Final Analysis STAT 209

Nam Nguyen & Hailey Nguyen

16 May, 2024

Contents

1 Introduction

1.1 What our research about?

In 2019, pollution accounts for approximately 9 million deaths, equivalent to a staggering 1 out of 6 fatalities, and this number has not seen any significant change since then. We aim to answer our research questions: "How do various environmental exposures impact health burden measured by Disability-Adjusted Life Years (DALYs) across different socio-economic settings?".

It is common knowledge that environmental exposures could affect physical health, but there has not been much discourse on the mental distress caused by pollution. This paper would add to the literature of pollution

exposures' effect on both non-communicable and communicable diseases. On communicable diseases, Young-Min Kim et al investigates the effect of chemicals generated by waste incinerators in Korea on exposure to airborne diseases, and finds out that NO2 ranked highest in all-cause mortality, and this effect is amplified when population exposure is accounted for. Regarding non-communicable diseases. Paul A Peters, Zhiqiang John Zhai discovered that exposure to pollutants can increase the risk for neuro-developmental disorders. Through this project, we hope to utilize our results to make suggestions to public health policy and global health evaluations. We also hope to gain more insights into:

1. How does mental illness and energy populations contributes to DALYs?
2. Does socio-ecomonic and geographic factors contributes to DALYs?

2 Data

2.1 Data Source

Before dive into further analysis, we want to define DALYs as the total of years lost due to premature death or total years lost due to disabilities of a country. Our data is extracted from Kaggle and composes of three main datasets: Life Expectancy & Socio-Economic dataset, Mental Disorder dataset and Sustainable Energy dataset. Our first dataset contains both demographic (Region; Income Level) and socioeconomic attributes (GDP; Health & Education Expenditure; Unemployment; Percentage of undernourished population; Life Expectancy). The second dataset focuses on non-communicable diseases, containing information on Percentage of the population who have certain mental health disorders such as Schizophrenia, Bipolar, or Depression. The last dataset contains key information on environmental exposures and energy consumption Factors: amount of electricity generated from fossil fuels, renewable energy, or nuclear; Percentage of population got access to electricity/cooking fuel; Energy amount consumed. Since DALYs as we define heavily related on the population of a country, we want to redefine our variable of interest as percentage of DALYs due to a specific reason to remove the skewness of DALYs due to population. Hence, our new variable of interests are:

$$\%DALY_{Noncomm} = \frac{DALYs\ due\ to\ Noncommunicable\ Disease\ (in\ Years)}{Total\ DALYs\ (in\ Years)}$$

$$\%DALY_{Comm} = \frac{DALYs\ due\ to\ Communicable\ Disease\ (in\ Years)}{Total\ DALYs\ (in\ Years)}$$

2.2 Data Summary

We did an inner join of all the datasets and ended up with 1,944 observations with 34 attributes from 112 countries, with a time span from 2001 to 2019. Table 1 provides a comprehensive guide of the meaning and units of measurement for each variable.

Table 1: Summary Statistic Table

Variable	Statistical Summary			
	Median	Mean	Std. Dev.	Count
Mental Illness (Continuous)				
Schizophrenia Disorder(% Pop)	0.3	0.3	0.0	1,944
Bipolar Disorder(% Pop)	0.6	0.7	0.3	1,944
Eating Disorder(% Pop)	0.2	0.2	0.2	1,944
Anxiety Disorder(% Pop)	4.1	4.4	1.4	1,944
Substance Use Disorder(% Pop)	0.6	0.7	0.4	1,944
Depressive Disorder(% Pop)	4.0	4.0	0.9	1,944
Alcohol Use Disorder(% Pop)	1.6	1.7	1.0	1,944
Country Info (Continuous)				
Population density per square kilometer (Pop/Km2)	83.0	146.0	208.6	1,944
Total land area in square kilometers (Km2)	176,215.0	637,013.4	1,417,276.8	1,944
Annual GDP growth rate based on local currency (%)	3.7	3.8	4.1	1,944
Gross domestic product per person (GDP/capita)	4,843.7	13,454.2	19,309.6	1,944
Undernorished(% Pop)	5.8	9.7	9.7	1,944
Health expenditure (% of GDP)	6.1	6.3	2.4	1,944
Education expenditure (% of GDP)	4.5	4.6	1.8	1,944
Labor force that is without work (% labor force)	5.8	7.5	5.7	1,944
Access to Electricity(% Pop)	98.5	79.5	30.0	1,944
Access to Clean Cooking Fuel(% Pop)	85.4	65.9	37.9	1,944
Life Expectancy (Year)	73.1	70.7	9.2	1,944
Percentage of DALYs due to Communicable diseases (%)	16.1	26.5	24.8	1,944
Percentage of DALYs due to Non-Communicable diseases (%)	69.8	62.7	23.1	1,944
Energy (Continuous)				
Carbon dioxide emissions (kiloton)	17,020.0	167,394.9	810,829.0	1,944
Generating capacity of Renewable electricity (W/capita)	10.7	84.2	182.8	1,944
Fossil Fuels Electricity (TWh)	4.5	76.3	353.5	1,944
Nuclear Electricity (TWh)	0.0	7.8	29.5	1,944
Renewable Electricity (TWh)	3.2	26.7	113.1	1,944
Renewable Energy in total energy consumption (% Energy)	29.0	34.0	27.4	1,944
Electricity from low-carbon sources (% Energy)	38.0	41.7	33.8	1,944
Energy consumption per person (kWh/person)	14,376.8	24,612.5	30,306.1	1,944
Energy use per unit of GDP at purchase power parity (PPP GDP)	4.2	5.0	3.0	1,944
FDI for clean energy (USD)	10,000.0	54,179,953.7	219,206,327.8	1,944
Country Info (Categorical)				
Continient Location				
East Asia & Pacific	-	-	-	257
Europe & Central Asia	-	-	-	580
Latin America & Caribbean	-	-	-	380
Middle East & North Africa	-	-	-	148
South Asia	-	-	-	101
Sub-Saharan Africa	-	-	-	478
Income Group				
High	-	-	-	691
Low	-	-	-	227
Lower middle	-	-	-	512
Upper middle	-	-	-	514
Energy (Categorical)				
Has FDI for clean energy				
No	-	-	-	935
Yes	-	-	-	1,009
Has Renewable electricity generator				
No	-	-	-	589
Yes	-	-	-	1,355

