Pierre Mishra 12/13/2019
Lannette Rangel ENV 710

A Statistical Analysis of Variables Influencing Global Lung Cancer Incidence

I. Introduction

Lung cancer is the most commonly diagnosed cancer worldwide (Hart, 2016). According to the most recent global estimates, there were over two million new lung cancer cases diagnosed in 2018 (Globocan, 2019).

Overall, exposure to environmental air pollution leads to elevated levels of disease and death (Cohen, et al, 2017). Air pollution was recently proclaimed to be carcinogenic to humans by the International Agency for Research on Cancer (Hart, 2016). With recent increases in development, industrialization, and urbanization, people around the world are being exposed to increasing levels of air pollution. Over the past 25 years, low to middle income nations have experienced higher levels of air pollution as they continue to develop and move their societies into the 21st century. This development has contributed significantly to global disease rates of late. Recent studies have explored the spatial and temporal trends of how air pollution has contributed to death rates at country-wide, regional, and global scales (Cohen et al, 2017).

One of the most significant contributors to a nation's air pollution is coal-fired power plants. While in the past studies primarily looked only at particulate matter (PM 2.5) produced by these plants, recent research has begun to look beyond this metric. It has been found that lung cancer risk is greater within a nation if it relies more heavily on coal to produce its electricity (Lin et al, 2019). In addition to electricity generated from coal, indoor air pollution for domestic cooking and heating poses a public health risk, as nearly three billion people around the globe are exposed to it. Indoor air pollution from coal combustion was recently deemed a human carcinogen, and biomass combustion is likely carcinogenic to humans (Hosgood et al, 2010).

While air pollution is a contributing factor to global lung cancer rates, smoking tobacco remains the primary factor driving lung cancer cases (Islami et al, 2015). Changes in smoking habits in developed countries has led to an increase or leveling-off of lung cancer diagnoses in women, whereas male smoking and lung cancer incidence continues to decline. In developing nations, smoking habits and cancer rates vary greatly, and environmental exposure may be a greater lung cancer risk factor (Barta et al, 2019).

Since lung cancer is having such adverse public health impacts worldwide, work must be done to examine its drivers in tandem, in order to determine the most effective strategies for addressing it. Previous studies look at the major drivers, smoking and air pollution, separately. Therefore, we decided to examine these two factors within the same statistical model. We also wanted to account for national levels of socio-economic development. Thus, the following three statistical models explore seven potential explanatory variables driving lung cancer incidence globally. We choose variables associated with air pollution, environmental quality, personal habits, and socio-economic development. Specifically, our independent variables are:

1. *Human Development Index*: a composite measurement of key factors of human development, including a life expectancy index, education index, and Gross National Income (GNI) per capita (United Nations Development Program, n.d.)
2. *Smoking prevalence*: as a percentage of a nation's total population
3. *CO_2 emissions:* measured in the number of tons per capita
4. *Environmental Performance Index (EPI):* a score that measures a nation's environmental health and ecosystem vitality, based on 24 performance indicators, to see their level of proximity to established environmental policy goals (Yale, 2019).
5. *Coal Use:* The percentage of a nation's energy produced from coal
6. *Solid fuel use:* measured as the percentage of a nation's population that uses solid fuel. Solid fuel includes coal and biomass such as charcoal, wood, plant matter, or waste. It's used in household combustion to produce energy (WHO, 2004).
7. *Interaction effect*: between CO_2 emissions and EPI score (CO_2: EPI)

Through the creation of three Multiple Linear Regression models, we analyzed the aforementioned variables to answer our research question: if and to what extent do these different factors influence lung cancer rates?

Although we had initially included a direct measure of particulate matter, PM 2.5, as one of the explanatory variables for our models, there is a significant amount of temporal and spatial variability that exists for this variable. For instance, there may be much higher levels of it in the middle of an urban area, while insignificant levels are detected in a nation's countryside or forests. Thus, we ultimately decided to exclude it from our final models. Instead we introduced an interaction term of CO_2 per capita with EPI scores as a proxy for air pollutants. Since CO_2 is not an accurate measure of air pollution and other environmental contaminants, as countries with higher CO_2 emissions do have low air quality index (AQI) and cleaner environments, we suspected that for countries with higher EPI scores, CO_2 would have a lesser effect on lung cancer rates than for countries with low EPI scores. Therefore, the effect of CO_2 per capita on lung cancer rates would depend on a nation's EPI score.

Upon beginning this project, we suspected that countries with higher levels of coal and solid fuel use and their associated CO_2: EPI interaction would have higher lung cancer rates because of higher levels of air pollution and other environmental contaminants. We also anticipated that nations with higher levels of smoking prevalence would have elevated rates of lung cancer. Lastly, we hypothesized that nations with higher HDIs would be associated with lower lung cancer rates, as we assumed they would have improved access to healthcare.

Data Sources and Strategy

We sourced our data from a number of agencies and organizations. All of the datasets were publicly available. Lung cancer incidence rates, smoking prevalence and solid fuel use data was retrieved from *The Cancer Atlas*, which is produced by the American Cancer Society, the International Agency for Research on Cancer, and the Union for International Cancer Control. . Human Development Index (HDI) data was sourced from the RAND corporation as a part of its Food-Energy-Water Security Index. Carbon dioxide emission data was sourced from the World

Bank. Environmental Performance Index scores were retrieved from Yale University. Coal usage data was sourced from International Energy Agency.

The data was collected by performing internet searches using relevant terms such as “global coal use data” and “smoking prevalence data by country”. Once found, we downloaded the relevant datasets from their respective websites, examined the data in Excel. We then transferred it to R to clean, organize, and perform the appropriate statistical analyses. The data used herein does not represent random sampling, as for datasets at a global level, nations with similar geographic proximity tend to have similar residuals. Geography influences observations and therefore, global data is not considered random sampling.

We performed the entirety of our data analysis using R statistical software. Three ordinary least squares (OLS) multiple linear regression models were generated. We began by examining all of the countries in our datasets in our first model. We then proceeded to create two subsequent models that examined high and low income nations separately, to determine if factors and their associated influence is due to a country’s level of economic development.

II. Exploratory Data Analysis and Descriptive Statistics

The unit of our analysis was countries. After compiling and cleaning data, our final data frame comprised of 156 countries, out of which 152 were used for developing multiple linear regression (MLR) models. The exclusion of four countries from the final list was due to missing values for some of their independent variables that were included in our models.

Lung cancer rates ranged from 0.93 people to 118.80 people per 100,000 for 152 countries. We explored the distribution by constructing histograms and boxplots for logged and unlogged lung cancer rates. Unlogged cancer rates were positively skewed (Figure 1) while logged rates turned out to be negatively skewed (Figure 2). The boxplot for the logged lung cancer rates had the median more centrally located (Figure 4) between the first and third quartile compared to that of the unlogged lung cancer rates (Figure 3). We explored the correlations of the unlogged and logged response variable with several independent variables via scatterplot matrices. We discovered that logged lung cancer rates showed higher linear correlations with our explanatory variables than unlogged lung cancer rates. Thus, we decided to examine the explanatory variables with logged lung cancer rates. The final version of our scatterplot matrix can be found in the appendix (Figure 7). A summary of the descriptive statistics of the explanatory and response variables is provided on the following page (Table 1).

Variables	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard deviation
				Response variable			
Lung Cancer Rates (people per 100,000)	0.93	8.60	25.00	31.88	52.80	118.80	25.39
				Explanatory variables			
HDI (out of 1)	0.34	0.56	0.72	0.68	0.81	0.94	0.16
Smoking prevalence (% population)	1.70	17.95	30.30	30.46	40.58	72.80	15.19
CO2 (metric tons per capita)	0.05	0.80	2.63	4.36	5.90	34.16	5.25
EPI Score (out of 100)	27.43	45.34	57.06	56.43	64.83	87.42	13.21
Coal use (% electricity)	0.00	0.34	1.71	14.57	16.51	99.51	0.43
Solid fuel use (% population)	0.016	0.87	12.50	31.14	59.00	99.00	35.07

Table 1. Descriptive summary statistics for model variables

We log transformed HDI to improve its linear relationship with logged lung cancer rates. We also tried log transforming coal use and CO_2 per capita because their distribution was positively skewed (Figure 7). However, log transformation increased multicollinearity between the explanatory variables. This goes against one of the assumptions of linear regression models. Therefore, we proceeded with the unlogged coal use and CO_2 per capita data. Based on the scatterplot matrix (Figure 7), except for coal use and CO_2 per capita, all the explanatory variables had a somewhat strong linear relationship with logged lung cancer rates. See figure 7 for correlation coefficients. In order to test their significance, we performed statistical analyses, which we discuss in the following section.

III. Statistical Analysis

In order to predict log lung cancer rates based on several factors related to air quality, healthcare access, and personal habits, we developed three MLR models by using the ordinary least squares (OLS) method to estimate the beta coefficients of our explanatory variables. The first model incorporated all the countries in our dataset, while the second and third models were specific to low-income and high-income nations. Countries with a gross domestic product (GDP) per capita less than USD $15,000 were grouped as low-income countries while those with GDP per capita equal to or more than USD $15,000 were grouped as high-income countries. We used a threshold of USD $15,000 as it included the world's major emerging economies such as Chile, China and Russia. Moreover, some economists consider a range of USD $12,000 to $15,000 GDP per capita to be sufficient to qualify as a "developed" country (Investopedia, n.d). By subsetting countries based on GDP per capita, we attempted to capture the effects of our explanatory variables on lung cancer rates based on different social dynamics of countries to see if the significance of certain variables would change or not.

Hypotheses

Our hypotheses for all three of the models were the same. The null hypothesis states that there will be no significant prediction of log lung cancer rates by log HDI, smoking prevalence, coal use, indoor pollution and the interaction of CO_2 per capita with EPI score. This means that the coefficient parameters of all the explanatory variables will be zero. The alternative hypothesis states that there will be a significant prediction of log lung cancer rates by at least one of the explanatory variables. In other words, the beta coefficient of at least one of the explanatory variables would be a non-zero value. We set a significance level of 5% for the models. In notation form, our hypotheses are given below:

$$H_0 : \beta_{logHDI} = \beta_{coal} = \beta_{smoking} = \beta_{indoor} = \beta_{CO2:EPI} = 0$$

H_A: $\beta_j \neq 0$, for at least one j, where j = explanatory variables

OLS Models

All the models are summarized in Table 2. This section does not discuss some of the unusual findings, which require further explanation. Such findings are examined in detail in the discussion section.

