping each column.

Table 5: Number of missing values in currect dataset

number_of_NAs
0

After imputation, we find the number of missing values is zero.

Create new variable

To do further data exploration on different types of plots, we need both numeric and categorical 'Total expenditure' and 'Income composition of resources'. Converting current numeric variables to categorical variables helps us on stacked histograms, statistical summary graph, and etc. In many statistical research on social factors, health expenditure and HDI are always represented by different levels.

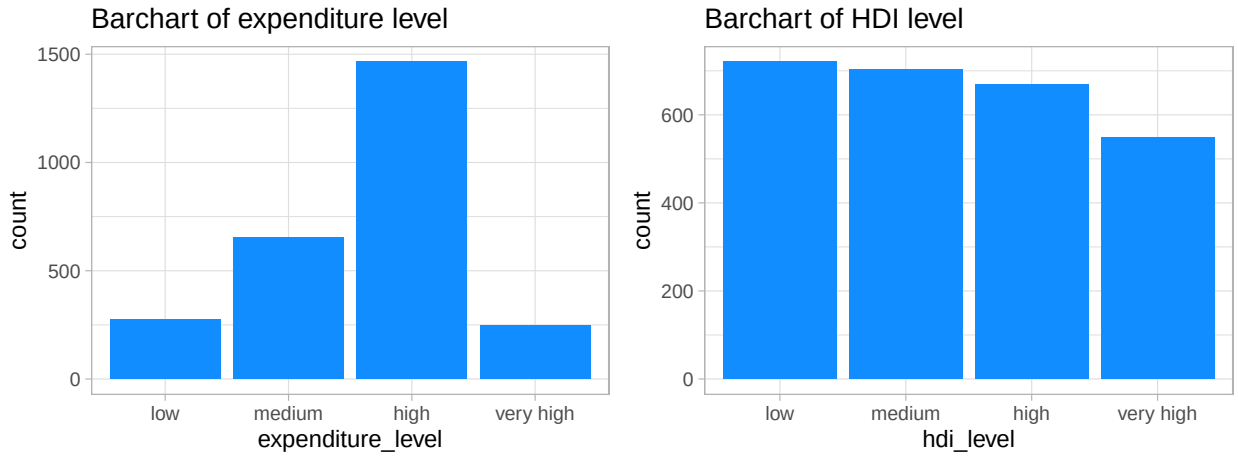
Create a new categorical variable named "expenditure_level" using total expenditure on health of a country. (rare total expenditure < 3; low total expenditure 3-5; mild total expenditure 5-9; high total expenditure > 9) and a new categorical variable named "hdi_level" indicating level of income composition of resources of countries(low income composition < 0.55; medium income composition 0.55-0.7; high income composition 0.7-0.8; very high income composition > 0.8). Additionally, we should use factor() function to give our levels an order for future convenience.

Table 6: Summery table of min total expenditure, max total expenditure, and number of observations for each level of total expenditure

expenditure_level	min_exp	max_exp	count
low	0.37	2.98	276
medium	3.00	5.00	652
high	5.10	8.99	1464
very high	9.10	14.39	248

Table 7: Summery table of min income composition of resources, max income composition of resources, and number of observations for each level of HDI

hdi_level	min_exp	max_exp	count
low	0.253	0.548	720
medium	0.550	0.700	703
high	0.701	0.800	668
very high	0.801	0.948	549



For most of the observations, they spend high level of health expenditure. There are fewer observations have low and very high health expenditure. However, for HDI level, most of the counties have low HDI level and the number of observations for each level do not have big gap.

Preliminary Results Section

Summary statistics of key predictors to life expectancy

Table 8: Summery table of min life expectancy, max life expectancy, mean life expectancy, and standard deviation of life ecpectancy for each expenditure level

expenditure_level	min_life_ex	max_life_ex	mean_life_ex	sd_life_ex
low	44.3	89	70.18442	9.796961
medium	44.0	87	66.33543	8.474075
high	36.3	89	69.98208	8.804188
very high	39.0	89	75.12212	10.639661

According to the table above, there are no trend of any statistics of life expectancy across each level of health expenditure, which means we may not have a strong linear relationship between total expenditure and life expectancy. Additionally, the standard deviation of life expectancy for each level is pretty big, indicating that for each level we may have a big range of life expectancy which also shows weak relationship between those two variables.

