Defining the Key Predictors of Losses in Healthy Years of Life: A Cross-Country Investigation

Kyselova, Vladyslava and Popovych, Tetiana and Buryhina, Khrystyna and Topchii, Sofiia

Kyiv School of Economics

15 April 2024

1

Defining the Key Predictors of Losses in Healthy Years of Life: A Cross-Country Investigation

Kyselova, Vladyslava and Popovych, Tetiana and Buryhina, Khrystyna and Topchii, Sofiia

Kyiv School of Economics

15 April 2024

[1] This study was part of the Regression Analysis course at Kyiv School of Economics, conducted under the curriculum designed for second-year students.

Defining the Key Predictors of Losses in Healthy Years of Life: A Cross-Country Investigation

15 April 2024

Vladyslava Kyselova
vkyselova@kse.org.ua

Tetiana Popovych
tpopovych@kse.org.ua

Khrystyna Buryhina
kburyhina@kse.org.ua

Sofiia Topchii
stopchii@kse.org.ua

1. Abstract

This study identifies and analyzes the key environmental factors contributing to the loss of healthy years of life, as quantified by Disability-Adjusted Life Years (DALYs), across various regions worldwide. Using data from The Organisation for Economic Co-operation and Development (OECD) we employ an ordinary least squares (OLS) regression model to estimate the impacts of several predictors on DALYs. Our analysis indicates a positive relationship between lead exposure, air pollution, second-hand smoke, and DALYs, while the quadratic effect of air pollution negatively impacts the number of healthy life years lost. The findings show variations in the influence of these factors among different global regions, highlighting the highest level of DALY in areas with environmental issues like Africa and the smallest in well-regulated environmental regions such as Oceania. This research supports targeted public health interventions and policies aimed at mitigating environmental risks, particularly in vulnerable populations, to enhance global health outcomes.

Keywords: Disability-Adjusted Life Years (DALYs), Cross-Country Analysis, Determinants of DALYs, Public Health Interventions

Contents

2. Introduction

The purpose of our study is to analyze how various factors affect the DALY indicator, which demonstrates the number of healthy years of life lost due to a number of environmental problems and threats. This analysis will allow us to understand the threat to human health and life posed by such environmental factors as water and air pollution, climate, waste disposal, and others. The importance and relevance of this study is to get a clear understanding of how the invisible dangers we face every day become destructive factors for our health. The study of the relationship between various environmental hazards and DALYs will be able to attract the attention of the government, healthcare and research institutions in order to reduce the impact of these indicators and make changes in public policy to focus on solving the problems of pollution, harmful emissions, waste, etc.

This study should show how environmental pollution is a serious threat that reduces life expectancy. And it can become the basis for making important decisions aimed at solving environmental problems and, as a result, preserving people's health and lives.

This study focuses on a universally compelling narrative: the impact of the environment on human health, as measured by the DALYs. It is fascinating to explore the invisible, often overlooked connections between daily environmental exposures (such as air and water pollution) and their long-term impact on human well-being. This study is particularly compelling because it highlights the potential for critical health and environmental policies to reduce these risks. It underscores the importance of collective action and awareness to inspire sustainable practices and policy changes that promote healthy environments. The convergence of individual health outcomes and global environmental challenges makes this topic not only interesting, but also critical to society's efforts to address and adapt to these challenges.

The value of this study is that it can help to significantly reduce risks or avoid morbidity. The results obtained can contribute to the role of social responsibility in people, should encourage the public to be aware of their own actions and the impact they may have on others, to use environmentally friendly technologies and reduce emissions.

3. Data and Methods Description

3.1 Data Sources

Data for this study was obtained from the Organisation for Economic Co-operation and Development - Mortality, morbidity and welfare cost from exposure to environment-related risks (oecd.org). We used the OECD source because of its accuracy and reliability, and because the organization provides a platform for comparing policy experiences, finding solutions to common problems, identifying best practices, and working to coordinate domestic and international policies. In particular, for the study of disability-adjusted life years, data for the period of 2019 were taken from GBD (2019), Global Burden of Disease Study 2019 Results, to measure the average value of DALYs for all countries during the selected period. This allowed us to avoid errors and inaccuracies in the indicators and obtain accurate and reliable data.

3.2 Dependent variable

Our dependent variable, expressed as a % of total DALYs. One DALY represents the loss of the equivalent of one year of full health. In the context of DALYs, full health is defined as a state in which a person does not have a serious illness, injury, or disability that prevents them from living a full and productive life without any limitations. Using DALYs, the burden of diseases that cause premature death but little disability (such as drowning or measles) can be compared to that of diseases that do not cause death but do cause disability (such as cataract causing blindness).

3.4 Independent variables

The independent variables considered in this study are regarding Region, Air pollution, High temperature, Lead emissions, Unsafe water, Unsafe sanitation and Second-hand smoke. Each variable is measured as follows:

1. **Region:** North America, South America, Oceania, Asia, Africa and Europe.
2. **Air pollution:** Air pollution from solid fuels is estimated based on the proportion of households using solid cooking fuels. The definition of solid fuel includes coal, wood, charcoal, dung, and agricultural residues.
3. **High temperature:** High temperature is quantified as a variable representing the relative mortality risk associated with exposure to varying degrees of average daily temperatures and distinct temperature zones, measured in degrees Celsius (°C) or Fahrenheit (°F) and categorized within specific ranges or 'zones' that correlate with increased health risks.
4. **Lead:** Lead is defined in two different ways according to the currently known pathways of health loss. Acute lead exposure, relevant to disease burden through IQ loss in children, is measured as the micrograms of lead per decilitre of blood (μg/dL). Long-term lead exposure, relevant to disease burden in adults given the manifestation of health impact through increased systolic blood pressure and hence a decline of cardiovascular health, is measured as the accumulation of lead in the bone as micrograms of lead per gram of bone (μg/g).
5. **Unsafe water:** Unsafe water source is expressed as the proportion of individuals with unimproved water source prevalence or household water treatment.