Regression Model Results

	Dependent variable:		
	Log Lung Cancer Rates		
	All countries	High-Income Countries	Low-Income Countries
	(1)	(2)	(3)
Log HDI	2.420^{***}	1.620	2.505^{***}
	(0.550)	(1.137)	(0.527)
Smoking Prevalence	0.026^{***}	0.025^{***}	0.027^{***}
	(0.003)	(0.004)	(0.004)
CO_2	-0.151^{**}	-0.234^{**}	-0.109
	(0.068)	(0.107)	(0.220)
EPI Score	-0.002	-0.009	0.002
	(0.008)	(0.012)	(0.011)
Coal consumption	0.338^{***}	0.305^{**}	0.399^{**}
	(0.083)	(0.142)	(0.158)
Solid Fuel Use	-0.002	-0.018	-0.0003
	(0.003)	(0.018)	(0.003)
CO_2 : EPI Score	0.002^{**}	0.003^{**}	0.002
	(0.001)	(0.002)	(0.004)
Constant	3.393^{***}	3.805^{***}	3.112^{***}
	(0.594)	(0.943)	(0.727)
Observations	152	63	89
R^2	0.833	0.778	0.765
Adjusted R^2	0.825	0.750	0.745
Residual Std. Error	0.435 (df = 144)	0.319 (df = 55)	0.504 (df = 81)
F Statistic	102.557^{***} (df = 7; 144)	27.553^{***} (df = 7; 55)	37.692^{***} (df = 7; 81)
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 2. Summary of OLS multiple linear regression models

OLS Model 1: All Countries (n = 152)

The model including all the countries is formulated as:

$$\text{Log lung cancer rates} = 3.393 + (2.420*\text{Log HDI}) + (0.026*\text{Smoking}) – (0.151*CO_2) – (0.002*\text{EPI}) + (0.338*\text{Coal}) – (0.002*\text{Solid Fuels}) + (0.002*CO_2*\text{EPI})$$

According to our model, log HDI, smoking prevalence, coal consumption and the CO_2:EPI interaction had a significant effect in predicting log lung cancer rates. Log HDI, smoking prevalence and coal consumption for electricity use were positively associated with the log lung cancer rates. The F-statistic for the model is $F_{(7, 144)} = 102.55$ and p-value < 2.2e-16. Therefore, we rejected the null hypothesis in favor of the alternative hypothesis. The adjusted R^2 is 0.825.

Now we interpret the MLR equation of our model:

Holding all other variables constant, a 1% increase in HDI was associated with a 2.42% increase in lung cancer rates, at a significance level of 0.01 (t = 4.40, p = 2.06e-05). It was surprising to see that if we increase HDI, we would expect to see an increase in lung cancer rates.

Holding all other variables constant, a 1% increase in smoking tobacco was associated with a 2.6% increase in lung cancer rates, at a significance level of 0.01 (t = 8.87, p = 2.53e-15).

Holding all other variables constant, a 1% increase in coal consumption for electricity production was associated with a 33.8% increase in lung cancer rates, at a significance level of 0.01 (t = 4.07, p = 7.59e-05).

The interaction between CO_2 and EPI score had a significant effect on log lung cancer rates, at an alpha level of 0.05 (t = 2.09, p = 0.038). In order to interpret the interaction effect on predicting the response, we created a marginal effects plot (Figure 5). For a lower EPI score, an increase in CO_2 was associated with a decrease in log lung cancer rates. For a high EPI, an increase in CO2 did not affect log lung cancer rates.

Assumptions:

To test the linearity assumption, we created a residual vs fitted plot for the error terms (Figure 8). We found that the conditional mean of residuals almost fell in a straight line. In the QQ plot, we found that most of the quantiles of our error terms were near enough to the theoretical quantiles of the model derived from normal distribution, implying a normal distribution of error terms (Figure 9).

We performed the Breusch-Pagan test to examine the homoscedasticity assumption. We computed a p-value of less than 0.05 (Figure 10). Thus, we rejected the null hypothesis of the Breusch-Pagan test. Therefore, our error terms did not have a constant variance. We failed to meet the homoscedasticity assumption of OLS. Thus, we used robust standard errors to determine the significance of the explanatory variables. The model presented in Table 2 incorporates robust standard errors.

Next, we developed a Cook's distance plot and found several outliers (Figure 11). However, we decided not to remove outliers as countries vary greatly in terms of socio-political, economic and environmental conditions and removing outlier nations based on Cook's threshold may not result in a realistic model.

Lastly, we tested for multicollinearity by calculating variance inflation factors (VIF) for each term in our model (Figure 12). All of the VIF values for our variables were less than 10, except for CO_2 and the interaction effect between CO_2 and EPI. The high VIF for the interaction effect was intuitive because it was an interaction term. Since we are not examining the main effect of CO_2 individually, it is acceptable that CO_2 per capita had a high VIF.

OLS Model 2: High-Income Countries (n = 63)

$$\text{Log lung cancer rates} = 3.805 + (1.620*\text{Log HDI}) + (0.025*\text{Smoking}) – (0.234*CO_2) – (0.009*\text{EPI}) + (0.305*\text{Coal}) – (0.018*\text{Solid Fuels}) + (0.003*CO_2*\text{EPI})$$

According to our model for high-income nations, smoking prevalence, coal consumption and the CO_2:EPI interaction had a significant effect in predicting log lung cancer rates. Smoking prevalence and coal consumption were positively associated with the log lung cancer rates. The F-statistic for the model was $F_{(7, 55)} = 27.55$ and p-value < 7.955e-16. Therefore, we rejected the null hypothesis in favor of the alternative. The adjusted R^2 was 0.749.

Now we interpret the MLR equation of our model:

Holding all other variables constant, a 1% increase in smoking tobacco was associated with an increase of 2.5% in lung cancer rates, at a significance level of 0.01 (t = 7.923, p = 1.16e-10).

Holding all other variables constant, a 1% increase in coal consumption for electricity was associated with a 30.5% increase in lung cancer rates, at a significance level of 0.05 (t = 3.174, p = 0.0025).

The interaction between CO_2 and EPI score had a significant effect on log lung cancer rates, at an alpha level of 0.05 (t = 3.15, p = 0.0026). In order to interpret the interaction effects on predicting the response, we created a marginal effects plot (Figure 6). For a lower EPI score, an increase in CO_2 was associated with a decrease in log lung cancer rates. While for a high EPI, an increase in CO2 was associated with an increase in log lung cancer rates.

Assumptions:

To test the linearity assumption, we created a residual vs fitted plot for the error terms (Figure 14). We found that the conditional mean of residuals did not fall in a straight line. The linearity assumption was not met. In the QQ plot, we found that most of the quantiles of our error terms were near enough to the theoretical quantiles of the model derived from normal distribution. However, a few of the error terms were off, implying a weak normal distribution of error terms (Figure 15).

To test the homoscedasticity assumption, we performed a Breusch-Pagan test. We got a p-value of less than 0.05 which meant that we rejected the null hypothesis of Breusch-Pagan test. Therefore, our error terms did not have a constant variance (Figure 16). We failed to meet the homoscedasticity assumption of OLS. Thus, we used robust standard errors to determine significance of the explanatory variables. The model presented in Table 2 incorporates robust standard errors.

Next, we developed a Cook's distance plot and found several outliers as in the previous model (Figure 17). Lastly, we tested for multicollinearity by calculating variance inflation factors for each term in our model (Figure 18). We did not get a VIF of more than 4 for any term except the interaction term, which was intuitive and not problematic, and CO_2. Although CO_2 per capita had a high VIF, we were not looking at the main effect of CO_2 individually, so it was acceptable.

OLS Model 3: Low-Income Countries (n = 89)

Log lung cancer rates = 3.112 + (2.505*Log HDI) + (0.027*Smoking) – (0.109*CO_2) –
(0.002*EPI) + (0.399*Coal) – (0.0003*Solid Fuels) + (0.002*CO_2*EPI)

According to our model for low-income countries, smoking prevalence and coal consumption had a significant effect in predicting log lung cancer rates. Both variables were positively associated with the log lung cancer rates. The F-statistic for the model was $F_{(7, 81)}$ = 37.69 and p-value < 2.2e-16. Therefore, we rejected the null hypothesis in favor of alternative hypothesis. The adjusted R^2 was 0.744.

Now we interpret the MLR equation of our model:

Holding all other variables constant, a 1% increase in HDI was associated with a 2.50% increase in lung cancer rates, at a significance level of 0.01 (t = 4.75, p = 8.65e-06).

Holding all other variables constant, a 1% increase in smoking tobacco was associated with an increase of 2.7% in lung cancer rates, at a significance level of 0.01 (t = 8.87, p = 5.67e-09).

Holding all other variables constant, a 1% increase in coal consumption was associated with a 39.88% increase in lung cancer rates, at a significance level of 0.05 (t = 2.52, p = 0.0137).

Assumptions:

We created a residual vs fitted plot for the error terms to test the linearity assumption (Figure 20). We found that the conditional mean of residuals did not fall in a straight line. The linearity assumption was not met. In the QQ plot, we found that most of the quantiles of our error terms were near enough to the theoretical quantiles of the model derived from normal distribution. However, a few of the error terms were off, implying a weak normal distribution of error terms (Figure 21).

To test the homoscedasticity assumption, we performed a Breusch-Pagan test. We got a p-value greater than 0.05. Thus, we accepted the null hypothesis of the Breusch-Pagan test as our error terms had a constant variance (Figure 22). Next, we developed a Cook's distance plot and found several outliers as in the previous models (Figure 23). Lastly, we tested for multicollinearity by calculating variance inflation factors for each term in our model (Figure 24). We did not get a VIF of more than 5 for any term except the interaction term, which was intuitive and not problematic, and CO_2. CO_2 per capita had a high VIF but since we were not looking at the main effect of CO_2 individually, it was acceptable.

IV. Discussion

Smoking Prevalence

We found that smoking tobacco was the strongest predictor of lung cancer rates in all the three models. It had a strong positive association with lung cancer rates. Our result aligns with the popular cancer research where it is an accepted fact that smoking tobacco is the leading cause of lung cancer and a decrease in smoking is linked to a decrease in lung cancer rates (American Cancer Society, n.d.; Islami et al. 2015)

Coal Consumption

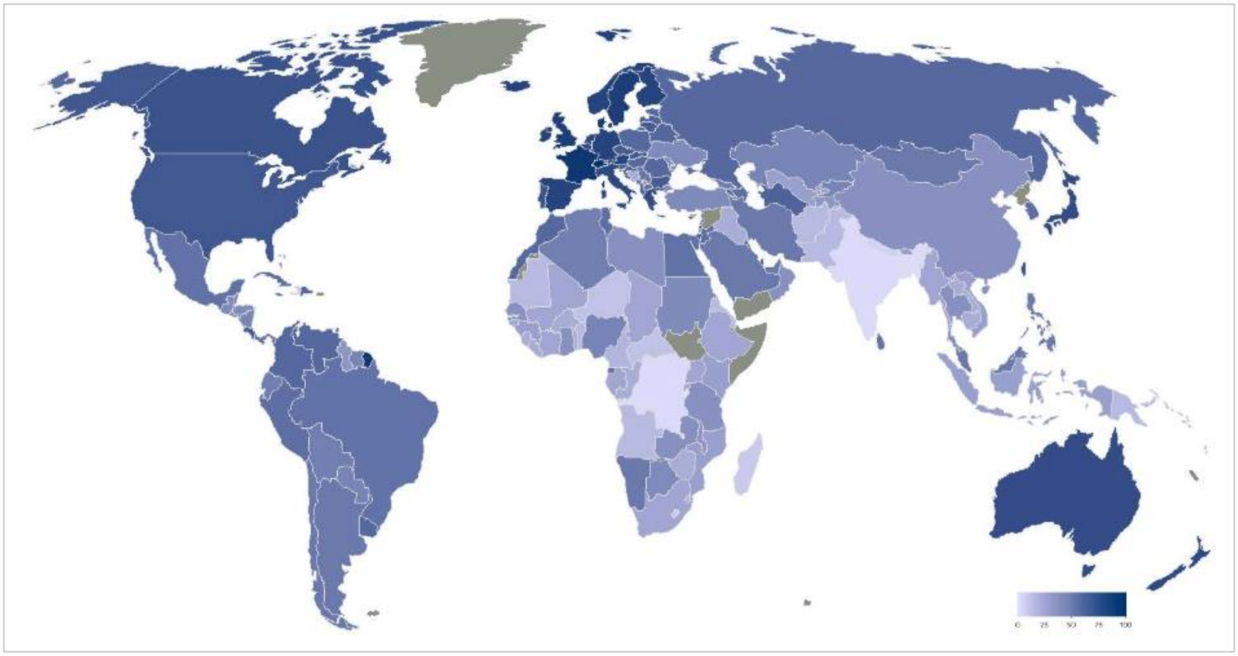
We found that for all three models, the percentage of electricity produced by coal had a strong positive association with lung cancer rates. Burning of coal produces a higher concentration of various air pollutants than other fossil fuels, which are proven to have a link with lung cancer in studies. A study done by Lin et al. (2019) also found that increasing reliance on coal for power generation increases the risk of lung cancer.

