Multiple Regression Analysis: Atmospheric CO_2 Concentration

Gregory Cameron, Edwin Aguirre, Maruthi Mutnuri, Atlanta Liu

Data 604: Statistical Modeling With Data

University of Calgary

Fall 2019

Contents

List of Figures

List of Tables

Introduction

Climate change has been investigated quite extensively in scientific literature, with many researchers attempting to find ways to reduce carbon emissions at the global level. To reinforce the importance of addressing climate change globally, the concept of a "carbon budget" is often utilized by researchers to further develop the climate change narrative (Dahlstrom, 2014). In this context, carbon budgets are defined as the total tolerable amount of carbon emissions that can enter the atmosphere while keeping global warming below a specified level (ie. 1.5 °C or 2.0 °C) (Tokarska, et al, 2019). To explore this narrative in our statistical analysis, we will be utilizing carbon budgets as a tool to communicate the implications behind increasing atmospheric CO_2 concentrations in relation to rising global surface temperatures (see Figure 1).

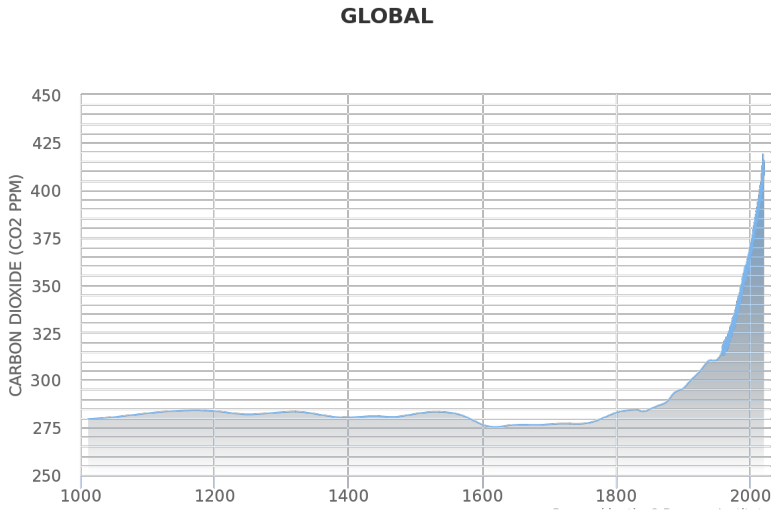


Figure 1: Global CO_2 Levels

Earth has recently surpassed the 400ppm mark and is on the path to increase beyond 450ppm in the upcoming years if carbon emission levels remain unchanged (Hausfather, 2018). There are serious consequences if Earth's atmospheric carbon dioxide concentration remains at 450ppm or greater for too long. Several tipping points have been identified at 450ppm CO_2. Ocean acidification will increase and result in drops of pH levels by more than 0.2 units, bringing about substantial losses in marine life ecology (McNeil & Matear, 2008). This produces a massive cascading effect which could involve the loss of coral reefs, changes in ocean food webs, as well as overall declines in both the fishery and tourism industries. Sustaining CO_2 levels beyond 450ppm will also result in greater ice loss in the arctic regions and many other positive feed backs leading to increasing overall surface temperature (Rogelj, Shindell, & Jiang, 2018).

This report aims to explore the extent that atmospheric CO_2 levels is affected by factors such as time, population count, energy consumption and GDP per capita at the global level. Energy consumption will be measured through fossil fuels and renewable energy sources in terawatt hours. Examples of included fossil fuels include coal, oil, and gas. Examples of renewable energy sources include hydropower, wind, solar, as well as 'other renewables' (geothermal, marine, and biofuels) in terawatt hours. In total, this will amount to 10 explanatory variables altogether. These variables will be used to predict the atmospheric CO_2 levels in 2020, 2030, and 2040.

Methodology

Data Source

We first collected our data in a CSV format from Our World In Data (OWID), a website that hosts open-source data. Additionally, we found projected population growth and renewable energy consumption from IEA Renewables 2019 and World Oil Outlook 2040. The data is then loaded into RStudio, where we wrangled it extract the data required for our analysis. The data we needed for our regression analysis included the dependent variable CO_2 emissions(ppm) as a function of 10 independent variables, including: Coal, Oil, Gas, Hydropower, Solar, Wind, Other Renewables, GDP per capita, Population and Year. The datasets were first read into R using the function "read.csv", creating a separate dataframe for each CSV file.

The range of years of measurement was different between files. Also, in addition to the global measurements, several files contained measurements for individual countries. We filtered the data to a consistent range of years and to only include the global measurements. Additionally, we got rid of columns and rows that were not needed or usable by using the "-c" command. The outcome of the filter stage in the data wrangling was to produce new dataframes that were consistent for each dataset. We ended up with five dataframes which only included the necessary data, global values from the year 1965 to 2016.

The second and final stage of data wrangling was to produce a master dataframe from the previously defined dataframes. The main function used was the "cbind" function, which appends one dataframe to another. Finally, we created a master all dataframe from the individual dataframes including all the dependent and independent variables as columns.

Variable Explanations and Data Assumptions

The data for atmospheric CO_2 levels (response variable) was recorded from Mauna Loa, in the Hawaiian Islands. CO_2 data have been collected here since 1958. Mauna Loa is an ideal location to measure CO_2 for several reasons. It is far from big population centres. Also at night the prevailing winds blow from the land out to sea, bringing clean air from high in the atmosphere to the observatory. The CO_2 levels are measured daily. We are using an annual average of these daily measurements.

Our data on global fossil fuel consumption (coal, oil, and gas) and renewable energy (solar, wind, hydro, and other) is aggregated by multiple countries across the world in terawatt hours. The data represents primary energy consumption rather than final energy consumption, as such it

accounts for the total energy demands of the country instead of simply energy consumed by end-users. OWID has normalized the data into terawatt hours, by converting it from million tonnes of oil equivalent (Mtoe). A potential source of error in these datasets may include the inaccuracy of the reporting. Coal, oil, and gas bear 5 to 10 percent uncertainties in developed countries and much more in many under-developed countries (Wang et al, 2017).

