Regression and Forecast Analysis of UNSDG Data

Emmanuelle Lamarche
and Beatrice Lee

Overview

- Intro:
  - Analyzing the ONU Sustainability of Countries Development dataset, which contains important data to assess the sustainability of a country's development, and can be used to track whether countries are achieving their sustainable development goals.
- Motivation:
  - Both care a lot about sustainability and want to learn more about what type of factors impact sustainability on a large scale
- Methods:
  - We hope to predict greenhouse gas emissions based on various factors after performing simple linear regression, multiple linear regression, and time series analysis.

Dataset Description: Data

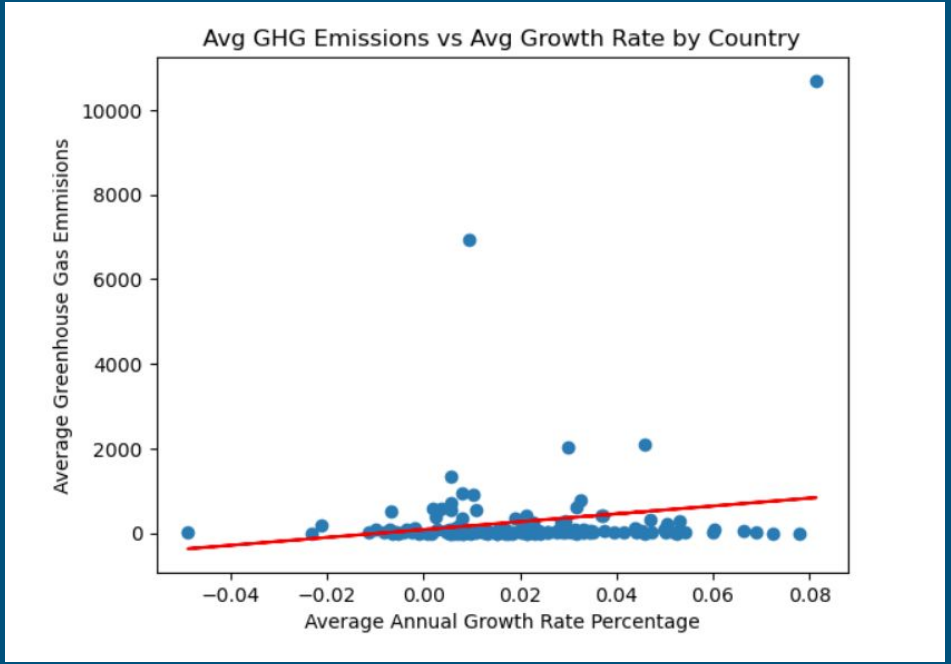
This dataset contains data from the United Nations Sustainable Development Goals Indices for many countries over the time period from 2002 to 2021.

Relevant Columns include:

- Greenhousegas_emissione_mtco2equivalent, which represents total greenhouse gas emissions (GHG) per year in tonnes of carbon dioxide equivalent
- Annual_growth_rate_perc, which is the annual growth rate of real gdp per capita
- Proportion_of_population_with_primary_reliance_on_clean_fuels_a, proportion of population with primary reliance on clean fuels and technology is calculated as the number of people using clean fuels and technologies for cooking, heating and lighting divided by total population reporting that any cooking, heating or lighting, expressed as percentage
- Mortality_rate_perc, mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease
- Level_of_development, country ONU level of development
- Region, country ONU region
- Year, year data was collected

SLR

- Independent variable: Average Annual growth rate percent (real gdp per capita) by country from data from 2002-2021
- Dependent variable: Average greenhouse gas emissions (GHG) per year in tonnes of carbon dioxide equivalent by country from data from 2002-2021
- We decided to do the average by country from 2002-2021 to have cleaner data and account for missing data.



SLR Significance

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	82.8333	130.520	0.635	0.527	-175.261	340.927
annual_growth_rate_perc	9279.4103	4426.342	2.096	0.038	526.624	1.8e+04

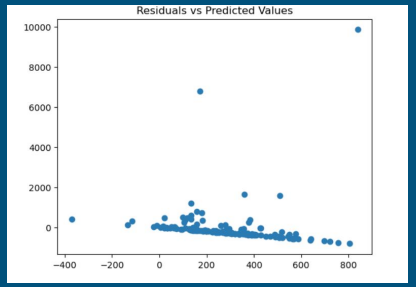
- For a country with an average growth rate of 0.00%, the estimated average greenhouse gas emissions per year would be 82.833 tonnes of carbon dioxide equivalent. Since P > |t| = 0.527 > .05, there is not enough evidence to suggest that this intercept is significant.
- For an increase in growth rate of 1%, the expected increase in greenhouse gas emissions per year would be 9229.4103 tonnes of carbon dioxide equivalent. Since P > |t| = 0.038 < .05, there is evidence to suggest a significant relationship between average growth rate and average greenhouse gas emissions.
- 139 countries and one parameter -> df = 137.

