Balint Thaler – 2300541
GitHub repo: https://github.com/balint-thaler/Data-Analysis-2-term-project

Data Analysis 2 Project

Introduction

The purpose of the current project is to analyse the association between income inequality and life expectancy.

Intuitively, we might think that to get a clear picture on the association between the two variables we need to control for the effects of the differences in the healthcare practices of the countries which might influence the dependent variable. To do so, I will introduce two control variables which measure two things related to healthcare: how much is spent on it and how effective it actually is. The expectation is that expenses are to be positively associated with life expectancy but the addition of an effectiveness measure is also required to account for the effect of effective but underfunded and ineffective but expensive healthcare systems.

I set out to answer the question "Do countries with higher (or lower) income inequality rates have higher or lower life expectancy on average after we control for the effects of their healthcare?"

The variables

To measure the aforementioned constructs, I will use the Gini index, life expectancy, per capita health expenses and lastly the so-called Health Access and Quality index.

The Gini index is a variable that measures inequality as the difference between the actual distribution of wealth to a perfectly equal distribution. The scale goes from 0 to 100 where 0 represents a situation in which every individual in the country possesses the exact same wealth (i.e. perfect equality), while 100 represents a case where the total wealth of the country is held by one single person (i.e. perfect inequality). Data on the Gini index is downloaded from the World Bank.

Life expectancy is measured by the expected length of ones life at birth in a specific year. The data is downloaded from the website of Our World in Data, however the original source of the data is the UN World Population Prospects report from 2022.

For control variables I introduce a financial measure of per capita healthcare expenses and a more qualitive measure of the Health Access and Quality index (HAQ). The former measures the annual per person expenses related to healthcare in international dollars at purchasing power parity. The latter measures death rates from 32 causes of death that could be avoided by timely and effective medical care. The HAQ index is measured on a scale of 0-100 where 0 is the worst, 100 is the best. Data on the control variables is downloaded from the website of Our World in Data, however the original source is the Global Burden of Disease study by the Institute for Health Metrics and Evaluation.

Analytical choices

To conduct the analysis, I filter out any missing values and concentrate on one singular year of 2015. However, due to the lack of consistent availability on Gini data, for this variable only I will use the latest estimation regardless of the timing of the data. Given that the Gini index is a fairly stable construct[1] that changes little in a few years and given that this increases the available observations greatly, the benefits of this choice outweigh the weakening of the robustness due to the unaccounted change of time in the data.

The data

After cleaning and joining the different tables, there are 150 non missing and non-zero value observations in the dataset. This will likely lead to the model being only capable of showing only relatively strong effects only, it will likely be limited by the relatively small sample size.

The values of the Gini index range from 23.2 for Slovakia to 63.0 for South Africa. Given the choice about using the latest available data for this variable, the timescale of the variable varies between 2010 and 2022. Life expectancy ranges between 51.1 years for Lesotho and 83.9 years for Japan with an average of 71.3 years. Health expenses range between 35.94 USD for the Democratic Republic of Congo to 9392.1 USD in the United States. Lastly, the HAQ index in the dataset ranges from 35.7 for Lesotho to 93.6 for Iceland.

The model

Upon inspecting the distribution of the variables, we see that Gini is distributed with a right tail but it looks close to a normal distribution with a mode between 30-40. Life expectancy shows similar distribution but with a longer left tail and a mode around 75 years. Health expenses show a log normal-like distribution with a long right tail and few extreme high values at the end. It is likely that a linear regression model will not be a good fit for this variable without some transformation. Finally, the HAQ index has a right tail with a mode around 50.