Even though the population dataset downloaded from OWID, the data itself is derived from the United Nations: World Population Prospects. While population projections from the UN have been rather accurate when compared to recent years, they have acknowledged that the largest source of variation occurs in Asia, Africa, and Latin America. Census data in those regions are expected to be less complete, resulting in lower levels of precision. Regardless of this, we will be using this dataset as it provides the most complete information in an open-source format. Population count is measured in millions of people globally.

Gross Domestic Product (GDP) is market value of all the goods and services produced by a country over a certain time period, in this case a year. The global GDP is the cumulative sum of GDP over all the countries across the world. The global GDP per capita (the value used for our modelling) is the global GDP divided by global population. This is measured in 2011 USD currency.

Year was included as an independent variable to account for cyclical changes over time.

The following is a complete list of variables used in our modelling process. All variables were reported annually at a global level (units shown in parenthesis):

1. CO_2 - Atmospheric CO_2 measurement (CO_2 parts per million (ppm)) *Dependent variable
2. Coal - Coal Consumption (Terawatt Hour (TWh)) *Independent Variable
3. Oil - Oil Consumption (Terawatt Hour (TWh)) *Independent Variable
4. Gas - Natural Gas Consumption (Terawatt Hour (TWh)) *Independent Variable
5. Hydropower - Hydro-Electricity Generation (Terawatt Hour (TWh)) *Independent Variable
6. Wind - Wind Electricity Generation (Terawatt Hour (TWh)) *Independent Variable
7. Solar - Solar Electricity Generation (Terawatt Hour (TWh)) *Independent Variable
8. Year - Year of measurement (Year) *Independent Variable
9. Other Renewables - Other sources of renewable energy, including modern biofuels, geothermal, wave and tidal (Terawatt Hour (TWh)) *Independent Variable
10. Population - Global Population (people) *Independent Variable
11. GDP per Capita - Global GDP per Capita (2011 USD) *Independent Variable

Datasets for Prediction (Oil and Other Renewables)

Two datasets will be used to make predictions regarding projected atmospheric carbon dioxide concentration (ppm). The first includes a oil dataset from the World Oil Outlook 2017. For oil consumption to be projected, several assumptions were made based on the previous technological achievements in the past few decades (ie. directional drilling and hydraulic fracturing). It is expected that this trend in technological developments will continue to improve across time. Less developed countries are expected to become potential markets to meet energy demands through petrochemicals. As a result, global oil consumption is expected to increase over time. (Table 2).

The second dataset includes the projected renewable electricity generation by technology in TWh as shown in Table 3. This dataset is retrieved from IEA's article on Renewables 2019: Analysis and Forecast to 2024. Multiple assumptions have been made for each country of interest to create the projected rates. One key assumption includes higher renewable energy penetration markets for countries across Europe, which will foster further technological growth to enhance renewable electricity efficiency gains and widespread usage. In particular, the authors forecast several countries will see a shift in reducing oil dependency, as biofuels become more adopted in road transport and aviation. Since the chart only extrapolated the values to 2024, we utilized the provided compound average annual growth rate to project beyond it.

Modelling Plan

We plan to approach this project using the methods we have learned in Data 603. We will first run a linear regression model using all predictors and test the variables for multicolinearity. We are concerned that since our data are time-series (values for each year between 1965 and 2016) that we will have high multicolinearity between many variables. Once we have removed the variables with high multicolinearity, we will use step-wise regression to recommend a model of main effects. We will then perform a partial f-test to compare our full model and reduced model.

Once we are satisfied with our main effects, we will use the individual t-test to check for significant higher-order terms and interactions. We intend to test this model with another f-test to evaluate if the higher order terms and interactions are significant. Any significant higher-order or interaction terms will be added to our main effects to produce our final model. Our model will then test for the following 6 assumptions as shown below:

1. Linearity Assumption - Review residual plots
2. Independence Assumption - Review residual against year (time)
3. Normality Assumption - Using Shapiro-Wilk normality test
4. Equal Variance Assumption (heteroscedasticity) - Using Breusch-Pagan test
5. Multicollinearity - Using variance inflation factors (VIF)
6. Outliers - check Cook's distance and leverage

If our model does not satisfy any of these assumptions, we will review our workflow above to see if any improvements can be made. Once we find our model to satisfy all of the assumptions, we will use the model to predict future atmospheric CO_2 levels.

Results

Variable Selection Procedures

We built a first-order model that consisted of every explanatory variable as a base comparison as shown below. This will be helpful as we continue to select different variables through various selection procedures.

First order Model

$$\widehat{Y_{CO_2}} = \beta_0 + \beta_1 X_{Oil} + \beta_2 X_{Gas} + \beta_3 X_{Coal} + \beta_4 X_{Hydropower} + \beta_5 X_{Wind} + \beta_6 X_{Solar} + \beta_7 X_{Year} + \beta_8 X_{OtherRenewables} + \beta_9 X_{Population} + \beta_{10} X_{GDP\ per\ Capita}$$

To begin the variable selection procedure, we decided to drop our predictive variables based on their VIF values rather than utilizing the stepwise regression and all possible regressions procedure to screen for significance. The justification for this step lies in the inherent multicolinearity of our variables. With extremely large VIF values all across our predictive variables, running a stepwise regression may eliminate important predictors due to their multicolinearity. So we decided to run several VIF tests and individually remove predictors with the highest VIF values as having higher VIF values invalidates the model and its p-values.

After multiple iterations of checking and dropping predictors based on their VIF values, we arrived at the final model that included: Oil, Year, and Other Renewables. The respective VIF values is: Oil = 14.4345, Year = 22.5088, Other = 6.9391. We decided to keep year in the model despite its high VIF value because it helps us explain the seasonality trend in atmospheric CO_2 concentrations. Additionally, keeping year as a predictive variable in the model helps us pass the linearity assumption.