6. **Unsafe sanitation:** Unsafe sanitation is expressed as the proportion of individuals without sewer connection or improved sanitation.
7. **Second-hand smoke:** Second-hand smoke is expressed as the proportion of the population exposed to second-hand smoke at home, work or in other public places.

3.5 Model estimation methodology

The model will be evaluated through ordinary least squares (OLS) regression analysis. This OLS approach will determine the values of the coefficients. T-tests will be used to evaluate the significance of these coefficients, while the model's overall effectiveness will be calculated by using the coefficient of determination (R^2) and the adjusted R^2. Additionally, the utility of each independent variable will be examined using the F-statistic to construct the most explanatory model. Ultimately, this model will be used for forecasting purposes and to realize the key factors influencing the dependent variable.

3.6 The model equation

The regression model equation can be represented as follows:

DALYs = b0 + b1*Region + b2*Air_pollution + b3*High_temperature + b4*Lead + b5*Unsafe_water_source + b6*Unsafe_sanitation + b7*Second_hand_smoke + e

Where:

• "β0" is the intercept (constant term).

• "βi" are the coefficients for the corresponding predictor variables.

• "ε" is the error term (representing the difference between the actual and predicted values of the dependant variable).

3.7 Prior expectations

The above variables were selected based on research conducted by the World Health Organisation, and the Environmental Protection Agency (EPA).

Considering the existing theories, the expectations regarding the influence of each predictor on the DALYs can be formulated as follows:

Hypothesis 1: DALYs are expected to be highest in countries in the Africa region and lowest in countries in the Europe region. This is because countries in the Africa region have the least resources and capacity to control the factors that cause death, and countries in the Europe region have the most resources and capacity.

Hypothesis 2: Air pollution is expected to have a positive effect on DALYs, as it can lead to respiratory and cardiovascular diseases, as well as cancer.

Hypothesis 3: Extremely high temperatures are expected to have a positive impact on DALYs because they can lead to increased illness and death from heat, cold, and extreme weather events.

Hypothesis 4: Lead emissions are expected to have a positive effect on DALYs because they lead to serious health problems such as high blood pressure and lung cancer.

Hypothesis 5: Unsafe water indicators are expected to have a positive effect on DALYs as they lead to the spread of diseases such as malaria.

Hypothesis 6: Unsafe sanitation indicators are expected to have a positive effect on DALYs as they lead to the spread of diseases such as hepatitis and cholera.

Hypothesis 7: Second-hand smoke risks are expected to have a positive effect on DALYs because they can lead to respiratory diseases, cardiovascular diseases, and lung cancer, and specific diets can lead to diabetes and cancer.