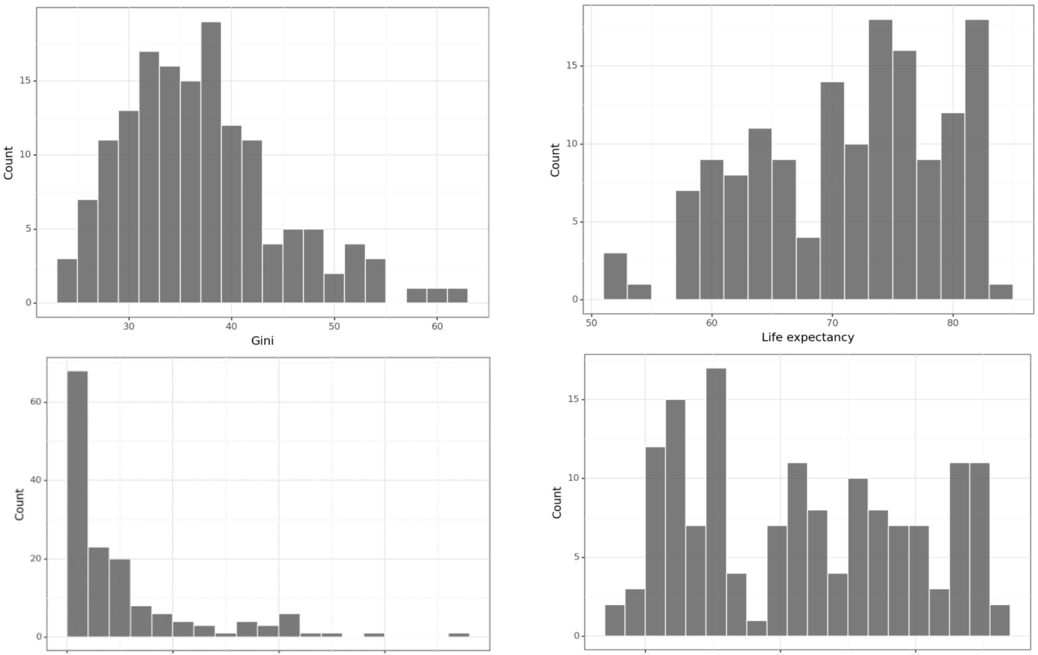
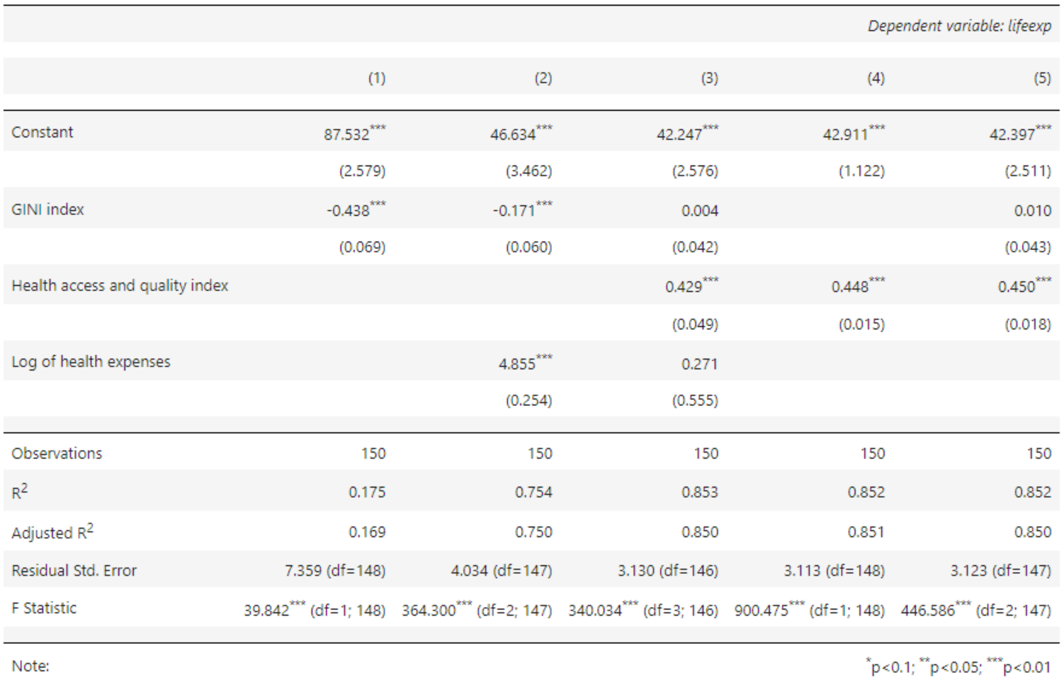


Figure 1: histograms of the variables

[1] To check this, I have randomly selected some countries and have visualised the change over time on the website of the World Bank. For the results, see Figure 1 in the Appendix.