Independent Variable	Oil(VIF value)	Gas(VIF value)	Coal(VIF value)	Hydropower(VIF value)	Wind(VIF value)	Solar(VIF value)	Year(VIF value)	Other Renewables(VIF	Population(VIF	GDP per Capita(VIF
Stage 1	114.10736	1232.37	190.2563	281.41529	364.99718	71.77171	20846.65	548.50573	25012.789	930.33339
Stage 2	112.45587	1011.31	190.1882	276.66037	336.14375	71.76767	608.8018	215.45772		735.86833
Stage 3	87.96069		167.3119	275.90286	330.34461	71.71897	394.7048	137.18313		735.84248
Stage 4	19.16007		59.84481	274.92261	199.95806	48.63308	394.2036	135.98313		
Stage 5	18.38754		59.44218		147.69674	46.06194	104.8979	12.28028		
Stage 6	15.812312		31.28694			6.940387	35.26547	59.26302		
Stage 7	15.231675					5.107297	35.25879	25.05411		
Stage 8	14.434475						22.50883	6.939145		

Figure 2: Table of VIF values as variables are removed sequentially

Hypothesis Statement for Individual T-tests:

$$H(0) : \beta_i = 0$$
$$H(A) : \beta_i \neq 0$$
$$i = \textit{Oil, Other Renewables, and Year}$$

Main Effects Individual T-tests:

$$Oil : t = -1.89,\ p = 0.0648$$
$$Year : t = 38.26,\ p < 0.001$$
$$OtherRenewables : t = 20.62,\ p < 0.001$$

Individual T-tests were also used in our variable selection to determine the best predictors based on a significance level of $\alpha = 0.05$. From the results of these tests, we would reject the null hypothesis in favor of the alternative. This suggests that year and other renewables are significant predictors of atmospheric CO_2 concentration on their own. In addition, while oil is slightly above our specified α, we decided to keep it in the model because the the p-value is close (0.0648) and we believe oil to be an important predictor for atmospheric CO_2. For this reason, these variables will be added to our model for further comparison between interaction and higher order terms. Our main effects model is shown below:

$$\widehat{Y_{C02}} = \beta_0 + \beta_1 X_{Oil} + \beta_2 X_{Year} + \beta_3 X_{OtherRenewables}$$

Hypothesis Statement for Individual T-tests (Higher Order Terms):

$$H(0) : \beta_i = 0$$
$$H(A) : \beta_i \neq 0$$
$$i = Oil^2,\ OtherRenewables^2,\ and\ Year^2$$

When checking for higher-order terms, all of the values appeared to be significant based on the t-test. But since they were not significant when interaction terms were accounted for, they were not kept in our model. A partial F-test will be later conducted in comparison to our best-fitted model to justify this removal.

$$\widehat{Y_{C02}} = \beta_0 + \beta_1 X_{Oil} + \beta_2 X_{Year} + \beta_3 X_{OtherRenewables} + \beta_4 X_{Oil}^2 + \beta_5 X_{Year}^2 + \beta_6 X_{OtherRenewables}^2$$

Higher Order Individual T-tests:

$$Oil^2 : t = 2.250,\ p = 0.0294$$
$$Year^2 : t = 2.214,\ p = 0.0320$$
$$OtherRenewables^2 : t = 2.399,\ p = 0.0207$$

Looking into the presence of possible interaction effects between our predictive variables, two out of three significant interaction terms emerged. This includes oil and year, as well as oil and other renewables. After removing the non-significant interaction term from the individual t-test(Year $*$ Other: p = 0.6688) and re-running a summary of individual t-test, we are left with the results below:

Hypothesis Statement for Individual T-tests (Interaction Terms):

$$H(0) : \beta_i = 0$$
$$H(A) : \beta_i \neq 0$$
$$i = Oil * Year,\ Oil * OtherRenewables,\ OtherRenewables * Year$$

Interaction Term T-tests:

$$Oil * Year : t = 6.089,\ p < 0.001$$
$$Oil * OtherRenewables : t = 3.366,\ p = 0.0016$$

Since these two interaction terms are significant predictors of atmospheric CO_2 levels, they will be added to our model. This also makes practical sense since oil consumption changes by the year, when there is recession less Oil is consumed compared to year when economy is booming. Similarly 'other renewable' energy consumption impacts the use of oil, the more renewable energies are used the less oil is needed to be used to meet the energy demand. The interaction model is shown below:

$$\widehat{Y_{CO_2}} = \beta_0 + \beta_1 X_{Oil} + \beta_1 X_{Year} + \beta_3 X_{OtherRenewables} + \beta_4 X_{Oil \times Year} + \beta_5 X_{Oil \times OtherRenewables}$$

Hypothesis Statement for ANOVA Test:

$$H(0) :\ \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0 : Higher\ order\ terms\ are\ not\ significant$$
$$H(A) :\ at\ least\ one\ \beta_p \neq 0 : At\ least\ one\ higher\ order\ term\ is\ significant$$

We conducted a final ANOVA test to ensure that no higher order terms are significant in the presence of interaction variables. To do this, we compared our reduced model (main effect + interaction) with the full model (main effect + interaction + higher order). From the results of the ANOVA (F = 0.9843, p = 0.4091), we failed to reject the null hypothesis. This indicates that the higher order terms do not significantly predict atmospheric CO_2 concentration. As a result, they will be left out from our model. Table 1 below summarizes the results of the partial F-test.

Source of Variation	Df	Sum of Squares	Mean Squares	F-Statistic	P value (>F)
Regression	43	0.75329	0.017518372	0.9843	0.4091
Residual	3	10.96971	3.65657		
Total	46	11.723			

Table 1: ANOVA Table

Best Fitted Model: Included Interaction Effects

$$\widehat{Y_{CO_2}} = \beta_0 + \beta_1 X_{Oil} + \beta_1 X_{Year} + \beta_3 X_{OtherRenewables} + \beta_4 X_{Oil \times Year} + \beta_5 X_{Oil \times OtherRenewables}$$

Multiple Regression Assumptions

The sections below will address how we tested our model to meet various assumptions associated with running multiple regression. These assumptions must be tested, to ensure that our model results are, to an extent, trustworthy.

Linearity Assumption

Our model relies on the assumptions that the true relationship between our predictors and response variables are linear in nature. Using residual plots as shown in Figure 3, we check to see if there are any discernable patterns that are non-linear. From the plot, we see that there is no prominent patterns showing in the trend of our data, suggesting that it passes the linearity assumption.