4. Empirical Analysis and Results

4.1 Visualizing the relationship

We start start the analysis by exploring the relationship in our data visually

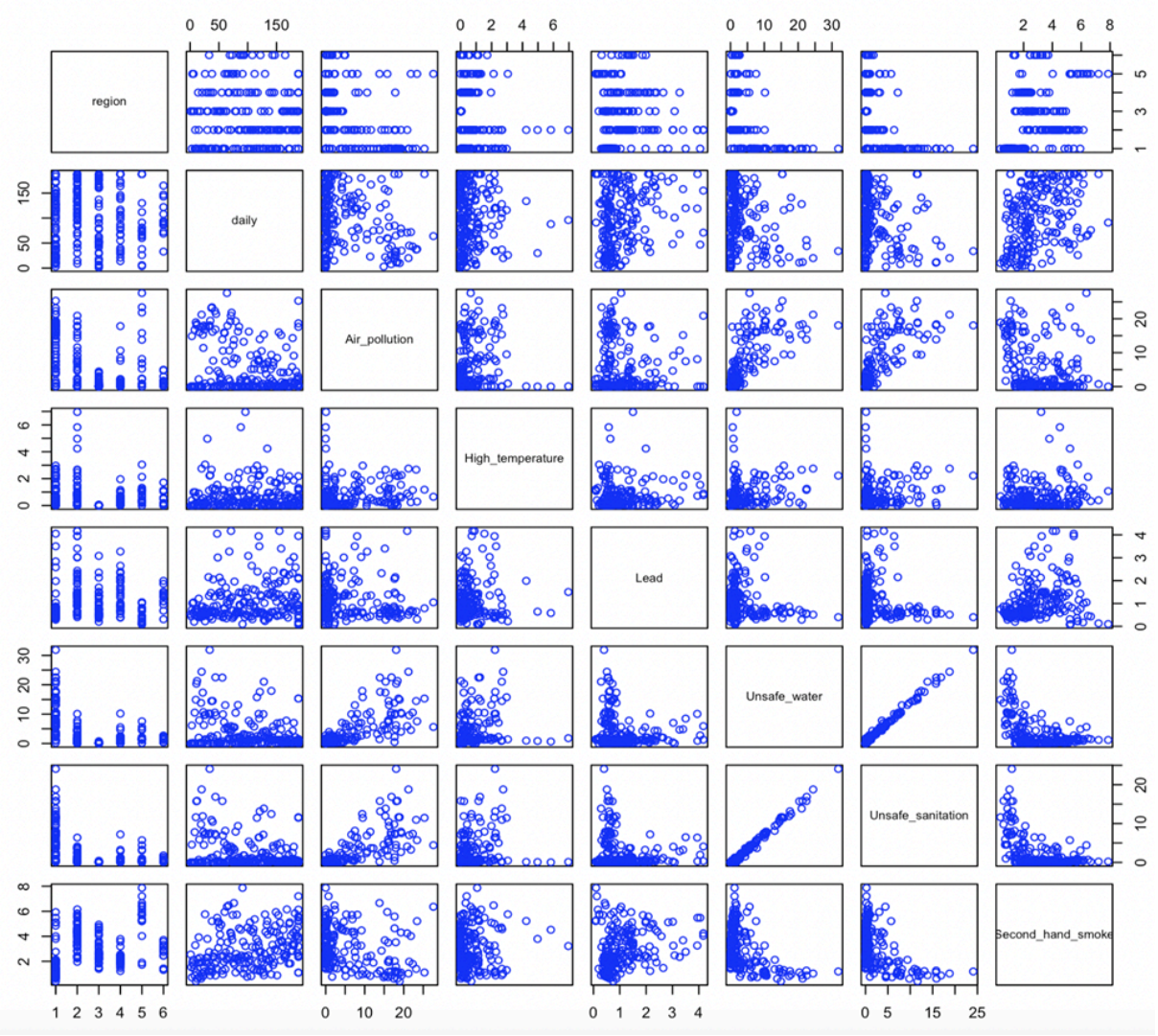


Figure 1: Plot of the relationships between all the variables within the dataset

The scatterplots show a strong positive relationship between the variables Unsafe_water and Unsafe_sanitation, as the points form an almost straight line. For the other variables, the points do not suggest obvious linear relationships, and require further investigation.

4.2 Exploring correlations

These preliminary conclusions can be supported when calculating correlations between variables in our dataset.

	daily	Air_pollution	High_temperature	Lead	Unsafe_water	Unsafe_sanitation	Second_hand_smoke
daily	1.00000000	-0.29500502	-0.11824688	0.32761174	-0.28358132	-0.2844197	0.390449605
Air_pollution	-0.29500502	1.00000000	0.13499761	-0.05772097	0.72885627	0.7334178	-0.226439969
High_temperature	-0.11824688	0.13499761	1.00000000	0.08600661	0.24877080	0.2194743	0.112442103
Lead	0.32761174	-0.05772097	0.08600661	1.00000000	-0.14141521	-0.1574252	0.253892184
Unsafe_water	-0.28358132	0.72885627	0.24877080	-0.14141521	1.00000000	0.9949537	-0.459735510
Unsafe_sanitation	-0.28441967	0.73341780	0.21947432	-0.15742522	0.99495367	1.0000000	-0.467908866
Second_hand_smoke	0.39044960	-0.22643997	0.11244210	0.25389218	-0.45973551	-0.4679089	1.000000000
regionAsia	0.16177289	-0.06134948	0.26580653	0.29847189	-0.18444888	-0.1997515	0.384320117
regionEurope	0.23562194	-0.35877970	-0.36243019	-0.11414578	-0.34730133	-0.3119931	-0.008564486
regionOceania	-0.17794299	0.15322610	0.04148348	-0.18900647	-0.06791605	-0.0726835	0.448229829
regionNorth America	-0.09258753	-0.16126423	-0.05909320	0.16654429	-0.12658862	-0.1413875	-0.193499337
regionSouth America	-0.04515627	-0.14422612	-0.03588305	0.03782273	-0.11376362	-0.1210101	-0.049130724
regionAfrica	-0.17405496	0.49442002	0.11477843	-0.21417713	0.69005481	0.6900611	-0.456145282

regionAsia	regionEurope	regionOceania	regionNorth America	regionSouth America	regionAfrica
0.16177289	0.235621941	-0.17794299	-0.09258753	-0.04515627	-0.1740550
-0.06134948	-0.358779700	0.15322610	-0.16126423	-0.14422612	0.4944200
0.26580653	-0.362430192	0.04148348	-0.05909320	-0.03588305	0.1147784
0.29847189	-0.114145775	-0.18900647	0.16654429	0.03782273	-0.2141771
-0.18444888	-0.347301331	-0.06791605	-0.12658862	-0.11376362	0.6900548
-0.19975152	-0.311993125	-0.07268350	-0.14138749	-0.12101011	0.6900611
0.38432012	-0.008564486	0.44822983	-0.19349934	-0.04913072	-0.4561453
1.00000000	-0.297948146	-0.16248390	-0.21390028	-0.14956878	-0.3544901
-0.29794815	1.000000000	-0.14839262	-0.19534996	-0.13659756	-0.3237473
-0.16248390	-0.148392617	1.00000000	-0.10653271	-0.07449250	-0.1765533
-0.21390028	-0.195349958	-0.10653271	1.00000000	-0.09806490	-0.2324218
-0.14956878	-0.136597555	-0.07449250	-0.09806490	1.00000000	-0.1625198
-0.35449013	-0.323747271	-0.17655327	-0.23242176	-0.16251984	1.0000000

Table 1: The correlation matrix of all the variables

The correlation matrix shows that Unsafe_water and Unsafe_sanitation have a high positive correlation coefficient of 0.9949357, indicating an almost perfect positive linear relationship between these variables.

