What You Sea Is What You Get: The Relationship Between Population Health and Environmental Change

Jeffrey Chang, David Liang, Kimberley Yu, Phillip Yu

December 9, 2016

1 Introduction and Research Question

1.1 Background

The effects of numerous environmental factors on human health have been well-documented. For example, PM 2.5, a fine particulate matter that can be inhaled and subsequently cause health problems, has been shown to be associated with mortality rates even at low concentrations. [**?**] Other common measures of pollution, such as CO2 emissions, have also been shown to have an association with air pollution-related mortality rates. [**?**] We are interested in seeing the relationship between population health and environmental change on a per-country basis.

It is unclear whether any observed associations will continue to hold up when adjusting for economic prosperity. A country's economic status has been shown to have a complicated relationship with its environmental health, and there is no scientific consensus regarding the exact model for this relationship. The environmental Kuznets curve (EKC), a once-popular model associating income with pollution, proposes that an increase in wealth results in greater environmental pollution for less-developed countries but less environmental pollution for more-developed countries. However, this simplified model has lost ground among a growing number of scientists, due to the complexity of the issue and abundance of other income and environmental health factors. [**?**]

Similarly, a complicated relationship exists between a country's economic prosperity and the health of its population. Studies have shown a somewhat positive correlation between a country's wealth and health for less-developed and developing countries, although this correlation weakens significantly for developed countries. [**?**] Notably, other economic factors like income inequality likely play important roles in this association, and the specific relationship between health and wealth is still unclear. [**?**]

1.2 Research Question and Hypothesis

In our project, we seek to answer two important research questions: First, is a country's population health associated with environmental change? Second, will this association hold when we condition on per-capita GDP?

In general, we hypothesize that the data will show that as environmental change gets worse, so too will population health. More specifically, since our predictors are measures of how *un*healthy the environment is, we hypothesize a negative correlation between our predictors and life expectancy: as pollution (PM2.5 and CO2) increases, life expectancy decreases, and as the number of threatened species (proportional to land area) increases, life expectancy decreases. After conditioning on per-capita GDP, we hypothesize that environmental change will still be negatively correlated with population health. However, we hypothesize that this relationship will not be as prominent, since a number of confounding variables related to economic prosperity will have been accounted for in the per-capita GDP.

We conducted a variety of different analyses in order to determine the relationship between environmental and population health, and the effect of GDP on this relationship. We determined specific facets of our overarching research question to investigate, and the specific questions we hoped to answer and their motivations are as follows:

- What is the relationship between life expectancy and GDP?

  - This was the most general analysis we conducted. Because the relationship between GDP and life expectancy is well-studied and significant, we wanted to be able to control for GDP's effects in future tests. From this analysis, we were able to use the base relationship solely between these two factors to find any additional or opposite relationships between life expectancy and other predictor variables.
  - We hypothesized this relationship would be positive.

- What is the relationship between life expectancy and pollution? What happens when you control for GDP?
  - Pollution, specifically air pollution, was the first proxy of environmental health we analyzed. Pollution seemed like the a general way to effectively quantify environmental change, as all living organisms are affected in some way by pollution.
  - Furthermore, according to the World Health Organization, air pollution is "a major environmental risk to health," with outdoor air pollution estimated to have caused 3 million premature deaths worldwide in 2012. [**?**]
  - We hypothesized the relationship between pollution and life expectancy would be negative.

- What is the relationship between life expectancy and biodiversity? What happens when you control for GDP?
  - Biodiversity was the second proxy of environmental health we analyzed, which we measured with the land density of threatened species. We believed biodiversity would be a good way to quantify environmental change, as biodiversity reflects the environment's ability to support and provide for a variety of inhabitants.
  - Biodiversity and human health are also closely related, with biodiversity affecting the water, food, and fuel used in order to sustain human life.
  - We hypothesized that the relationship between the threatened species and life expectancy would be negative.

2 Methods

2.1 Dataset Overview

We are using the World Bank's World Development Indicators dataset. [**?**] The dataset contains over a thousand indicators of economic development from hundreds of countries. Within this dataset, we are particularly interested in indicators in the subfields of Climate Change, Economy, Environment, and Health.

The following table contains all of our variables of interest.

Variable Name	Description	Units
SP.DYN.LE00.IN	Life expectancy at birth, total	years
NY.GDP.PCAP.CD	GDP per capita	current USD
EN.ATM.PM25.MC.M3	PM2.5 air pollution, mean annual exposure	micrograms per cubic meter
EN.ATM.CO2E.KT	CO2 emissions	kilotonnes (kt)
SP.POP.TOTL	Population, total	#
EN.BIR.THRD.NO	Bird species, threatened	#
EN.FSH.THRD.NO	Fish species, threatened	#
EN.HPT.THRD.NO	Plant (higher) species, threatened	#
EN.MAM.THRD.NO	Mammal species, threatened	#
AG.LND.TOTL.K2	Land area	sq. km

In our analysis, we use life expectancy at birth as our proxy for population health, as life expectancy provides a quantifiable and comparable statistic for the health of each country. For environmental change, we use two major categories of proxies - pollution and biodiversity. The remaining variables are either categorized into one of these types or used to normalize variables.

2.2 Data Processing

In its original form, the data set comprised over 5.6 million rows, with each statistic for each country and variable on a separate row. Thus, it required some pre-processing in order to filter out extraneous data and restructure the data into a more usable form for statistical analysis. To accomplish this, a Python script using the pandas library was written (included in Appendix) to create our final data set. After processing the entire original data set, the final data set was structured with the statistics of interest for each country on a single row, making it significantly easier to analyze.

Additionally, we created new variables that we believed would be more interesting and applicable predictors of life expectancy. One big consideration was normalizing our data. We calculated CO2 per capita because we believed it was a better reflection of a country's pollution contributions relative to its number of people. We also aggregated our species threatened data into one overall metric by adding the number of threatened Bird, Fish, Plant, and Mammal species and dividing by the country's land area, since we believed that a country's species diversity should be calculated based off the amount of land they have to hold species. This is of course not a perfect metric, since land varies widely in type, from desert to forest to tundra.

A summary of our added variables:

Variable Name	Description	Units
threat.prop	$1000 \cdot \frac{\text{Bird + Fish + Plant (higher) + Mammal species, threatened}}{\text{Land area}}$	# per sq. km
CO2_KT_percapita	CO2 emissions / Population, total	kilotonnes / # of people

3 Results and Analysis

3.1 Simple and Multiple Linear Regression

3.1.1 GDP

We believed that linear modeling is a good fit for our data because the relationships are generally linear or easily transformable to be linear. First, we studied life expectancy at birth as a function of GDP per capita. After checking assumptions, we found that taking the log of the GDP improved satisfaction of our assumptions. After running simple linear regression, we found that log(GDP per capita) is a significant positive predictor of life expectancy with a regression coefficient of 4.71 and a p-value of $< 2e-16$ (Figure 1). Additionally, the adjusted R-squared was 0.657.

3.1.2 Air Pollution

Note: For this section, we used the log of every variable except life expectancy.

We next began studying how various environmental factors relate to life expectancy, and how GDP might play a role in those relationships. To address our questions on the relationship between life expectancy and pollution, we used simple linear regression to model how each of GDP per capita, PM2.5 air pollution, and CO2 emissions predict life expectancy at birth. We found that PM2.5 air pollution significantly negatively correlates with life expectancy (coefficient = -2.15, p-value = 0.03, adjusted R^2 = 0.016) (Figure 2a).

We found that CO2 emissions significantly positively correlates with life expectancy (coefficient = 1.06, p-value = $5.8e-09$, adjusted R^2 = 0.14). When adjusting for total population, we found that CO2 (per capita) was an even better predictor with a stronger correlation, larger coefficient, and smaller p-value (coefficient = 3.96, p-value = $< 2e-16$, adjusted R^2 = 0.57) (Figure 2b).

Finally for simple linear regression, we decided to see whether PM2.5 predicts GDP and whether CO2 emissions predicts GDP. We found a significant negative correlation with PM2.5 (coefficient = -0.49, p-value = 0.0045, adjusted R^2 = 0.032) (Appendix) and a significant positive correlation with CO2 emissions per capita (coefficient = 0.802, p-value = $< 2e-16$, adjusted R^2 = 0.798) (Appendix), which was as expected. Seeing this was useful in allowing us to confirm that GDP could affect correlations between pollution and life expectancy. As a result, we decided to conduct multiple linear regression.

Using multiple linear regression, we wanted to see whether PM2.5 concentration and CO2 emissions still predict life expectancy after adding GDP as another predictor. For Life Expectancy $\sim$ log(GDP) + log(PM2.5), we found that PM2.5 concentration was no longer a significant predictor, with a p-value

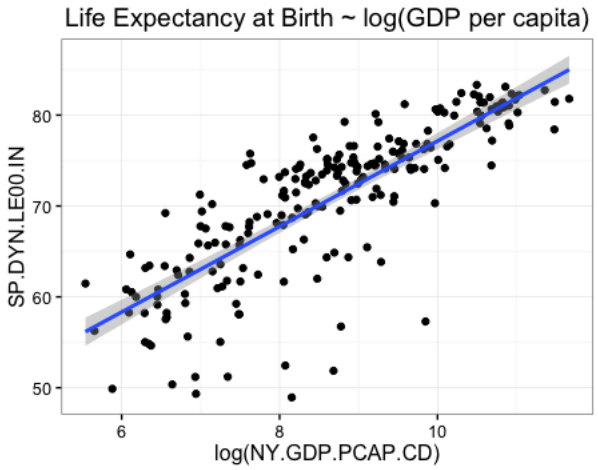
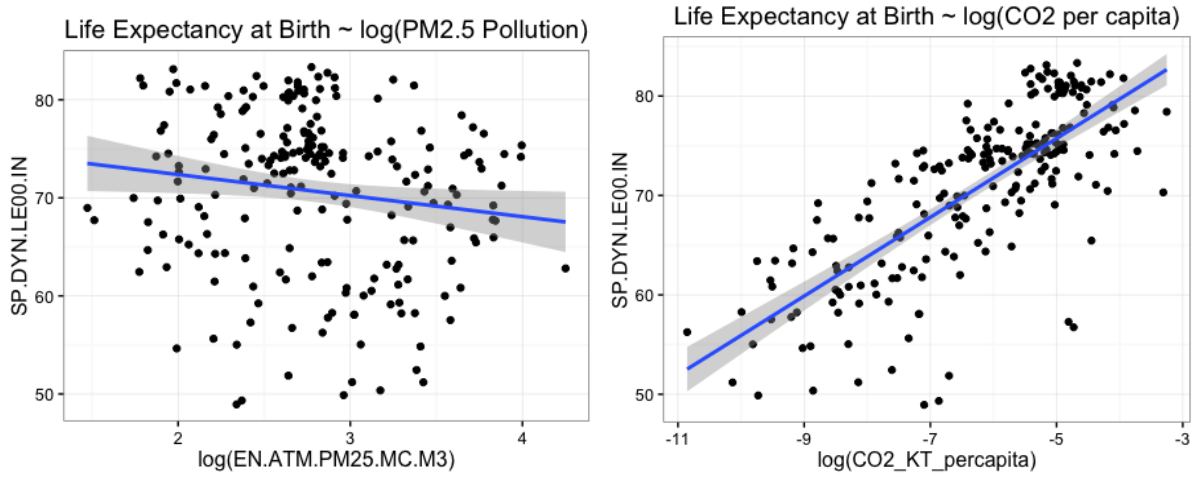


Figure 1: log(GDP per capita) significantly positively predicts life expectancy at birth.



(a) (b)

Figure 2: Life expectancy is negatively predicted by log(PM2.5 air pollution) while positively predicted by CO2 emissions per capita.

of 0.775 (Appendix). For Life Expectancy $\sim$ log(GDP) + log(CO2/total population), we see that CO2 emissions also loses significance with a p-value of 0.552 (Appendix).