After observing the scatterplots[2] of the explanatory and control variables with life expectancy, and adding a lowess non-parametric regression to the visualisations, I transform the health expenses variable with a log transformation. With this, all variables in a model are fairly well fitted with a linear regression[3]. The summary of the regressions can be seen in table 1 below.



					Dependent variable: lifeexp
	(1)	(2)	(3)	(4)	(5)
Constant	87.532***	46.634***	42.247***	42.911***	42.397***
	(2.579)	(3.462)	(2.576)	(1.122)	(2.511)
GINI index	-0.438***	-0.171***	0.004		0.010
	(0.069)	(0.060)	(0.042)		(0.043)
Health access and quality index			0.429***	0.448***	0.450***
			(0.049)	(0.015)	(0.018)
Log of health expenses		4.855***	0.271		
		(0.254)	(0.555)		
Observations	150	150	150	150	150
R^2	0.175	0.754	0.853	0.852	0.852
Adjusted R^2	0.169	0.750	0.850	0.851	0.850
Residual Std. Error	7.359 (df=148)	4.034 (df=147)	3.130 (df=146)	3.113 (df=148)	3.123 (df=147)
F Statistic	39.842*** (df=1; 148)	364.300*** (df=2; 147)	340.034*** (df=3; 146)	900.475*** (df=1; 148)	446.586*** (df=2; 147)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1: regression table of different models with heteroskedasticity robust standard errors in parenthesis

Interpretation of the results

Model 1 (shown in the 1st column) shows that the relationship between Gini and life expectancy is a negative one. Specifically, countries with a 1-point higher Gini index have on average a life expectancy of 0.44 years less. The coefficient is significant at 1% and already this singular variable explains 17.5% of the variance in life expectancy.

In model 2, the health expenses variable is added. The model now explains more than 75% of the variation in life expectancy, and the coefficient of Gini is changed so that countries with a 1-point higher Gini index but with equal health expenses have on average a life expectancy of 0.17 years less. Countries with the same Gini index but with 10% higher health expenses per capita have 0.49 years higher life expectancy on average. Both Gini and the health expenses variables are significant at 1%.

Model 3 includes all variables which greatly changes the picture: with the addition of the HAQ index, we see that both the Gini index and the health expenses variable becomes insignificant while the model fit (measured by the R squared) increases to 0.853. The coefficient of the HAQ index shows that countries with a 1-point higher HAQ index and with equal Gini figures and equal per capita health

[2] See Figure 2 in Appendix

[3] For the HAQ index and Gini as well, higher order polynomials were also considered based on the observation of the lowess graph, but they all produced lower R squared than the linear model. See details in Table 2 and 3 in the Appendix.

expenses, have a life expectancy of 0.43 years higher on average. We also see that the standard error of the health expenses variable has more than doubled versus model 2.

The results mean that there is no statistically significant association between wealth inequality and life expectancy after we control for the countries' healthcare effects by adding the HAQ and the health expenses variables.

The results also suggest that despite the initial idea of the HAQ index and the health expenses being two separate constructs, they overlap to some extent. Therefore, adding the health expenses to the model on top of HAQ does not add much to the model (see Model 5 vs 3 in table 1). However, adding the HAQ on top of the health expenses, increases the explanatory power and makes the latter insignificant (see model 3 vs 2 in table 1). We also see that the HAQ on its own explains more than 85% of the variation in life expectancy (see model 4). The two explanatory variables also have a correlation of 0.78 also confirming the presence of multicollinearity. There are also very few countries that have low health expenses but high HAQ values or vice versa as we see in the scatterplot[4].

Causality

The statistical methods presented in this analysis are incapable of confirming causality. More data and deeper analysis are required to establish a direct causal link between HAQ and life expectancy. It seems plausible that better access to quality healthcare does indeed lead to higher life expectancy, but it is also possible that there are more variables (not accounted for in the current model) that are playing roles in this relationship.

Conclusion

Overall, the analysis has found that countries' wealth inequality is irrelevant for life expectancy if the healthcare systems' differences are also accounted for. The current analysis is however rather limited by the relatively small sample size and the missing Gini values for 62 countries. It is important to note that there is a rather large number of very small economies (such as Oceania, Caribbean nations) with missing Gini values that might influence the results. It is possible that the associations presented show a different picture depending on the specifics (e.g. size, advancement etc.) of the economies. The current dataset is not capable of showing these subtle differences, therefore its external validity is limited.