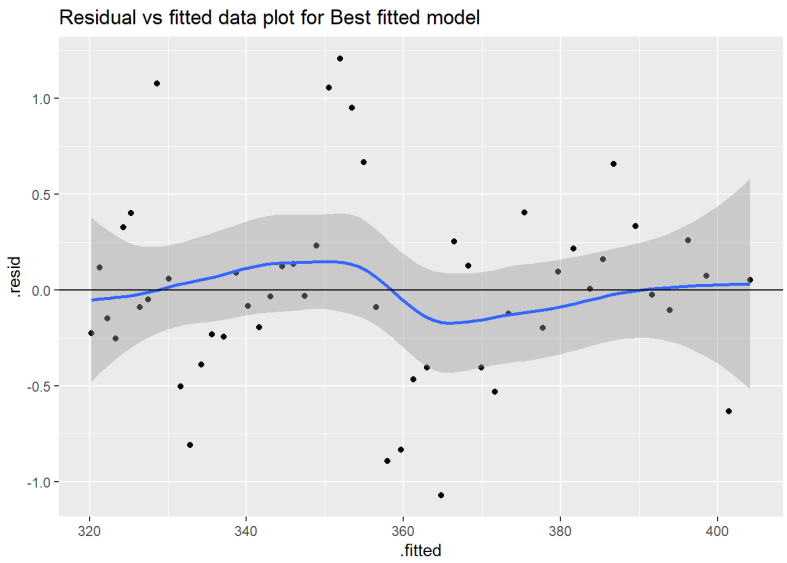


Figure 3: Plot to check for linearity

Independence Assumption

When we time plot the residuals against the years (time) as shown in Figure 4, we can observe a pattern over 25 years between 1975 to 2010. In some sections of the plot the ϵ_i is providing little information about the sign of ϵ_{i+1}. So, the assumption of independent errors is not completely satisfied because successive errors appear to be slightly correlated across certain years (ie. 1965-1975, 1975-1985). This occurs because the data for the atmospheric CO_2 concentrations (dependent) and all predictive variables are observed sequentially over a period of time from 1965 to 2016. Despite the slight trends occurring throughout these years, we will be using this model for predictions in future years.

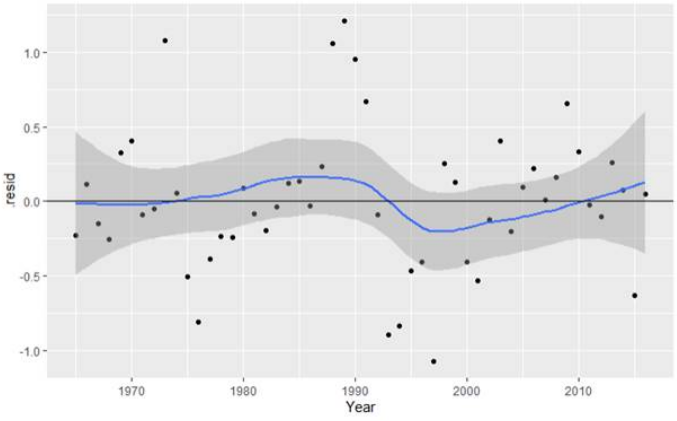


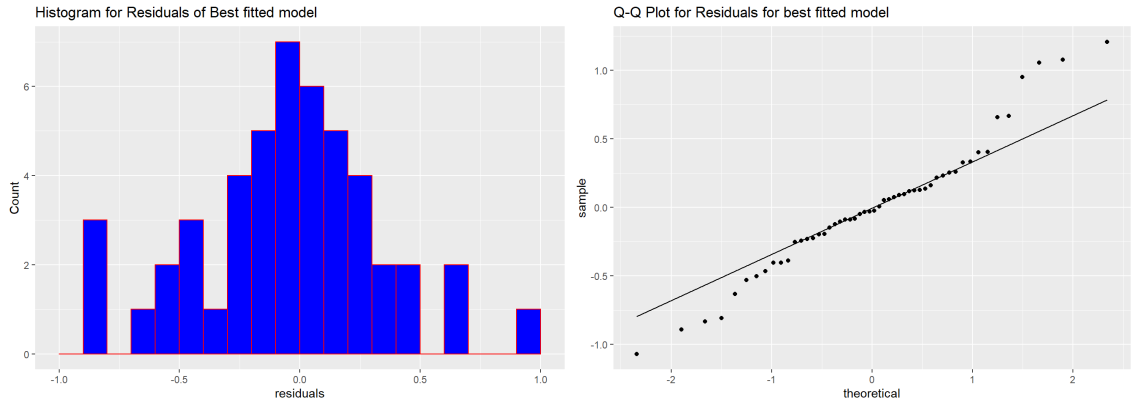
Figure 4: Plot to check for independence of error terms

Normality Assumption

In order for our multiple regression analysis be held valid, our residuals must be normally distributed (Figure 5). To test this, we can see that the distribution of the residuals in the histogram follows a fairly normal trend with some data points occurring near the tail ends. Additionally, a normal probability plot of residuals is provided. Again, we see that most of the data points approximate the normal line, however, there are a few points flaring outwards near the tails indicating the presence of possible outliers.

Null Hypothesis H(0): The sample data is normally distributed
Alt. Hypothesis H(A): The sample data is not normally distributed

This suspicion is further confirmed in our Shapiro-Wilk normality test. Based on $\alpha = 0.05$, the results of the Shapiro-Wilks test (W = 0.9653, p = 0.1329) reveal that our data is indeed normally distributed. Since the p-value is greater than 0.05, we fail to reject the null hypothesis. Overall, our data successfully meets the normality condition.



(a) Histogram of Residuals

(b) Quantile-Quantile Plot

Figure 5: Plots to check for normal distribution

Equal Variance Assumption

Null Hypothesis H(0): Heteroscedascity is not present
Alt. Hypothesis H(A): Heteroscedascity is present

Next, we tested to see if our data is homoscedasctic through a plot of fits to residuals as well as the studentized Breusch-Pagan test. Looking at the plot of fits to residuals, we see that there is an inverted wedge-like pattern to the plotted data as the fitted values increase. This is an indicator that the data is not have common variance. From the results of the Breusch-Pagan test (BP = 13.409, p = 0.01983), we would reject the null hypothesis in favor of the alternative, suggesting that our model fails to be homoscedastic.