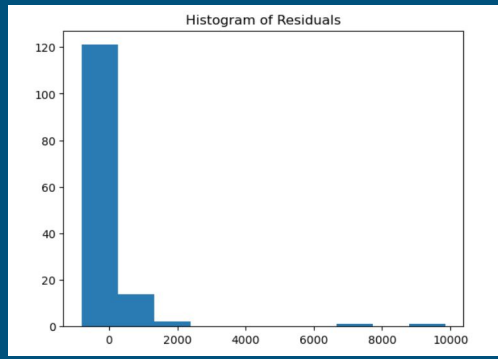
SLR Diagnostics

- Normality - Shapiro-Wilk Test
  - Very small, significant p value, which indicates the model does not have normally distributed residuals
- Mean of Residuals = 0
  - Close to 0
  - However, histogram is not normally distributed
- Constant Variance
  - The plot of residuals vs predicted regression values (right) has a negative linear trend, which indicates non-constant variance.
- Independence
  - For k= 1 and n = 137 at $\propto$ = .05, dU ~ 1.7
  - The Durbin Watson test results in a value of 2.020, which is > 1.7 so fail to reject the hypothesis that the error terms are not autocorrelated
- Overall, not a very good fit for the model

```
ShapiroResult(statistic=0.3347156047821045, pvalue=1.9587312778086958e-22)
```

```
Mean of residuals: -1.7012131112299377e-13
```



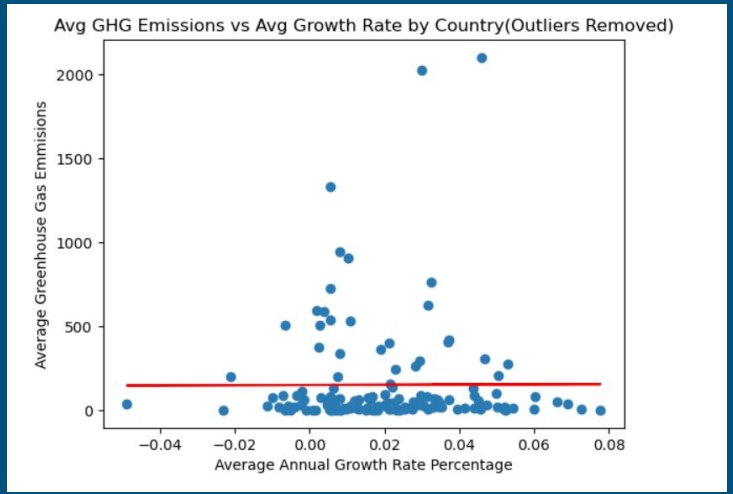



```
Durbin-Watson:                   2.020
```

Remove Outliers

	student_resid	unadj_p	bonf(p)
China	15.409677	1.256344e-31	1.746318e-29
United States of America	7.333236	1.822859e-11	2.533775e-09

- There seem to be a few unusual observations that are separated from the data or have irregularly large residuals, which may be the cause of some of the misfit of the data. So, we tested the studentized residuals of the data and found the data points with SRES >2.
- We remove these outliers to see if it improves the regression relation.
    - The regression changes significantly, including no longer having a significant relationship between average growth rate and average greenhouse gas emissions.



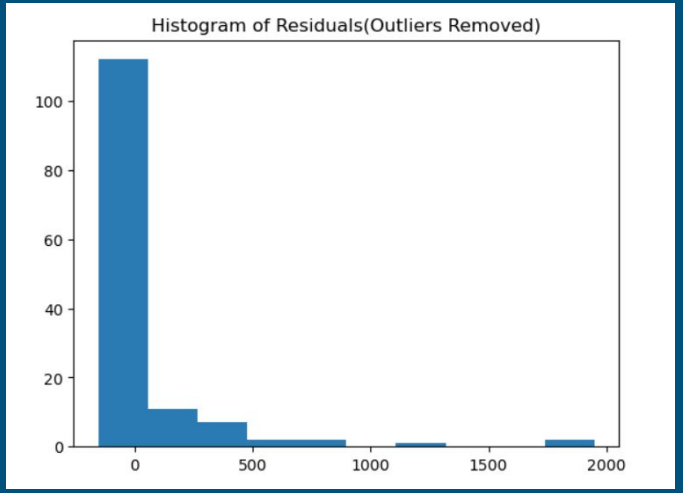
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	149.4619	38.615	3.871	0.000	73.093	225.831
annual_growth_rate_perc	62.2156	1337.816	0.047	0.963	-2583.572	2708.003

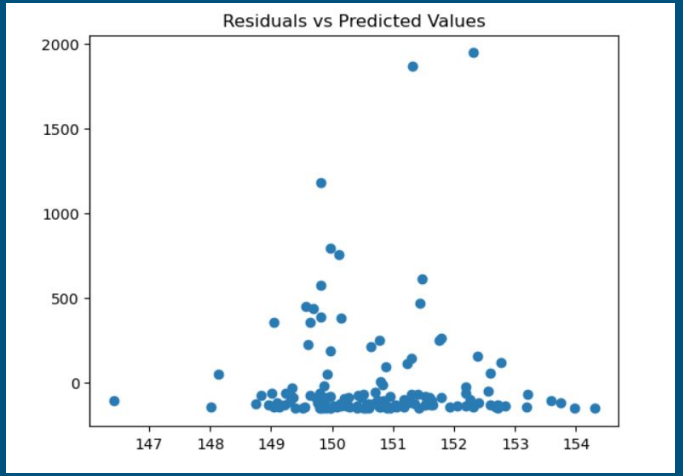
Rerun Diagnostics

Still indicates that model is not a good fit

```
Mean of residuals: 8.256817776131967e-14
```

```
ShapiroResult(statistic=0.4985344409942627, pvalue=1.1269621110286968e-19)
```





SLR Other Independent Variables

- Tested other variables as a predictor of greenhouse gas emissions
  - Renewable energy share on total energy consumption
  - Proportion of population with primary reliance on clean fuels
  - Mortality Rate percentage
- All also not significant, no need to further verify diagnostics