In conclusion, increasing life expectancy (and with it likely the volume of active physically fit workers in a country) is of paramount importance. Administrations should focus to understand the ways of achieving this. A more detailed and thorough deep dive is needed confirm a causal relationship which may serve as the basis for policy related decisions.

[4] See Figure 3 in Appendix

Balint Thaler – 2300541
GitHub repo: https://github.com/balint-thaler/Data-Analysis-2-term-project

Appendix

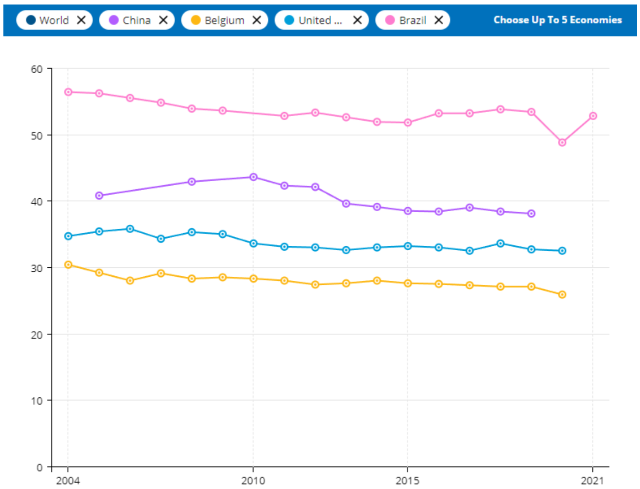


Figure 1: changes of 4 randomly selected countries' Gini indices.