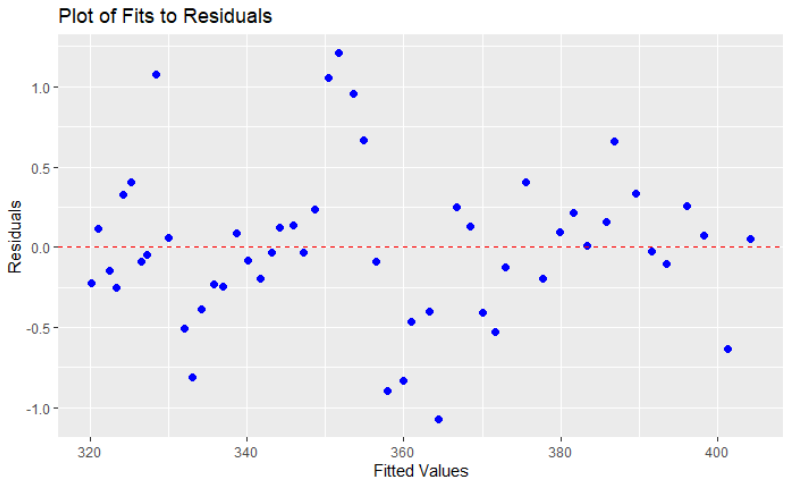


Figure 6: Plot to check for homoscedascity

Multicolinearity Tests

To test for multicolinearity in our models, we looked at multiple variance inflation factors (VIF) to determine which variables should remain in our best fitted model. After several tests, our final model included population and 'other renewables' as they had the least VIF values (Oil = 14.4345, Year = 22.5088, Other Renewables = 6.9391) when compared to the other variables from our dataset. In addition, we also ran a ggpairs function to ensure that there were no extremely high correlations (r > 0.80) in our model (Figure 7).

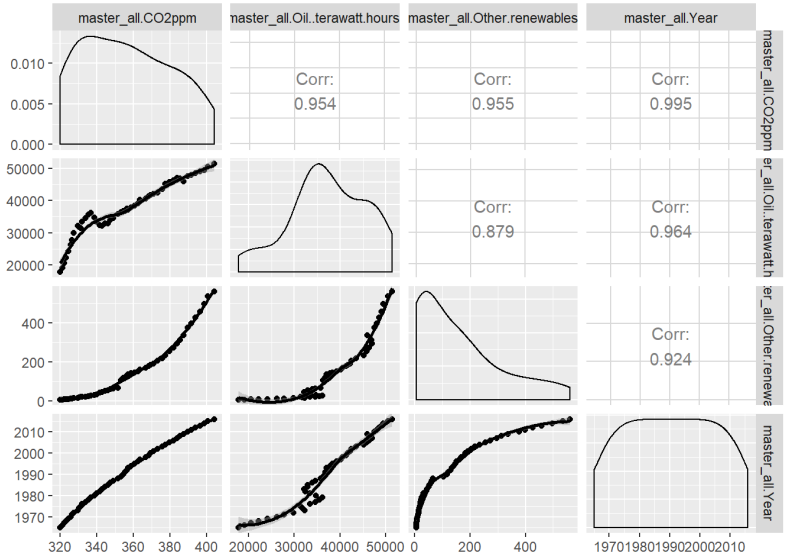


Figure 7: Plots to check for multicolinearity

Influential Points and Outliers

Influential subjects can have a large effect on our model, to check for this we plot the values against Cook's distance which is shown as a red dashed line (Figure 10). From the residuals vs leverage plot on the left, we see that there are no points beyond Cook's distance. This signifies that there are no influential points that have a disproportional effect on our regression results.

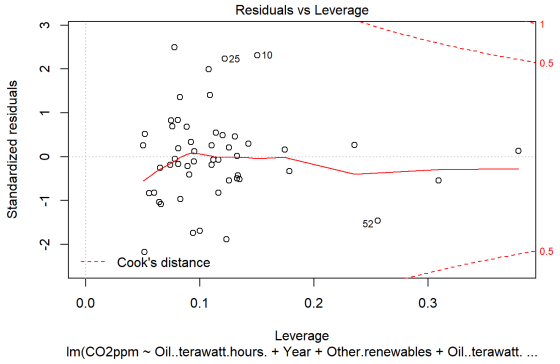
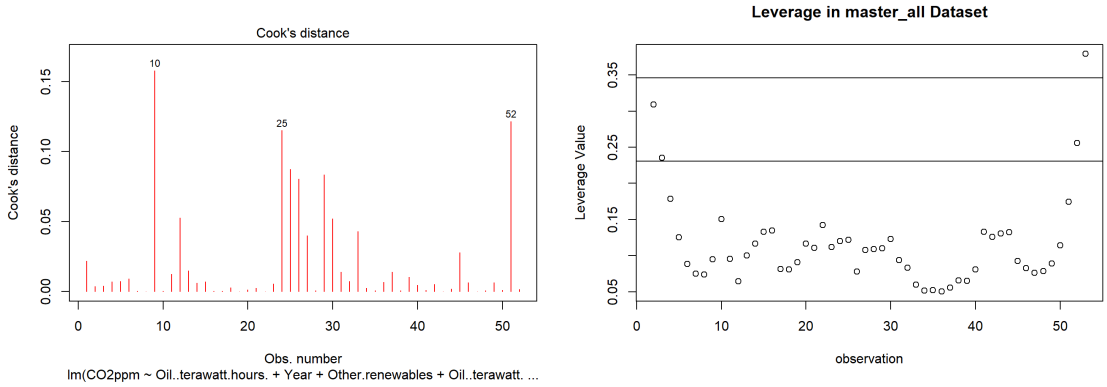


Figure 8: Plot to check for influential points

The plot on the left shows Cook's distance plotted for each observation (Figure 9). This plot helps us indicate the overall influence the outlier points have on our regression by clearly identifying the observation number and the extent of its effect. The most prominent points of interest include observation number 10, 25, and 52 as they show the highest Cook's distance. However, their Cook's Distance value is all less than 0.5, so they are not influential. Next we used the leverage plot on the right to remove outliers beyond 2p/n and 3p/n thresholds. Our model was then refitted for both of these thresholds, but there were no substantial changes to our R^2_{adj} value.