3 Analytic Framework

In order to check for statistical significance and multicollinearity, we utilized a world heatmap (Figure 1) for the two types of diseases and a correlation table (Figure 2) for all the explanatory variables. This would give us some insights into the distribution of diseases, possible explanations for such patterns, as well as weeding out irrelevant variables.

3.1 DALYs Distribution

Figure 1 gives us a visualization of DALYs distribution, broken down to communicable diseases and non-communicable diseases. Our assumption is that Region is a significant factor in affecting the percentage of DALYs, and that DALYs vary between continents. Upon inspection, we can see that Sub-Sahara and Africa areas observed a much higher distribution of DALYs due to communicable disease. This might be due to the lack of health resources and protocols as well as higher prevalence of deadly diseases (malaria,tuberculosis) due to specific geographical and environmental diseases. This would result in higher fatality for communicable diseases. On the contrary, we noticed a higher concentration of DALYs due to non-communicable diseases in Europe and North America.

3.2 DALYs relationship with other variables

Figure 2 gives us the correlation between each type of disease and variables of interest. The strongest correlation for non-communicable diseases are: Schizophrenia (positive), LifeExpectancy(positive), Undernourished (negative), Electricity Access (positive), Cooking Fuel Access (positive), Has Renewable Energy Generator or not (negative). The strongest correlation for non-communicable diseases are: Schizophrenia (negative), LifeExpectancy(negative), Undernourished (positive), Electricity Access (negative), Cooking Fuel Access (negative). Later on, we will implement our supervised learning strategies to confirm these findings.

3.3 Model Implementation

In order to identify statistically significant variables to utilize for our regressions, we used Lasso and OLS. The research employs Ordinary Least Squares (OLS) for its simplicity and efficiency in estimating linear regression model parameters, useful for determining the strength and nature of relationships between dependent and independent variables. Lasso Regression is utilized as a regularization technique to enhance prediction accuracy and interpretability by shrinking the coefficients of less important variables to zero, which comes in handy for our models with many predictors.

We will conduct our model implementation to both DALYs due to noncommunicable diseases and communicable disease simultaneously. To confirm our hypothesis, our LASSO models will initially use all explanatory variables as predictors with large λ to reveal standout variables to DALYs due to (non)communicable disease. After that, we want to quantify these variables' impact using OLS regression. Other than significant variables shown in the LASSO models, we want to include time effect (Year) and regional effect (Region) in the model as predictors.