Human Development Index

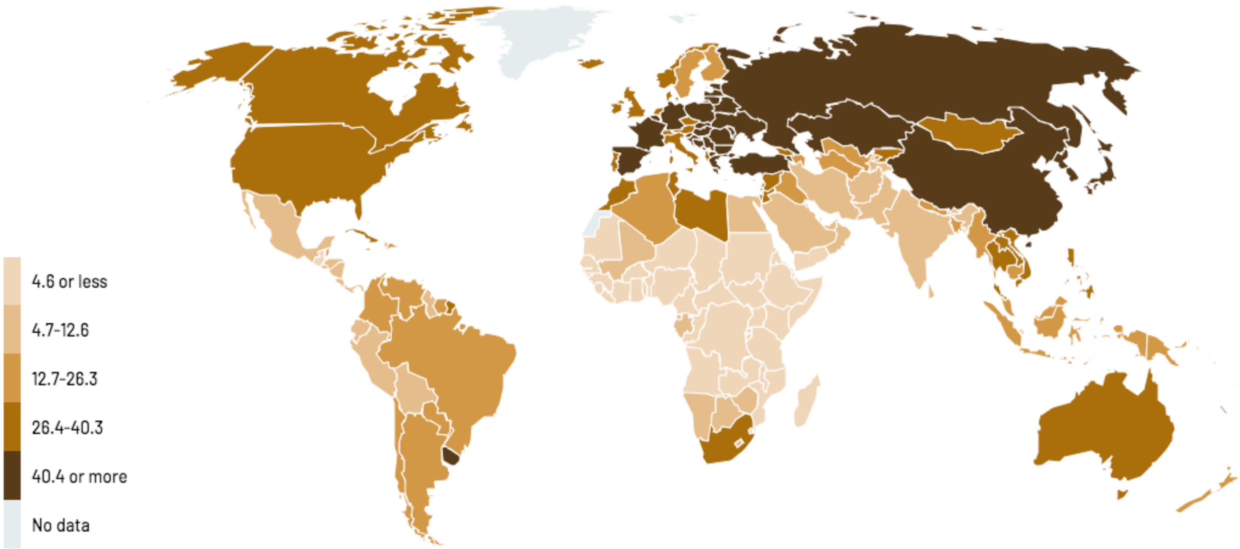
An increase in HDI was associated with an increase in lung cancer rates for the models incorporating all countries and low-income countries. For high income countries, HDI did not have a significant effect on lung cancer rates. In other words, for OLS Model 1 and 2, an increase in HDI was associated with an increase in lung cancer rates. While this might not seem intuitive, there could be several explanations for our findings. As we discussed before, HDI incorporates life expectancy, per capita income and education level. First, longer life expectancy could be positively related to lung cancer rates as there is a higher possibility of developing lung cancer for a person who lives 80 years compared to someone who lives for 50 years due to those additional years of life. In fact, studies have found that older populations are more susceptible to lung cancers (Rossi et al 2005; Venuta et al. 2016). Also, better healthcare access means that there is generally a better chance of diagnosis of lung cancer. So, in countries with low HDI, there might be underreporting of lung cancer rates simply because fewer people could access healthcare. If lung cancer is not even diagnosed in the first place, it goes unreported. In fact, a study investigating this issue also found under-reporting in Africa and several low-income economies (Sartorius & Sartorius 2016).

CO_2 : EPI Interaction

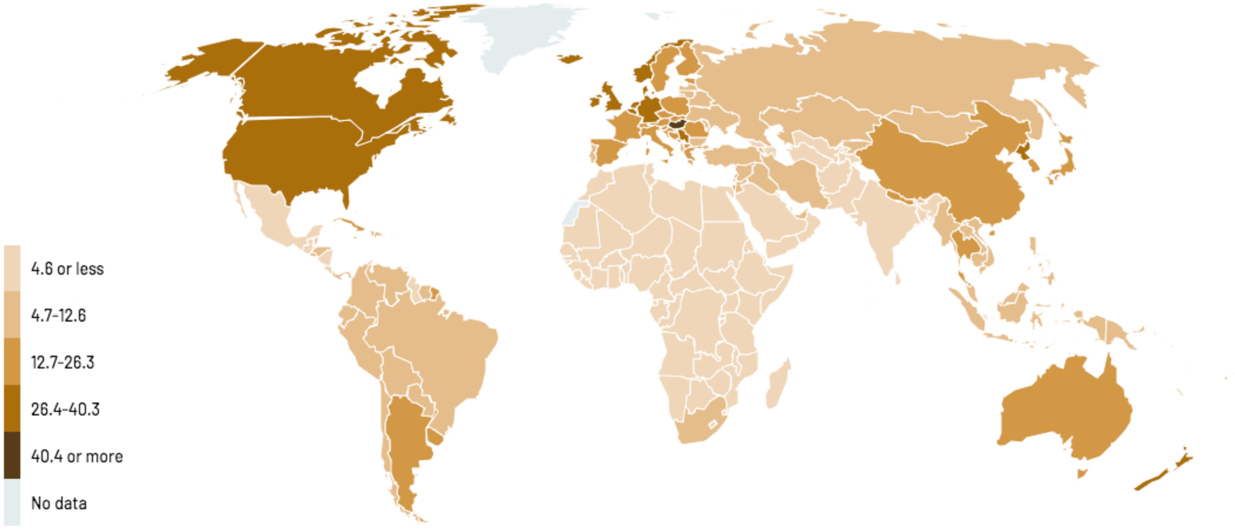
The interaction effect was significant for the models incorporating all countries and high-income countries. Our results for the interaction effect of CO_2 : EPI did not align with our expectations. For instance, for the OLS model 1, we saw that as EPI increased, increasing CO_2 decreased lung cancer rates which does not sound intuitive. However, if we were to assume underreporting of lung cancer rates in low-income nations, it is easier to make sense of the unusual finding. If we look at figure 5, we observe that for a high EPI, we no longer observe an unusual decrease in lung cancer rates for increasing $CO_{2.}$ The slope seems to be constant meaning that for countries with higher EPI, CO_2 per capita did not have any effect on lung cancer rates. However, for the subset of high-come countries, high EPI resulted in a positive effect of CO_2 on lung cancer (Figure 6). Generally speaking, if we look at the following maps of global EPI scores and lung cancer rates, we tend to see that countries with higher EPI also have higher lung cancer rates. These nations tend to be the more advanced economies of the world with higher CO_2 per capita. Underreporting in low-income and emerging economies due to lower HDI (lower access to health care) could be one of the possible explanations for such unusual findings.



Lighter shade translates to lower EPI while darker shades translate to higher EPI scores. Gray color signifies missing data (Source: Yale University)



Lighter shades translate to lower lung cancer rates while darker shades translate to higher lung cancer rates for males. Gray color signifies missing data (Source: Cancer Atlas; Ferlay et al., n.d.)



Lighter shades translate to lower lung cancer rates while darker shades translate to higher lung cancer rates for females. Gray color signifies missing data (Source: Cancer Atlas; Ferlay et al., n.d.)

Indoor pollution

The influence of indoor pollution in the form of solid fuel use on lung cancer rates was not as we expected. In all of our models, indoor pollution did not have any significance in predicting lung cancer rates. However, various studies have found that burning solid fuels could increase the risk of lung cancers due to the release of carcinogens (Hosgood et al. 2010). Future studies should look into different types of solid fuels instead of grouping them as one, as different sources could have different levels of indoor emissions.

V. Limitations

This study takes into account some, but not all of the potential drivers of lung cancer. Factors including family health history and lung cancer diagnoses; the exposure to harmful chemicals such as radon, arsenic, asbestos; the use of marijuana or e-cigarettes; and exposure to previous radiation treatments were omitted from this analysis. These omissions lead to omitted variable bias, creating biases in our coefficient estimates and potentially introducing endogeneity. Moreover, we failed to meet certain assumptions. For OLS Models 2 & 3, we did not meet the linearity assumption. While for all the models, our QQ plot of normality signified weak normal distribution of residuals. We also had several outliers in all of our models. Therefore, our equations are inherently suspect, and we can’t definitively declare that lung cancer incidence is being influenced by the analyzed variables alone. Performing additional analyses that take these variables into account, given data availability, would be a valuable next step in determining how to best prevent future lung cancer incidence.

Our scope of inference is the majority of the world’s nations. However, since our models analyze data at a global scale, they result in generalized results. Therefore, we wouldn’t be able to directly apply the results of our models at a national level to drive public policy that would result in the construction of fewer coal fire powered plants, or public awareness campaigns to decrease smoking prevalence. The models presented here would be useful when paired with other statistical models incorporating spatial and temporal analysis performed at a national or regional level. Focusing on a specific geographical location would also eliminate potential differences in data collection methodologies for different nations and reduce the effects of possible under reporting of lung cancer rates. Therefore, if we had additional resources and time, we would perform future analyses on one specific country and/or region to compare and contrast their results to our global analyses.

If we were to perform this study again, we would look into including panel data, given its availability, since it examines data over time. This is important when studying lung cancer incidence, as it tends to increase approximately 20-30 years after one’s exposure to elements, such as tobacco and air pollution. Individual smoking habits are another limitation to this study, and other studies that don’t use panel data, as the smoking habits of someone today may be drastically different from their smoking habits in a decade. By only capturing data from a short snapshot in time, we fail to account for changes in personal habits. On that same note, given the world’s ongoing transition towards renewable energy sources, a lack of panel data on national coal consumption may bias results.

VI. References

American Cancer Society. (n.d.). Risks of Tobacco. Retrieved from: https://canceratlas.cancer.org/risk-factors/risks-of-tobacco/

Barta, J.A., Powell, C.A. and Wisnivesky, J.P., (2019). Global Epidemiology of Lung Cancer. *Annals of Global Health*, 85(1), p.8. DOI: http://doi.org/10.5334/aogh.2419

Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K.,…Forouzanfar (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet, 389,* 1907-1918. Doi: http://dx.doi.org/10.1016/S0140-6736(17)30505-6

Ferlay J, Ervik M, Lam F, et al.. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Retrieved from: https://gco.iarc.fr/today.