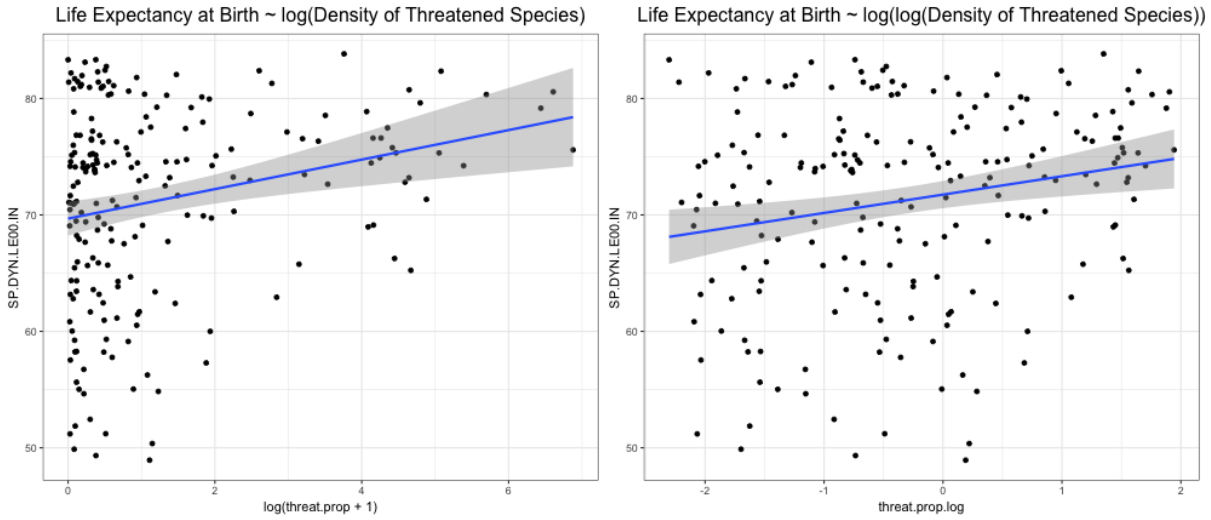
We next tested for interaction effects between PM2.5 and GDP as well as between CO2 and GDP. With PM2.5, we found there was not a significant interaction between PM2.5 concentrations and GDP. GDP continued to be the only significant predictor (p-value = 0.00013) of life expectancy, and neither the main effect of PM2.5 nor the interaction effect were significant (p-values 0.80 and 0.81, respectively). On the other hand, we see that CO2 again becomes a significant positive predictor (coefficient = 3.09, p-value = 0.02) along with GDP (coefficient = 2.18, p-value 0.04) after the interaction effect is added. The interaction effect is not significant (but almost, with a p-value of 0.0741.)

3.1.3 Biodiversity

We used the density of threatened species as our metric of biodiversity. A greater density of threatened species should indicate less biodiversity. For the density predictor variable, `treat.prop`, we added the number of threatened bird, fish, plant, and mammal species in a country, and divided by the area of the country, and scaled by 1000. We noticed that the density was skewed right. We had to use the logarithm transformation twice to get a normal-looking distribution (We added by a constant to account for non-positive values). For the first transformation, we did `log(threat.prop+1)`,

and for the second transformation we defined a new attribute for convenience `threat.prop.log = log(log(threat.prop+1)+0.1)`. We will refer to these as the log(density) and log(log(density)) variables, respectively.

We ran a linear regression of life expectancy against the density, as well as log(density) and log(log(density)). For all the linear regressions, we get a significant value for the β_1 the slope. These slope values are also all positive. However, the value of the slope for all three models is small relative to the spread of the residuals. In fact, the adjusted R^2 value never exceeds 0.06. (Figure 3)



(a) (b)

Figure 3: Linear Regression of Life Expectancy vs. log(density) (a) and vs. log(log(density)) (b)

We also ran multiple regressions of the two transformations of life expectancy vs. log(density) and the per capita GDP and also log(per capita GDP). We did the same for log(log(density)). When we ran against the per capita GDP, the slopes for both predictor variables were both significant, with the p-value for GDP significantly lower than the p-value for log(density) and log(log(density)) in each model. The R^2 value also rose to about 0.36. When we run against the log(GDP) instead of GDP, the R^2 increases even more to 0.65. Meanwhile the slope for log(density) is no longer significant, and the p-value for the slope for log(log(density)) is 0.044, so almost not significant.

3.1.4 Assumptions of all models

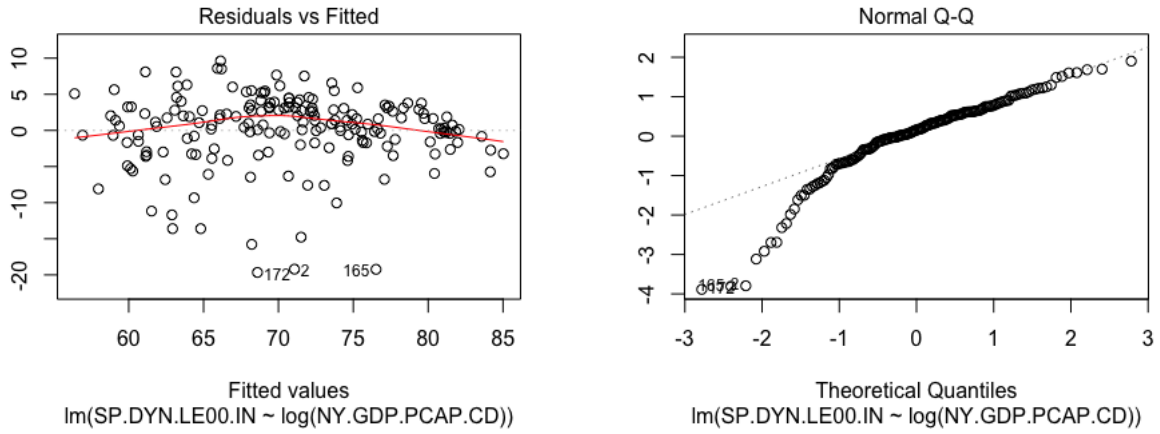
Before we conducted our modeling, we needed to be sure that each regression satisfied the assumptions for linear regression. To correct for any violations, we applied log transformations to our data until we were satisfied that they met all assumptions. The assumptions we analyzed are as follows:

- Linearity: We analyzed the distribution of the data as well as the residual plot versus the fitted data to identify any non-linear trends.
- Constant variance: We analyzed the residual plot versus the fitted data to see if the residuals were randomly distributed around 0.
- Independence of errors: We reasoned about the independence of the data samples from our data set.
- Normality of errors: We analyzed the QQ plot for any major skewness in our data.

A few general trends emerged as we were evaluating the assumptions for all our models. First, many of our models' data were slightly skewed right, as our regression plots showed a tightly clustered collection of points at lower x-values before transformation. After transformation, our QQ plot reveals a skew left. This could be a factor of the log transformation being too extreme. This may also show the underlying effect of GDP on other predictors we used, as many other models shared the same QQ plot shape as the basic model of life expectancy vs GDP. We discuss this more later in the report. Second, the data were

generally equally distributed with constant variance, as the residuals vs. fitted plot demonstrated. This allowed us to more confidently apply our linear models. Lastly, we do believe there is some amount of dependence in the data points, as many of our predictors are not evenly split among country lines (ex. biodiversity is not determined by country borders). There may be some more general geographic effect that causes these points to be slightly correlated. However, as mentioned earlier, we attempt to address this issue by normalizing our data - either by dividing by land area or population. Overall, we believe that each of our tests satisfy the linear regression assumptions, and our results are generally indicative of true results. An example of representative results can be seen in the life expectancy vs GDP plots.

Life expectancy vs. log(GDP per capita)



We also include the residual vs. fitted plot and QQ plot for each test in our appendix.

3.2 Model Selection

After selecting specific variables to test against as specified in part 3.1, we then performed backwards, forwards, and stepwise model variable selection, with all variables and their interaction effects for both pollution and biodiversity. We then compared each of these models with **Model A**, the model including all predictors, to determine the best overall model for life expectancy. Finally, we ran a cross-validation (CrVa) study for each of the three models to find which one had the least error (minimized SSE).

The following table summarizes each of the models, where **Model B** is the backwards-selected model and **Model C** is the forwards-selected and stepwise-selected model (both methods yielded the same model). See appendix for additional model specifics.

Model	**Method**	**Variables**	**AIC**	**BIC**	**CrVa SSE**
Model A	All predictors	Life exp. ~ log(GDP) + log(PM2.5) + log(log(threat.prop)) + log(CO2 per capita)	599.1	615.2	24.6
Model B	Backwards	Life exp. ~ log(GDP) + log(PM2.5) + log(log(threat.prop)) + log(CO2 per capita) + log(GDP)*log(PM2.5) + log(PM2.5)*log(CO2 per capita) + log(log(threat.prop))*log(CO2 per capita)	596.1	621.9	23.4
Model C	Forwards/ Stepwise	Life exp. ~ log(GDP) + log(log(threat.prop)) + log(CO2 per capita) + log(log(threat.prop))*log(CO2 per capita)	595.7	611.8	24.2

Based off of these metrics, Model C overall performed the best (lowest AIC and BIC, second-lowest SSE). Model C includes the predictors: log(GDP), which has a positive slope of 3.7878 and a p-value of < 0.0001; log(log(threat.prop)), which has a negative slope of -1.7378 and a p-value of 0.24; log(CO2 per capita), which has a positive slope of 0.6115 and a p-value of 0.26; and the interaction between log(log(threat.prop)) and log(CO2 per capita), which had a negative slope of -0.4333 and a p-value of

0.06. Model B and C had adjusted R^2 values of 0.6634 and 0.6589. This is reverse from the result of the AIC and BIC, which can be a result of how the adjusted R^2 values don't penalize as much as AIC and BIC for adding more predictors. We also note that the adjusted R^2 values are very close to each other.

4 Conclusion

We found that the amount of pollution (measured in PM2.5) in a country's air may be a weak negative predictor of life expectancy, while the amount of pollutants emitted (measured in CO2) is a strong positive predictor. Accounting for economic prosperity (GDP) removed the significance of PM2.5 air pollution as a predictor, while CO2 remains a significant predictor of life expectancy if we also consider the interaction effect between GDP and CO2 emissions. This indicates that economic prosperity perhaps is a confounding variable between air pollution and human health. Meanwhile, no significant correlations between biodiversity and human health were found. Model selection using stepwise and forward variable selection methods did not significantly improve the model.

5 Discussion

Our first model (Life expectancy vs. GDP) confirmed our hypothesis that a higher GDP is associated with longer life expectancy (and vice versa), which makes sense because greater economic development means more resource to be able to sustain life. The model had a good adjusted R-squared, indicating that it was good at maximizing the variance explained while minimizing the variance of the residuals – i.e. our model had a strong correlation.

Our second aim to study the relationship between life expectancy and pollution involved two parts: Life expectancy vs. PM2.5 air pollution and Life expectancy vs. CO2 emissions. To distinguish these models, we note that the PM2.5 metric measures the concentration of PM2.5 in that country's atmosphere, whereas the CO2 emissions metric measures how much a country pollutes. PM2.5 refers to airborne molecules with diameter smaller than 2.5 micrometers. It is an especially deadly air pollutant because of its tiny size and consequent ease in penetrating deeper into the lungs [**?**], and is therefore a good metric for atmospheric pollution. CO2 emissions is a good metric for the contributions of a country towards global pollution.

For life expectancy vs. PM2.5, our model may suggest a negative correlation, which we predicted because a greater concentration of deadly PM2.5 in a country's atmosphere means greater exposure to its residents and consequently negative health impacts. However, this correlation was very weak with a very low R^2 value, indicating the relationship is weak if it exists at all. This may be because of opposing trends: while PM2.5 likely is associated with lower life expectancy in a vacuum, countries that tend to have less PM2.5 also tend to be less-developed and thus have shorter lifespans. It is interesting to note that there appears to be a cluster in the data of countries with high life expectancy and medium levels of PM2.5. These are probably developing countries that have high life expectancies due to their development, but still have pollutant-heavy and inefficient processes to produce energy and other resources.