Table 9: Summery table of min life expectancy, max life expectancy, mean life expectancy, and standard deviation of life ecpectancy for each human development index level

hdi_level	min_life_ex	max_life_ex	mean_life_ex	sd_life_ex
low	36.3	73.0	57.99639	6.466973
medium	46.0	79.0	69.30128	5.372311
high	64.4	79.6	73.91820	2.652999
very high	71.1	89.0	79.87632	3.538607

The table we have for HDI level shows a increasing trend of statistics of life expectancy across levels except standard deviation, which may shows a positive relationship between HDI and life expectancy. Also, compare to standard deviation we have in table of expenditure, that of HDI is much lower, indicating a more concentrated values of life expectancy for each level.

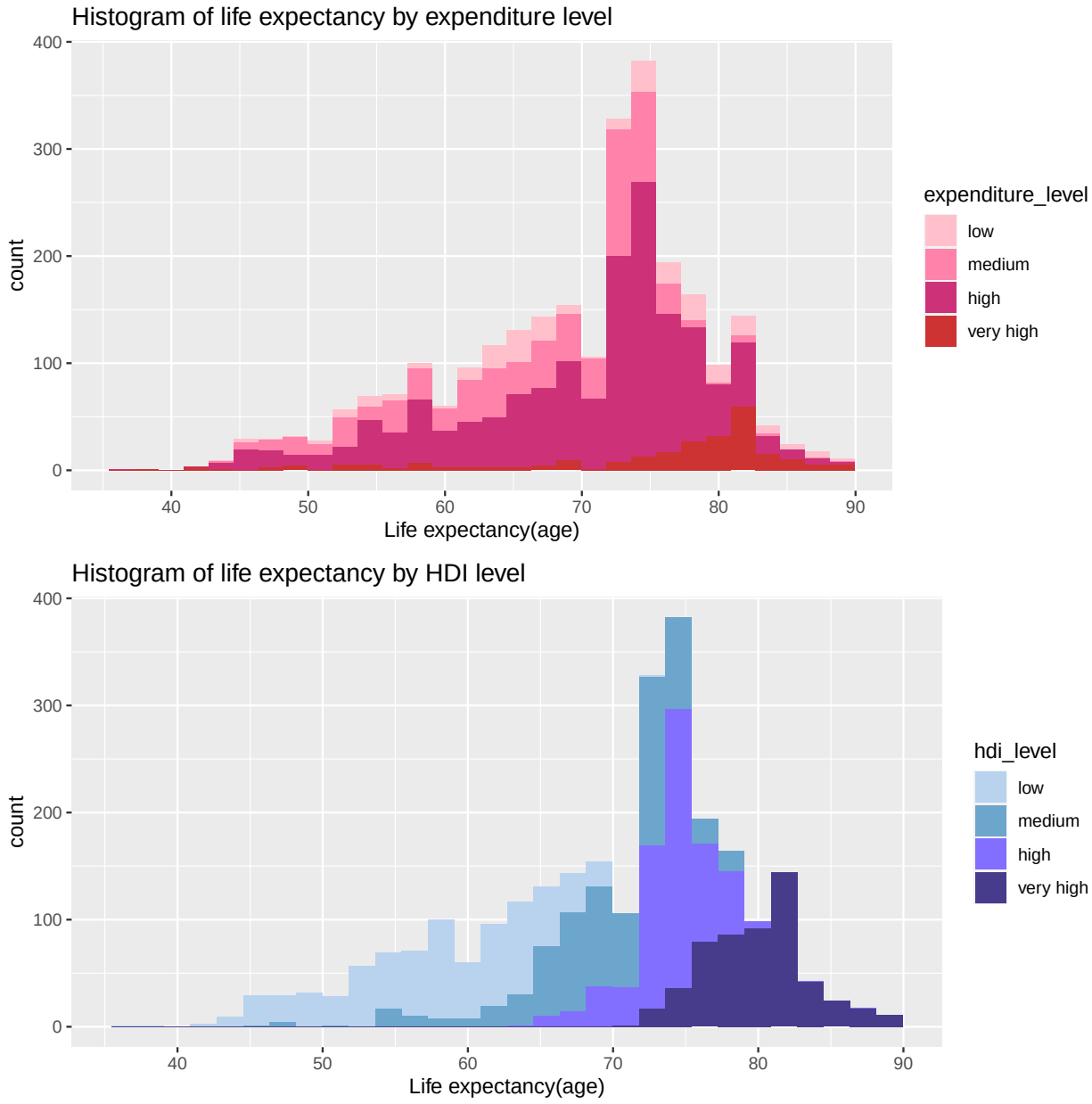
Table 10: Summery table of min life expectancy, max life expectancy, mean life expectancy, and standard deviation of life ecpectancy for developed and developing countries

Status	min_life_ex	max_life_ex	mean_life_ex	sd_life_ex
Developed	69.9	89	79.26573	4.042134
Developing	36.3	89	67.52130	8.815044

The table we have for development status reflects that more developed countries have much higher stable(due to standard deviation) estimate on life expectancy compare to that of developing countries.