Globocan. (2019). All Cancers. [Online factsheet]. *The Global Cancer Observatory*. Retrieved from: http://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf

Hart, J.E. (2016). *Air pollution affects lung cancer survival*. *Thorax,* 71 (10), 875-876. Doi: http://dx.doi.org/10.1136/thoraxjnl-2015-207927

Hosgood, H.D., Boffetta, P. Greenland, S., Lee, Y.C., McLaughlin, J., Seow, A.,…Lan, Q. (2010). In-Home Coal and Wood Use and Lung Cancer Risk: A Pooled Analysis of the International Lung Cancer Consortium. *Environmental Health Perspectives, 118* (12). https://doi.org/10.1289/ehp.1002217

Islami, F., Torre, L. A., & Jemal, A. (2015). Global trends of lung cancer mortality and smoking prevalence. Translational lung cancer research, 4(4), 327–338. doi:10.3978/j.issn.2218-6751.2015.08.04

Lin, C., Lin, R., Chen, T. *et al.* A global perspective on coal-fired power plants and burden of lung cancer. *Environ Health* 18, 9 (2019) doi:10.1186/s12940-019-0448-8

Rossi A, Maione P, Colantuoni G, Guerriero C, Ferrara C, Del Gaizo F, Nicolella D and Gridelli C. (2005). Treatment of small cell lung cancer in the elderly. The Oncologist 10: 399-411

Sartorius, B., & Sartorius, K. (2016). How much incident lung cancer was missed globally in 2012? An ecological country-level study. *Geospatial Health, 11* (2), 396 https://doi.org/10.4081/gh.2016.396

Tan, K.S. (2019). Misclassification of the actual causes of death and its impact on analysis: A case study in non-small cell lung cancer. Lung Cancer, 134, 16-24. Doi: 10.1016/j.lungcan.2019.05.016

United Nations Development Programme. (n.d.). Human Development Index. Retrieved from: http://hdr.undp.org/en/content/human-development-index-hdi

Venuta, F., Diso, D., Onorati, I., Anile, M., Mantovani, S., & Rendina, E. A. (2016). Lung cancer in elderly patients. *Journal of thoracic disease*, *8*(Suppl 11), S908–S914. Doi:10.21037/jtd.2016.05.20

World Health Organization (WHO). (2004). Indoor smoke from solid fuels: Assessing the environmental burden of disease. Retrieved from: https://www.who.int/quantifying_ehimpacts/publications/9241591358/en/

Yale University. (2019). Environmental Performance Index. Retrieved from: https://epi.envirocenter.yale.edu/