For life expectancy vs. CO2, we found a positive correlation. This is contrary to our hypothesis where we thought this would be negatively correlated. However on further thought, we realized that this makes sense because countries that are industrialized and developed both pollute more and have longer-living citizens (e.g US and China). CO2 emissions reflects how much they contribute to global environmental change, not just that of their own country.

After running simple linear regression, we wanted to test if GDP could change the relationships we observed if added as another predictor. This is an interesting question because it asks if air pollution affects life expectancy regardless of a country's economic development. If two countries have similar economic statuses, does the one with a higher PM2.5 concentration or fewer CO2 emissions still have a lower life expectancy as previously predicted? (Note that GDP continues to be a significant predictor for almost all multiple regressions conducted.)

For both CO2 and PM2.5, adding GDP as another predictor resulted in loss of significance. This suggests that PM2.5 and CO2's previous correlations with life expectancy may just be because of their correlations with GDP. At the same GDP, differences in PM2.5 and CO2 do not make a significant difference in life expectancy, which leads us to think that there are other aspects of GDP (such as poverty level) that affect life expectancy more significantly. As a result, if the GDP is held constant, any differences in PM2.5 and CO2 can make minimal effect.

Finally, we tested for interaction effects between GDP and PM2.5 as well as GDP and CO2. For PM2.5, neither the interaction nor PM2.5's main effect were significant. For CO2, we interestingly found that adding the interaction effect between CO2 and GDP changed things. While insignificant with GDP alone, CO2's main effect became significant once again when the interaction effect between GDP and CO2 was added. This suggests that different levels of GDP may differ in how and the extent to which they correlate with CO2 emissions per capita. By accounting for this interaction effect, we address the problem discussed previously where any differences in CO2 could be overshadowed by metrics of GDP. Here, the interaction effect takes into account how GDP and CO2 vary together, and as a result, we are able to parse out the main effect of CO2 as well.

The result from the biodiversity test did not confirm our hypothesis, since the slope values were positive instead of negative. Less biodiversity (and more threatened species) would predict a longer life expectancy. However, we note that this is not a strong correlation. The slope is rather small. So an increase in one unit of log(log(density)) would require a more significant increase in density of threatened species in order to affect the predicted life expectancy. Furthermore, the correlation is very weak because of the really small R^2 value. When we ran the multiple regression with log(GDP), the R^2 value did not change from the simple linear regression of life expectancy vs log(GDP), so the information about density of threatened species did not add any information. We note that this data itself is rather hard to collect, as it is hard to monitor all the plant and animal life in many countries, so much of the data might be incomplete or not completely accurate itself.

Finally, we performed model selection to determine the best-performing model in predicting life expectancy, with both biodiversity and air pollution as possible predictor variables. After running backwards, forwards, and stepwise variable selection, and comparing the AIC, BIC, and Cross-Validation SSE for each of the three models, we determined that the forwards/stepwise-selected model (**Model C**) slightly outperformed other models. Model C had a lower AIC and BIC than Model A (all predictors) and Model A (backwards selection). Interestingly, it also had a slightly higher Cross-Validation SSE than Model B, but this is likely due to the fact that it has significantly fewer predictors than Model B, and SSE fails to penalize a model for an increased number of predictors.

Model C appears to pass the necessary assumptions, although its distribution is somewhat left-skewed (see Appendix). For life expectancy vs. CO2 per capita, we found a positive correlation, indicating that when adjusting for other predictors, countries with more CO2 emissions tend to have higher life expectancy. This result, which is in accordance with our air pollution-only models, is contrary to our hypothesis and is discussed in depth above.

For life expectancy vs. threatened species, we interestingly found a negative correlation, indicating that when adjusting for other predictors, countries with a higher proportion of threatened species tend to have lower life expectancy. This result is consistent with our hypothesis; controlling for GDP, countries with more at-risk biodiversity tend to also display other 'bad traits' like lower life expectancy. Interestingly, this result contrasts with our result for biodiversity alone (which has a positive slope), possibly because of the added pollution variables in this model explaining some of the same variability (alternatively, and perhaps more likely, the association between threatened species and life expectancy could simply just not be significant).

This final model also included an interaction effect between threatened species and CO2 per capita, indicating that the effect of CO2 on life expectancy varies across countries with different proportions of threatened species. Notably, however, none of these predictors individually are quite significant, although the overall model is highly significant (p-value $< 2e-16$). This is likely due to multicolinearity among the predictors (ex. countries with more threatened species likely have fewer CO2 emissions). Because of this, in addition to the fact that AIC and BIC were only very slightly improved, we should be cautious about drawing conclusions from this model; this model did not significantly improve on our previous models above.

5.1 Limitations and Challenges

While we did a significant amount of work finding the most appropriate statistics to use in our analysis, there are still several limitations of our data. First, our species-based biodiversity statistic, which uses the number of threatened species as a proxy for environmental health, is not a perfect statistic. Even though we attempt to normalize the data by dividing by land area, our statistic does not fully take into account the significant biodiversity differences that occur around the world (for example, deserts versus rainforests), and thus would be biased. This may explain why we did not find significance. It would be useful to have a baseline number of species in each country (perhaps at an earlier time point with less

environmental change) for comparison, so we can calculate how biodiversity has changed relative to each country's unique environment.

Second, we had to deal with missing values in some of our data, since the World Bank lacked statistics for some countries for some variables. We dealt with this issue by removing countries on an as-needed basis; that is, if some country lacked data for a specific variable, then we failed to include that country in the analysis for that variable. In general, this method tended to eliminate very small countries like Monaco, Andorra, and San Marino, as well as some very poor and/or recently-formed countries like South Sudan. Thus, because eliminated countries tended to be of a certain type (either very small or very poor), our metrics may be somewhat biased. As a result, we should be hesitant about extending our results to conclusions about these kinds of countries. It would be useful to have data for these states to increase the generalizability of our data.

Acknowledgments

We would like to thank Prof. Kevin Rader and Ms. Katy McKeough for supervising us throughout this project (and this class).

References

[1] Shi, L., Zanobetti, A., Kloog, I., Coull, B. A., Koutrakis, P., Melly, S. J., & Schwartz, J. D. (2016). Low-concentration PM2. 5 and mortality: Estimating acute and chronic effects in a population-based study. *Environmental health perspectives*, 124(1), 46.

[2] Jacobson, M. Z. (2008). On the causal link between carbon dioxide and air pollution mortality. *Geophysical Research Letters*, 35(3).

[3] Stern, D. I. (2004). The rise and fall of the environmental Kuznets curve. *World development*, 32(8), 1419-1439.

[4] World Bank, World Development Report 1993 (New York: Oxford University Press, 1993).

[5] Pop, I. A., van Ingen, E., & van Oorschot, W. (2013). Inequality, wealth and health: is decreasing income inequality the key to create healthier societies?. *Social Indicators Research*, 113(3), 1025-1043.

[6] WHO — 7 million premature deaths annually linked to air pollution. (2014, March 25). Retrieved December 05, 2016, from http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

[7] The World Bank. 2012. "World Development Indicators." URL: http://data.worldbank.org/data-catalog/world-development-indicators.

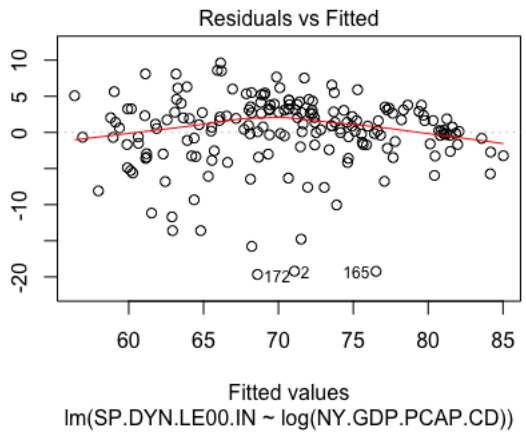
[8] Davidson, C. I., Phalen, R. F., & Solomon, P. A. (2005). Airborne particulate matter and human health: a review. *Aerosol Science and Technology*, 39(8), 737-749.

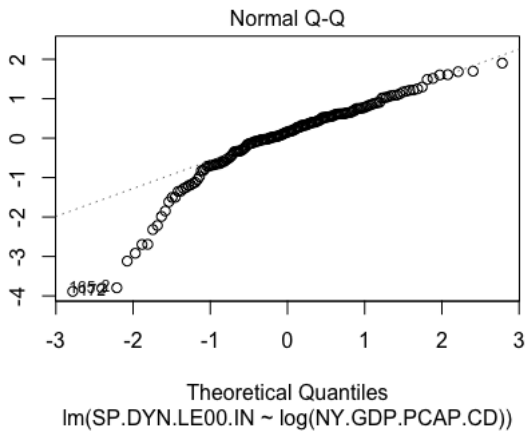
Appendix

5.2 Air Pollution

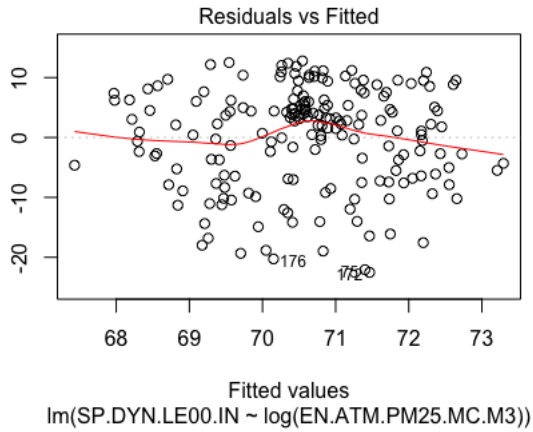
5.2.1 Assumption checking

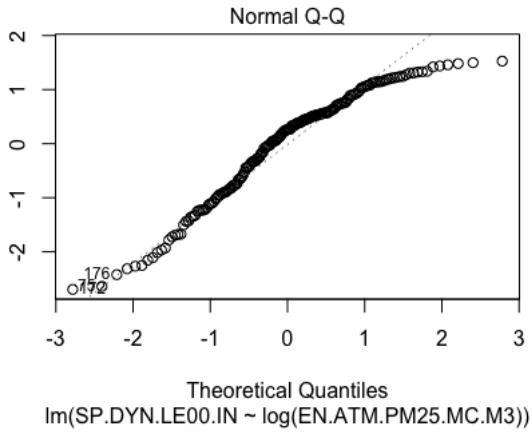
Life expectancy vs. log(GDP per capita)



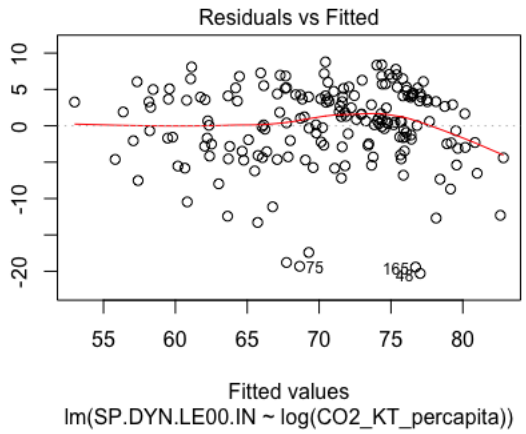


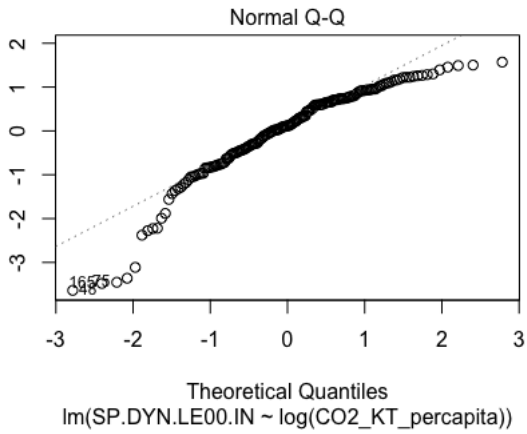
Life expectancy vs. log(PM2.5 air pollution)