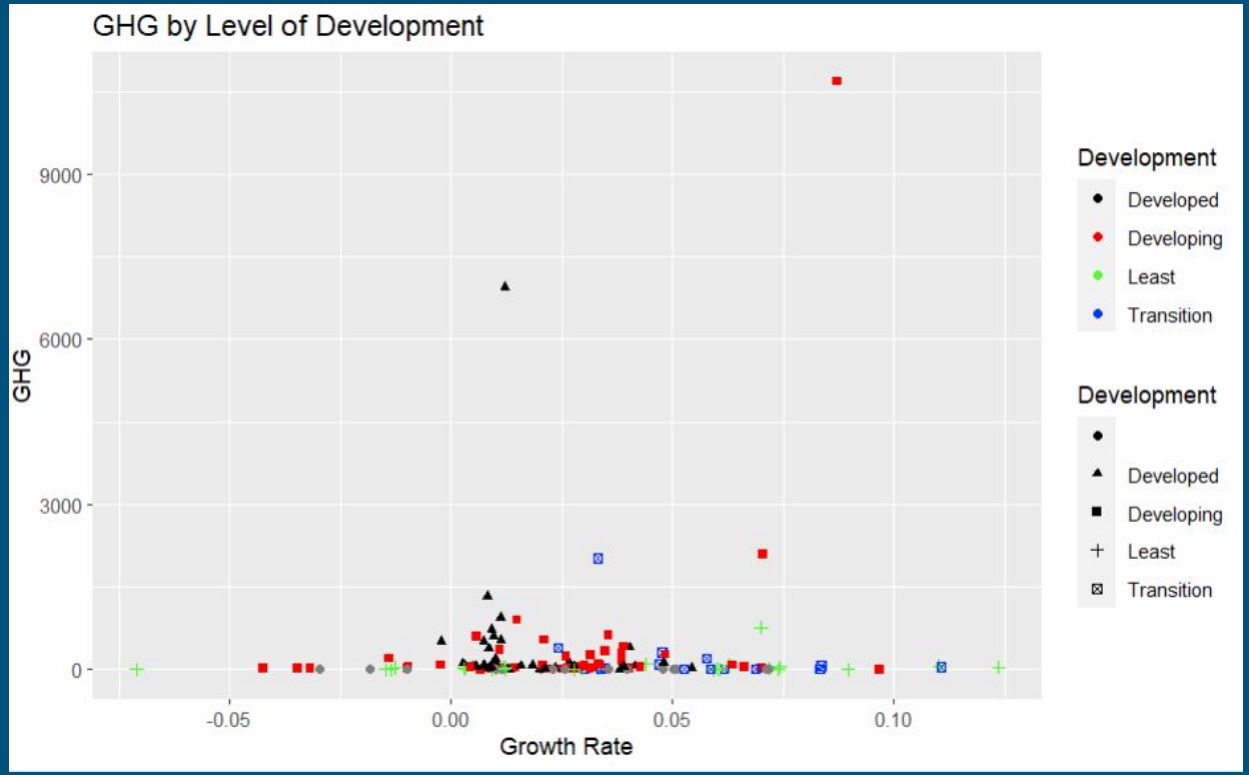
	coef	std err	t	P>|t|
Intercept	405.4529	136.704	2.966	0.004
renewable_energy_share_on_the_total_energy_consumption	-4.5711	3.535	-1.293	0.198

Intercept	160.3542	222.721	0.720	0.473
proportion_of_population_with_primary_reliance_on_clean_fuels_a	1.7243	2.875	0.600	0.550

Intercept	528.7871	277.818	1.903	0.059
mortality_rate_perc	-1157.3703	1279.917	-0.904	0.368

MLR with Qualitative Variables

- First wanted to see if there was an impact by level of development that could make the model more accurate
- Created dummy variables for this qualitative category