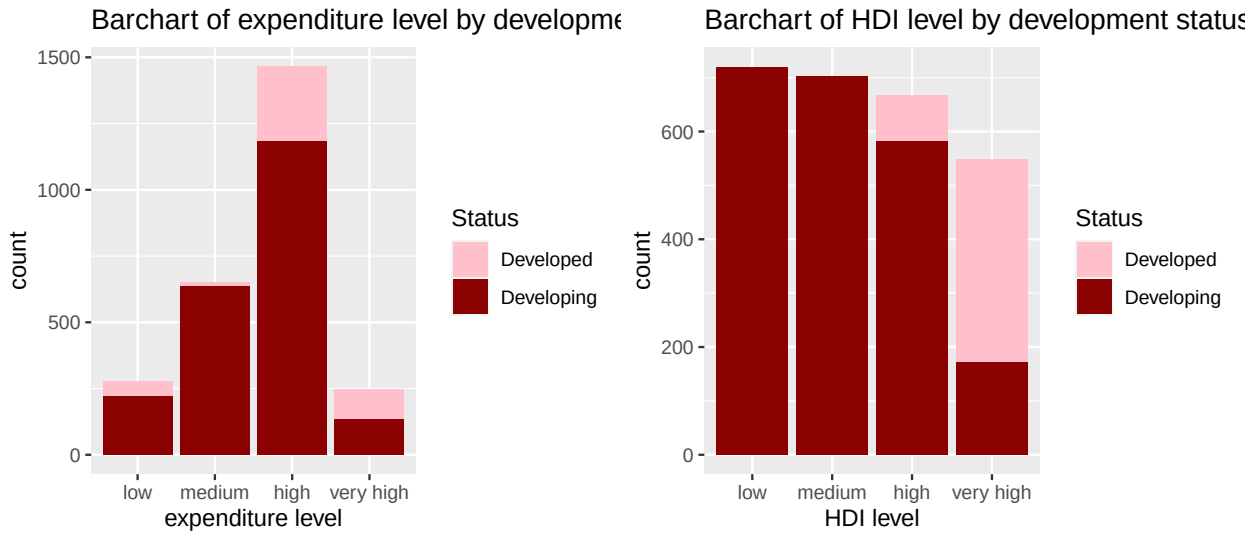
Visualizations

Histograms

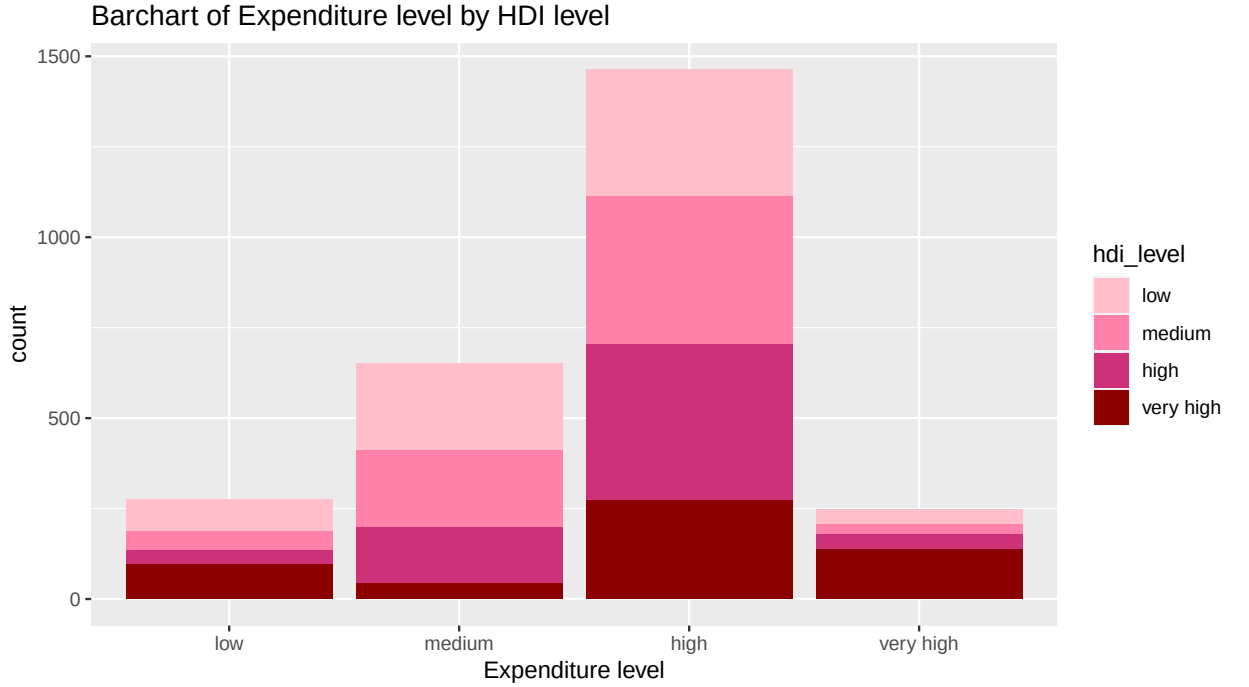


The first plot we have is the stacked histograms of life expectancy by expenditure level and HDI level. The proportion of each level of expenditure does not make a big difference across different ages. However, in the stacked histogram for HDI level, it is very clear that higher HDI level become more concentrated on the right, which is higher age. For different range of age, there always have a dominated HDI