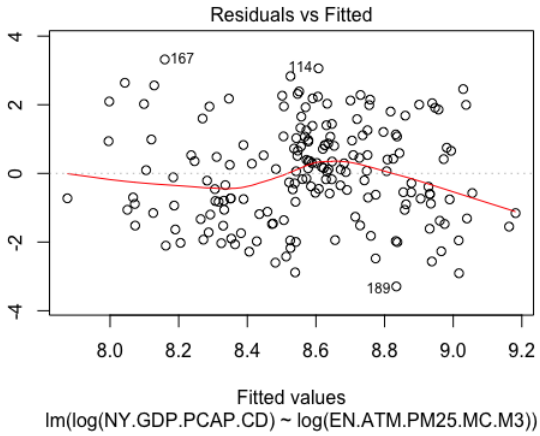


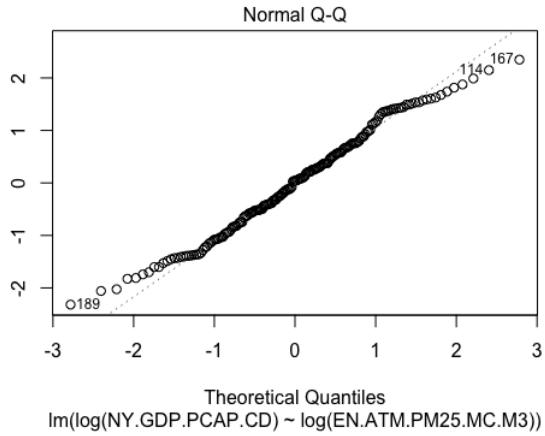
Life expectancy vs. log(CO2 emissions per capita)



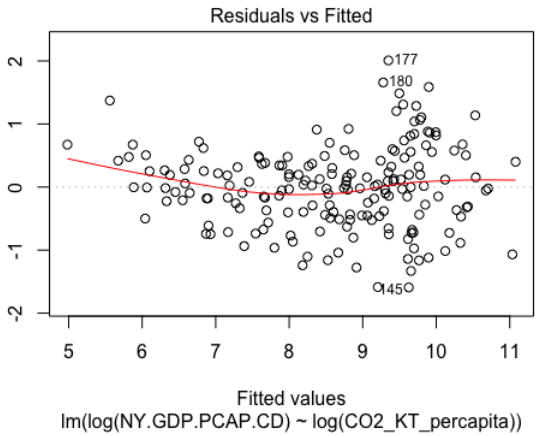


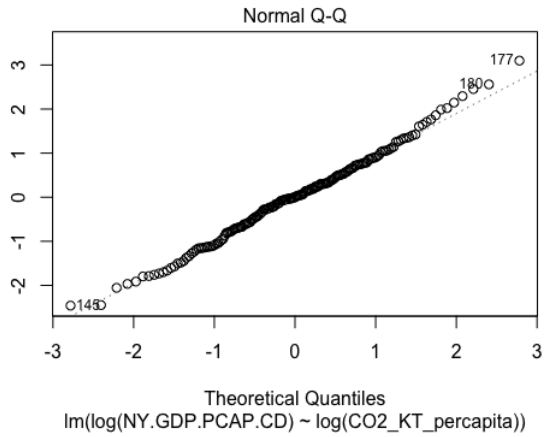
log(GDP) vs. log(PM2.5)



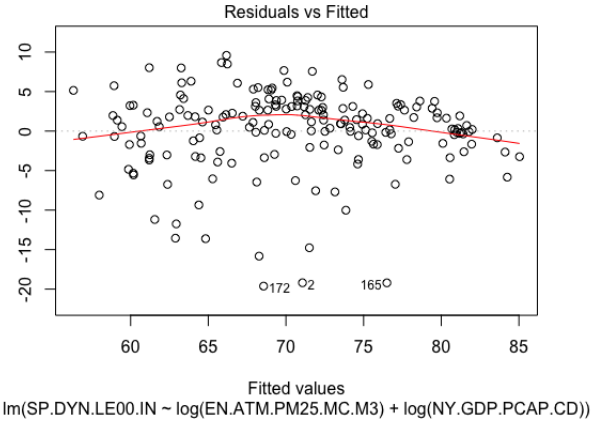


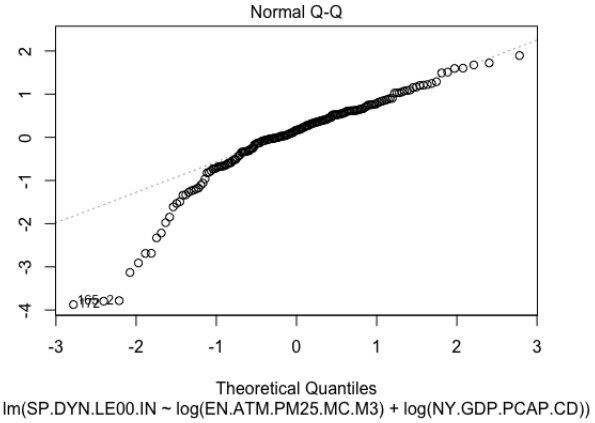
log(GDP) vs. log(CO2 per capita)



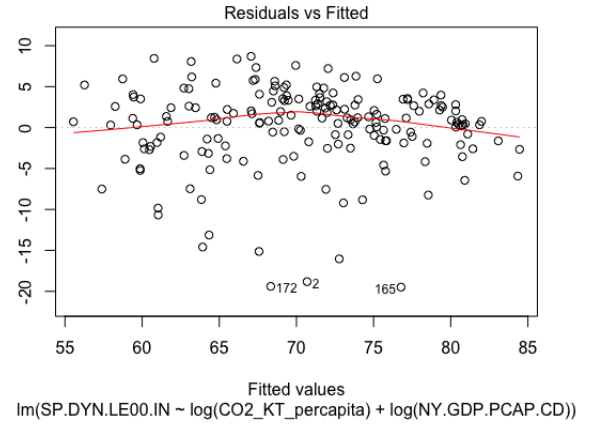


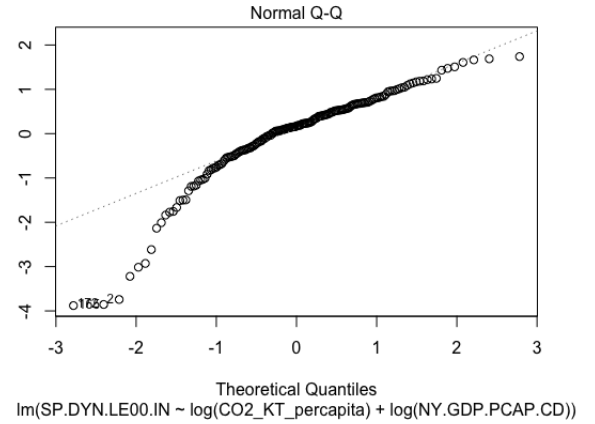
Life expectancy ~ log(GDP per capita) + log(PM2.5 air pollution)



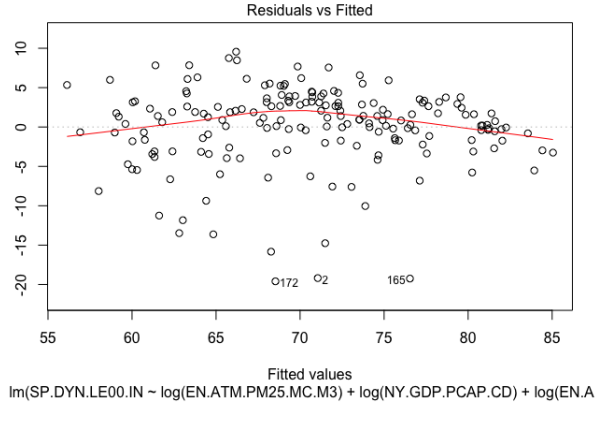


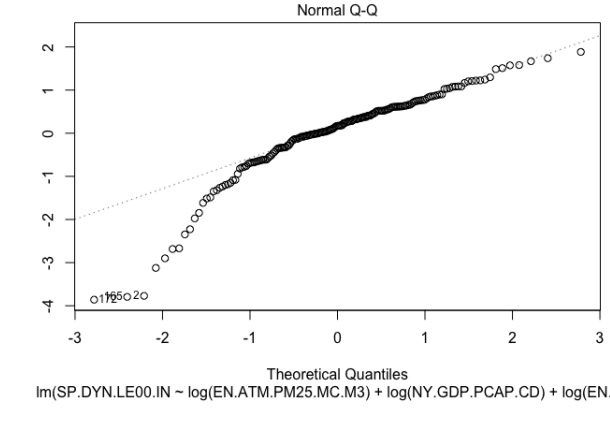
Life expectancy ~ log(GDP per capita) + log(CO2 emissions per capita)



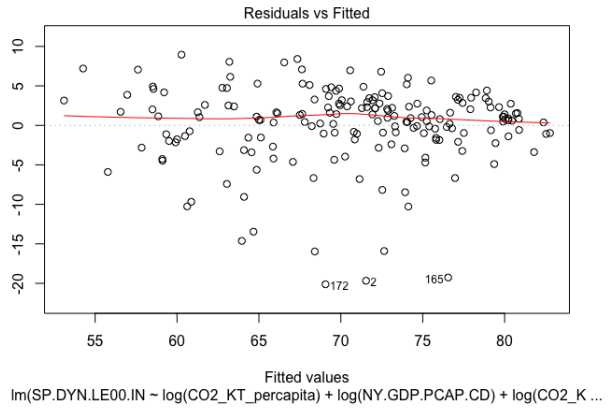


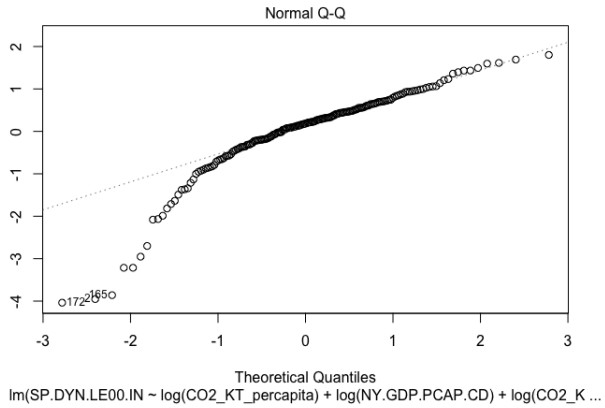
Life expectancy ~ log(GDP per capita) + log(PM2.5 air pollution) + log(GDP)*log(PM2.5)





Life expectancy ~ log(GDP per capita) + log(CO2 emissions per capita) + log(GDP)*log(CO2)

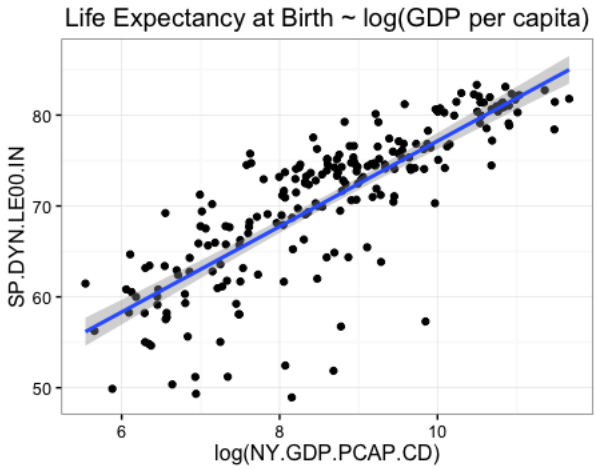




5.2.2 Simple linear regression

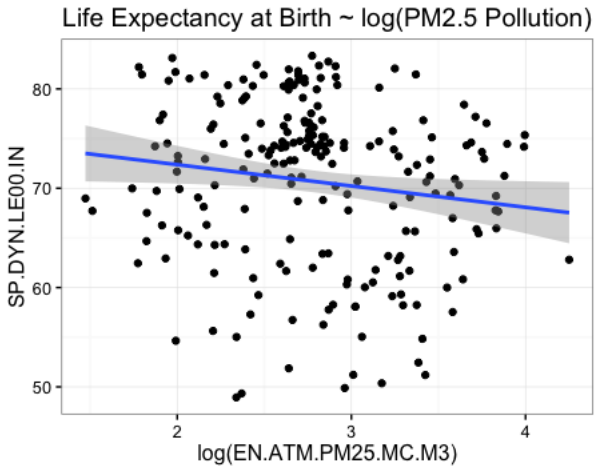
Life Expectancy vs. log(GDP)