VII. Appendix

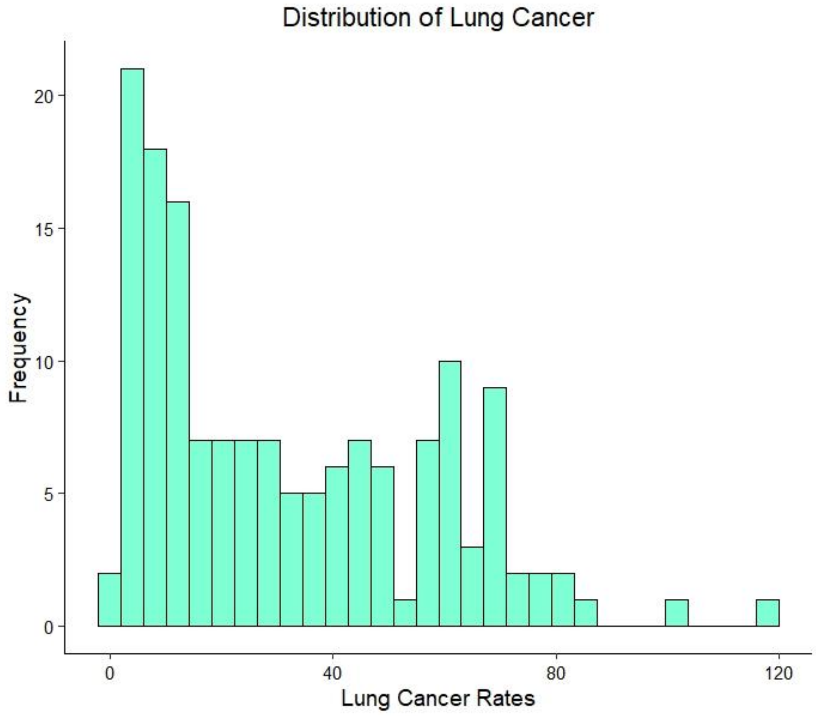


Figure 1. Distribution of lung cancer rates (n=152) using a histogram

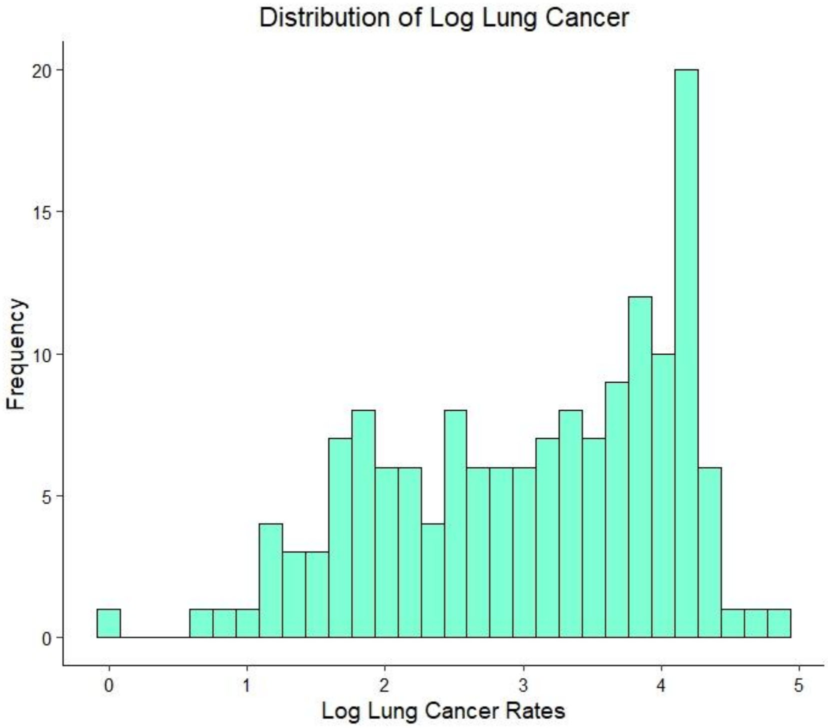


Figure 2. Distribution of log lung cancer rates (n=152) using a histogram

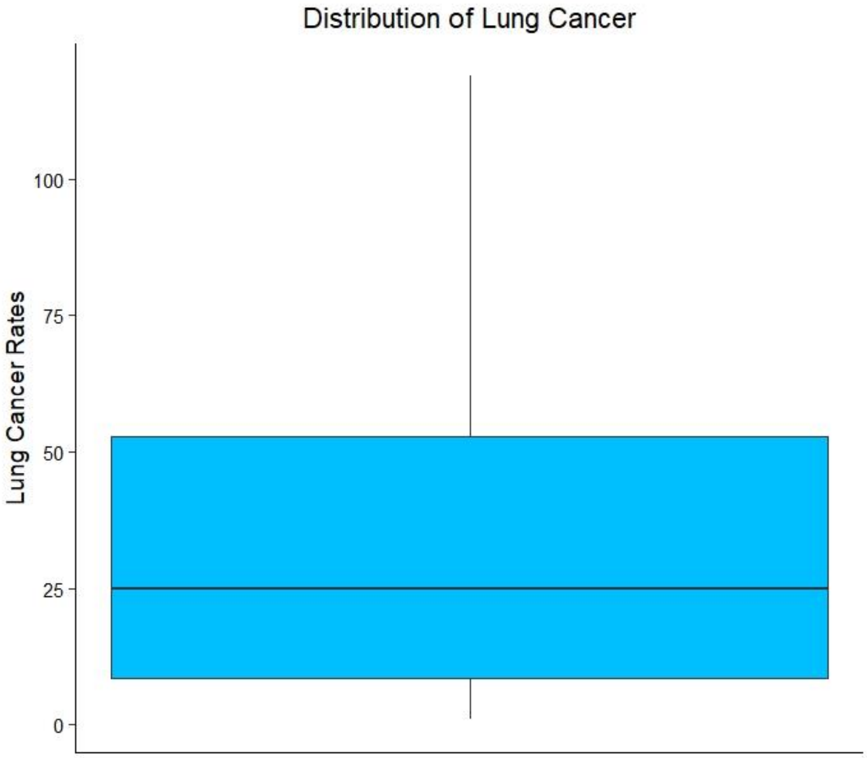


Figure 3. Distribution of lung cancer rates (n=152) using a boxplot

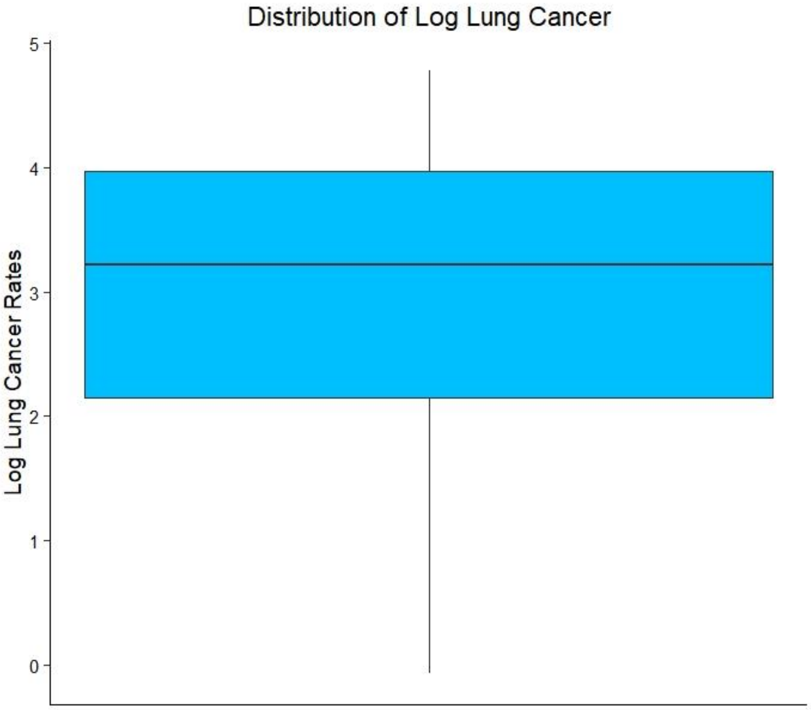


Figure 4. Distribution of log lung cancer rates (n=152) using a boxplot

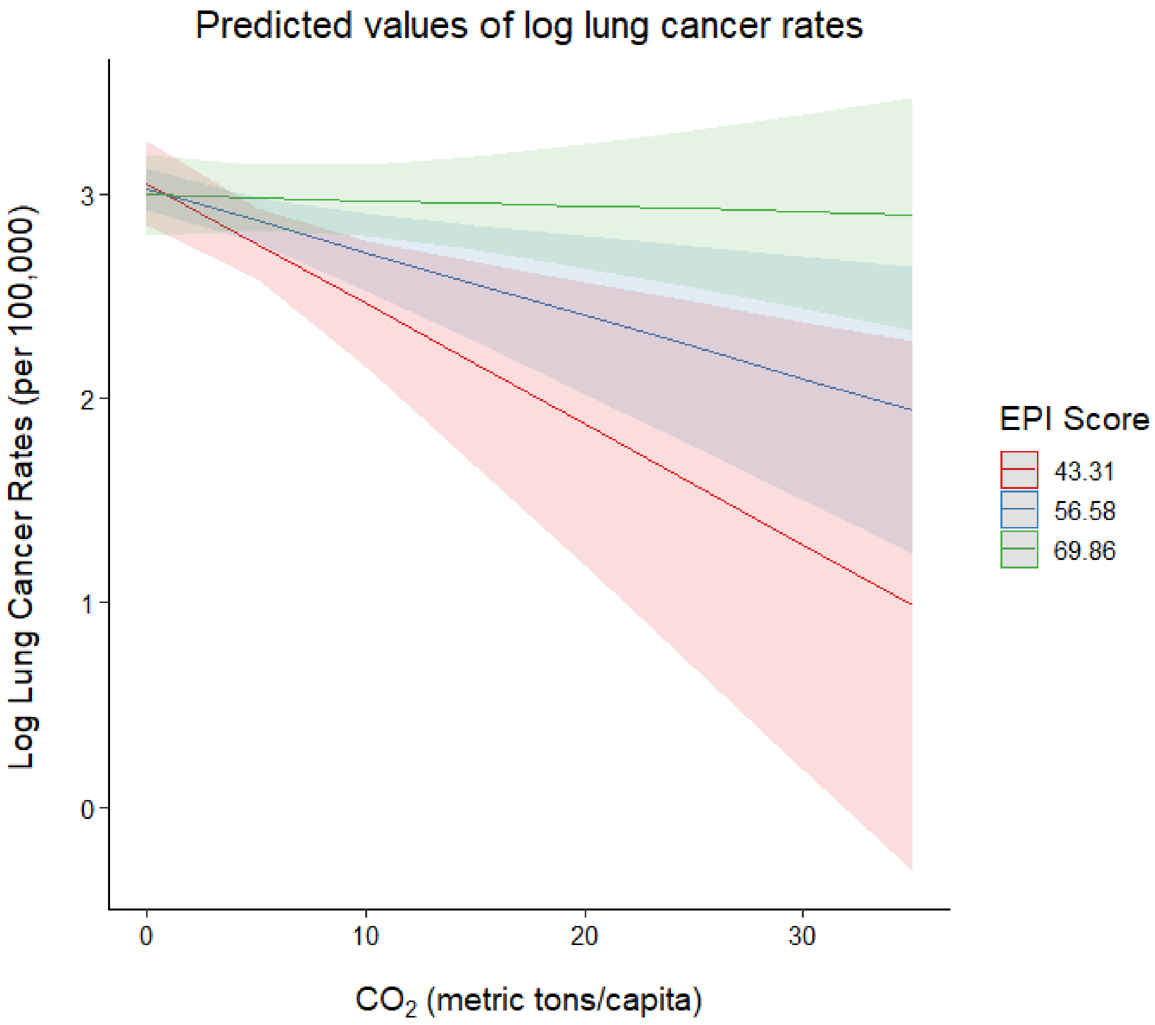


Figure 5. Marginal effects plot to interpret the effect of interaction CO$_2$:EPI Score on log lung cancer rates for all countries

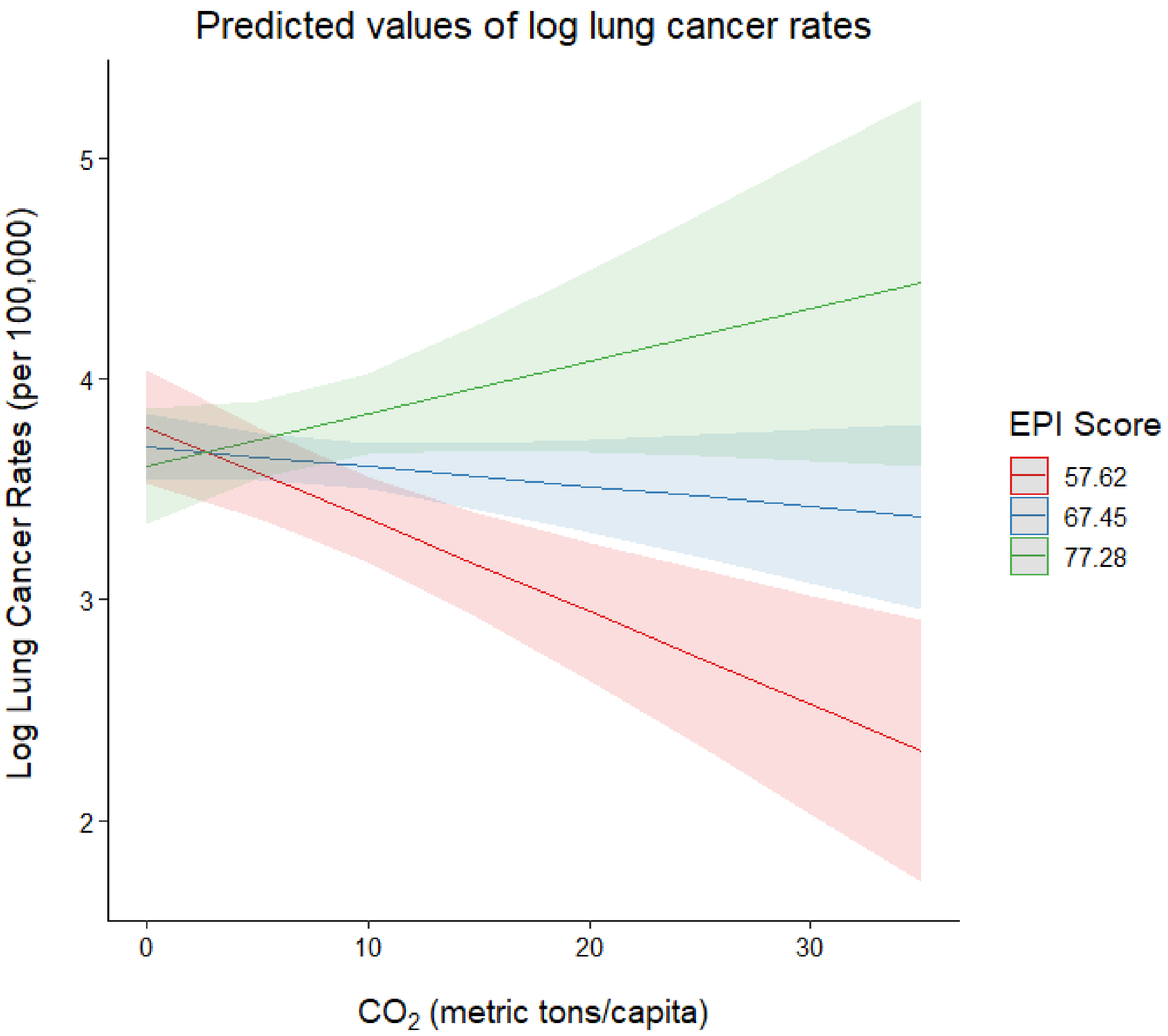


Figure 6. Marginal effects plot to interpret the effect of interaction CO_2:EPI Score on log lung cancer rates for high-income countries

OLS Model 1

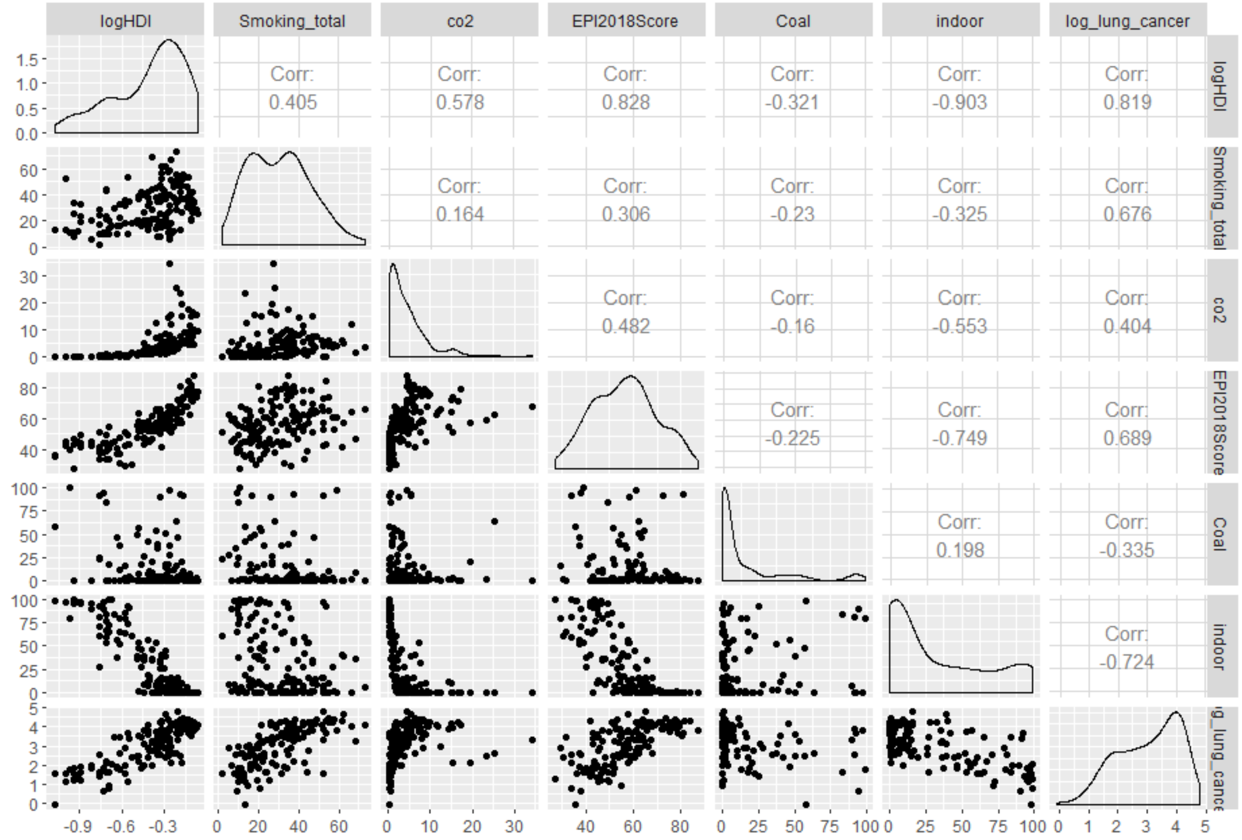


Figure 7. Scatterplot matrix of correlation between response and explanatory variables and among explanatory variables for OLS Model 1 (all countries)

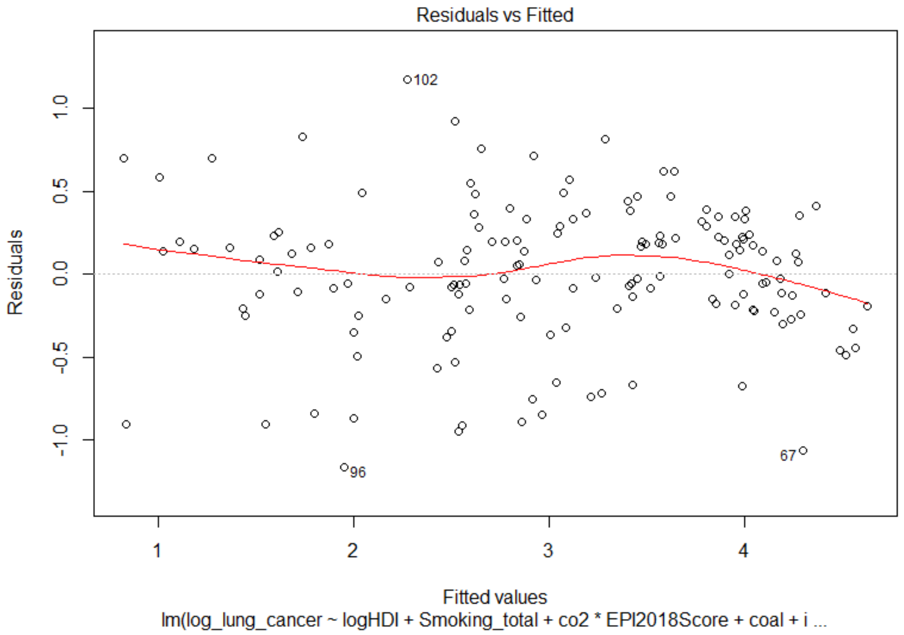


Figure 8. Residual Vs Fitted plot for OLS Model 1 (all countries)