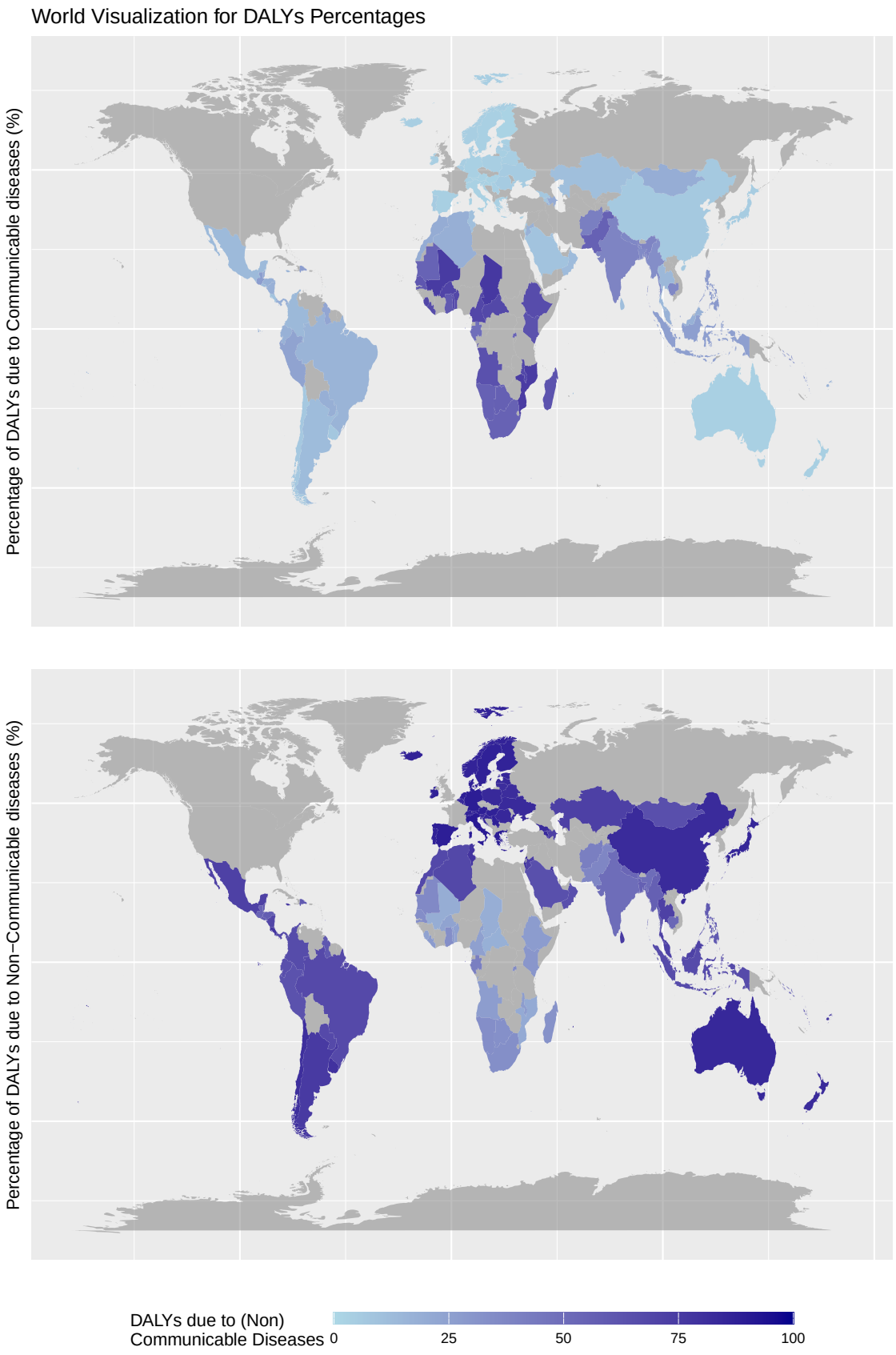

Figure 1: DALYs due to (Non)communicable distribution across continents in 2013

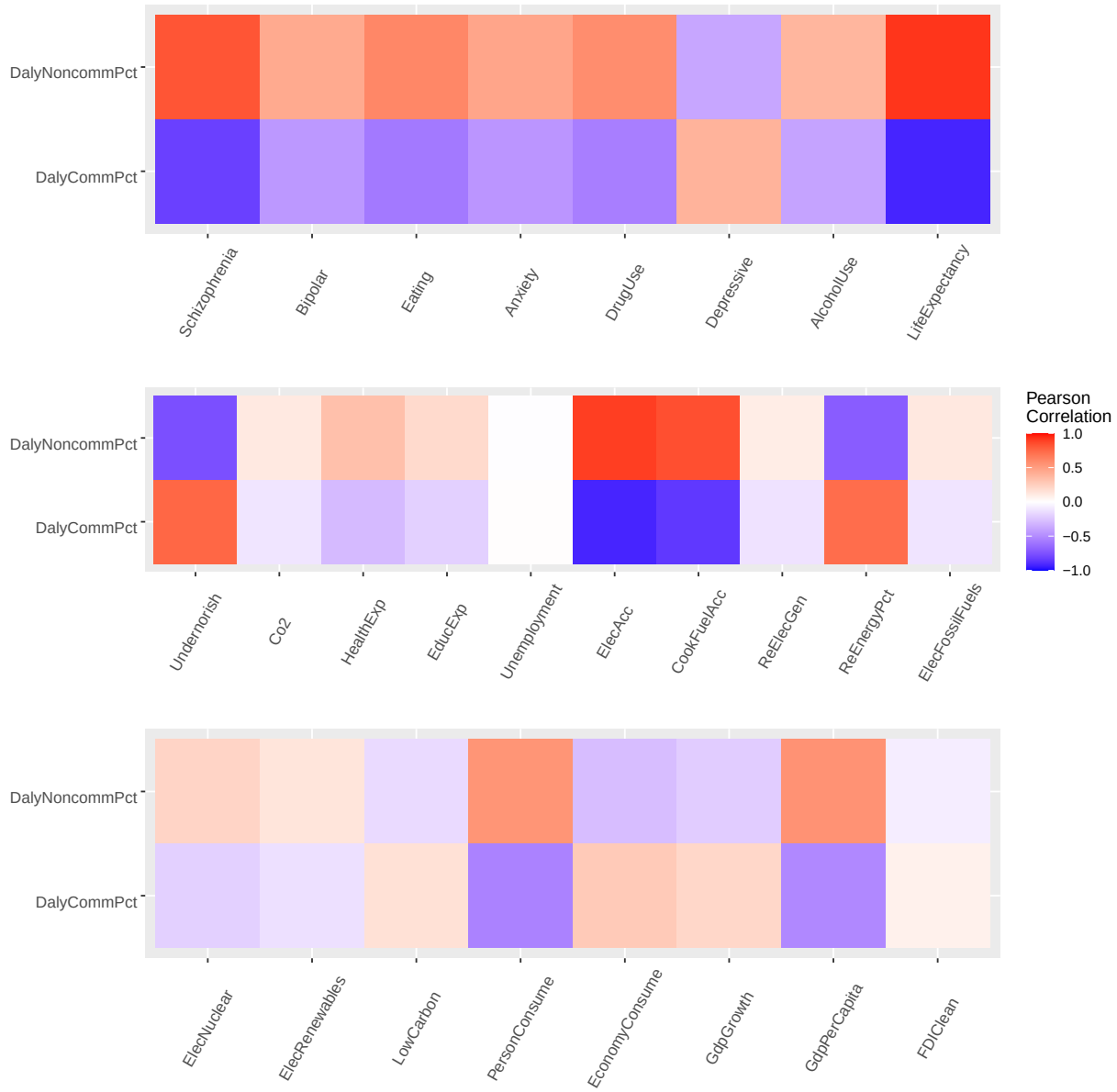

Figure 2: DALYs correlation with other variables

4 Results

4.1 Feature Selection

From our initial LASSO model, we chose $log(\lambda) = 1$ for both models and identified 6 variables that impact noncommunicable diseases DALYs (Figure 3) and 9 variables impact communicable diseases (Figure 4). Choosing a relatively high λ comes with some drawbacks on our cross-validated MSE, but it also gives us a better understanding of what variable to focus on in later model as well as improving the robustness of the model in general.

4.2 Final LASSO & OLS Model

For our non-communicable diseases LASSO model, we were able to narrow down to: Schizophrenia (% population- continuous), and Undernourished (% population- continuous). It can be deduced that mental illnesses that are characterized by irrational actions and high risk level (Schizophrenia), and undernourishment, are significant drivers of non-communicable DALYs. For our communicable model, we identified Electricity Access (% population, continuous), Cooking Fuel Access (% population, continuous), Has Renewable Energy Generator or not (binary), Has investment for FDI on renewable energy or not (binary). This model agrees with previous literature on the effects of byproducts through energy consumption on physical diseases. Life Expectancy (number of years, continuous), Time (Years, continuous), and Continent (categorical) are statistically significant explanatory variables in both models. This finding shows strong evidence on how DALYs is impacted regionally. Other evidences shows regional influences on DALYs are interactions between Bipolar with Sub-Sahara Africa region (Figure) in the non-communicable diseases model and interaction between Electricity Accessibility with Europe-Central Asia. Therefore quantify these impact, we finalize our OLS model with variables appears in individually and in interaction terms. We want to look at how each of these variables contributes to DALYs separately; therefore, our OLS model will only consist their fixed effect terms

$$DALY_{Noncomm} = Year + LifeExpectancy + Region + Schizophrenia + Bipolar + Undernourished$$

$$DALY_{Comm} = Year + LifeExpectancy + Region + ElecAcc + CookFuelAcc + RenewGen + Invested$$

Table 2 and 3 shows the coefficients for both LASSO and OLS models in comparison. We can see that the OLS model showcase stronger coefficients for both non-communicable and communicable diseases model. This is because LASSO put a heavy penalty on the coefficient, which help to regularize the coefficient and somewhat hinder our interpretation for variables' effects. Another disadvantage from the LASSO model we can see from the communicable disease model is that many coefficients is close to zero, which does not helpful for our investigation eventhough they are standout compared to other variables. Therefore, we will only focus on interpreting the effect of predictors on the OLS model.