(a) Cook's Distance

(b) Leverage Values vs Observation Plot

Figure 9: Plots to check for outliers

Our final, best-fitted model includes main effects and interactions. We can express the final model in 3 different ways, as shown below.

(1) Final model expanded with all terms

$$\begin{aligned}\widehat{Y_{CO_2}} &= \beta_0 + \beta_1 X_{Oil} + \beta_2 X_{Year} + \beta_3 X_{OtherRenewables} + \beta_4 X_{Oil\times Year} + \beta_5 X_{Oil\times OtherRenewables} \\ &= -1927.6009 - 0.0202(X_{Oil}) + 1.1461(X_{Year}) + 0.0424(X_{OtherRenewables}) + 0.000010163(X_{Oil\times Year}) \\ &\quad + 0.0000010589(X_{Oil\times OtherRenewables})\end{aligned}$$

(2) Final model with Oil terms collected

$$\begin{aligned}\widehat{Y_{CO_2}} &= \beta_0 + \beta_2 X_{Year} + \beta_3 X_{OtherRenewables} + (\beta_1 + \beta_4 X_{Year} + \beta_5 X_{OtherRenewables})(X_{Oil}) \\ &= -1927.6009 + 1.1461(X_{Year}) + 0.0424(X_{OtherRenewables}) + (-0.0202 + 0.000010163(X_{Year}) + \\ &\quad 0.0000010589(X_{OtherRenewables}))(X_{Oil})\end{aligned}$$

(3) Final model with Year and Other Renewables terms collected

$$\begin{aligned}\widehat{Y_{CO_2}} &= \beta_0 - \beta_1 X_{Oil} + (\beta_2 + \beta_4 X_{Oil})(X_{Year}) + (\beta_3 + \beta_5 X_{Oil})(X_{OtherRenewables}) \\ &= -1927.6009 - 0.0202 X_{Oil} + (1.1461 + 0.000010163 X_{Oil})(X_{Year}) \\ &\quad + (0.0424 + 0.0000010589 X_{Oil})(X_{OtherRenewables})\end{aligned}$$

R^2adj and RMSE of Best Fitted Model

$R^2_{adj} = 0.9996$, this value indicates that 99.96 percent of the variation of the response variable atmospheric CO_2 concentrations is explained by the final model containing the predictors oil, year, other renewables as well as the interactions between $oil \times year$, $oil \times other\ renewable$.

$RMSE = 0.5048$, this value indicates that the standard deviation of the unexplained variation in estimation of response variable atmospheric CO_2 concentrations is 0.5048 ppm.

Interpreting Coefficients

There are a total of three β_i (i = Oil, Year, Other Renewables) coefficients in our final model. These are interpreted below in relation to atmospheric CO_2 concentration levels.

$$\begin{aligned}\widehat{\beta_{Oil}} &= -0.0202 + 0.000010163(X_{Year}) + 0.0000010589(X_{OtherRenewables}) \\ &= -0.02021 + 0.000010163 * (2020) + 0.0000010589 * (1) \\ &= 0.000314\end{aligned}$$

This value suggests that the effect of Oil consumption on CO_2 concentration in the atmosphere (in ppm) changes by year and amount of other renewables consumption. For year 2020, when other renewables consumption is 1 terawatt-hour, then increasing oil consumption by 1 terawatt-hour leads to an increase in CO_2 presence in atmosphere by 0.000314 ppm.

$$\begin{aligned}\widehat{\beta_{Year}} &= 1.146 + 0.000010163(X_{Oil}) \\ &= 1.146 + 0.000010163 * (1) \\ &= 1.1460\end{aligned}$$

This value suggests that the effect of year on CO_2 concentration in the atmosphere (in ppm) changes by oil consumption. Which means, when other renewables consumption is held constant and Oil consumption is 1 terawatt-hour, then every subsequent Year leads to an increase in CO_2 presence in atmosphere by 1.1460 ppm.

$$\begin{aligned}\widehat{\beta_{Other}} &= -0.0424 - 0.0000010589(X_{Oil}) \\ &= -0.0424 - 0.0000010589 * (1) \\ &= -0.0424\end{aligned}$$

This value suggests that the effect of other renewables consumption on CO_2 concentration in the atmosphere (in ppm) changes by oil consumption. Which means, when year is held constant and oil consumption is 1 terawatt-hour, then increasing other renewable consumption by 1 terawatt-hour leads to an decrease in CO_2 presence in atmosphere by 0.04241 ppm

Predicted Atmospheric CO_2 Levels

To predict future atmospheric CO_2 concentration, we used the projected data from the population and renewable energy dataset to extrapolate beyond the year of 2016. In 2020, the expected atmospheric CO_2 level is 419.4836ppm. In 2030, the expected atmospheric CO_2 level is 440.9800ppm. And finally in 2040, the expected atmospheric CO_2 level is 460.9179ppm. The results of our predictions are shown below:

$$\begin{aligned}\widehat{Y_{CO_2|Year=2020}} &= -1927.6009 - 0.0202(X_{Oil}) + 1.1461(X_{Year}) + 0.0424(X_{Other}) + 1.0163 \times 10^{-5} * (X_{Oil \times Year}) \\ &+ 1.0588 \times 10^{-6}(Oil \times Other) \\ &= -1927.6009 - 0.0202 * (57272) + 1.1461 * (2020) + 0.0424 * (739) \\ &+ 1.0163 \times 10^{-5} * (115689743) + 1.0588 \times 10^{-6}(42324119) \\ &= 419.4836 ppm\end{aligned}$$