level. For example, for life expectancy less than 60, low HDI level dominates. Hence, according to these histograms, income composition of resources have a stronger relationship with life expectancy.

Bar Chart

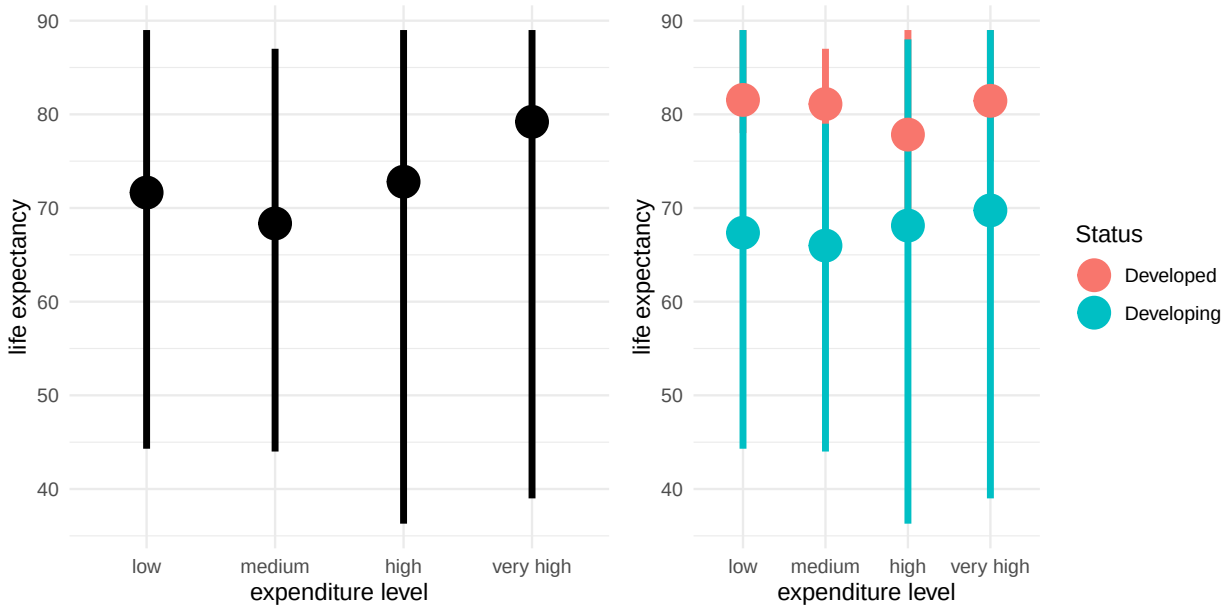


According to the barcharts we have for categorical variables expenditure level and HDI level by development status, developed countries tend to have higher expenditure and income composition of resources. However, the trend between HDI level and development status is stronger.