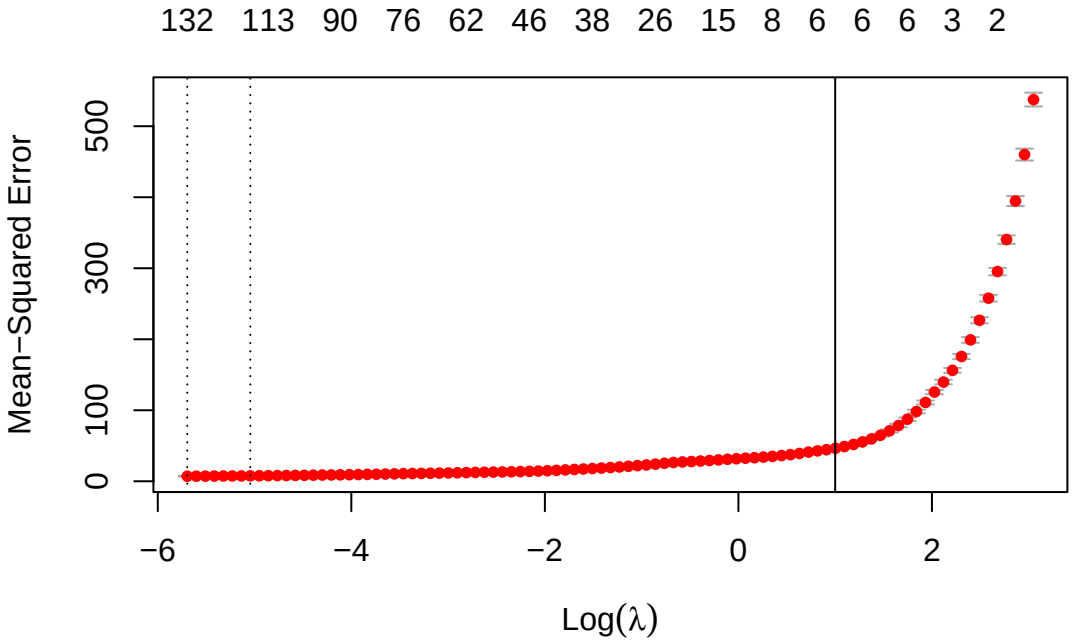

Figure 3: Noncommunicable disease LASSO models decision for chosen λ

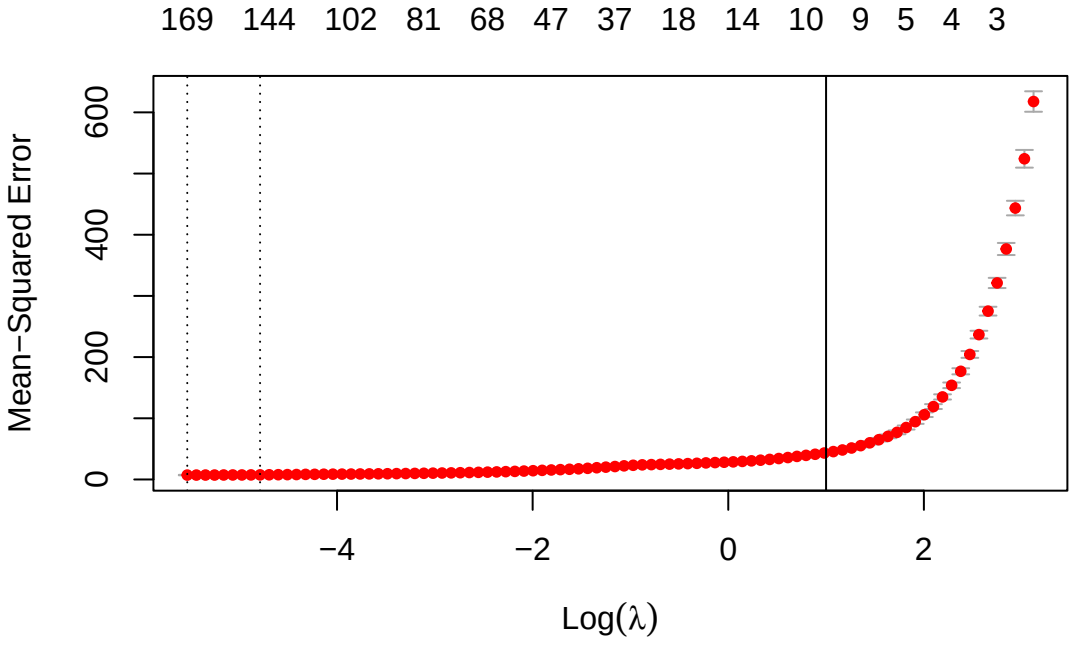

Figure 4: Communicable disease LASSO models decision for chosen λ

4.3 Model Interpretation

4.3.1 Non-communicable Disease Model

For non-communicable disease OLS model, we found that Schizophrenia is the most impactful predictors among mental health and socio-economic factors. Given in Table 2, a percentage point increase of Schizophrenia population leads to an increase of 94.584 percentage point of DALYs due to non-communicable disease. Although the coefficient is large, we can see that from Table **??**, Schizophrenia does not have wide spread (St. Dev. Schizophrenia ≈ 0). In other words, Schizophrenia is a still good predictor of DALYs due to non-communicable diseases but the lack of variation poses an issue, in particular a relatively "constant" effect for non-communicable disease DALYs. On the other hand, a percentage point increase in Bipolar population leads to 5.324 decrease of percentage point in DALYs due to non-communicable diseases. This finding contradicts with what we hypothesized earlier (positive correlation).

Regarding region, our model also shows the coefficient's difference between regions' coefficients, where Europe and Central Asia (coef = 11.668) has greater shares of non-communicable diseases than Sub-Saharan Africa or South Asia (-9.95 and -8.994, respectively). We can see that as Life Expectancy of a country increases, non-communicable DALYs also increase. This suggests that as people live longer, we would expect better treatment of communicable disease due to advanced health technology. However, access to technology and automation isn't always a good thing, as it could lead to an increased portion of non-communicable disease (eg: processed food, social meida, etc). Figure 5 portraits the relative distribution of both life expectancy and non-communicable disease, showing strong evidence of positive correlation across regions.