More specifically, we can see that DALYs correlate with the chosen predictors in our dataset. Hence, it makes sense to fit a regression model to our data.

4.3 Initial full model

We begin our regression analysis by including all of the predictors that we collected based on the theory.

```
Call:
lm(formula = daily ~ region + Air_pollution + High_temperature +
    Lead + Unsafe_water + Unsafe_sanitation + Second_hand_smoke,
    data = merged_with_dummies)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2037 -1.5326 -0.3556  1.2451  7.6125

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.16350    0.77181   0.212  0.83248
regionEurope        1.45120    0.57957   2.504  0.01320 *
regionOceania      -3.15495    0.79729  -3.957  0.00011 ***
regionNorth America 0.07378    0.65142   0.113  0.90996
regionSouth America -0.09164   0.75838  -0.121  0.90396
regionAfrica        1.13457    0.60430   1.877  0.06211 .
Air_pollution      -0.06387    0.03685  -1.733  0.08480 .
High_temperature   -0.38526    0.17977  -2.143  0.03349 *
Lead                0.65595    0.23230   2.824  0.00530 **
Unsafe_water        0.28885    0.31973   0.903  0.36755
Unsafe_sanitation  -0.32641    0.42817  -0.762  0.44688
Second_hand_smoke   1.04362    0.16676   6.258 2.91e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.2 on 175 degrees of freedom
Multiple R-squared:  0.4026,    Adjusted R-squared:  0.365
F-statistic: 10.72 on 11 and 175 DF,  p-value: 5.748e-15
```

We use the backward selection procedure to select only those predictors that impact DALYs in a statistically significant way.

4.4 Reduced models and interactions

Our reduced model can be presented as follows:

```
Call:
lm(formula = daily ~ region + Air_pollution + High_temperature +
    Lead + Second_hand_smoke, data = merged_with_dummies)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2180 -1.5239 -0.3099  1.2842  7.6356

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.37880    0.74796   0.506 0.613179
regionEurope        1.34273    0.56231   2.388 0.018000 *
regionOceania      -3.11819    0.79479  -3.923 0.000125 ***
regionNorth America 0.09292    0.64950   0.143 0.886407
regionSouth America -0.07865   0.75692  -0.104 0.917361
regionAfrica        1.31974    0.56025   2.356 0.019585 *
Air_pollution      -0.04983    0.02924  -1.704 0.090120 .
High_temperature   -0.31817    0.16982  -1.874 0.062643 .
Lead                0.67357    0.23071   2.920 0.003961 **
Second_hand_smoke   1.00448    0.16034   6.265 2.76e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.196 on 177 degrees of freedom
Multiple R-squared:  0.3979,    Adjusted R-squared:  0.3673
F-statistic:    13 on 9 and 177 DF,  p-value: 8.161e-16
```

In this model, we retain only region, Air_pollution, High_temperature, Lead, Second_hand_smoke. According to their p-value we should omit the Air_pollution, as its p-value is more than 0.05, but our further investigation shows that including the interaction between region and Air_pollution increases the model quality a lot:

```
Call:
lm(formula = daily ~ region:Air_pollution + Air_pollution + Lead +
    Second_hand_smoke + region, data = merged_with_dummies)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6792 -1.2102 -0.2396  0.9869  6.5982

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.75093    0.74968   2.336 0.020660 *
Air_pollution                    -0.03855    0.04934  -0.781 0.435712
Lead                              0.55771    0.22585   2.469 0.014505 *
Second_hand_smoke                 0.58873    0.16317   3.608 0.000404 ***
regionEurope                     -0.03754    0.59104  -0.064 0.949430
regionOceania                    -2.83808    0.90026  -3.153 0.001908 **
regionNorth America              -0.11691    0.68513  -0.171 0.864711
regionSouth America              -0.39439    0.94616  -0.417 0.677315
regionAfrica                      1.45783    0.66227   2.201 0.029041 *
regionEurope:Air_pollution        1.71979    0.28039   6.134 5.67e-09 ***
regionOceania:Air_pollution       0.02505    0.07732   0.324 0.746340
regionNorth America:Air_pollution -0.12511    0.10835  -1.155 0.249829
regionSouth America:Air_pollution -0.01924    0.36755  -0.052 0.958317
regionAfrica:Air_pollution        -0.09702    0.06541  -1.483 0.139789
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.013 on 173 degrees of freedom
Multiple R-squared:  0.5059,    Adjusted R-squared:  0.4688
F-statistic: 13.63 on 13 and 173 DF,  p-value: < 2.2e-16
```

The summary shows that multiple R-squared is 10% bigger for the model with interaction.

```
Analysis of Variance Table

Model 1: daily ~ region + Air_pollution + High_temperature + Lead + Unsafe_water +
    Unsafe_sanitation + Second_hand_smoke
Model 2: daily ~ region + Air_pollution + High_temperature + Lead + Second_hand_smoke
Model 3: daily ~ region:Air_pollution + Air_pollution + Lead + Second_hand_smoke +
    region
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    175 847.32
2    177 853.96 -2    -6.633 0.8188    0.4427
3    173 700.76  4   153.201 9.4554 6.17e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We use the F-test to compare the full and reduced models. As p-value for the third model is less than 0.05, we reject the null hypothesis in favor of accepting the alternative one.

4.5 Exploration of quadratic effects

The analysis proceeds by checking whether there is any quadratic effect in our model. Further testing showed that one of our predictors indeed develops a nonlinear relationship to the response variable.