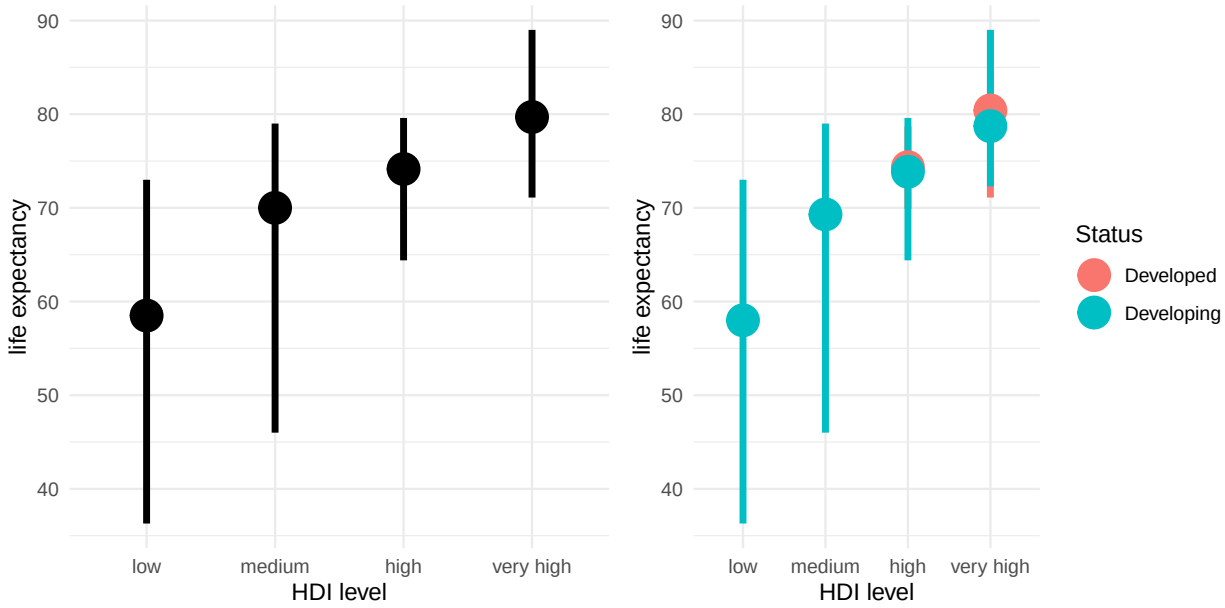


According to the bar chart we get for expenditure level by HDI level, the proportion of low and high HDI level is the largest in the low expenditure level. The proportion of medium HDI level is the largest in medium expenditure level and also for high and very high level. Hence, we could say counties with high expenditure level have higher probability to have high HDI level. Predictors expenditure and HDI may have a linear relationship.