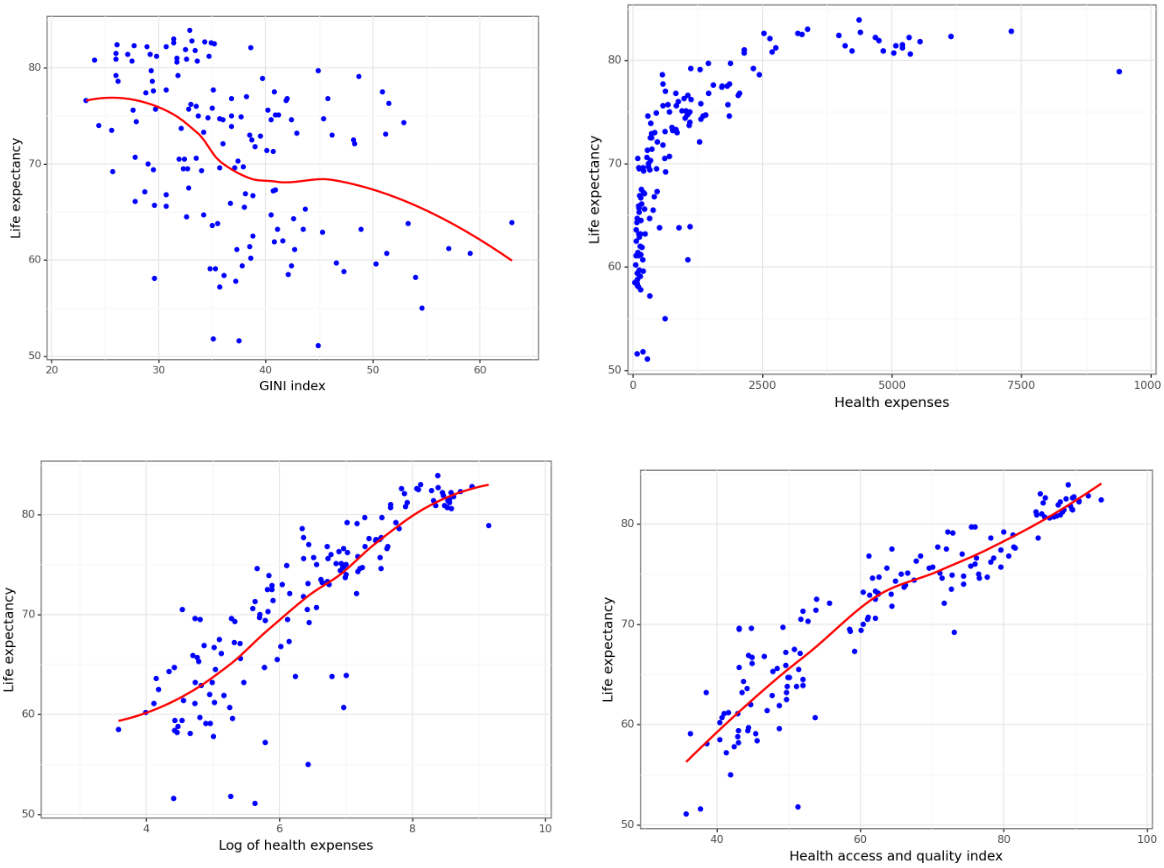
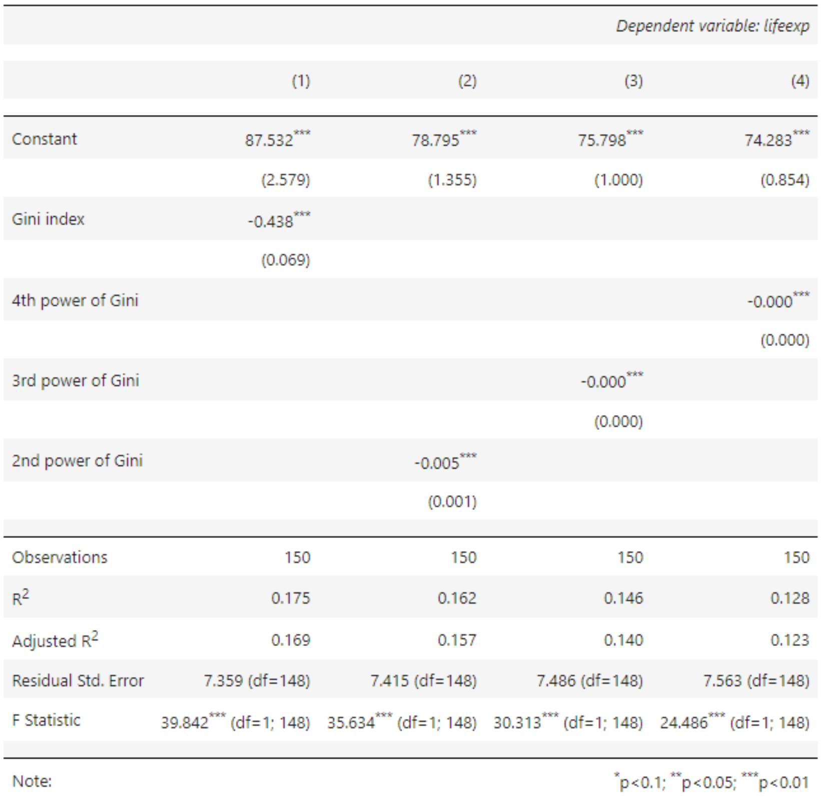
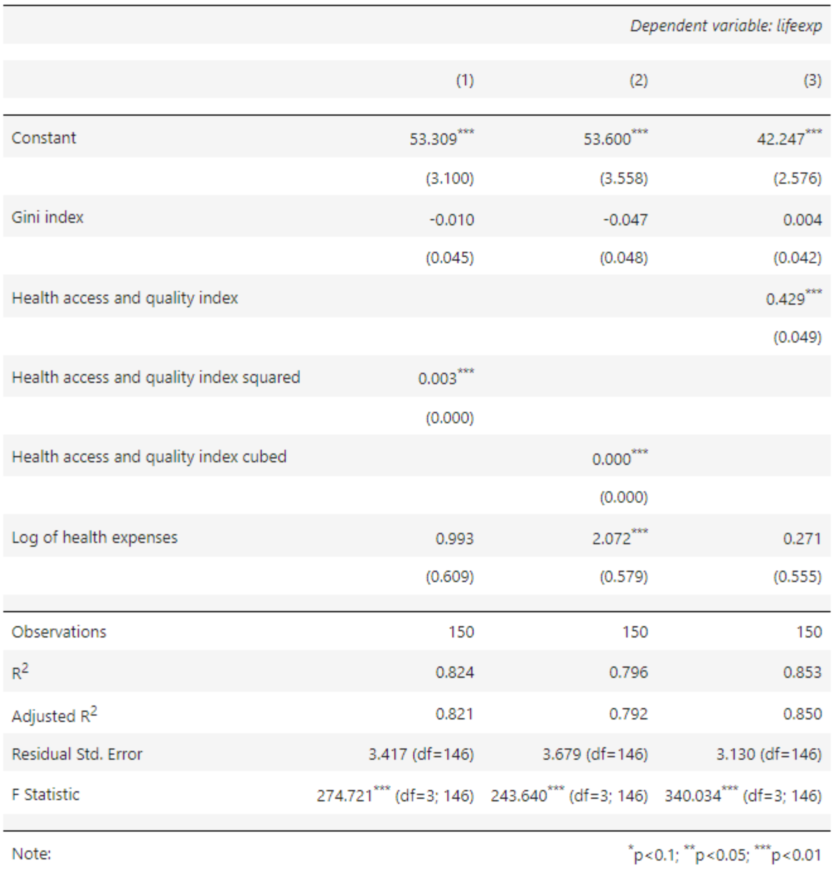