```
Call:
lm(formula = daily ~ region:Air_pollution + I(Air_pollution^2) +
    Air_pollution + Lead + Second_hand_smoke + region, data = merged_with_dummies)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6192 -1.2950 -0.3041  1.1250  6.5613

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      1.269610   0.770596   1.648 0.101267
I(Air_pollution^2)              -0.009971   0.004393  -2.270 0.024465 *
Air_pollution                    0.130976   0.089198   1.468 0.143828
Lead                             0.555614   0.223191   2.489 0.013746 *
Second_hand_smoke                0.662304   0.164472   4.027 8.47e-05 ***
regionEurope                     0.261583   0.598753   0.437 0.662747
regionOceania                   -3.378348   0.920939  -3.668 0.000325 ***
regionNorth America              0.071264   0.682107   0.104 0.916913
regionSouth America             -0.170672   0.940187  -0.182 0.856166
regionAfrica                     1.173625   0.666337   1.761 0.079963 .
regionEurope:Air_pollution       1.541744   0.287971   5.354 2.73e-07 ***
regionOceania:Air_pollution      0.104023   0.083957   1.239 0.217032
regionNorth America:Air_pollution -0.142126  0.107336  -1.324 0.187220
regionSouth America:Air_pollution -0.117519  0.365794  -0.321 0.748395
regionAfrica:Air_pollution       -0.051961   0.067615  -0.768 0.443257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.989 on 172 degrees of freedom
Multiple R-squared:  0.5203,    Adjusted R-squared:  0.4813
F-statistic: 13.33 on 14 and 172 DF,  p-value: < 2.2e-16
```

Also, the quality of the model has slightly improved, according to the R-squared and adjusted R-squared values. We proceed to the comparison of the two models: with interaction and with quadratic effect.

```
Analysis of Variance Table

Model 1: daily ~ region:Air_pollution + Air_pollution + Lead + Second_hand_smoke +
    region
Model 2: daily ~ region:Air_pollution + I(Air_pollution^2) + Air_pollution +
    Lead + Second_hand_smoke + region
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    173 700.76
2    172 680.38  1    20.378 5.1517 0.02447 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value states that there is a statistically significant difference between the models at 0.05 significance level. Thus, we chose the last model as the one that predicts the results better.

Before commencing to interpreting the results, we need to check whether the model meets all the assumptions for the regression analysis.

4.6 Testing for collinearity

```
> library(faraway)
> vif(model4)
                I(Air_pollution^2)                       Air_pollution                                Lead
                         18.231546                          18.999523                            1.566244
             Second_hand_smoke                         regionEurope                       regionOceania
                          3.006356                           2.849768                            2.776982
          regionNorth America                 regionSouth America                         regionAfrica
                          2.372540                           2.509489                            4.213693
  regionEurope:Air_pollution       regionOceania:Air_pollution regionNorth America:Air_pollution
                          1.670545                           4.142411                            1.646004
regionSouth America:Air_pollution       regionAfrica:Air_pollution
                          2.181705                           8.650695
```

To test whether there is a collinearity problem in our model, we apply the Variance Inflation Factor (VIF) test. In this output, we see that some predictors have a VIF greater than 5, which is not a problem of multicollinearity, but the result of the interaction between air pollution and region, and quadratic effect.

4.7 Testing for regression analysis assumptions

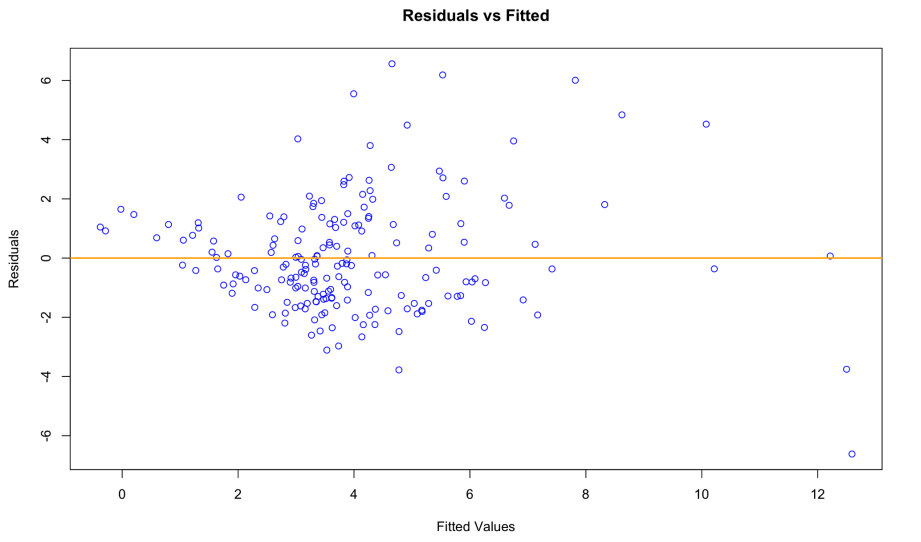


Figure 2: The fitted & residuals plot of the final model

Here, we see two things very clearly:

1) For any fitted value, the residuals seem roughly centered at 0. Hence, the linearity assumption is met.