	Estimate	Std. Error	p-value
log(NY.GDP.PCAP.CD)	4.7061	0.2311	$< 2e-16$
Adjusted R-squared: 0.657			



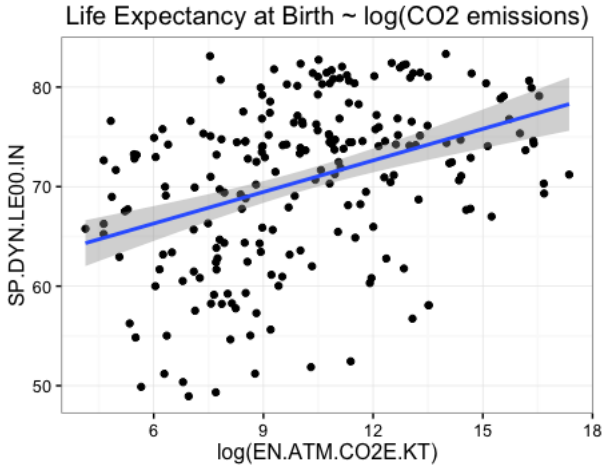
Life Expectancy vs. log(PM2.5)

	Estimate	Std. Error	p-value
log(EN.ATM.PM25.MC.M3)	-2.148	1.001	0.033
Adjusted R-squared: 0.01641			



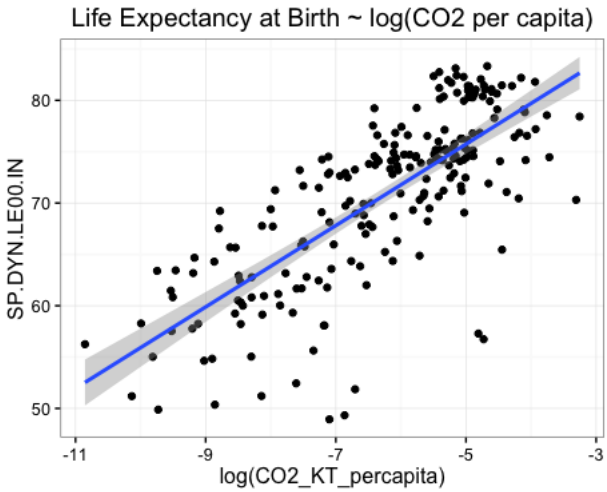
Life Expectancy vs. log(CO2)

	Estimate	Std. Error	p-value
log(EN.ATM.CO2E.KT)	1.0562	0.1741	$5.84e-09$
Adjusted R-squared: 0.1422			



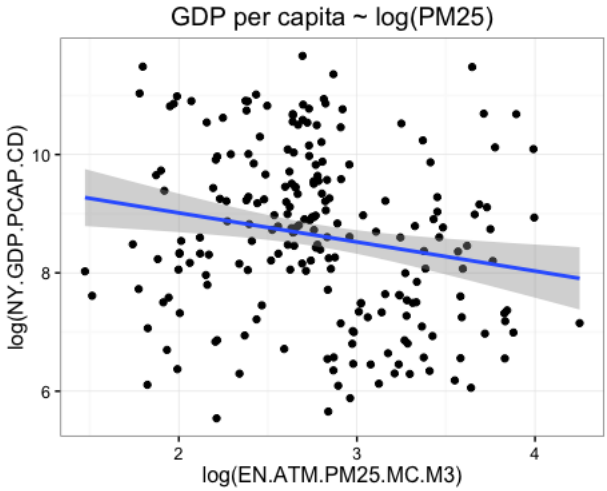
Life Expectancy vs. log(CO2/total population)

	Estimate	Std. Error	p-value
log(CO2_KT_percapita)	3.9643	0.2344	$< 2e-16$
Adjusted R-squared: 0.5688			



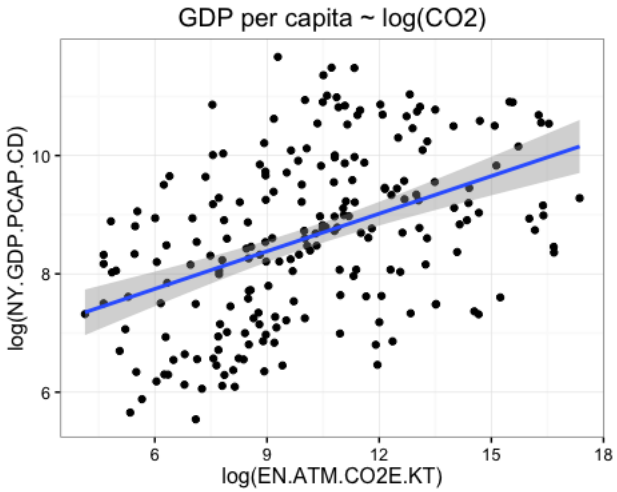
log(GDP) vs. log(PM2.5)

	Estimate	Std. Error	p-value
log(EN.ATM.PM25.MC.M3)	-0.4917	0.1712	0.00448
Adjusted R-squared: 0.03248			



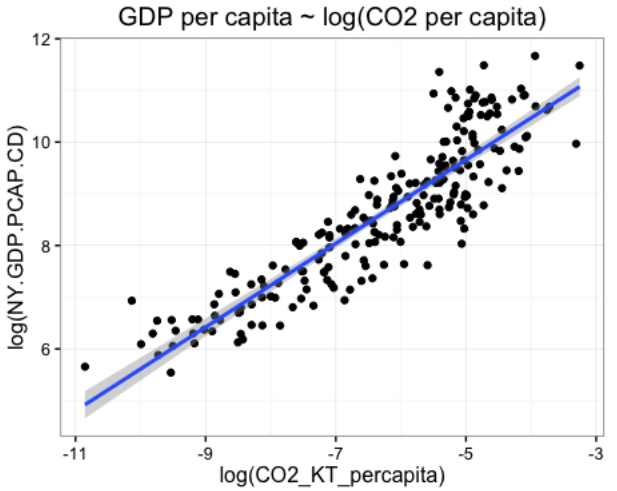
log(GDP) vs. log(CO2)

	Estimate	Std. Error	p-value
log(EN.ATM.CO2E.KT)	0.2119	0.0291	$6.14e-12$
Adjusted R-squared: 0.1941			



log(GDP) vs. log(CO2/total population)

	Estimate	Std. Error	p-value
CO2_KT_percapita	0.8020	0.0297	$< 2e-16$
Adjusted R-squared: 0.7983			



5.2.3 Multiple linear regression

LE ~ log(GDP) + log(PM2.5)

	Estimate	Std. Error	p-value
log(EN.ATM.PM25.MC.M3)	0.1729	0.6036	0.775
log(NY.GDP.PCAP.CD)	4.7191	0.2360	$< 2e-16$
Adjusted R-squared: 0.6555			

LE ~ log(GDP) + log(CO2)

	Estimate	Std. Error	p-value
log(EN.ATM.CO2E.KT)	0.0734	0.1231	0.552
log(NY.GDP.PCAP.CD)	4.6376	0.2584	$< 2e-16$
Adjusted R-squared: 0.656			

LE ~ log(GDP) + log(CO2/total pop)

	Estimate	Std. Error	p-value
log(CO2_KT_percapita)	0.7814	0.4650	0.0944
log(NY.GDP.PCAP.CD)	3.9339	0.5139	$6.57e-13$
Adjusted R-squared: 0.6599			

5.2.4 Interaction effects

LE ~ log(GDP) + log(CO2/total pop) + log(GDP)*log(CO2/total pop)

	Estimate	Std. Error	p-value
log(CO2_KT_percapita)	3.0942	1.3374	0.0218
log(NY.GDP.PCAP.CD)	2.1765	1.0706	0.0435
log(CO2_KT_percapita):log(NY.GDP.PCAP.CD)	-0.3119	0.1736	0.0741
Adjusted R-squared: 0.6475			

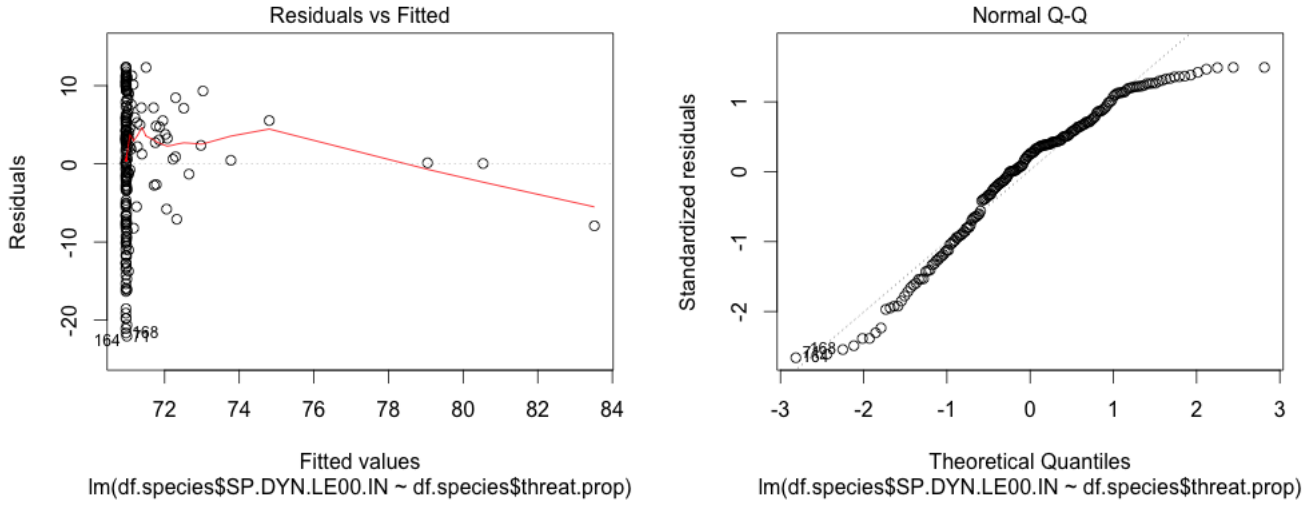
LE ~ log(GDP) + log(PM2.5) + log(GDP)*log(PM2.5)

	Estimate	Std. Error	p-value
log(EN.ATM.PM25.MC.M3)	0.9961	3.8354	0.795378
log(NY.GDP.PCAP.CD)	4.9813	1.2754	0.000132
log(EN.ATM.PM25.MC.M3):log(NY.GDP.PCAP.CD)	-0.1082	0.4512	0.810679
Adjusted R-squared: 0.6357			