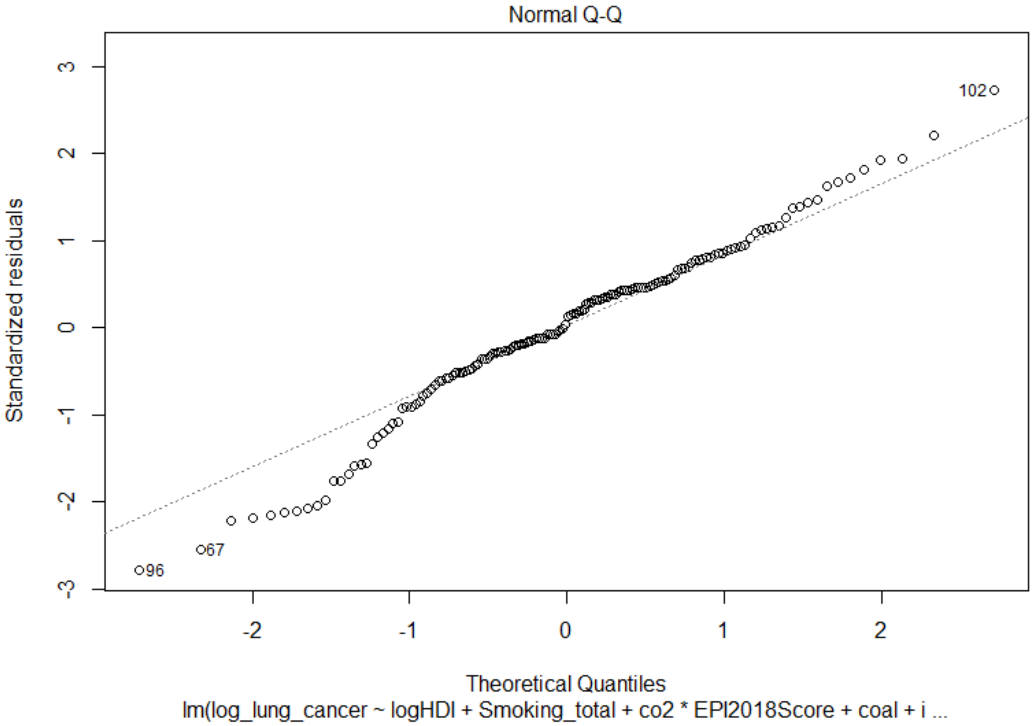
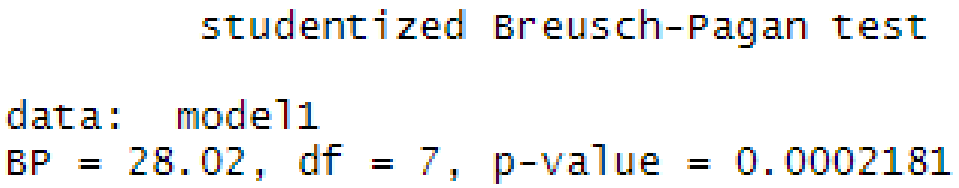


Figure 9. QQ plot to test the normality of error terms for OLS Model 1 (all countries)



```
        studentized Breusch-Pagan test

data:  model1
BP = 28.02, df = 7, p-value = 0.0002181
```

Figure 10. Results of Breusch-Pagan test of homoscedasticity for OLS Model 1 (all countries)

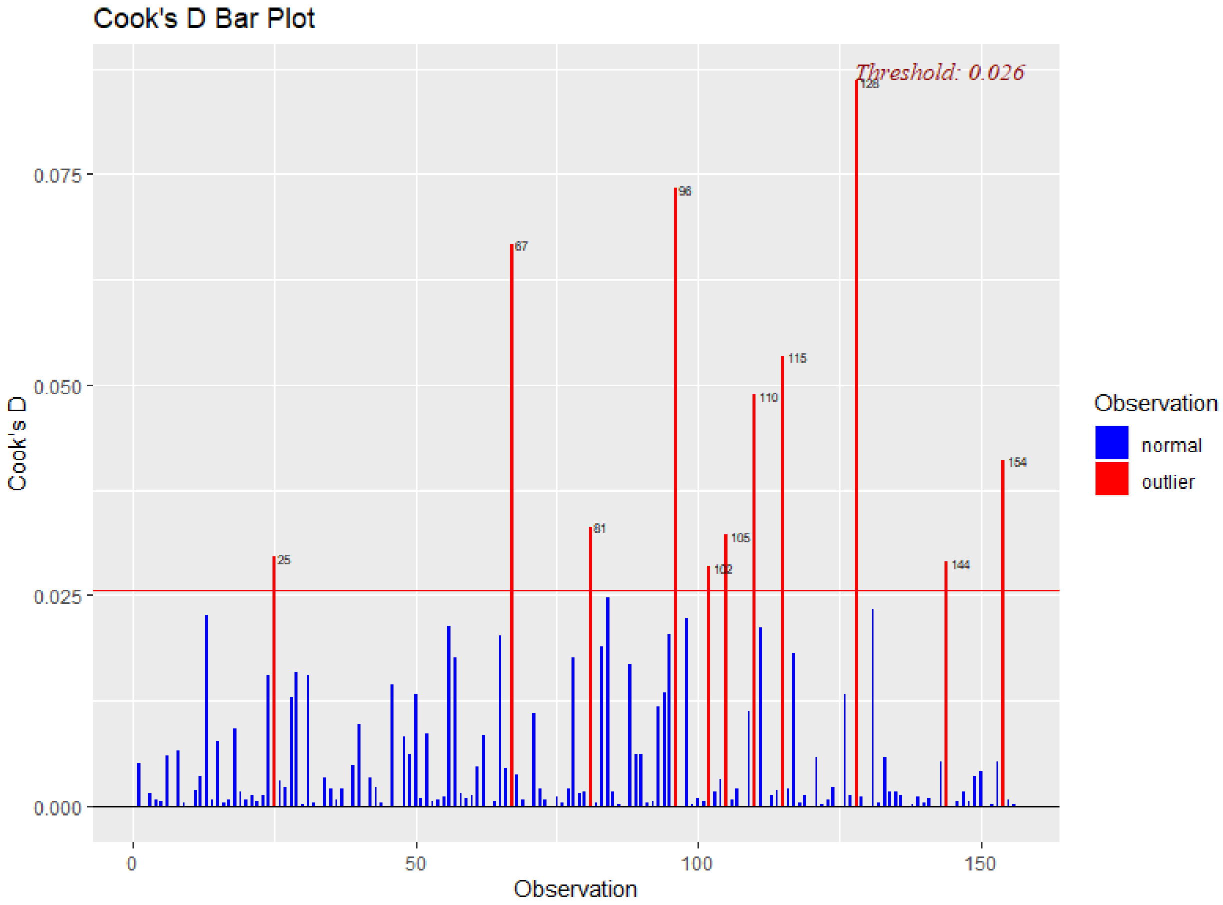
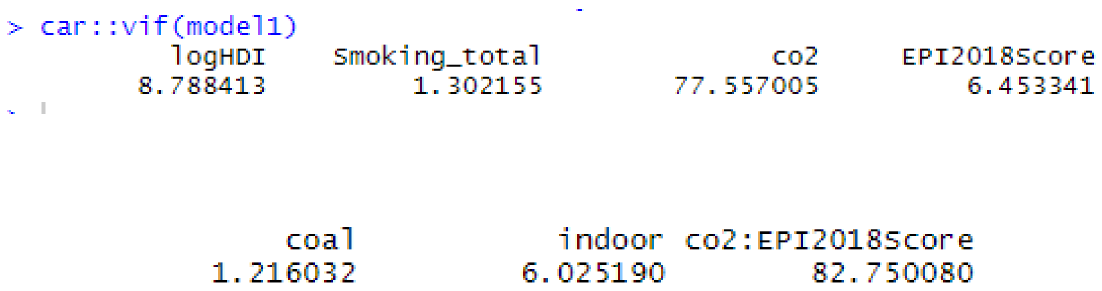


Figure 11. Cook’s Distance plot to determine outlier countries in the OLS Model 1 (all countries)



```
> car::vif(model1)
          logHDI    Smoking_total            co2     EPI2018Score
        8.788413         1.302155      77.557005         6.453341


           coal           indoor  co2:EPI2018Score
       1.216032         6.025190          82.750080
```

Figure 12. Variance Inflation Factors for each of the explanatory variables in OLS model 1 (all countries)

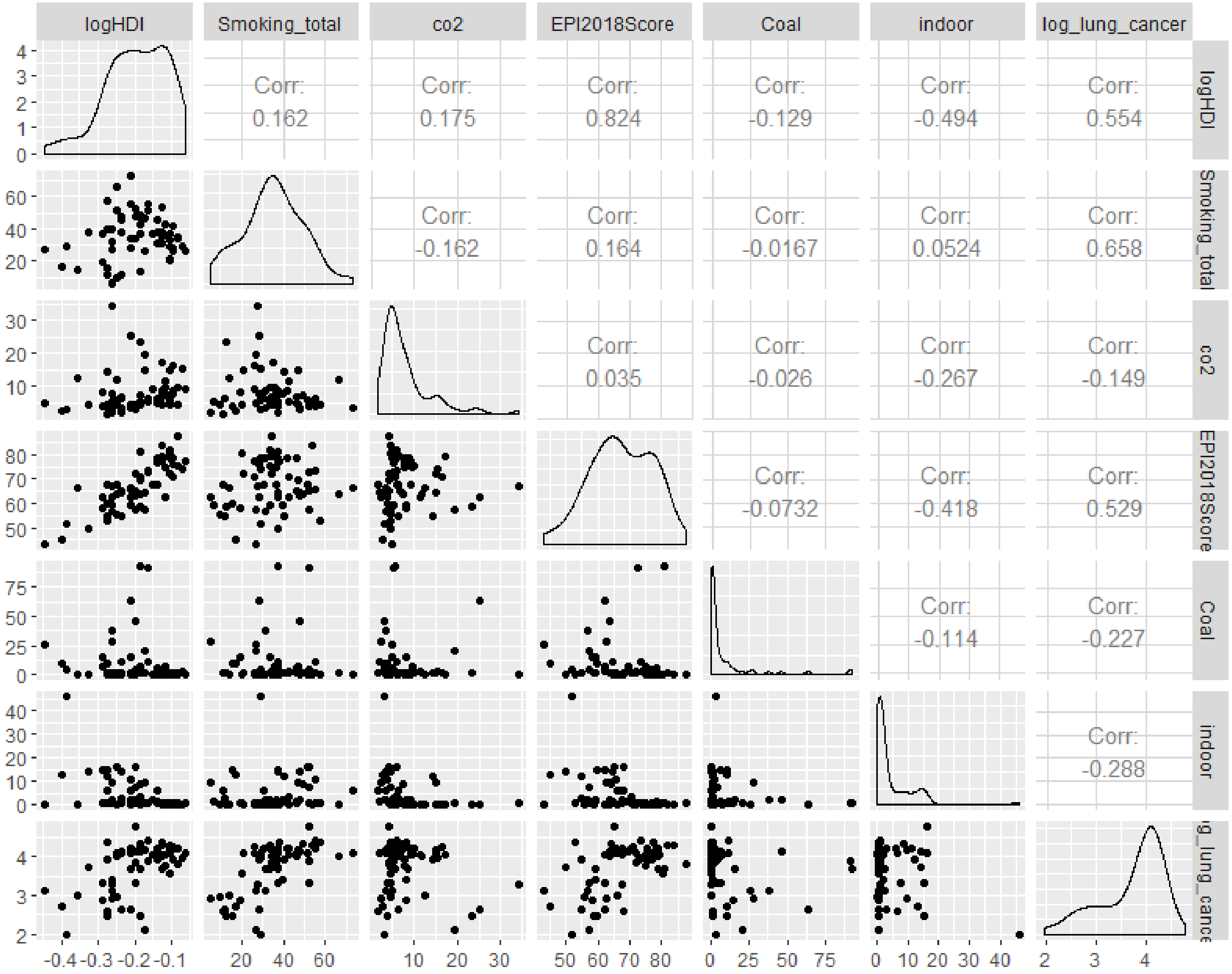


Figure 13. Scatterplot matrix of correlation between response and explanatory variables and among explanatory variables for OLS Model 2 (high-income countries)