2) We also see that for larger fitted values, the spread of the residuals is larger. Hence, the constant variance assumption or the normal distribution of errors assumption is violated.

We used the Breusch-Pagan test, to analytically test the assumption of constant variance of the residuals. Hypotheses:

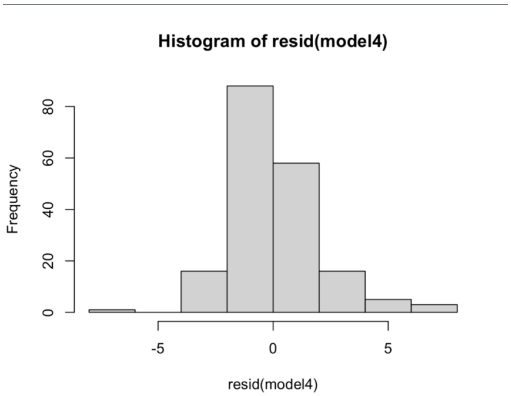
•H0: Homoscedasticity - The residuals have constant variance about the true model.

•H1: Heteroscedasticity - The residuals have non-constant variance about the true model.

The p-value of the Breusch-Pagan test statistic is 7.089e-05, which is below the significance level 0.05. Therefore, we reject the null hypothesis that the residuals have constant variance in favor of the alternative hypothesis that the residuals have nonconstant variance.We have the case of heteroscedasticity.

```
        studentized Breusch-Pagan test

data:  model4
BP = 43.513, df = 14, p-value = 7.089e-05
```



According to the histogram, we can see that the distribution of residuals is not normal.

```
        Shapiro-Wilk normality test

data:  resid(model4)
W = 0.95307, p-value = 7.512e-06
```

We used the Shapiro-Wilk test, to analytically test the assumption of normal distribution of residuals. Hypotheses:

H0: The distribution of residuals is normal

H1: The distribution of residuals is non-normal

Since the p-value for W-statistic is smaller than the significance level of 0.05, we reject the null hypothesis in favor of the alternative one.

Hence, we can conclude that the assumption of normality is violated.

To meet the assumptions of normality, we can introduce the logarithmic transformation into our model (Annex 1). According to the following investigation regarding positive changes in Breusch-Pagan and Shapiro-Wilk tests, there is no significant difference between model with quadratic effect and model with logarithm. The model fit has not improved. So, we proceed to the conclusion that our data was sampled from a NON-normal distribution.

4.8 Exploring the model quality

The following conclusions have been driven from the final model summary and can be found in Annex 2.

Since the p-value of the F-statistic is less than the significance level of 0.05, we can conclude that at least one of the predictors in our model has a non-zero relationship with daly.

Based on R-squared, we can say that modeling daly through a linear relationship with region, air pollution, lead, second-hand smoke and the interaction between regions and high temperature allows us to explain 52.03% of the variation in daly percentages in our sample.

Furthermore, explaining DALYs in Europe, Oceania, North America, South America and Africa through a linear relationship with region, air pollution, lead, second-hand smoke and the interaction between regions and high temperature results in the error of 1.989 percent, which is a relatively small number.

Therefore, we keep the model:

DALYs = 1.27 - 0.01*(Air_pollution)^2 + 0.13*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + 0.26*regionEurope - 3.38*regionOceania + 0.07*regionNorth America - 0.17*regionSouth America + 1.17*regionAfrica + 1.54*regionEurope:Air_pollution + 0.1*regionOceania:Air_pollution - 0.14*regionNorthAmerica:Air_pollution - 0.12*regionSouth America:Air_pollution - 0.05*regionAfrica:Air_pollution + e

4.9 Model interpretation

The fitted regression line for Europe:

DALYs = 1.53 - 0.01*(Air_pollution)^2 + 0.67*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + e

In Europe, there is 1.53 % of total DALYs, if all other predictors equal zero.

A one unit increase in the air pollution in Europe results in an estimated average increase by 0.67 percent in DALYs, holding other predictors constant.

The fitted regression line for Oceania:

DALYs = -2.11 - 0.01*(Air_pollution)^2 + 0.23*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + e

In Oceana, there is -2.11 % of total DALYs, if all other predictors equal zero.

A one unit increase in the air pollution in Oceania results in an estimated average increase by 0.23 percent in DALYs, holding other predictors constant.

The fitted regression line for North America:

DALYs = 1.34 - 0.01*(Air_pollution)^2 - 0.01*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + e

In North America, there is 1.34 % of total DALYs, if all other predictors equal zero.

A one unit increase in the air pollution in North America results in an estimated average decrease by 0.01 percent in DALYs, holding other predictors constant.

The fitted regression line for South America:

DALYs = 1.1 - 0.01*(Air_pollution)^2 + 0.01*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + e

In South America, there is 1.1 % of total DALYs, if all other predictors equal zero.

A one unit increase in the air pollution in South America results in an estimated average increase by 0.01 percent in DALYs, holding other predictors constant.

The fitted regression line for Africa:

DALYs = 2.44 - 0.01*(Air_pollution)^2 + 0.08*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + e

In Africa, there is 2.44 % of total DALYs, if all other predictors equal zero.

A one unit increase in the air pollution in Africa results in an estimated average increase by 0.08 percent in DALYs, holding other predictors constant.