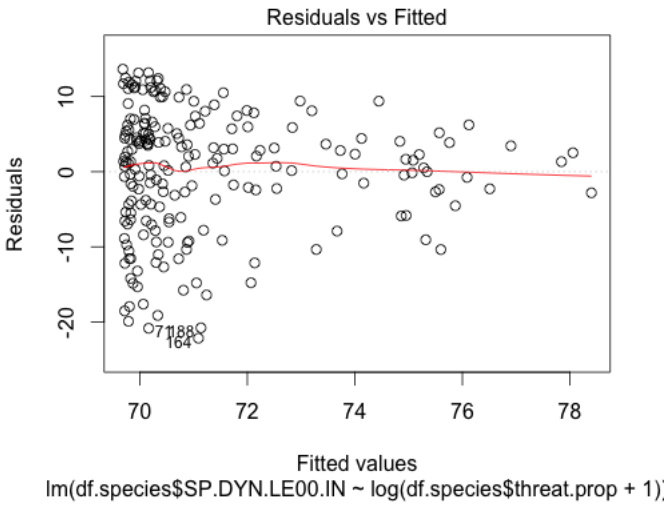
5.3 Biodiversity

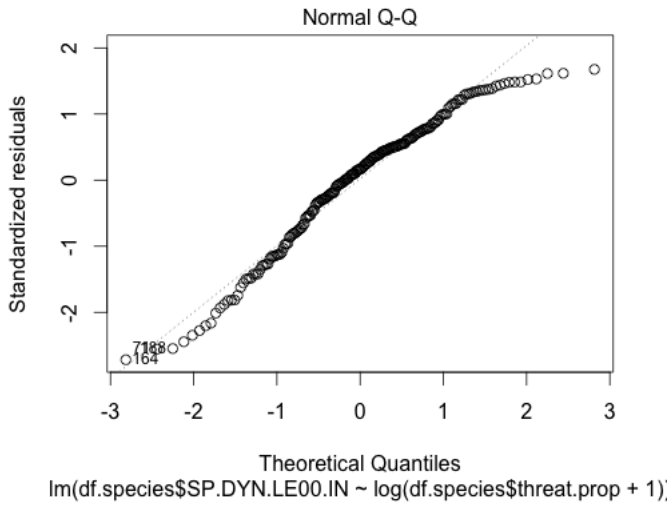
5.3.1 Assumption checking

Life expectancy vs. Threatened species

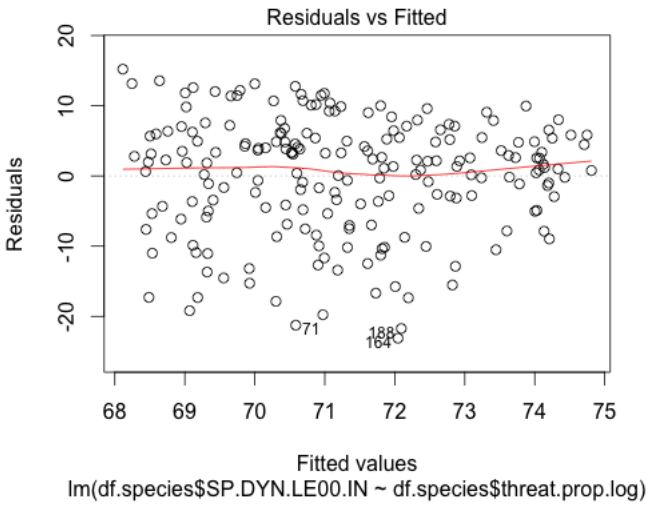


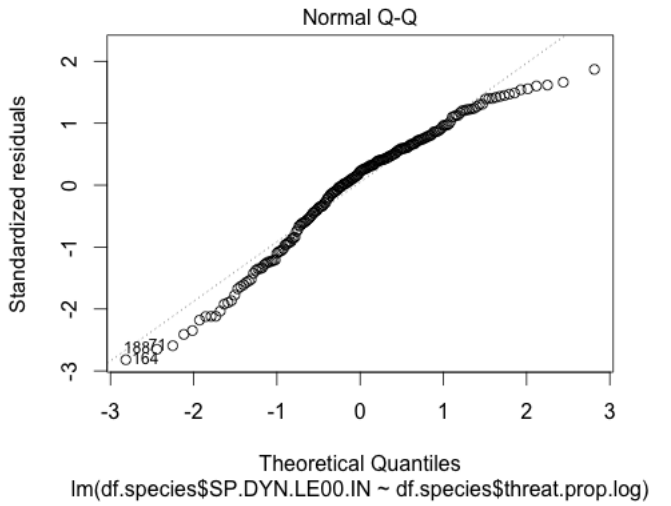
Life expectancy vs. log(Threatened species + 1)



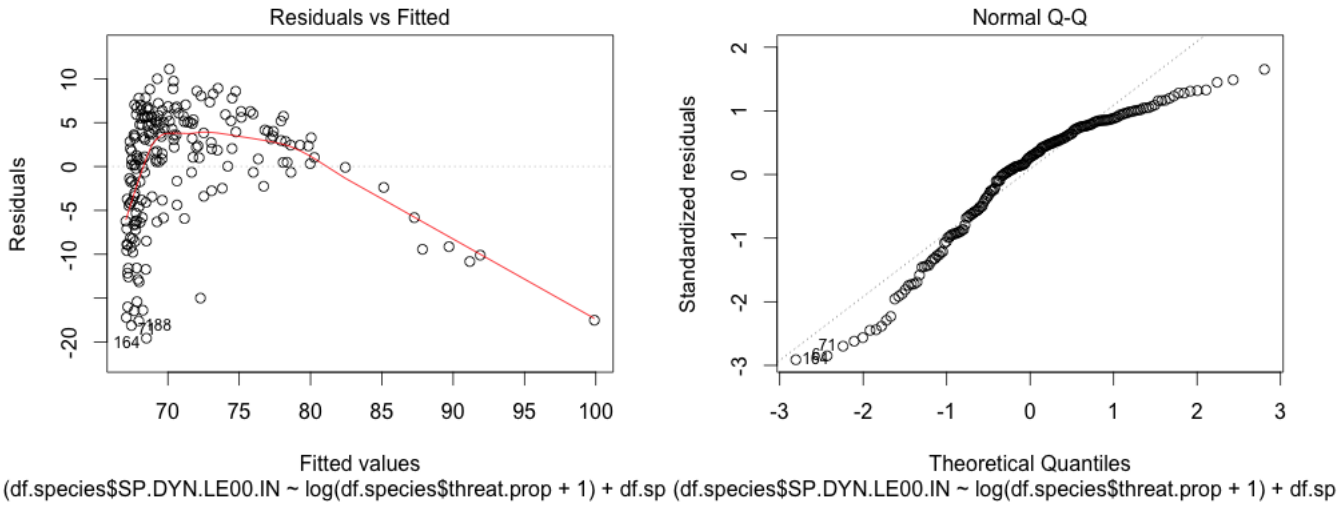


Life expectancy vs. log(log(Threatened species + 1) + 0.1)

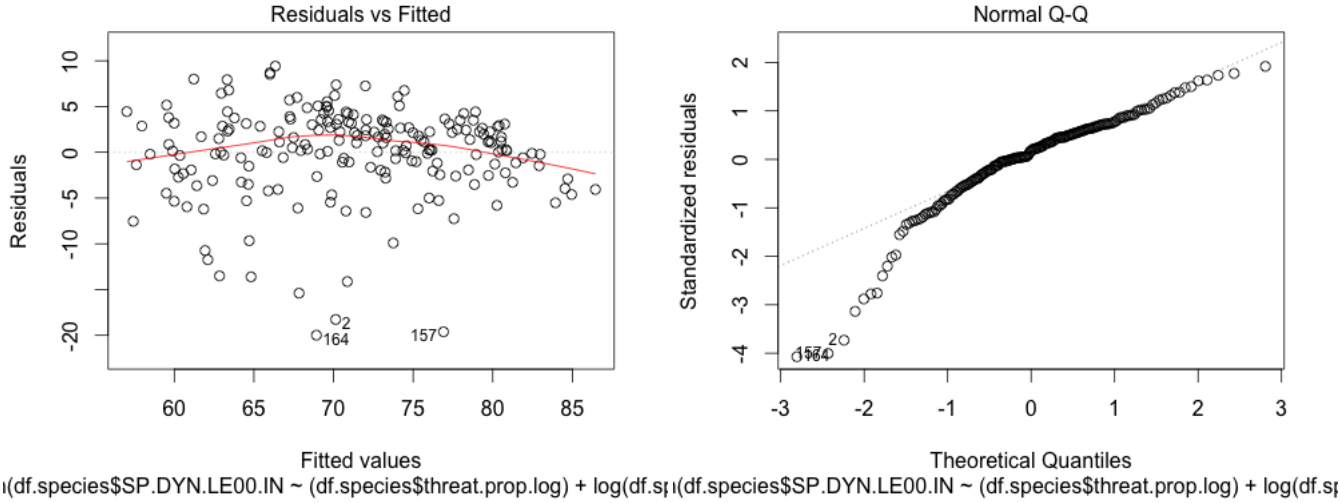




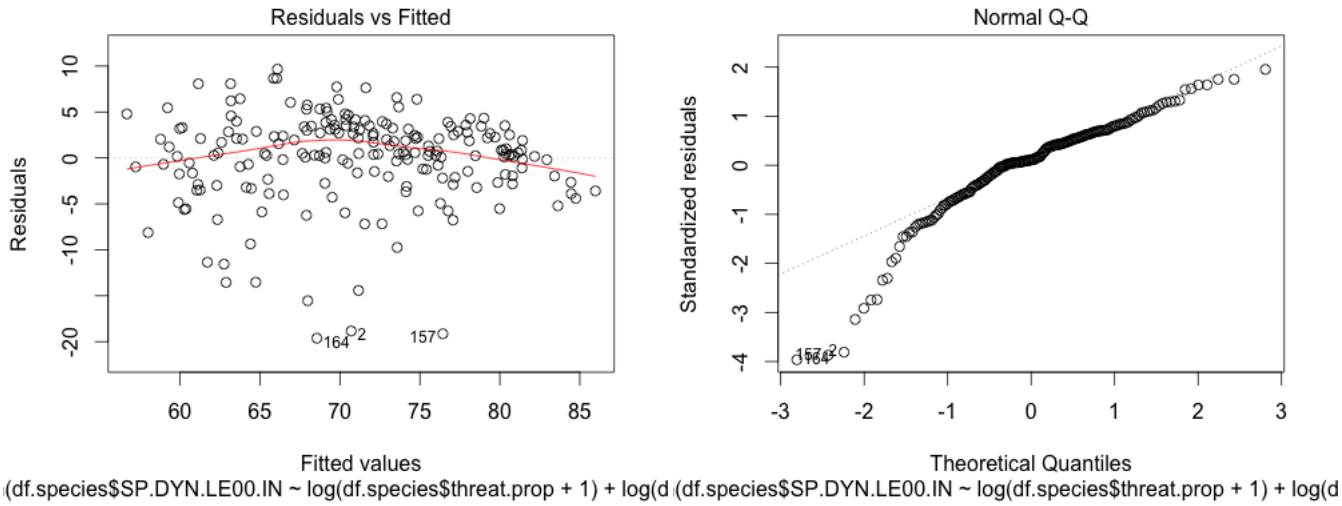
Life expectancy vs. log(Threatened species + 1) + gdp



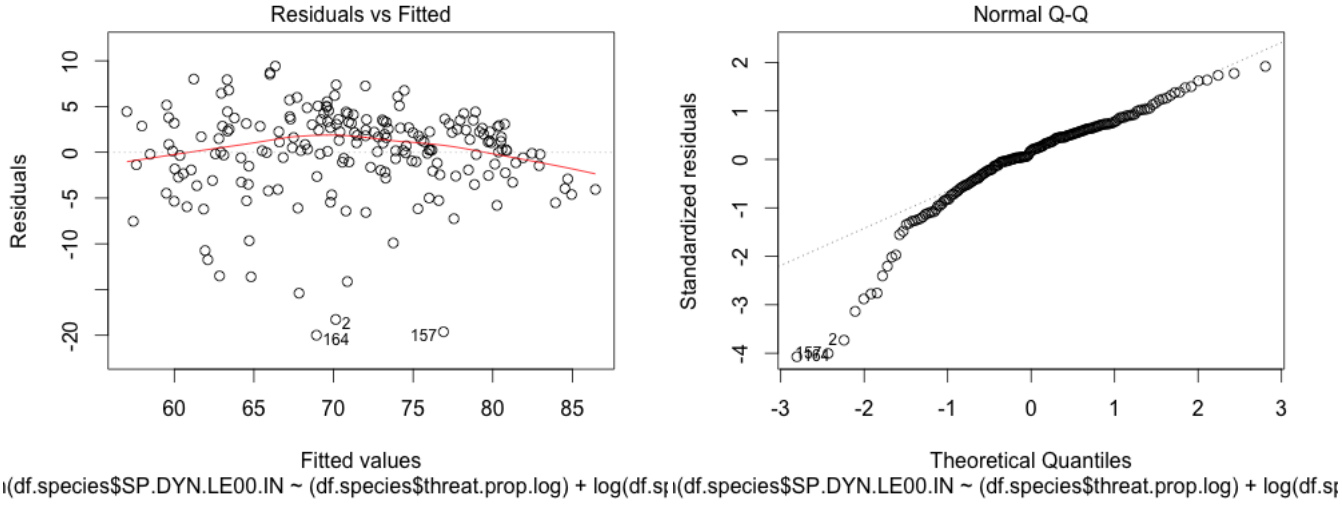
Life expectancy vs. log(log(Threatened species + 1) + 0.1) + gdp