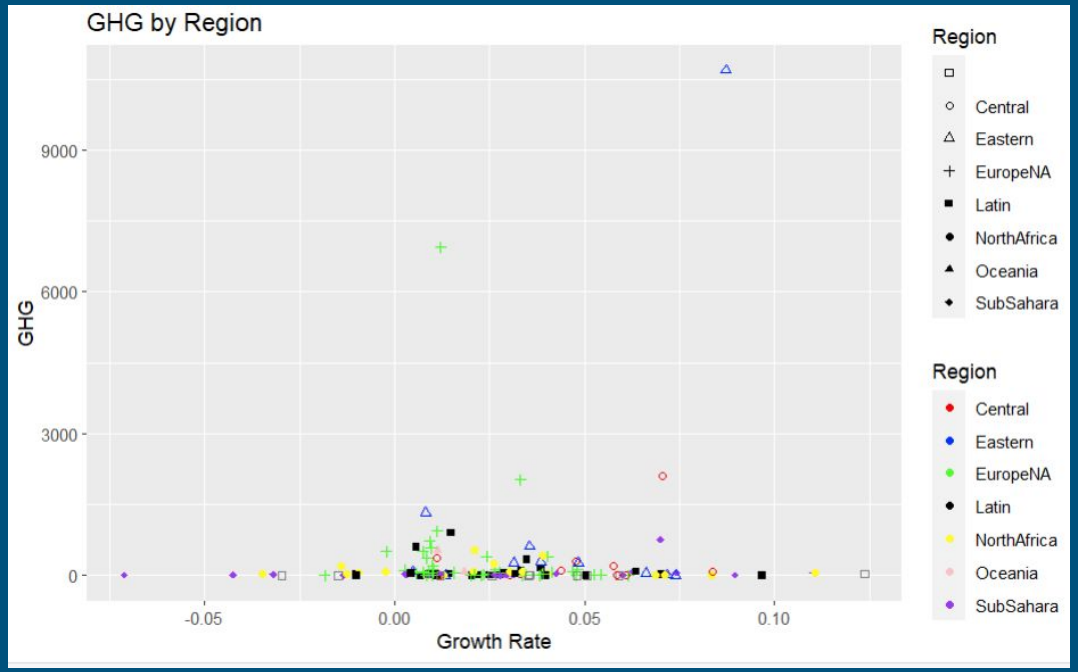


MLR Significance

```
Call:
lm(formula = greenhousegas_emissione_mtco2equivalent ~ annual_growth_rate_perc +
    Developed + Developing + Transition + Least, data = byCountry)

Residuals:
   Min     1Q Median     3Q    Max
-819.7 -345.9 -188.9   66.6 9918.4

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            -128.94     272.06  -0.474   0.6363
annual_growth_rate_perc 5427.49    3253.25   1.668   0.0976 .
Developed               424.21     318.33   1.333   0.1849
Developing              425.83     308.05   1.382   0.1692
Transition               44.91     397.16   0.113   0.9101
Least                   -14.16     339.51  -0.042   0.9668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1101 on 133 degrees of freedom
Multiple R-squared:  0.04547,   Adjusted R-squared:  0.009582
F-statistic: 1.267 on 5 and 133 DF,  p-value: 0.2819
```

No significant relationship between any of the levels of development, as shown by the Pr(>|t|)

MLR with Qualitative Variables




```
Call:
lm(formula = greenhousegas_emissione_mtco2equivalent ~ annual_growth_rate_perc +
    Central + Eastern + NorthAfrica + SubSahara + Oceania + EuropeNA +
    Latin, data = unsdg)

Residuals:
    Min      1Q  Median      3Q     Max
-1235.9  -330.4  -225.5     6.6 11062.8

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)               10.92     202.13   0.054   0.9569
annual_growth_rate_perc   -50.23     697.28  -0.072   0.9426
Central                   142.99     235.21   0.608   0.5433
Eastern                  1229.95     238.52   5.157 2.95e-07 ***
NorthAfrica               113.11     226.10   0.500   0.6170
SubSahara                  65.03     238.04   0.273   0.7848
Oceania                   295.73     267.57   1.105   0.2693
EuropeNA                  371.19     204.50   1.815   0.0698 .
Latin                     218.28     226.80   0.962   0.3360
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1060 on 1178 degrees of freedom
  (3989 observations deleted due to missingness)
Multiple R-squared:  0.05152,   Adjusted R-squared:  0.04508
F-statistic: 7.999 on 8 and 1178 DF,  p-value: 1.524e-10
```

Dummy Variables for Region of the World

Also not Significant, except for Eastern Asia (likely due to China outlier, as seen in SLR)

Stepwise MLR

Decided to include all potentially relevant quantitative variables and do stepwise regression to see if any combination of them would provide a significant relationship

Original:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1441.450	753.335	1.913	0.0584	.
annual_growth_rate_perc	3717.734	3102.064	1.198	0.2335	
renewable_energy_share_on_the_total_energy_consumption	-12.117	6.216	-1.949	0.0540	.
proportion_of_population_with_primary_reliance_on_clean_fuels_a	-7.522	5.542	-1.357	0.1777	
mortality_rate_perc	-1693.405	1562.662	-1.084	0.2810	

Stepwise MLR

```
Start:  AIC=1536.24
greenhousegas_emissione_mtco2equivalent ~ annual_growth_rate_perc +
    renewable_energy_share_on_the_total_energy_consumption +
    proportion_of_population_with_primary_reliance_on_clean_fuels_a +
    mortality_rate_perc

                                                      Df Sum of Sq       RSS    AIC
- mortality_rate_perc                                  1   1483440 132858203 1535.5
- annual_growth_rate_perc                              1   1814402 133189165 1535.7
- proportion_of_population_with_primary_reliance_on_clean_fuels_a  1   2326883 133701645 1536.2
<none>                                                            131374762 1536.2
- renewable_energy_share_on_the_total_energy_consumption  1   4799487 136174249 1538.2

Step:  AIC=1535.47
greenhousegas_emissione_mtco2equivalent ~ annual_growth_rate_perc +
    renewable_energy_share_on_the_total_energy_consumption +
    proportion_of_population_with_primary_reliance_on_clean_fuels_a

                                                      Df Sum of Sq       RSS    AIC
- proportion_of_population_with_primary_reliance_on_clean_fuels_a  1   1131691 133989893 1534.4
- annual_growth_rate_perc                              1   1504860 134363063 1534.7
<none>                                                            132858203 1535.5
- renewable_energy_share_on_the_total_energy_consumption  1   3891081 136749284 1536.6

Step:  AIC=1534.39
greenhousegas_emissione_mtco2equivalent ~ annual_growth_rate_perc +
    renewable_energy_share_on_the_total_energy_consumption

                                                    Df Sum of Sq       RSS    AIC
- annual_growth_rate_perc                            1   1925224 135915117 1534.0
<none>                                                          133989893 1534.4
- renewable_energy_share_on_the_total_energy_consumption  1   3144457 137134351 1534.9

Step:  AIC=1533.95
greenhousegas_emissione_mtco2equivalent ~ renewable_energy_share_on_the_total_energy_consumption

                                                    Df Sum of Sq       RSS    AIC
<none>                                                          135915117 1534.0
- renewable_energy_share_on_the_total_energy_consumption  1   2913846 138828963 1534.3
>
```