$$\begin{aligned}\widehat{Y_{CO_2|Year=2030}} &= -1927.6009 - 0.0202(X_{Oil}) + 1.1461(X_{Year}) + 0.0424(X_{Other}) + 1.0163 \times 10^{-5} * (X_{Oil \times Year}) \\ &+ 1.0588 \times 10^{-6}(Oil \times Other) \\ &= -1927.6009 - 0.0202 * (60746.95) + 1.1461 * (2030) + 0.0424 * (740.3) \\ &+ 1.0163 \times 10^{-5} * (123316309) + 1.0588 \times 10^{-6}(44970967) \\ &= 440.9800 ppm\end{aligned}$$

$$\begin{aligned}\widehat{Y_{CO_2|Year=2040}} &= -1927.6009 - 0.0202(X_{Oil}) + 1.1461(X_{Year}) + 0.0424(X_{Other}) + 1.0163 \times 10^{-5} * (X_{Oil \times Year}) \\ &+ 1.0588 \times 10^{-6}(Oil \times Other) \\ &= -1927.6009 - 0.0202 * (62484.35) + 1.1461 * (2040) + 0.0424 * (741.6) \\ &+ 1.0163 \times 10^{-5} * (127468074) + 1.0588 \times 10^{-6}(46338394) \\ &= 460.9179 ppm\end{aligned}$$

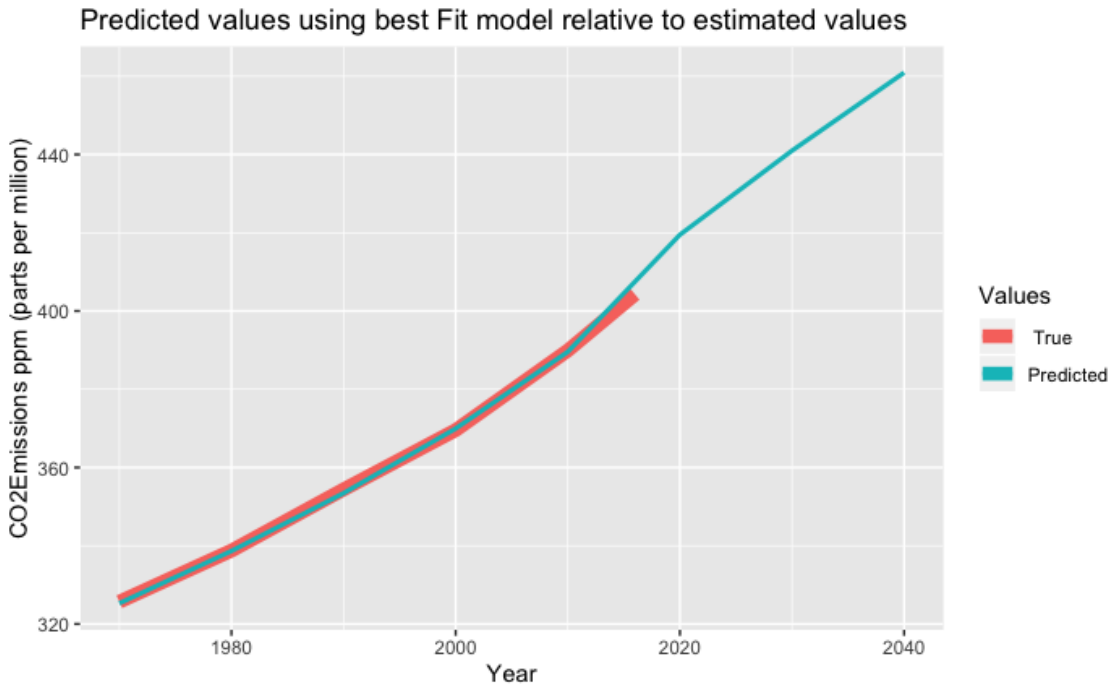


Figure 10: True vs Predicted Atmospheric CO_2 levels

The figure above shows the true atmospheric CO_2 levels (measured) compared to the results from our best fitted model. For historic data, our model results matches the measured data quite well. This is expected given our high R^2adj and low RMSE. The figure also shows our predictions for the future. We predict that the important threshold of 450 ppm atmospheric CO_2 will be crossed in the late 2030's, exhausting our carbon budget.

Conclusion

To summarize our findings from the analysis, the main effects of oil, year, and other renewable are shown be significant with less multicolinearity (VIF values) compared to the other predictor variables. Even though higher order terms were significant, they were borderline significant (p-values close to $\alpha = 0.05$). Interactions between oil and year, and oil and other renewable were

significant (based on individual coefficient test: p-values). After we combined the main effects, higher order terms, and interactions together, we found that all the terms in our model were not significant through the partial F-test. Our terms all had p-values greater than $\alpha = 0.05$. To fix this, we removed the higher order terms from our model and only kept the significant interactions. Overall, our best fitted model includes oil, year, other renewables, as well as the interactions between $oil \times year$, and $oil \times other\ renewables$.

To examine the practical implications of our model, we first must identify how each term affects CO_2 emissions. The inclusion of oil as a predictive variable confirms the generally accepted notion that higher oil consumption is linked to increases in CO_2 concentration in the atmosphere. Similarly, the variable 'year' implicitly accounts for various factors not kept in our model, such as population, GDP, and renewable energy consumption through its ability to consider changes across time and seasonality trends in CO_2 emissions. Finally, other renewable energy consumption explains the reduction in the need for fossil fuel consumption to meet the global energy demand, thereby reducing the rate of increase in atmospheric CO_2 concentration. Our interaction terms between oil consumption changes by the year also explains how recessions negatively affect oil consumption when compared to years of economic boom. Additionally, the interaction between oil usage and renewable energy consumption affects how energy demand is met in different countries around the world.

Discussion

While atmospheric carbon dioxide concentrations have been largely attributed to anthropogenic emissions, it can be difficult to narrow down the concept of carbon budgets and to include all the various factors involved. Our model considers energy consumption in the form of fossil fuels and renewable energy to explain atmospheric CO_2 emissions, however, it does not take into account methane emissions from agricultural livestock, deforestation, and other natural sources. Since we did not account for these factors, our model results in large levels of uncertainty in its prediction.

When we began this project, we expected to find that our strong predictors would be the most carbon-intensive sources of energy: coal, oil, and gas (in that order). We also expected population and GDP per capita to be strong predictors as there have been numerous studies relating these variables to atmospheric CO_2 levels.