Life expectancy vs. log(Threatened species + 1) +log(gdp)



Life expectancy vs. log(log(Threatened species + 1) + 0.1) + log(gdp



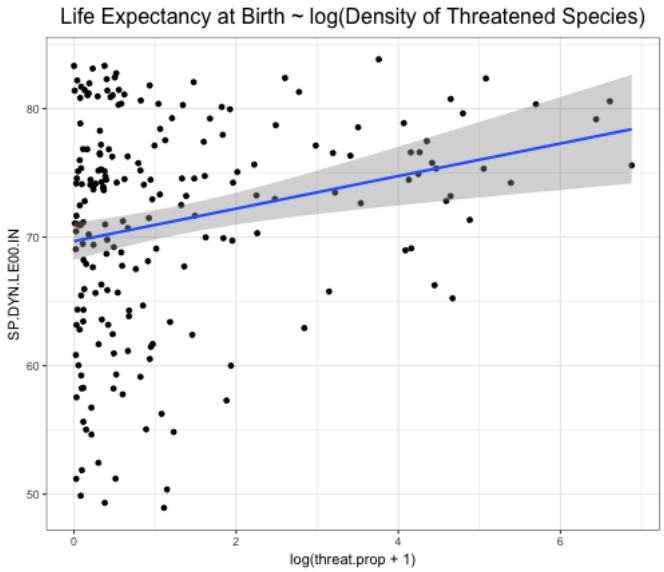
5.3.2 Simple linear regression

LE ∼ threat.prop

	Estimate	Std. Error	p-value
threat.prop	0.012938	0.005766	0.0259
Adjusted R-squared: 0.0194			

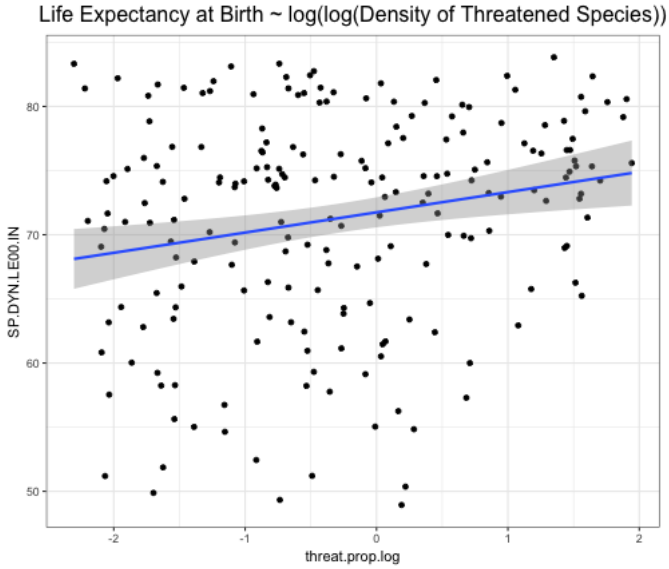
LE $\sim$ log(threat.prop + 1)

	Estimate	Std. Error	p-value
log(threat.prop + 1)	1.2670	0.3681	0.000702
Adjusted R-squared: 0.05048			



LE $\sim$ log(log(threat.prop + 1) + 0.1)

	Estimate	Std. Error	p-value
threat.prop.log	1.5769	0.5143	0.00247
Adjusted R-squared: 0.03955			



5.3.3 Multiple linear regression

LE $\sim$ log(threat.prop + 1) + GDP.percapita

	Estimate	Std. Error	p-value
log(threat.prop + 1)	$7.461e-01$	$3.345e-01$	0.0268
NY.GDP.PCAP.CD	$2.079e-04$	$2.071e-05$	$< 2e-16$
Adjusted R-squared: 0.3619			

LE $\sim$ log(log(threat.prop + 1) + 0.1) + GDP.percapita

	Estimate	Std. Error	p-value
threat.prop.log	1.083	$4.362e-01$	0.0139
NY.GDP.PCAP.CD	$2.104e-04$	$2.050e-05$	$< 2e-16$
Adjusted R-squared: 0.3657			

LE $\sim$ log(threat.prop + 1) + log(GDP.percapita)

	Estimate	Std. Error	p-value
log(threat.prop + 1)	0.2496	0.2493	0.318
NY.GDP.PCAP.CD	4.5340	0.2417	$< 2e-16$
Adjusted R-squared: 0.6539			

LE $\sim$ log(log(threat.prop + 1) + 0.1) + log(GDP.percapita)

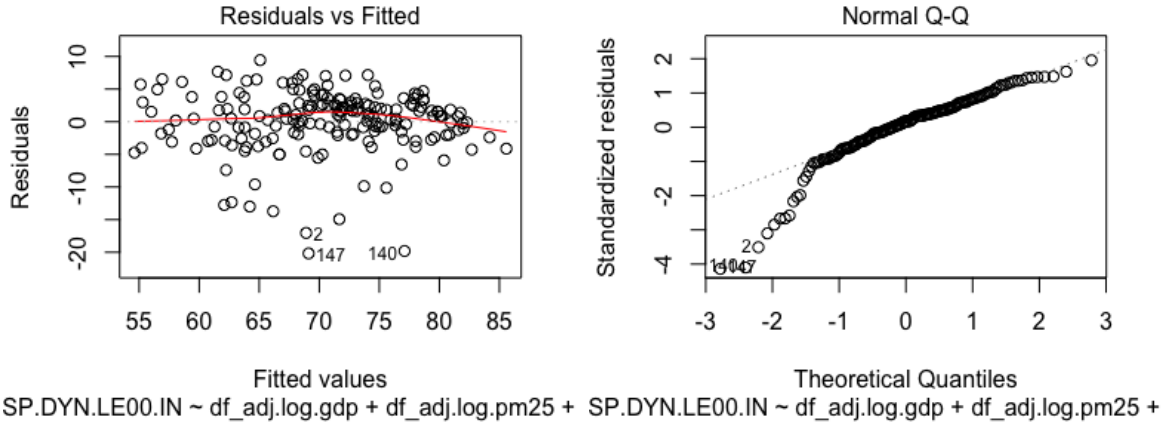
	Estimate	Std. Error	p-value
threat.prop.log	0.6511	0.3214	0.0441
NY.GDP.PCAP.CD	4.5219	0.2364	$< 2e-16$
Adjusted R-squared: 0.6593			

5.4 Model Selection + Cross-Validation

5.4.1 Model Results

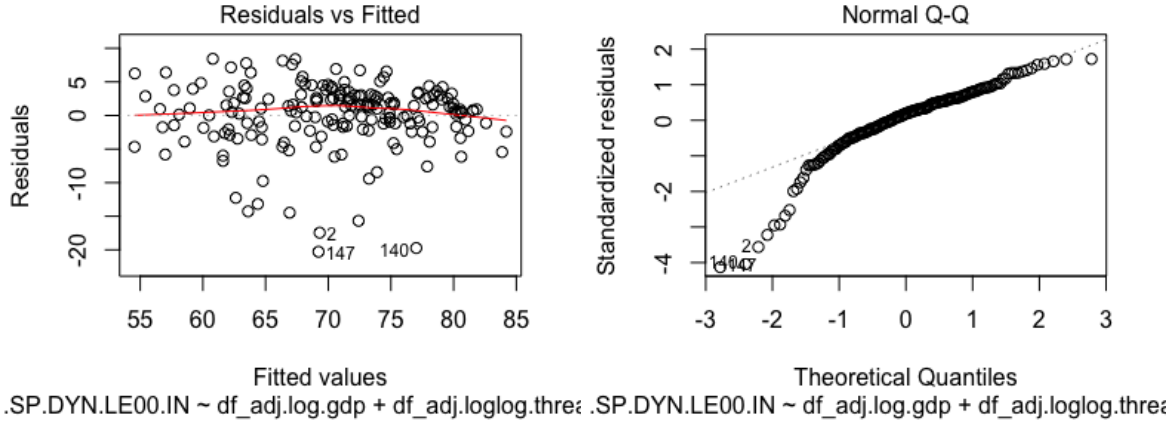
Results of Model B, backwards variable selection (AIC 596.1955):
LE $\sim$ log(GDP) + log(PM2.5) + log(log(threat.prop)) + log(CO2 per capita) + log(GDP)*log(PM2.5) + log(PM2.5)*log(CO2 per capita) + log(log(threat.prop))*log(CO2 per capita)

	Estimate	Std. Error	p-value
log.gdp	10.5143	3.1613	$< 2e-16$ ***
log.pm25	32.9721	15.9834	$< 2e-16$ ***
loglog.threat.prop	-1.0549	1.5706	$< 2e-16$ ***
log.co2percapita	-4.3778	2.8703	0.12899
log.gdp:log.pm25	-2.4589	1.1498	0.03385 *
log.pm25:log.co2percapita	1.8234	1.0265	0.07741
loglog.threat.prop:log.co2percapita	-0.3607	0.2485	0.14838
Adjusted R-squared: 0.6634			



Results of Model C, forwards and stepwise variable selection (AIC 595.7663):
LE ∼ log(GDP) + log(log(threat.prop)) + log(CO2 per capita) + log(log(threat.prop))*log(CO2 per capita)

	Estimate	Std. Error	p-value
log.gdp	3.7878	0.5680	3.07e-10 ***
loglog.threat.prop	-1.7378	1.4818	0.2424
log.co2percapita	0.6115	0.5454	0.2637
loglog.threat.prop:log.co2percapita	-0.4333	0.2330	0.0646
Adjusted R-squared: 0.6589			



5.4.2 Cross-Validation

Results of cross-validation study, where SSEA = SSE of model with all predictors, SSEB = SSE of backwards-selected model, SSEF = SSE of forwards-selected model, SSES = SSE of stepwise-selected model:

```
> c(mean(ssea),mean(sseb),mean(ssef),mean(sses))/(n-ntest)
[1] 24.66327 23.49480 24.22267 24.22267
```

5.5 Scripts

5.5.1 Data Processing

```
import pandas
data_frame = pandas.read_csv('Indicators.csv')
factors_of_interest \\ all data column names that are needed for data analysis
regions \\ non-country data points that should be excluded from data processing

data_dict = {}
# for index, row in data_frame.iloc[:200000].iterrows():
for index, row in data_frame.iterrows():
    if index % 25000 == 0:
        print "INDEX:", index
    if not row['CountryName'] in regions and row['IndicatorCode'] in factors_of_interest:
        if row['CountryCode'] in data_dict:
            data_dict[row['CountryCode']][row['IndicatorCode']] = row['Value']
        else:
            data_dict[row['CountryCode']] = {
                'CountryCode': row['CountryCode'],
                'CountryName' : row['CountryName'],
                row['IndicatorCode'] : row['Value']
            }

columns = ["CountryCode", "CountryName"] + factors_of_interest
formatted_data_frame = pandas.DataFrame(columns=columns)
for d in data_dict:
    dlist = [data_dict[d][c] if c in data_dict[d] else "NA" for c in columns]
    formatted_data_frame.loc[len(formatted_data_frame)]= dlist

formatted_data_frame.to_csv('formatted_data.csv')
```