Stepwise MLR

Step 1: Remove Mortality rate

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	856.465	525.896	1.629	0.1064	
annual_growth_rate_perc	3367.344	3087.727	1.091	0.2780	
renewable_energy_share_on_the_total_energy_consumption	-10.647	6.072	-1.754	0.0824	.
proportion_of_population_with_primary_reliance_on_clean_fuels_a	-4.567	4.829	-0.946	0.3465	

Step 2: Remove Proportion of population with primary reliance on clean fuels

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	392.325	188.862	2.077	0.0402	*
annual_growth_rate_perc	3771.985	3056.416	1.234	0.2199	
renewable_energy_share_on_the_total_energy_consumption	-6.352	4.027	-1.577	0.1177	

Step 3: Remove annual growth rate percentage

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	513.451	161.750	3.174	0.00196	**
renewable_energy_share_on_the_total_energy_consumption	-6.107	4.032	-1.515	0.13283	

MLR Conclusions

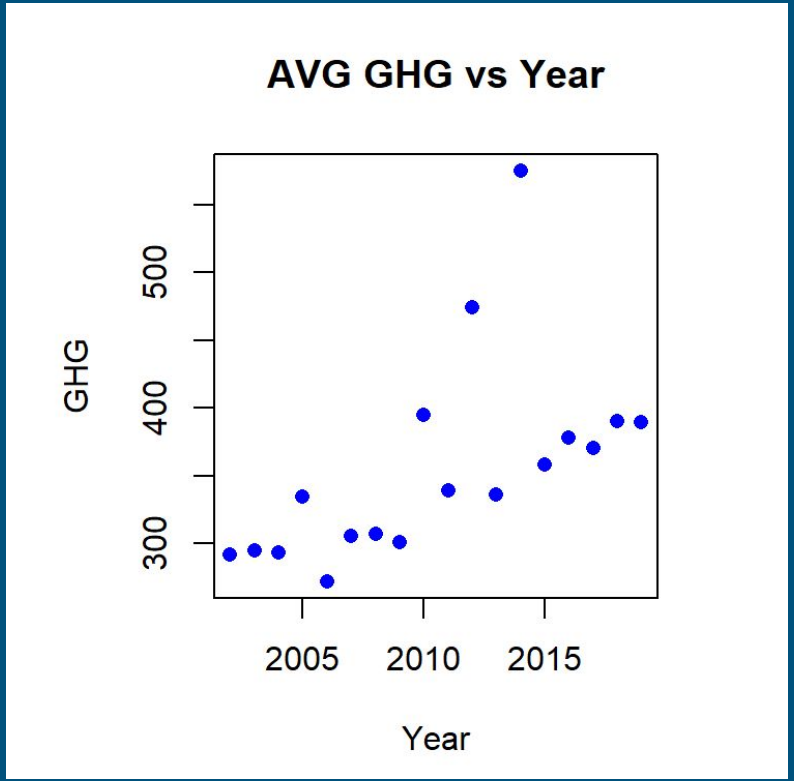
- Multiple Linear Regression also doesn’t seem to account for the variations in the data, when considering both quantitative and qualitative variables

Analysis of GHG Emissions vs Year

To perform forecasting analysis on time-related data, we opted to look at the average yearly GHG Emissions in mt co2 equivalent collected in the UN SDG data.

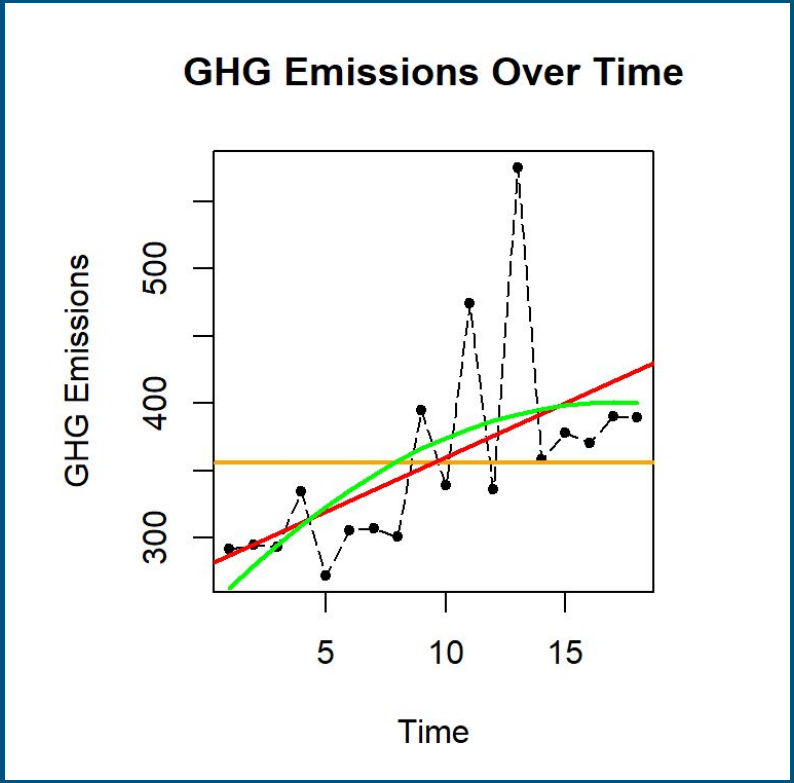
A preliminary plot of the data is shown to the left.