The fitted regression line for Asia:

DALYs = 1.27 - 0.01*(Air_pollution)^2 + 0.13*Air_pollution + 0.56*Lead + 0.66*Second_hand_smoke + e

In Asia, there is 1.27 % of total DALYs, if all other predictors equal zero.

A one unit increase in the air pollution in Asis results in an estimated average increase by 0.13 percent in DALYs, holding other predictors constant.

For all regions:

A one unit increase in the lead results in an estimated average increase of 0.56 percent in DALYs, holding other predictors constant.

A one unit increase in the second hand smoke results in an estimated average increase of 0.66 percent in the daly, holding other predictors constant.

The negative coefficient for the squared term of air pollution (-0.01) suggests that there is a non-linear relationship between air pollution and the daily variable. As air pollution increases, its effect on the daily variable increases at a decreasing rate.

4.10 Standardized coefficients

```
Call:
lm(formula = daily ~ region:Air_pollution + I(Air_pollution^2) +
    Air_pollution + Lead + Second_hand_smoke + region, data = merged_with_dummies)

Standardized Coefficients::
                      (Intercept)               I(Air_pollution^2)                     Air_pollution
                               NA                     -0.511808773                      0.338010751
                             Lead                Second_hand_smoke                       regionEurope
                      0.164530811                     0.368729703                      0.038948296
                    regionOceania             regionNorth America             regionSouth America
                     -0.322835980                     0.008498573                     -0.015186627
                     regionAfrica       regionEurope:Air_pollution      regionOceania:Air_pollution
                      0.190936301                     0.365438550                      0.133174532
regionNorth America:Air_pollution regionSouth America:Air_pollution       regionAfrica:Air_pollution
                     -0.089714990                    -0.025060645                     -0.119365598
```

We compare coefficients of different explanatory variables: Air_pollution, Lead, Second_hand_smoke, region: Air_pollution, I(Air_pollution^2), region in a regression analysis, by computing standardized coefficients.

Lead: A one standard deviation increase in lead display is associated with a 0.1645 standard deviation increase in the DALYs holding other variables constant.

Air_pollution: A one standard deviation increase in air pollution is associated with a 0.338 standard deviation increase in the DALYs, holding other variables constant.

Second-hand smoking: A one standard deviation increase in second-hand smoke exposure is associated with an increase of 0.3687 standard deviations in the daily variable, holding other predictors constant. This indicates a stronger positive influence compared to lead.

I(Air_pollution^2): The negative coefficient for the squared term of air pollution (-0.5118) suggests that there is a non-linear relationship between air pollution and the daily variable. As air pollution increases, its effect on the daily variable increases at a decreasing rate, indicating a potential diminishing return or negative acceleration in its effect.

RegionEurope:Air_pollution: For one-unit increase in Air_pollution, the estimated average measurement DALYs increases by 0.36 % of total DALYs, more for Europe than for other regions, holding other predictors constant.

Region Oceania:Air_pollution: For one-unit increase in Air_pollution, the estimated average measurement DALYs increases by 0.13 % of total DALYs, more for Oceania than for other regions, holding other predictors constant.

Region North America:Air_pollution: For one-unit increase in Air_pollution, the estimated average measurement DALYs decreases by 0.089 % of total DALYs, more for North America than for other regions, holding other predictors constant.

Region South America:Air_pollution: For one-unit increase in Air_pollution, the estimated average measurement DALYs decreases by 0.025 % of total DALYs, more for South America than for other regions, holding other predictors constant.

Region Africa:Air_pollution: For one-unit increase in Air_pollution, the estimated average measurement DALYs decreases by 0.12 % of total DALYs, more for Africa than for other regions, holding other predictors constant.

4.11 Comparing results to the expectations

Hypothesis 1 (Regions): the hypothesis is partially confirmed. As we thought, the region with the highest DALY rate was Africa, however, the region with the lowest rate was Oceania, not Europe, as we had assumed.

Hypothesis 2 (Air pollution): the hypothesis is partially confirmed, the air pollution indicator has a negative effect on the dependent variable in some regions, America and Africa .

Hypothesis 3 (High temperature): the hypothesis is not confirmed, the indicator of high temperature will not have a positive effect on DALYs.

Hypothesis 4 (Lead): the hypothesis is confirmed, the indicator of lead exposure has a positive effect on DALYs.

Hypothesis 5 (unsafe water): the hypothesis is not confirmed, the indicator is not significant for DALYs.

Hypothesis 6 (unsafe sanitation): the hypothesis is not confirmed, the indicator is not significant for DALYs.

Hypothesis 7 (second-hand smoke): the hypothesis is confirmed, the indicator of second-hand smoke has a positive effect on DALYs.

5. Conclusions

Based on the analysis conducted in this study, it is possible to draw conclusions about the significance of the impact of environmental pollution factors, according to regions of the world, on the indicator of years lost due to disability or death. We identified the most influential factors as air pollution, second-hand smoke and lead exposure. Another important factor appeared to be geographic region, with Africa, as expected, being the region with the highest DALY rate and Oceania the region with the lowest DALY rate. Another important factor for the analysis was air pollution in different regions, the most polluted air was in Europe, and the cleanest in Africa. Thanks to our model, we were able to successfully achieve our goal of investigating the impact of air pollution, high temperature, lead, poor water quality, poor sanitation, and second-hand smoke on DALYs. Based on our findings, we recommend that the governments of the most affected regions pay attention to the policy of preventing mortality and disability in the countries of Oceania.