5.5.2 Air Pollution regressions and assumption checking

```
library(scatterplot3d)
library(rgl)
library(ggplot2)

fname = "~/Stat139/Project/formatted_data_countries_only.csv"
df = read.csv(fname, header=T)

# for pollution
df_noNA_pol <- df[!is.na(df$SP.DYN.LE00.IN) & !is.na(df$EN.ATM.CO2E.KT) &
!is.na(df$EN.ATM.PM25.MC.M3) & !is.na(df$NY.GDP.PCAP.CD),]
## adding variables:
# per capita for PM2.5:
df_noNA_pol$PM25_percapita <- df_noNA_pol$EN.ATM.PM25.MC.M3/df_noNA_pol$SP.POP.TOTL
# per capita for CO2 emissions:
df_noNA_pol$CO2_KT_percapita <- df_noNA_pol$EN.ATM.CO2E.KT/df_noNA_pol$SP.POP.TOTL

# one example linear reg: Life expectancy vs. GDP
```

```
# log version
plot(y=df_noNA_pol$SP.DYN.LE00.IN, x=log(df_noNA_pol$NY.GDP.PCAP.CD))
ggplot(df_noNA_pol, aes(x=log(NY.GDP.PCAP.CD), y=SP.DYN.LE00.IN)) +
  geom_point() +
  geom_smooth(method=lm) +
  theme_bw() +
  ggtitle("Life Expectancy at Birth ~ log(GDP per capita)")
model1 <- lm(SP.DYN.LE00.IN ~ log(NY.GDP.PCAP.CD), data=df_noNA_pol)
summary(model1)
plot(y=model1$residuals, x=model1$fitted.values)
plot(model1)


#### air pollution and GDP ####
# LE vs. PM2.5
plot(y=df_noNA_pol$SP.DYN.LE00.IN, x=log(df_noNA_pol$EN.ATM.PM25.MC.M3))
ggplot(df_noNA_pol, aes(x=log(EN.ATM.PM25.MC.M3), y=SP.DYN.LE00.IN)) +
  geom_point() +
  geom_smooth(method=lm) +
  theme_bw() +
  ggtitle("Life Expectancy at Birth ~ log(PM2.5 Pollution)")
# logged version
model2 <- lm(SP.DYN.LE00.IN ~ log(EN.ATM.PM25.MC.M3), data=df_noNA_pol)
summary(model2)
plot(model2)
LEvlogPM25_cluster <- df_noNA_pol[log(df_noNA_pol$EN.ATM.PM25.MC.M3) < 3.0 &
log(df_noNA_pol$EN.ATM.PM25.MC.M3) > 2.5 & df_noNA_pol$SP.DYN.LE00.IN > 70,]
#Japan, Tunisia, Russian Federation, Brazil, Switzerland, Singapore for example

# LE vs. CO2
ggplot(df_noNA_pol, aes(x=log(EN.ATM.CO2E.KT), y=SP.DYN.LE00.IN)) +
  geom_point() +
  geom_smooth(method=lm) +
  theme_bw() +
  ggtitle("Life Expectancy at Birth ~ log(CO2 emissions)")
# logged version with CO2
summary(lm(SP.DYN.LE00.IN ~ log(EN.ATM.CO2E.KT), data=df_noNA_pol))
# per capita version
model3 <- lm(SP.DYN.LE00.IN ~ log(CO2_KT_percapita), data=df_noNA_pol)
summary(model3) # lost significance
plot(model3)
ggplot(df_noNA_pol, aes(x=log(CO2_KT_percapita), y=SP.DYN.LE00.IN)) +
  geom_point() +
  geom_smooth(method=lm) +
  theme_bw() +
  ggtitle("Life Expectancy at Birth ~ log(CO2 per capita)")

####mult with GDP
# logged versions
model4 <-lm(SP.DYN.LE00.IN ~ log(EN.ATM.PM25.MC.M3) + log(NY.GDP.PCAP.CD), df_noNA_pol)
summary(model4)
plot(model4)

model5<- lm(SP.DYN.LE00.IN ~ log(CO2_KT_percapita) + log(NY.GDP.PCAP.CD), df_noNA_pol)
summary(model5) # if control for GDP, PM2.5 becomes insignificant
plot(model5)

#3D scatterplot
```

```
mult_plot <- with(df_noNA_pol, {
  scatterplot3d(SP.DYN.LE00.IN, log(NY.GDP.PCAP.CD), log(EN.ATM.PM25.MC.M3),
                color="blue", pch=19, # filled blue circles
                #type="h",             # lines to the horizontal plane
                main="Life Expectancy ~ PM2.5 + GDP",
                xlab="PM2.5 air pollution (microgram per m^3)",
                ylab="GDP per capita",
                zlab="Life expectancy at birth (SP.DYN.LE00.IN)")
})
mult_plot$plane3d(PM25_model, lty.box = "solid")

# with interaction effects:
model6 <- lm(SP.DYN.LE00.IN ~ log(EN.ATM.PM25.MC.M3) + log(NY.GDP.PCAP.CD) +
log(EN.ATM.PM25.MC.M3)*log(NY.GDP.PCAP.CD), df_noNA_pol)
summary(model6)
plot(model6)
model7 <- lm(SP.DYN.LE00.IN ~ log(CO2_KT_percapita) + log(NY.GDP.PCAP.CD) +
log(CO2_KT_percapita)*log(NY.GDP.PCAP.CD), df_noNA_pol)
summary(model7)
plot(model7)

####simple with just pollution and GDP, no life expectancy
plot(y=log(df_noNA_pol$NY.GDP.PCAP.CD), x=log(df_noNA_pol$EN.ATM.PM25.MC.M3))
ggplot(df_noNA_pol, aes(x=log(EN.ATM.PM25.MC.M3), y=log(NY.GDP.PCAP.CD))) +
  geom_point() +
  geom_smooth(method=lm) +
  theme_bw() +
  ggtitle("GDP per capita ~ log(PM25)")
model8 <- lm(log(NY.GDP.PCAP.CD) ~ log(EN.ATM.PM25.MC.M3), data=df_noNA_pol)
summary(model8)
plot(model8)

plot(y=log(df_noNA_pol$NY.GDP.PCAP.CD), x=log(df_noNA_pol$CO2_KT_percapita))
ggplot(df_noNA_pol, aes(x=log(CO2_KT_percapita), y=log(NY.GDP.PCAP.CD))) +
  geom_point() +
  geom_smooth(method=lm) +
  theme_bw() +
  ggtitle("GDP per capita ~ log(CO2 per capita)")
model9 <-lm(log(NY.GDP.PCAP.CD) ~ log(CO2_KT_percapita), data=df_noNA_pol)
summary(model9)
plot(model9)
```

5.5.3 Biodiversity regressions and assumption checking

```
#the regular linear model
df.species = df[!is.na(df$SP.DYN.LE00.IN) & !is.na(df$EN.BIR.THRD.NO) &
!is.na(df$EN.FSH.THRD.NO) & !is.na(df$EN.HPT.THRD.NO) &
        !is.na(df$EN.MAM.THRD.NO) & !is.na(df$AG.LND.TOTL.K2),]
df.species$threat.prop = 1000 * (df.species$EN.BIR.THRD.NO + df.species$EN.FSH.THRD.NO +
        df.species$EN.HPT.THRD.NO + df.species$EN.MAM.THRD.NO)
        / df.species$AG.LND.TOTL.K2

model.life.species = lm(df.species$SP.DYN.LE00.IN ~ df.species$threat.prop)
summary(model.life.species)
plot(model.life.species)
ggplot(df.species, aes(y=SP.DYN.LE00.IN, x=threat.prop)) +
  geom_point() + geom_smooth(method=lm) + theme_bw() +
  labs(title="Life Expectancy at Birth ~ Density of Threatened Species") +
```

```
  theme(plot.title = element_text(size=16,hjust=0.5))

#linear model against log(species)
model.life.logspecies = lm(df.species$SP.DYN.LE00.IN ~ log(df.species$threat.prop + 1))
summary(model.life.logspecies)
plot(model.life.logspecies)
ggplot(df.species, aes(y=SP.DYN.LE00.IN, x=log(threat.prop+1))) +
  geom_point() + geom_smooth(method=lm) + theme_bw() +
  labs(title="Life Expectancy at Birth ~ log(Density of Threatened Species)") +
  theme(plot.title = element_text(size=16,hjust=0.5))

#linear model against log(log(species))
df.species$threat.prop.log = log(log(df.species$threat.prop + 1) + 0.1)
model.life.log2species = lm(df.species$SP.DYN.LE00.IN ~ df.species$threat.prop.log)
summary(model.life.log2species)
plot(model.life.log2species)
ggplot(df.species, aes(y=SP.DYN.LE00.IN, x=threat.prop.log)) +
  geom_point() + geom_smooth(method=lm) + theme_bw() +
  labs(title="Life Expectancy at Birth ~ log(log(Density of Threatened Species))") +
  theme(plot.title = element_text(size=16,hjust=0.5))
```

5.5.4 Model Selection

```
# model with all predictors
model_all = lm(df_adj.SP.DYN.LE00.IN ~ . ,data = df_bestmodel)

# models with all predictors and all interaction terms
model_interaction = lm(df_adj.SP.DYN.LE00.IN ~ .^2,data = df_bestmodel)

# model via backwards step selection
model_back = step(model_interaction, direction="backward")
summary(model_back)

# model with just intercept
model_cept = lm(df_adj.SP.DYN.LE00.IN~1,data=df_bestmodel)

# model via forwards step selection
model_fore = step(model_cept, scope=list(upper=model_interaction),
                  direction="forward")
summary(model_fore)

# model via stepwise selection
model_stepwise = step(model_cept, scope = list(lower = model_cept, upper = model_interaction),
                direction = "both")
summary(model_stepwise)

# compare AICs
extractAIC(model_all)
extractAIC(model_interaction)
extractAIC(model_cept)
extractAIC(model_back)
extractAIC(model_fore)
extractAIC(model_stepwise)

# compare BICs
extractAIC(model_all, k=log(n))
extractAIC(model_interaction, k=log(n))
extractAIC(model_cept, k=log(n))
```

```
extractAIC(model_back, k=log(n))
extractAIC(model_fore, k=log(n))
extractAIC(model_stepwise, k=log(n))

# compare residual standard error
summary(model_all)$sigma
summary(model_interaction)$sigma
summary(model_cept)$sigma
summary(model_back)$sigma
summary(model_fore)$sigma
summary(model_stepwise)$sigma

# cross-validation study
set.seed(500)
nsims=2000
n=nrow(df_bestmodel)
ntest=140
ssea=sseb=ssef=sses=rep(NA,nsims)

for(i in 1:nsims)
{
  reorder=sample(n)
  train=df_bestmodel[reorder[1:ntest],]
  test=df_bestmodel[reorder[ntest:n],]
  #models a, b, f, s aka all predictors, backward, forward, stepwise
  modela = lm(formula = df_adj.SP.DYN.LE00.IN ~ ., data = df_bestmodel)
  modelb = lm(formula = df_adj.SP.DYN.LE00.IN ~ df_adj.log.gdp + df_adj.log.pm25 +
                df_adj.loglog.threat.prop + df_adj.log.co2percapita + df_adj.log.gdp:df_adj.log.pm25
                df_adj.log.pm25:df_adj.log.co2percapita + df_adj.loglog.threat.prop:df_adj.log.co2pe
              data = df_bestmodel)
  modelf = lm(formula = df_adj.SP.DYN.LE00.IN ~ df_adj.log.gdp + df_adj.loglog.threat.prop +
                df_adj.log.co2percapita + df_adj.loglog.threat.prop:df_adj.log.co2percapita,
              data = df_bestmodel)
  models = lm(formula = df_adj.SP.DYN.LE00.IN ~ df_adj.log.gdp + df_adj.loglog.threat.prop +
                  df_adj.log.co2percapita + df_adj.loglog.threat.prop:df_adj.log.co2percapita,
                data = df_bestmodel)
  ssea[i]=sum(((test$df_adj.SP.DYN.LE00.IN)-predict(modela,new=test))^2)
  sseb[i]=sum(((test$df_adj.SP.DYN.LE00.IN)-predict(modelb,new=test))^2)
  ssef[i]=sum(((test$df_adj.SP.DYN.LE00.IN)-predict(modelf,new=test))^2)
  sses[i]=sum(((test$df_adj.SP.DYN.LE00.IN)-predict(models,new=test))^2)
}
# residual error for each model
c(mean(ssea),mean(sseb),mean(ssef),mean(sses))/(n-ntest)
```