Variable	OLS	LASSO
Schizophrenia	94.584	50.435
Bipolar	-5.324	
LifeExpectancy	1.207	1.235
Undernorish	-0.178	-0.138
Year	0.006	-0.015
Europe & Central Asia	11.668	6.796
Sub-Saharan Africa	-9.95	
South Asia	-8.994	
Latin America & Caribbean	1.807	
Middle East & North Africa	4.819	
Bipolar:Sub-Saharan Africa		-16.108
(Intercept)	-57.778	-7.807
MSE	30.442	47.366
R-square	0.941	0.908

Table 2: DALYs due to Noncommunicable Diseases Coefficients

4.3.2 Communicable Disease Model

Although we found more significant predictors for the communicable disease variables, no variables shows large impact similar to Schizophrenia in the other model for non-communicable disease model. Our two most significant environmental factors are Has Renewable Energy Generator or not and Has investment for FDI on renewable energy or not. This highlights the importance of the transitioning energy type. We also saw that both Electric and Cooking Fuel Accessibility has a negative coefficient with communicable disease DALYs. In contrast with non-communicable disease model, the regional effect on DALY is the greatest in Sub-Saharan Africa (coef: 17.65), followed by South Asia (coef: 9.938) and the least is Europe - Central Asia (coef: -2.918). These findings generally agrees with our findings for non-communicable diseases model. DALYs decrease with higher access to healthcare and clean energy. Once again, life expectancy has a negative correlation with communicable disease. This also show evidences that DALYs portions shift from communicable disease to non-communicable disease more as life expectancy increase.

Variable	OLS	LASSO
CookFuelAcc	-0.038	-0.055
ElecAcc	-0.248	-0.224
LifeExpectancy	-0.892	-0.936
RenewGenYes	0.409	0.427
InvestedYes	2.96	0.04
Year	-0.018	0.07
Europe & Central Asia	-2.918	-2.105
Sub-Saharan Africa	17.65	10.765
South Asia	9.938	
Latin America & Caribbean	2.132	
Middle East & North Africa	2.013	
ElecAcc:Europe & Central Asia		-0.001
Year:Sub-Saharan Africa		0
(Intercept)	141.611	-28.049
MSE	27.245	41.398
R-square	0.954	0.93