Figure 2: scatterplots and non-parametric regressions of Gini, health expenses, log of health expenses, HAQ index



				Dependent variable: lifeexp
	(1)	(2)	(3)	(4)
Constant	87.532***	78.795***	75.798***	74.283***
	(2.579)	(1.355)	(1.000)	(0.854)
Gini index	-0.438***			
	(0.069)			
4th power of Gini				-0.000***
				(0.000)
3rd power of Gini			-0.000***	
			(0.000)	
2nd power of Gini		-0.005***		
		(0.001)		
Observations	150	150	150	150
R^2	0.175	0.162	0.146	0.128
Adjusted R^2	0.169	0.157	0.140	0.123
Residual Std. Error	7.359 (df=148)	7.415 (df=148)	7.486 (df=148)	7.563 (df=148)
F Statistic	39.842*** (df=1; 148)	35.634*** (df=1; 148)	30.313*** (df=1; 148)	24.486*** (df=1; 148)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 2: regression table showing that Gini is best fitting with a linear regression



			Dependent variable: lifeexp
	(1)	(2)	(3)
Constant	53.309***	53.600***	42.247***
	(3.100)	(3.558)	(2.576)
Gini index	-0.010	-0.047	0.004
	(0.045)	(0.048)	(0.042)
Health access and quality index			0.429***
			(0.049)
Health access and quality index squared	0.003***		
	(0.000)		
Health access and quality index cubed		0.000***	
		(0.000)	
Log of health expenses	0.993	2.072***	0.271
	(0.609)	(0.579)	(0.555)
Observations	150	150	150
R^2	0.824	0.796	0.853
Adjusted R^2	0.821	0.792	0.850
Residual Std. Error	3.417 (df=146)	3.679 (df=146)	3.130 (df=146)
F Statistic	274.721*** (df=3; 146)	243.640*** (df=3; 146)	340.034*** (df=3; 146)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 3: regression table showing that HAQ is best fitting with a linear model

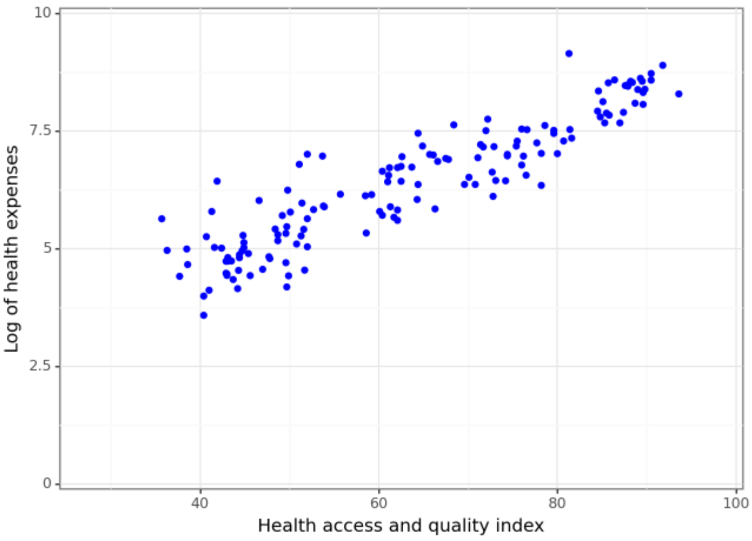


Figure 3: health expenses and HAQ: very few low HAQ but high expense (and vice versa) countries