We elected to try and forecast future data using a Time Series Regression Model and/or an Exponential Smoothing Model



Different Time Series Regression Analysis

- Ran Three Different Models to see what trend best fit the yearly GHG Emissions data:
  - No Trend (Orange)
  - Linear Trend (Red)
  - Quadratic Trend (Green)
- We can quickly see there is potential for either the linear trend model or the quadratic trend model to be better models to fit the data than no trend at all



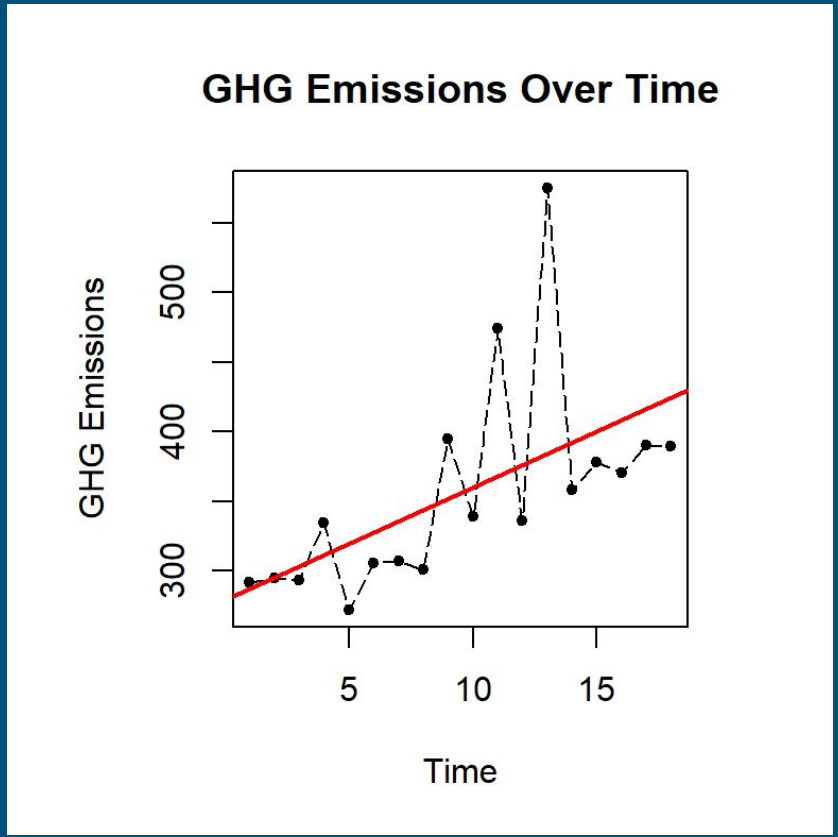
Different Time Series Regression Analysis

After running the models we found:

- No Trend Time Series Model: y = 355.9
  - The intercept value is significant
- Linear Trend Time Series Model: y = 279.027 + 8.092t
  - Both the intercept and slope values are significant
  - Adjusted R^2 of 0.2938
- Quadratic Trend Time Series Regression Model: y = 244.8416 + 18.3470t + -0.5398t^2
  - The intercept is significant, but the other two values are not, since their p values are too large

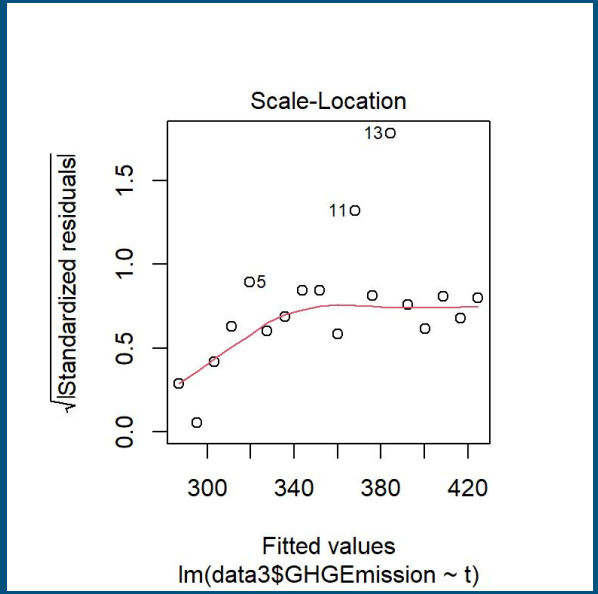
Best Time Series Model

- The Linear Trend model fits the data best
- All coefficients are significant
- This model is
  - Y= 279.027 + 8.092t
- There is a relatively small R^2 which indicates that there are more factors at play influencing the variability in the data than the change over time alone
- Next, we should check our underlying assumptions



Checking Assumptions of Time Series Regression

- We know from the provided data that he observations in the time series are collected in chronological order, and the sampling frequency is consistent, with observations collected yearly.
- The Cook's Distance plot shown to the left indicates that the point at year 13 is particularly influential to the fit of the data
  - The data used at year 13 is accurate, and, since we only have access to 18 years of data, we elected to leave this point in our data