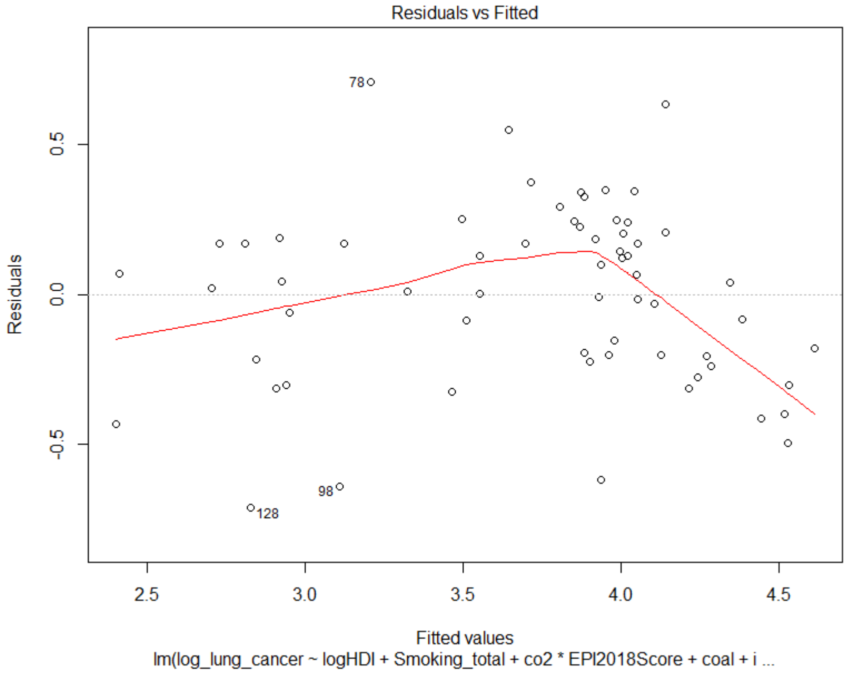


Figure 14. Residual vs fitted plot for OLS Model 2 (high-income countries)

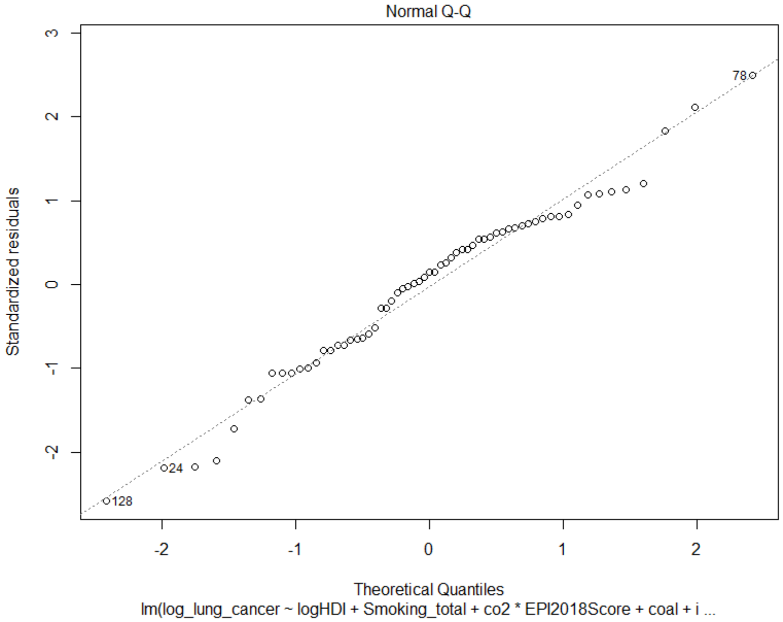
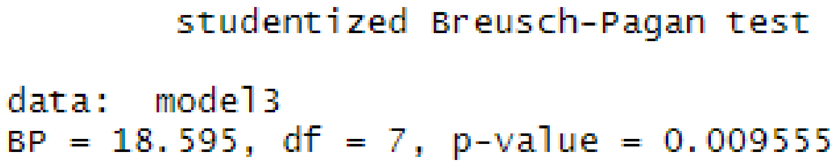


Figure 15. QQ plot to test the normality of error terms for OLS Model 2 (high-income countries)



```
        studentized Breusch-Pagan test

data:  model3
BP = 18.595, df = 7, p-value = 0.009555
```

Figure 16. Results of Breusch-Pagan test of homoscedasticity for OLS Model 2 (high income countries)

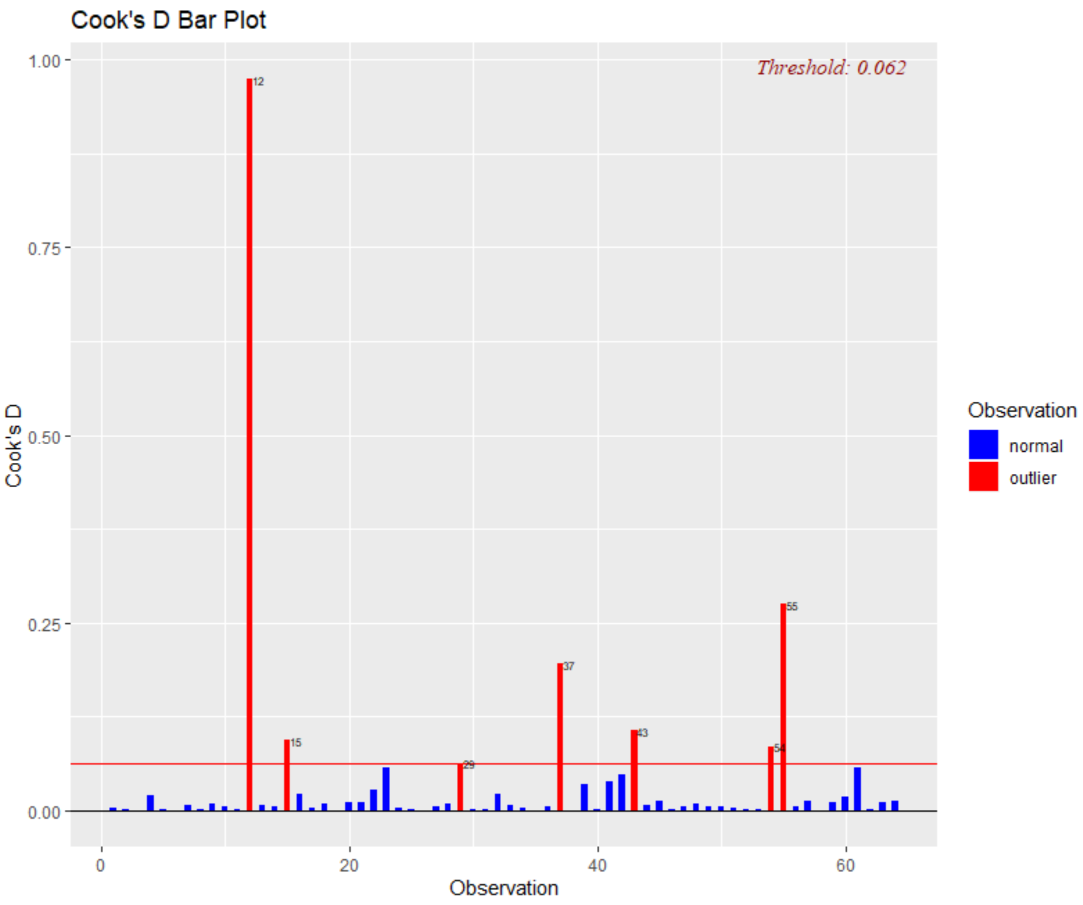
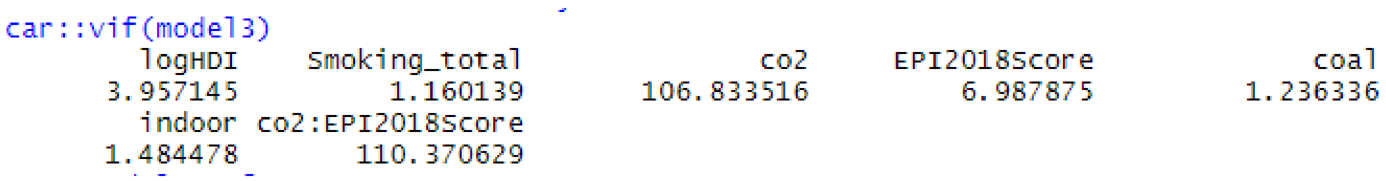


Figure 17. Cook's Distance plot to determine outlier countries in the OLS Model 2 (high income countries)



```
car::vif(model3)
       logHDI    Smoking_total              co2       EPI2018Score              coal
     3.957145         1.160139       106.833516           6.987875          1.236336
       indoor  co2:EPI2018Score
     1.484478       110.370629
```

Figure 18. Variance Inflation Factors for each of the explanatory variables in OLS model 2 (high income countries)

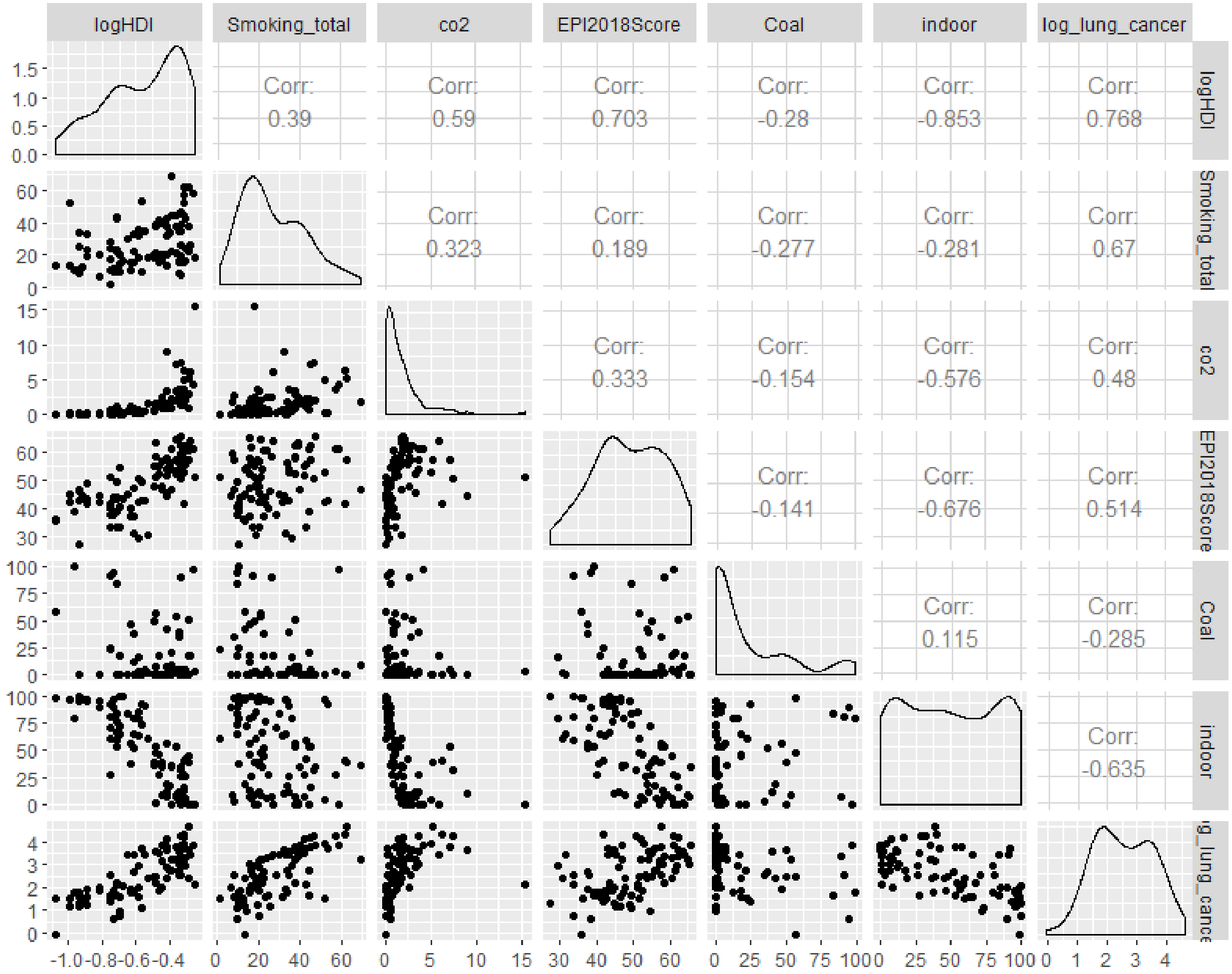


Figure 19. Scatterplot matrix of correlation between response and explanatory variables and among explanatory variables for OLS Model 3 (low-income countries)