Statistical summary graph

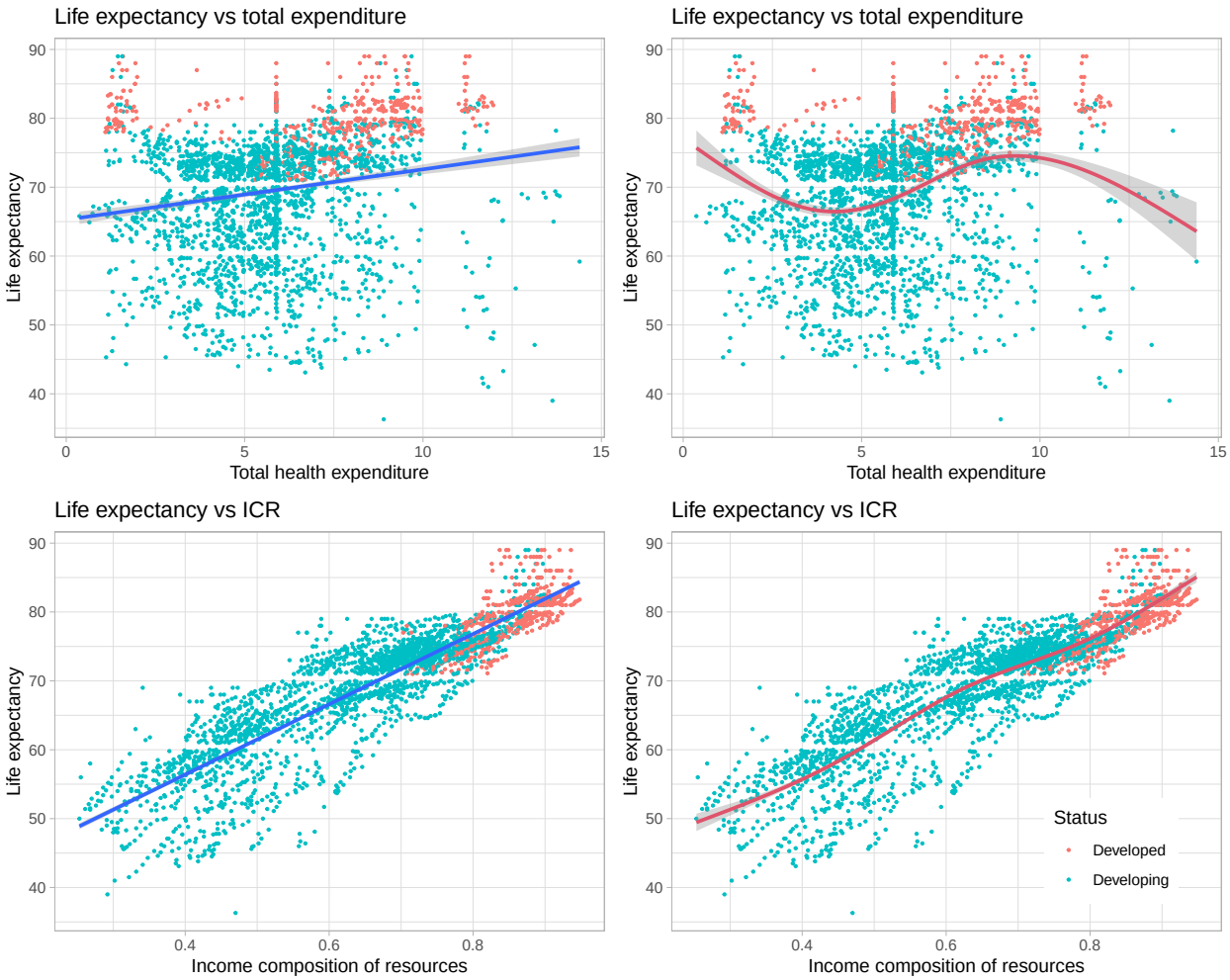


According to the statistical summary graph for expenditure level, though mean of life expectancy in low level of expenditure level is higher, we may have an increasing trend between expenditure level and life expectancy. However, if we adjusted by development status, we can see that the trend is clear for developing countries but not for developed countries. The higher mean of life expectancy in low level of expenditure was pulled up by the values of developed countries as the orange points shows. Additionally, the range of life expectancy for each expenditure level as the distance from min to max is large, which means our model may not fit tightly.



The statistical summary graph we have for HDI level shows a positive relationship between human development index and life expectancy. Adjusting by development status did not make a difference on our relationship. Additionally, the distance between min and max is much shorter than that in expenditure level which indicate a strong relationship and a tighter model fit.

Scatterplots



The scatter plot of live expectancy vs total health expenditure and income composition of resources clearly present what actual model fitting will be in our dataset. The two plots with two blue straight line on the left is the linear model fitted in each of the relationship. The two plots on the right is the cubic spline model we have, where the red curve is the fitted splines.

According to the plots we have for live expectancy vs total health expenditure, the linear model is not very fitted to our data. Spline model could explain more variation and yields better fit but the decreasing trend when total health expenditure is greater than 12.5 may comes from overfitting on the right-most points. Comparing to what we have in life expectancy vs income composition of resources, a positive linear trend is pretty clear. However, the fitted spline model does not make a big difference than the linear model. We need to further decide which model is better by adjusted R squared since spline model may have a higher adjusted R squared but the cost is overfitting.

Model Fitting

To compare which factor have stronger relationship with life expectancy, we use them as predictors and fit both linear and spline models. Since we also want to add consideration of development status, we will fit all the models and adjusted by status again. Additionally, we should also fit full models to see whether using both total expenditure and income composition of resources predict life expectancy better.

a) Models without adjusted by development status

Linear models:

Total expenditure as predictor: lm(life_exp ~ total_exp, data = ds)
Income composition of resources as predictor: lm(life_exp ~ income_com, data = ds)
Total expenditure and income composition as predictor: lm(life_exp ~ income_com + total_exp, data = ds)

Spline models:

Total expenditure as smooth terms: gam(life_exp~ s(total_exp,bs="cr",k=3),data=ds)
Income composition of resources as smooth terms: gam(life_exp~ s(income_com,bs="cr",k=3),data=ds)
Income composition of resources as smooth terms adjusted by total expenditure: gam(life_exp~ s(income_com,bs="cr",k=3) + total_exp,data=ds)
Total expenditure as smooth terms adjusted by income composition of resources: gam(life_exp~ s(total_exp,bs="cr",k=3) + income_com,data=ds)

b) Models with adjusted by development status

Linear models:
total expenditure and status as predictor: lm(life_exp ~ total_exp + status_ind, data = ds)
income composition of resources and status as predictor: lm(life_exp ~ income_com + status, data = ds)
total expenditure, income composition, and status as predictor: lm(life_exp ~ income_com + total_exp + status, data = ds)