The main challenge we encountered in our modeling was the high level of multicolinearity between most of our predictor variables. This is primarily because we are using time-series data. As the data were annual global measurements, the year-to-year trend was much stronger than the variation between variables. As such, variables that you might not expect to exhibit multicolinearity, such as coal and hydroelectricity, had extremely high values. We have not covered time-series analysis in Data 603. We would have benefited from using a time-series method to remove the time based effects on the data before running our linear regression model. This would have resulted in different and probably more useful model.

Multicolinearity can lead to unstable and biased standard errors, resulting in very unstable p-values (Vatcheva el al, 2018). Only after removing variables with the highest multicolinearity did we observe stable, realistic p-values. This led to our chosen method of removing variables one-by-one based on the multicolinearity values before confirming with a partial f-test and testing for the assumptions.

There are several ways we could improve our model. As mentioned above, these data would benefit from a time-series workflow. If we had more time to research these methods, we would incorporate this into our workflow. This would most likely result in a more useful model. We expect that once we removed the time-series effects, we would have much lower multicolinearity between variables. This would potentially allow for a more stable model that could be used to produce more useful prediction.

List of Tables

World primary energy demand by fuel type

	Levels *mboe/d*				Growth *% p.a.*
	2015	2020	2030	2040	2015–2040
Oil	86.5	92.3	97.9	100.7	0.6
Coal	78.0	80.7	85.8	86.2	0.4
Gas	59.2	65.2	79.9	93.2	1.8
Nuclear	13.5	15.8	20.1	23.8	2.3
Hydro	6.8	7.5	9.0	10.3	1.7
Biomass	28.0	30.1	34.0	37.3	1.2
Other renewables	3.8	6.6	12.9	20.0	6.8
Total world	**276.0**	**298.2**	**339.4**	**371.6**	**1.2**

Table 2: Projected Global Energy Demand

Table 6.5. Total renewable electricity generation by technology (TWh)

	2018e	2019	2020	2021	2022	2023	2024	CAAGR
Hydropower	4 203	4 258	4 385	4 483	4 537	4 591	4 648	2%
Pumped storage	115	125	129	134	139	144	149	4%
Bioenergy	546	599	640	683	715	746	761	6%
Wind	1 268	1 389	1 534	1 698	1 852	1 998	2 135	9%
Onshore wind	1 202	1 307	1 433	1 571	1 697	1 808	1 921	8%
Offshore wind	66	82	101	126	155	190	214	22%
Solar PV	585	720	864	1 005	1 151	1 309	1 480	17%
CSP	13	16	18	20	25	26	26	12%
Geothermal	90	94	98	101	106	111	116	4%
Marine	1	1	1	1	1	1	1	3%
Total	6 707	7 076	7 542	7 991	8 387	8 783	9 168	5%

Notes: TWh = terawatt hour. Generation data refer to gross electricity production and include electricity for own use. Renewable electricity generation includes generation from bioenergy, hydropower (including pumped storage), onshore and offshore wind, solar PV, solar CSP, geothermal, and ocean technologies. Generation from bioenergy includes generation from solid, liquid and gaseous biomass (including cofired biomass), and the renewable portion of municipal waste. The time series for onshore and offshore wind generation is estimated because wind generation data are only available at the aggregate level. CAAGR = compound average annual growth rate. Please refer to regional definitions in the glossary.For OECD member countries, 2018 generation data are based on IEA statistics published in Renewables Information 2019

Table 3: Projected Renewable Electricity Generation

References

Climate Change: Atmospheric Carbon Dioxide: NOAA Climate.gov. (2019, September 19). Retrieved from https://www.climate.gov/news-features/understanding-climate/climate-change-atmospheric-carbon-dioxide.

Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. Proceedings of the National Academy of Sciences, 111(Supplement-4), 13614–13620. doi: 10.1073/pnas.1320645111

Hausfather, Z. (2018). Analysis: Why the IPCC 1.5C report expanded the carbon budget. (2018, October 10). Retrieved from https://www.carbonbrief.org/analysis-why-the-ipcc-1-5c-report-expanded-the-carbon-budget.

IEA (2019), Renewables 2019: Analysis and forecasts to 2024, IEA, Paris, https://doi-org.ezproxy.lib.ucalgary.ca/10.1787/b3911209-en.

Malt, B. (n.d.). Educating Students on the Psychology of Sustainability. Retrieved from https://www.psychologicalscience.org/observer/educating-students-on-the-psychology-of-sustainability.

McNeil, B. I., & Matear, R. J. (2008, December 2). Southern Ocean acidification: A tipping point at 450-ppm atmospheric CO_2. Retrieved from https://www.pnas.org/content/105/48/18860.short.

Organization of the Petroleum Exporting Countries. (2017). *World Oil Outlook 2040.* Retrieved from https://www.opec.org/opec$_web/flipbook/WOO2017/WOO2017/assets/common/downloads/WOO$%202017.*pdf*

Rogelj, J., Shindell, D., & Jiang, K. (2018). *Global warming of 1.5 °C: an Ipcc special report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.* Geneva, Switzerland: Intergovernmental Panel on Climate Change.

Tokarska, K. B., Schleussner, C.-F., Rogelj, J., Stolpe, M. B., Matthews, H. D., Pfleiderer, P., & Gillett, N. P. (2019). Recommended temperature metrics for carbon budget estimates, model evaluation and climate policy. Nature Geoscience, 12(12), 964–971. doi: 10.1038/s41561-019-0493-5

Vatcheva KP, Lee M, McCormick JB, Rahbar MH. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. Epidemiology (Sunnyvale). 2016;6(2):227. doi:10.4172/2161-1165.1000227

Wang, Y., Broquet, G., Ciais, P., Chevallier, P., Vogel, F., Kadygrov, N., Wu, L., Yin, Yi., Wang, R. & Tao, S (2017) Estimation of observation errors for large-scale atmospheric inversion of CO_2 emissions from fossil fuel combustion, Tellus B: Chemical and Physical Meteorology, 69:1, DOI: 10.1080/16000889.2017.1325723