Table 3: DALYs due to Communicable Diseases Coefficients

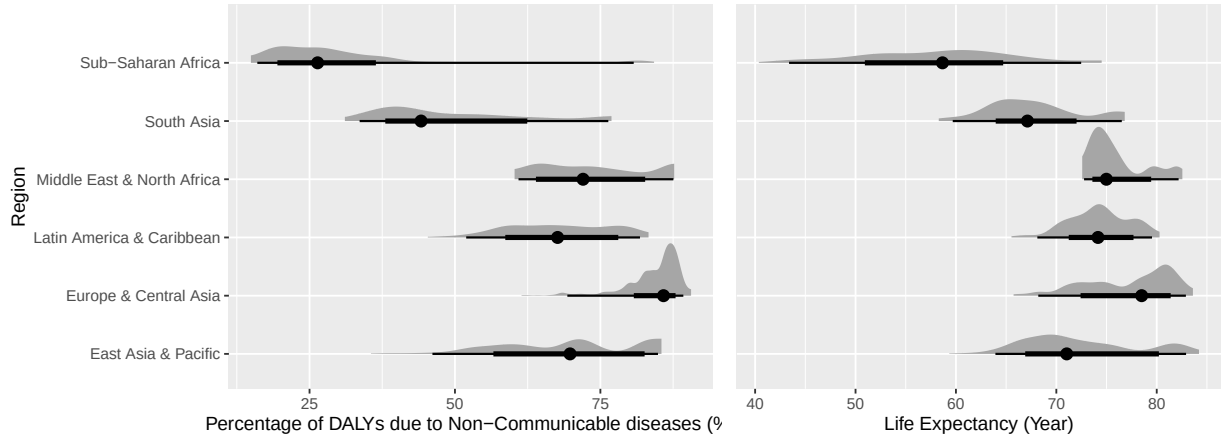

Figure 5: DALYs due to Noncommunicable Disease and Life Expectancy Distribution across regions

4.4 Evaluation

From table 2 and 3, we can see that our MSE for the model is around 50, which is high compared to a model with a response variable as percentage point. This implies some systematic errors on our models as we only evaluate a simple OLS regression. Despite a large MSE, our model got a significantly large R-square value across 4 models (> 0.9), suggesting that our predictors are unable to explain the variation from the response variable. On the other note, we can see from figure 6 and 7, both LASSO and OLS model produce an upward trend in the residual plot. This can mean that our model has yet truly capture some complex trend between DALYs and mental health, energy consumption, and socio-economic factor. We think that our model could be improved by using more advance model such as Mixed-Effect Model or difference-in-difference strategy to better quantify the regional effect on our response and predictor variables.