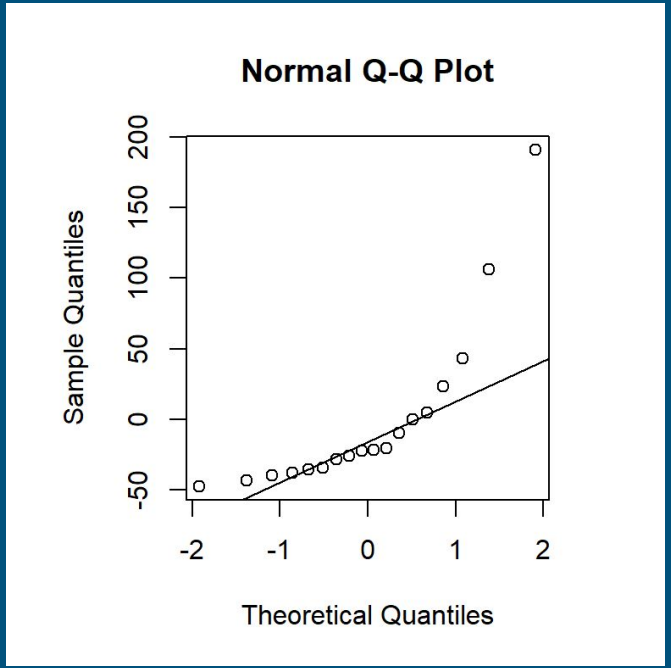


Checking Assumptions of Time Series Regression

- The Independence assumption, or checking for autocorrelation:
  - The Durbin-Watson test is used to check for the presence of autocorrelation in the residuals of a regression model
  - For the positive Autocorrelation Test the DW statistic is around 2.56, which is close to 2, suggesting little evidence of autocorrelation. The p-value is 0.8355, which is greater than the typical significance level of 0.05.
  - For the Negative Autocorrelation Test, the DW statistic is still around 2.56. The p-value is 0.1645, which is greater than 0.05.
  - In both cases, the high p-values suggest that there is not enough evidence to reject the null hypothesis of no autocorrelation, supporting the idea that the residuals do not exhibit a significant pattern of autocorrelation.

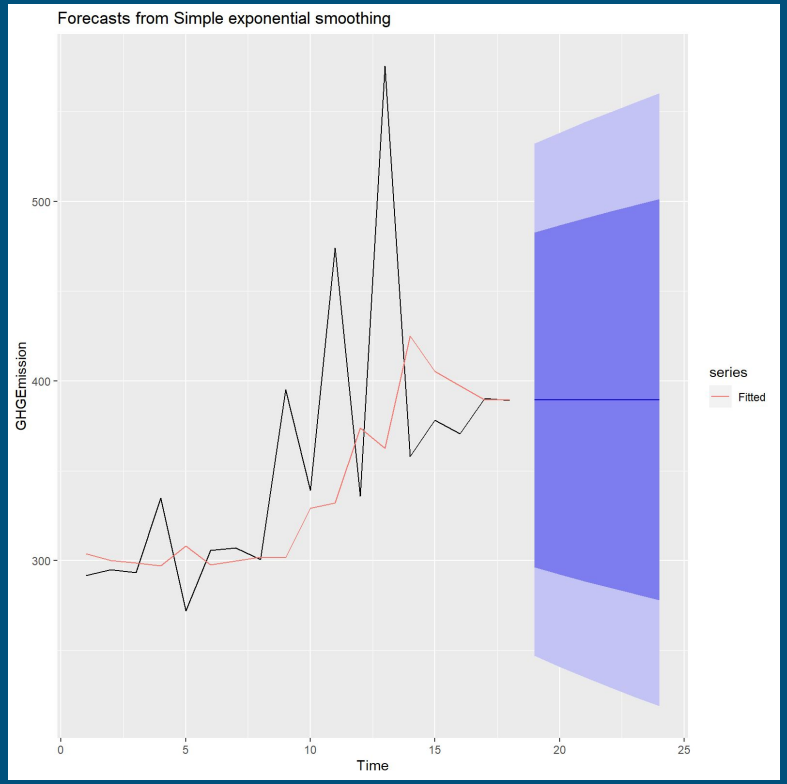
Checking Assumptions of Time Series Regression

- Normality of residuals/ underlying dataset
  - When running the Anderson-Darling normality test, the small p-value of 3.805e-05 suggests that we have sufficient evidence to reject the null hypothesis that the residuals of model2 follow a normal distribution.
  - We can also see from the normal Q-Q plot that the observed values in the dataset have higher values in the tails compared to what would be expected in a normal distribution.
- The mean of the Residuals is equal to 0



Simple Exponential Smoothing Model

- Results:
  - Alpha: 0.2937509
  - l (Initial Estimate): 303.7496488
- In the graph to the left there were 6 periods for forecasting, which are depicted in the blue area of the graph.
- The simple exponential smoothing model is used when forecasting a time series where the mean (level) of the time series is slowly changing over time, but there is no trend or seasonal factor.
- Smoothing models generally work well to remove rapid fluctuations in Time Series data