The environment in Ukraine has suffered significantly due to martial law, and the government's main goal is to overcome the consequences of full-scale invasion. After all, war crimes committed on the territory of Ukraine pose a much greater threat to the environment than the actions of civilians.

The next step in this study could be to analyze existing methods of combating health threats. In particular, the countries of Oceania, because, as already mentioned, they have the best indicators among all regions. An analysis of government policies and their effects would show which methods have the best impact on pollution and which should be followed by other countries. Such a study would allow for a better understanding of the factors that influence pollution and influence the development of strategies to combat them, which would lead to a reduction in DALYs in the future.

6. References

An introduction to statistical learning. (n.d.). An Introduction to Statistical Learning. https://www.statlearning.com/

Dataset-https://docs.google.com/spreadsheets/d/14oAM7tBi_qvW0HWhSNNG5iWr28HMBU12Wl8y62-T3CU/edit?usp=sharing

Dalpiaz, D. (n.d.). *Applied Statistics with R*. https://book.stat420.org/

Environmental Protection Agency. (n.d.). Search results. https://search.epa.ie/s/search.html?collection=epa-2021-search&query=DALYs

OECD. (n.d.). *Mortality, morbidity and welfare cost from exposure to environment-related risks*. © OECD. https://stats.oecd.org/Index.aspx?DataSetCode=EXP_MORSC#

WHO methods and data sources for global burden of disease estimates 2000-2019. World Health Organization. ghe2019_daly-methods.pdf (who.int)

7. Annexes

```
Call:
lm(formula = log(daily) ~ region:Air_pollution + I(Air_pollution^2) +
    Air_pollution + Lead + Second_hand_smoke + region, data = merged_with_dummies)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79688 -0.33289 -0.01701  0.37794  1.23445

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.390578   0.206070   1.895  0.05972 .
I(Air_pollution^2)               -0.003708   0.001175  -3.156  0.00189 **
Air_pollution                     0.058434   0.023853   2.450  0.01530 *
Lead                              0.116267   0.059685   1.948  0.05304 .
Second_hand_smoke                 0.202286   0.043982   4.599 8.19e-06 ***
regionEurope                     -0.038477   0.160116  -0.240  0.81038
regionOceania                    -1.154068   0.246274  -4.686 5.64e-06 ***
regionNorth America               0.108864   0.182406   0.597  0.55141
regionSouth America               0.129475   0.251421   0.515  0.60723
regionAfrica                      0.405330   0.178189   2.275  0.02416 *
regionEurope:Air_pollution        0.253528   0.077008   3.292  0.00121 **
regionOceania:Air_pollution       0.042093   0.022452   1.875  0.06251 .
regionNorth America:Air_pollution -0.041761   0.028703  -1.455  0.14751
regionSouth America:Air_pollution -0.054091   0.097819  -0.553  0.58100
regionAfrica:Air_pollution        -0.026649   0.018081  -1.474  0.14236
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5319 on 172 degrees of freedom
Multiple R-squared:  0.492,     Adjusted R-squared:  0.4507
F-statistic:  11.9 on 14 and 172 DF,  p-value: < 2.2e-16
```

Annex 1

```
Call:
lm(formula = daily ~ region:Air_pollution + I(Air_pollution^2) +
    Air_pollution + Lead + Second_hand_smoke + region, data = merged_with_dummies)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6192 -1.2950 -0.3041  1.1250  6.5613

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.269610   0.770596   1.648 0.101267
I(Air_pollution^2)               -0.009971   0.004393  -2.270 0.024465 *
Air_pollution                     0.130976   0.089198   1.468 0.143828
Lead                              0.555614   0.223191   2.489 0.013746 *
Second_hand_smoke                 0.662304   0.164472   4.027 8.47e-05 ***
regionEurope                      0.261583   0.598753   0.437 0.662747
regionOceania                    -3.378348   0.920939  -3.668 0.000325 ***
regionNorth America               0.071264   0.682107   0.104 0.916913
regionSouth America              -0.170672   0.940187  -0.182 0.856166
regionAfrica                      1.173625   0.666337   1.761 0.079963 .
regionEurope:Air_pollution        1.541744   0.287971   5.354 2.73e-07 ***
regionOceania:Air_pollution       0.104023   0.083957   1.239 0.217032
regionNorth America:Air_pollution -0.142126   0.107336  -1.324 0.187220
regionSouth America:Air_pollution -0.117519   0.365794  -0.321 0.748395
regionAfrica:Air_pollution        -0.051961   0.067615  -0.768 0.443257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.989 on 172 degrees of freedom
Multiple R-squared:  0.5203,    Adjusted R-squared:  0.4813
F-statistic: 13.33 on 14 and 172 DF,  p-value: < 2.2e-16
```

Annex 2

```
Descriptive statistics
=========================================================
Statistic          N  Mean  St. Dev.  Min   Max
---------------------------------------------------------
DALY             187 3.964  2.761   0.421 14.599
Air_pollution    201 5.595  7.220   0.000 27.600
High_temperature 201 0.731  1.021   0.000 6.960
Lead             201 1.114  0.836   0.080 4.170
Unsafe_water     201 3.842  5.612   0.010 31.870
Unsafe_sanitation 201 2.535  4.203   0.000 24.070
Second_hand_smoke 201 2.975  1.583   0.400 7.870
```

Annex 3