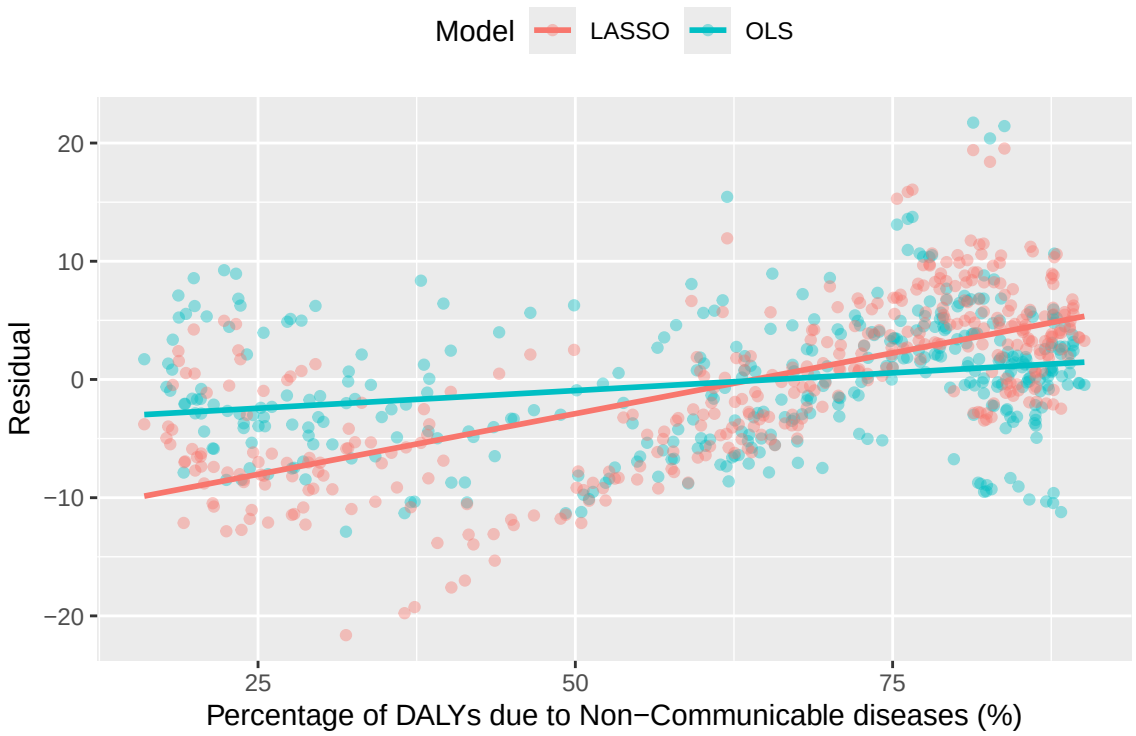

Figure 6: Residual vs. Actually Non-communicable Disease DALYs Plot

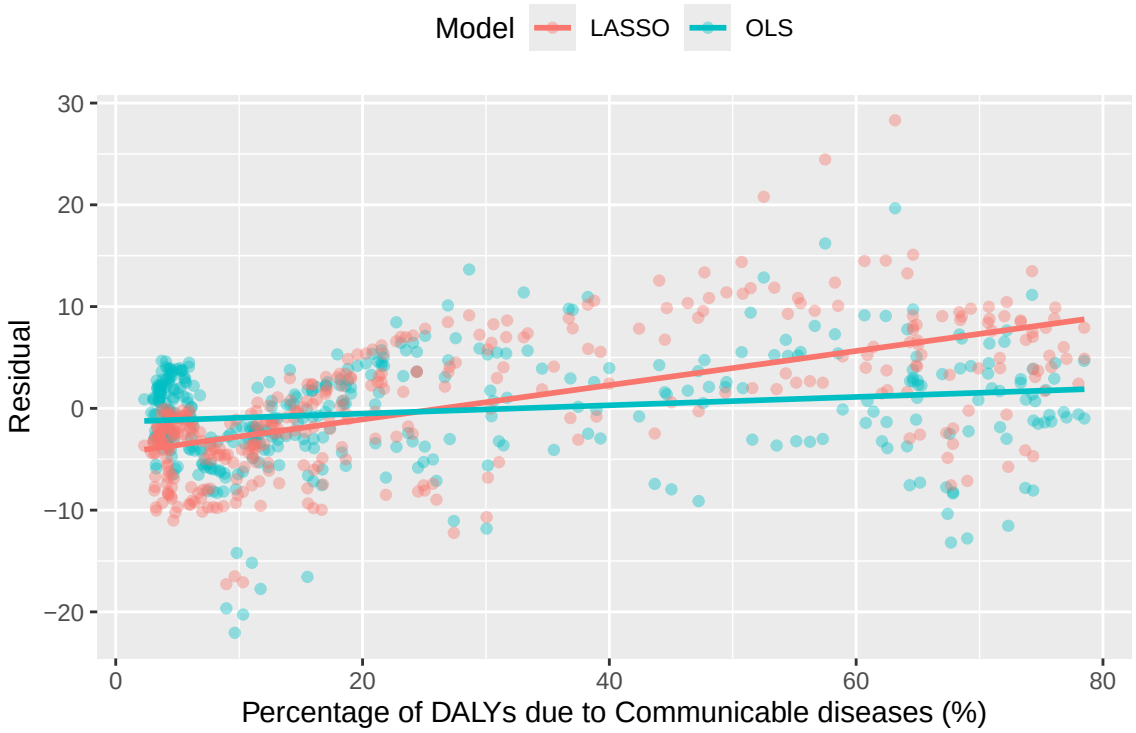

Figure 7: Residual vs. Actually Communicable Disease DALYs Plot

5 Implications

In summary, we think that our study has capture the essense of DALYs:

1. For mental health factor, schizophrenia shows the most impact to non-communicable disease DALYs.
2. For energy consumption, electricity and cooking fuel accessibility reduce communicable disease DALYs.
3. For socio-economic factor, life expectancy reduce communicable disease DALYs while increse non-communicable disease DALYs.
4. Regional location has impact on not just DALYs but also has complex relationship with socio-economic, mental healths and energy consumption factors

With these findings, we believe that DALYs caused by mental health issues, specifically those caused by neurodivergence, cannot be avoided in nature. There could be medications that address the issue, but that is only the tip of the iceberg. Our solution to reduce DALYs in general would be to allocate resources on addressing communicable diseases. By allocating emergency heathcare forces to areas of need and ensure clean, usable energy, we can gradually improve life expectancy and reduce chance of premature death due to communicable disease. It is important to note that our project focuses on studying the relative proportions of diseases in comparison to each other, rather than analyzing them individually. This means that if we can minimize total DALYs by reducing communicable DALYs.

Our research indicates that various regions should adopt tailored strategies to reduce their burden of disability-adjusted life years (DALYs). For instance, Sub-Saharan Africa faces a greater prevalence of communicable disease. For example, a few methods to combat malaria, one of the most deadly diseases in these areas, include but are not limited to: structured vaccination programs, mosquito net distribution, and antiretroviral therapy access.

Conversely, regions like Europe and Central Asia might focus on enhancing mental health care to address non-communicable disease DALYs. This could involve increasing access to counseling services, promoting awareness campaigns to reduce stigma around mental health issues, and integrating mental health screening into primary care settings.