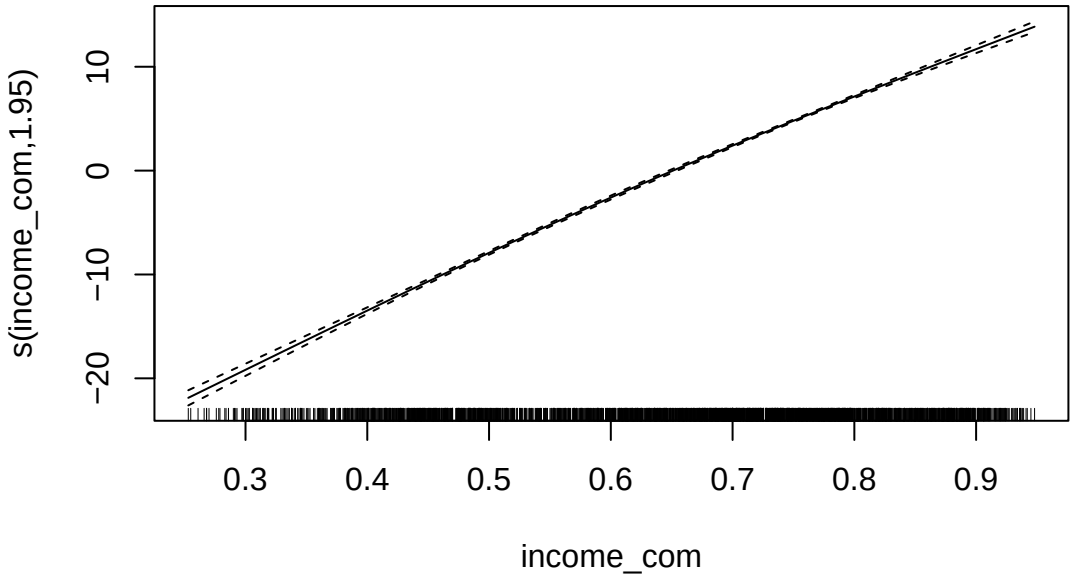
Spline models:
total expenditure as smooth terms adjusted by status: gam(life_exp~ s(total_exp, bs="cr",k=5) + status,data=ds)
income composition of resources as smooth terms adjusted by status: gam(life_exp~ s(income_com, bs="cr",k=3) + status, data=ds)
income composition of resources as smooth terms adjusted by total expenditure and status: gam(life_exp~ s(income_com,bs="cr",k=3) + total_exp + status, data=ds)
total expenditure as smooth terms adjusted by income composition of resources and status: gam(life_exp~ s(total_exp,bs="cr",k=3) + income_com + status, data=ds)

Conclusion and Discussion

Comparing all spline models

Table 11: Comparing all R squared of all spline models

models	R_square
total expenditure as smooth terms	0.0379077
income composition of resources as smooth terms	0.7903460
income composition of resources as smooth terms adjusted by total expenditure	0.7908591
total expenditure as smooth terms adjusted by income composition of resources	0.7897713
total expenditure as smooth terms adjusted by status	0.2422911
income composition of resources as smooth terms adjusted by status	0.7903013
income composition of resources as smooth terms adjusted by total expenditure and status	0.7907897
total expenditure as smooth terms adjusted by income composition of resources and status	0.7899147



The model of income composition of resources as smooth terms and adjusted by total expenditure has the highest R squared value, which is 0.79086.
Looking at the spline model we have, the trend is not curvy, which indicate that a liner model may be preferred to reduce overfitting.

Comparing all linear models

Table 12: Comparing all R squared of all linear models

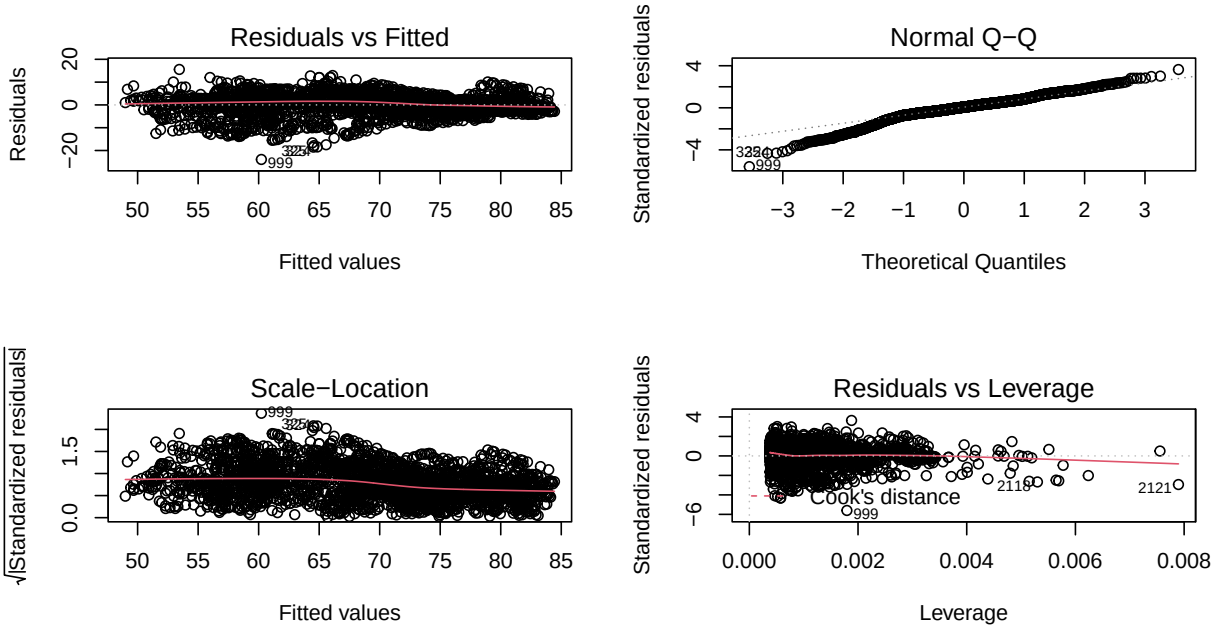
models	R_square
total expenditure as predictor	0.0326109
income composition of resources as predictor	0.7892844
total expenditure and income composition as predictor	0.7895989
total expenditure and status as predictor	0.2336129
income composition of resources and status as predictor	0.7894635
total expenditure, income composition, and status as predictor	0.7898889