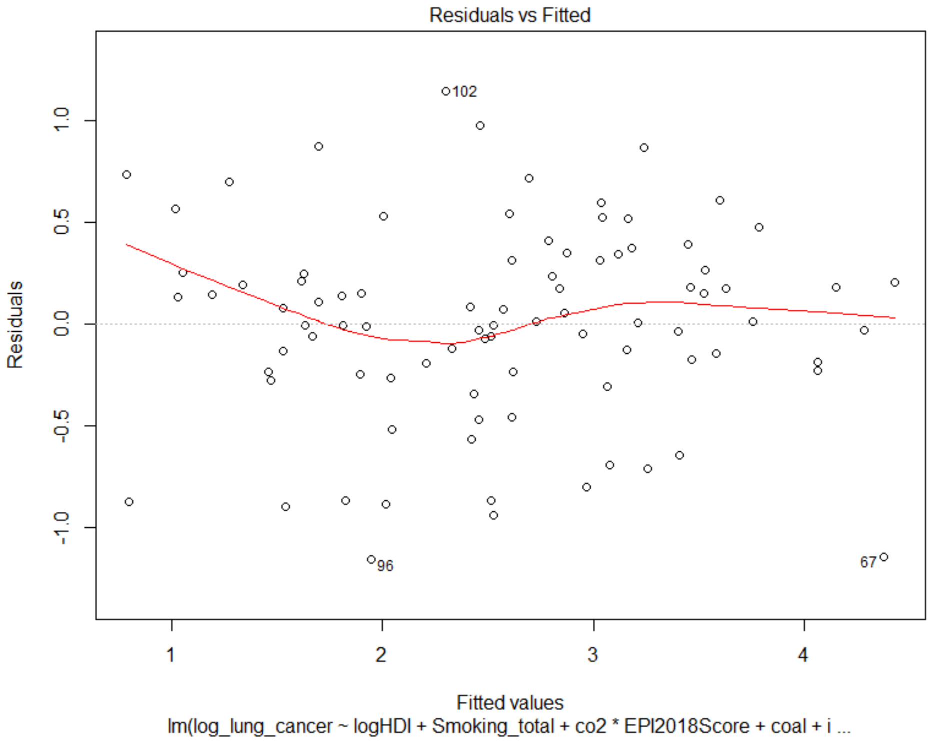


Figure 20. Residual Vs Fitted plot for OLS Model 3 (low-income countries)

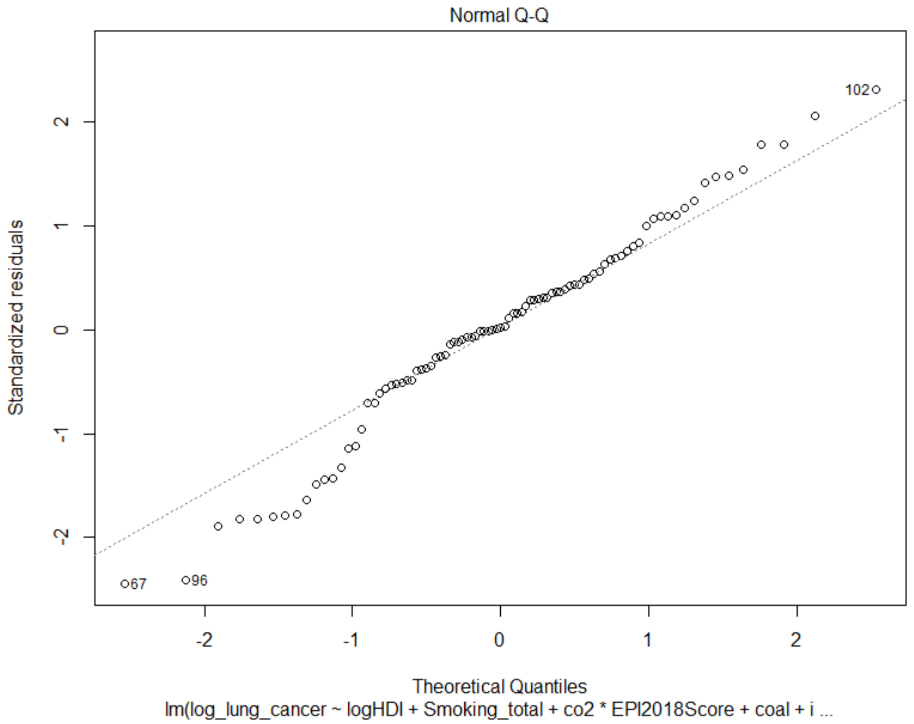
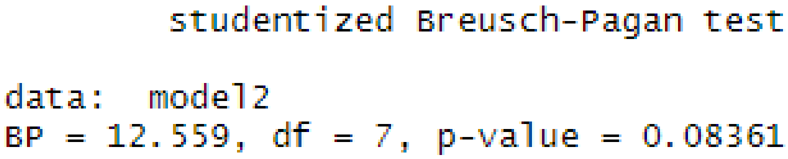


Figure 21. QQ plot to test the normality of error terms for OLS Model 3 (low-income countries)



```
          studentized Breusch-Pagan test

data:  model2
BP = 12.559, df = 7, p-value = 0.08361
```

Figure 22. Results of Breusch-Pagan test of homoscedasticity for OLS Model 3 (low-income countries)

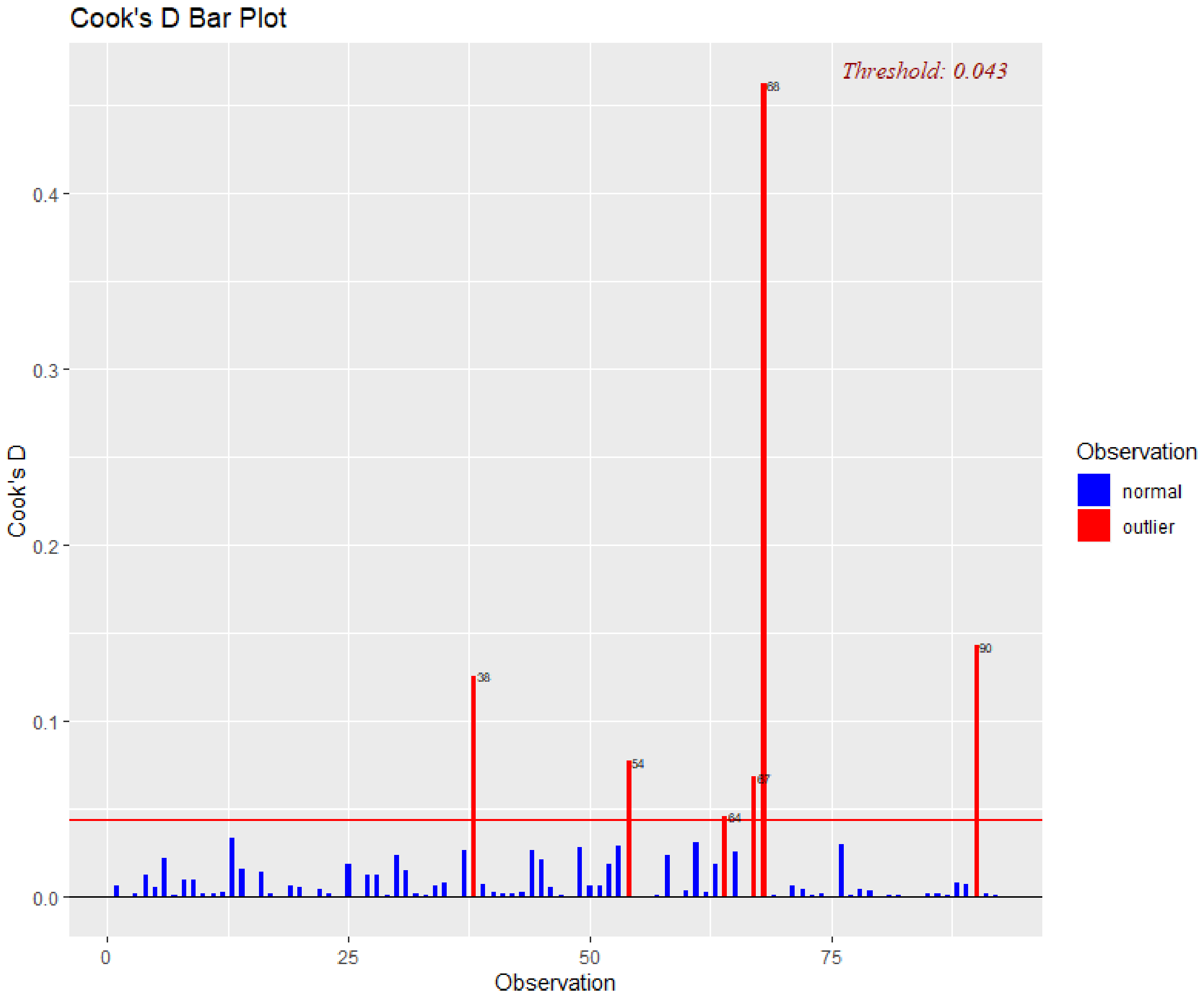
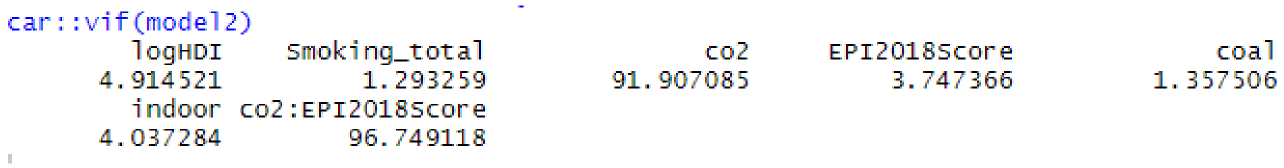


Figure 23. Cook's Distance plot to determine outlier countries in the OLS Model 3 (low-income countries)



```
car::vif(model2)
      logHDI    Smoking_total              co2      EPI2018Score             coal
    4.914521         1.293259        91.907085          3.747366         1.357506
      indoor co2:EPI2018Score
    4.037284        96.749118
```

Figure 24. Variance Inflation Factors for each of the explanatory variables in OLS model 3 (low-income countries)