Measures of Forecast Accuracy

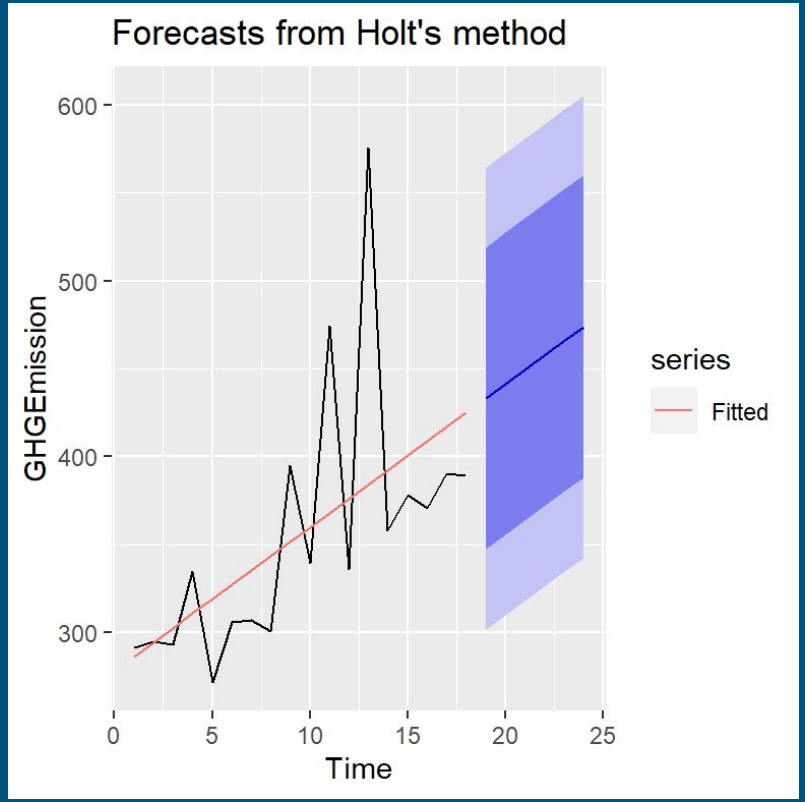
- When Running we get MAPE of 9.786
- A MAPE of 9.786% suggests a moderate level of forecasting error, meaning that the model's predictions are off by about 9.786% from the actual values.
- We get a MAD of 40.671
- A MAD of 40.671 means that the model's predictions deviate from the actual values by approximately 40.671 mt co2 equivalent on average.

Conclusion:

This model provides us with generally moderate accuracy. Ideally, the MAPE and MAD should be as small as possible.

Holt Exponential Smoothing Model

- Results:
  - Alpha: 1.000024e-04
  - Beta: 1.000003e-04
  - l (Initial Intercept Estimate) : 278.2201
  - B (Initial Slope Estimate): 8.156631
- In the graph to the left there are 6 periods for forecasting, which are depicted in the blue area of the graph.
- Holt's Model is used when both the level and growth rate are changing.
- Additive or multiplicative Holt-Winters Models are unnecessary since there are no seasons.
- NOTE: Since the Alpha and Beta parameters are close to 0, we have essentially re-achieved a Time Series Regression with trend.



Measures of Forecast Accuracy

- When Running we get MAPE of 10.368
- A MAPE of 10.368% suggests a moderate level of forecasting error, meaning that the model's predictions are off by about 10.37% from the actual values.
- We get a MAD of 40.987
- A MAD of 40.987 means that the model's predictions deviate from the actual values by approximately 40.987 mt co2 equivalent on average.

Conclusion:

This model provides us with generally moderate accuracy. Ideally, the MAPE and MAD should be as small as possible.

Checking Assumptions

Simple Linear Smoothing:

- Underlying data is the same as the time series data, so we know The underlying data passes the independence assumption, but has violations in the assumption that there are no outliers and the assumption that the residuals follow a normal distribution
- Checked smoothing parameters to ensure they remain constant over time
- Assumes no trend/ stationary data, which we can see from other models does not hold true. This assumption is violated.

Holt's Exponential Smoothing:

- Does not strictly require stationary data
- Underlying data is the same as the time series data, so we know The underlying data passes the independence assumption, but has violations in the assumption that there are no outliers and the assumption that the residuals follow a normal distribution
- Checked smoothing parameters to ensure they remain constant over time

Time Series Regression Model Conclusions

- Out of the potential Time Series Regression Models considered, the Linear trend model gave the most promising results.
- The underlying data passes the independence assumption, but contains other issues.
- The violation of the assumptions that there are no outliers and the assumption that the residuals do not follow a normal distribution indicate to us as data scientists that we should tread lightly in making broad conclusions and that we should consider trying other models to find a better fit.
- It may be valid to note the underlying data set is small with only 18 years of data.

Exponential Smoothing Conclusions

- If we were to use Exponential Smoothing, we should use Holt Model Exponential Smoothing
- The Holt Model Exponential Smoothing does not provide us with a better model than the one provided by Time Series Regression because the alpha and beta parameters are close to 0, making it essentially the same model as the Time Series model.
- We are able to test and see that there is moderate accuracy to this model's predictions.

General Conclusions

- Regression Techniques did not fit the data well
  - Could pitfalls be due to quality of the dataset?
    - Contained extensive data, but inconsistent with amount of reporting depending on years and country
- The relationship between the predictors we analyzed and GHG emissions were not significant, which went against what we originally predicted/expected
- Leads to follow-up questions of what may be actual significant predictors of GHG emissions by country?
  - Would require more research and other datasets

General Conclusions

- Running Time Series Regression Analysis and Exponential Smoothing Analysis did not give us particularly strong forecasting results, but we were able to find rudimentary potential models.
- If a data scientist were to use one of our models, they would carefully consider the intricacies of the data the models were built around and the potential inaccuracy of the models before using one of them.
- Generally, we could conclude there is a positive trend in the data indicating that average GHG emissions are increasing over the years. This indicates that future forecasts will indicate this trend will continue.