According to the all R squared value we have for all linear models, the one with all predictors have the highest R squared. However, linear model with total expenditure and income composition as predictor also have pretty high R squared. Hence, we need to conduct t test to see if any predictor is not significant.

Table 13: Significance of each predictors in full linear model

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	34.7666054	0.6102258	56.973349	0.0000000
income_com	51.4725787	0.6161445	83.539781	0.0000000
total_exp	0.0864233	0.0374102	2.310153	0.0209566
statusDeveloping	0.5093977	0.2670554	1.907461	0.0565695

According to the p-values we have for each predictor in the linear model, we can see that total expenditure and income composition are quite statistically significant variables in the model, if the significance level is 0.05. However, development status is not as significant as others. Hence, we could remove it to reduce complexity of our model.



The residuals vs fitted plot shows the linearity is not violated since no curving trend. The QQ plot shows that the normality may hold but a large deviation exist on the left tail. In the scale-location plot, we do not have a fanning pattern that indicate unequal variances. In the leverage plot, we do not have groups of influential points that may pull our model. Hence, we do have a good fit since assumptions are mostly satisfied.

Comparing picked linear model with picked spline model

Table 14: Comparing picked linear model with picked spline model

models	R_square
linear model	0.7895989
spline model	0.7908591

According to the table we have, the R squared value for those two models are pretty close. Though spline model yields better fit based on the score, a linear model may be better choice since the spline model we plotted is very close to a linear line. Choosing a linear model with almost the same wellness of fitting could reduce overfitting. In conclusion, the linear model with income composition of resources and total expenditure as predictors is our final model.

Answering research question

Research question: Is government health expenditure have higher impact on life expectancy than Human Development Index? Does it also depends on the development status of the country?

According to the data exploratory plots we have, the relationship between health expenditure and life expectancy is not strong. The models that only contains total expenditure and status have adjusted R squared less than 0.04, which means they fitted badly. However, in most of our plots, the relationship between HDI and life expectancy is strong. We also have pretty well fitted models with HDI as predictor have adjusted R squared around 0.79. Furthermore, adding development status into our model does not have any significant effect according to the model comparison results.

Hence the answer to our research question is no and development status does not play an important role in current situation with total health expenditure and HDI. However, the results from model selection process also shows that a linear model that include both total health expenditure and HDI will have better predict on life expectancy.

Discussion and limitation

According to the result we have, if the governments aim to increase life expectancy of the population, they should focus on factors that will increase the HDI. Not just spending too much money health development. In may cases, better health treatment and medical technologies does not benefit everyone. For most of the population, even poorer people, their health conditions need much longer time to respond to larger health expenditure than richer people. Additionally, people who would like to search for some counties to stay and try to live longer could choose countries with high HDI, rather than high health expenditure.

Limitations:
1. Since we impute NAs by mean value, we may result in biased standard error, variance, and sample mean. Our estimate may be pulled by other observations.
2. Since we have 16 years of observations per country and we investigate the dataset as a whole, our observations are not totally independent. Also, we ignored the structural difference between countries, like

race and climate.
3. Though in model comparison, some spline models have better performance than the linear model, picking a spline model also increase our risk on onverfitting.
4. We only fitted spline models with 3 knots. We should also vary it to compare spline models with different knot numbers.