Presented to: Mohsen Farhadloo, Ph.D.

BSTA 450 – Statistical Models for Data Analysis

John Molson School of Business

Analysis of Global Warming

Presented By:

Alain Euksuzian 40070126

April 17, 2022

Executive Summary

The focus of the project will be on analyzing and understanding the causes of global warming. To achieve said goal, explanatory variables such as Co2 emission, Transportation emission, total fossil fuel consumed by residential and commercial areas, renewable energy from residential and commercial areas, and the number of electric vehicles will be used to predict the average temperature in the United States.

Throughout the report, I will disprove the null hypothesis that global warming is sorely caused by commercial pollution. My report will show that other variables are more harmful to our environments, such as transportation emissions and CO2 emissions by the population.

My final report aligns with expert opinions, which expresses that the largest component of global warming is transportation (27%), as petroleum is the most popular compound used in cars, trucks, planes, etc.

Using various elimination methods (forward, stepwise, etc.), I will remove the least significant variables from the model and use the regression to predict the average temperature from a sample dataset.

Recommendation:

Hence, to mitigate the issue of global warming issue, an alternative should be discovered for transportation, energy production, and commercial use as they constitute the three largest shares of global warming.

Introduction

The subject chosen for my term project is the analysis of global warming in the United States. It has often been a popular topic during elections and has many contradicting views. As it is a topic that affects the world population, I will aim at understanding the subject using the statistical tools learned during this semester.

Throughout the project, I will rely on all explanatory variables mentioned to determine the average temperature per year as an accurate representation. Using the dataset and built models, I will predict sampled temperatures and compare them to expert predictions in my references. Due to the complexity of the topic and the variables used, the project will cover the United States rather than per state. Some of the variables were not available on a per-state basis, and some variables were published on a national level by the federal government.

Using various statistical tools available in SAS such as summary statistics, regression analysis and backward/forward selection, I will aim at answering if the chosen variables are statistically meaningful questions such as:

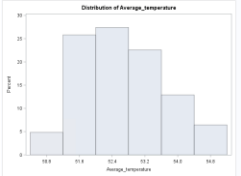
- Can the variables accurately predict future temperatures?
- Are the seven explanatory variables valid predictors for average temperature?
- Is the model a good fit? Which variables are best fitted in the model?
- Does the hypothesis: Global warming is caused sorely by commercial pollution, hold true?

Data Collection and Characterization

The datasets were collected from US government agencies such as the United States Energy Information Agency and the United States Environmental Protection Agency. The dependent variable will be the average annual temperature in the United States from 1960 to 2021. Various explanatory variables will also be used such as:

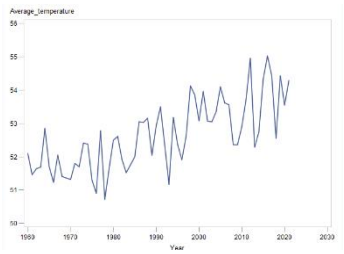
1) Yearly CO2 Emission (measured in trillion BTU)
2) Transportation Emission (measured in trillion BTU)
3) Total fossil fuel consumed by the residential area (Coal, natural gas, petroleum combined measured in trillion BTU)
4) Total fossil fuel consumed by the commercial area (coal, natural gas, petroleum combined measured in trillion BTU)
5) Total renewable energy consumed by the residential area (geothermal, solar, wood energy combined measured in trillion BTU)
6) Total renewable energy consumed by the commercial area (geothermal, solar, wood energy combined measured in trillion BTU)
7) Electric vehicle sales per year to measure possible negative effect on greenhouse gas emission

Summary Statistics & Correlation Analysis



Analysis Variable : Average_temperature				
Mean	Std Dev	Minimum	Maximum	N
52.6417742	1.0460394	50.7200000	55.0300000	62

With the analysis above, we can confirm that the average temperature dataset appears to be normally distributed with a mean of 52.64, and a standard deviation of 1.04.



Using a line chart, we do see a consistent increase in the average temperature in the last 60 years. Therefore, we can confirm that global warming is a real issue. Going forward, the purpose of the project will be to answer what variables are causing the increase in temperature.

Scatter Plot Analysis

Pearson Correlation Coefficients, N = 62 Prob > |r| under H0: Rho=0							
	Transportation_Emission	renewable_energy_commercial	total_fossil_fuel_residential	CO2 emission	electric_vehicule_sales	total_fossil_fuel_commercial	renewable_energy_residential
Average_temperature	0.71527	0.69723	-0.64991	0.57357	0.38041	0.14421	0.07000
	<.0001	<.0001	<.0001	<.0001	0.0023	0.2635	0.5888

With Transportation emission, renewable energy at the commercial level, and CO2 emission all having a very small P-value (p = <0.001) and positive correlation, we can confirm that they all have a linear relationship with the dependant variable average temperature, at 5% alpha using:

Ho: p= 0

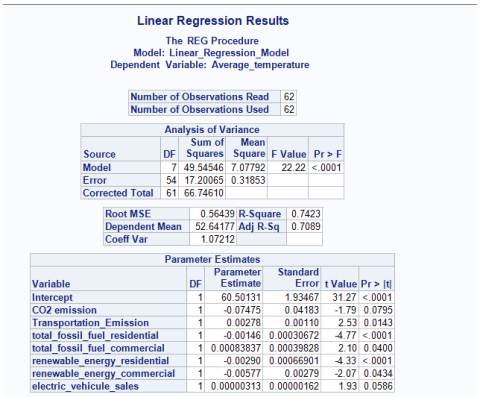
Ha: p =/= 0

p value <0001 = reject null, and linear relation between

Hence, we can confirm up to this point, that transportation emission is the strongest positively correlated with the output variable, meaning an increase in the explanatory variable would result in an increase in the output variable.

Statistical Analysis

For the upcoming part, I will conduct various statistical analyses on the dataset. The approach will be to first run a linear regression and obtain the regression equation for prediction purposes. Next, I will conduct a hypothesis analysis using an F-test to see if the model is a good fit. Lastly, I will conduct a forward selection, backward elimination, and stepwise regression to select the best-fitted variables.



Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Average_temperature

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	49.54546	7.07792	22.22	<.0001
Error	54	17.20065	0.31853		
Corrected Total	61	66.74610			

Root MSE	0.56439	R-Square	0.7423
Dependent Mean	52.64177	Adj R-Sq	0.7089
Coeff Var	1.07212		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60.50131	1.93467	31.27	<.0001
CO2 emission	1	-0.07475	0.04183	-1.79	0.0795
Transportation_Emission	1	0.00278	0.00110	2.53	0.0143
total_fossil_fuel_residential	1	-0.00146	0.00030672	-4.77	<.0001
total_fossil_fuel_commercial	1	0.00083837	0.00039828	2.10	0.0400
renewable_energy_residential	1	-0.00290	0.00066901	-4.33	<.0001
renewable_energy_commercial	1	-0.00577	0.00279	-2.07	0.0434
electric_vehicule_sales	1	0.00000313	0.00000162	1.93	0.0586

The regression equation for the dataset using the parameter estimates would be:

*Average temperature = 60.50 – 0.07*CO2 emission + 0.002*Transportation emission - 0.001* fossil fuel residential + 0.0008* fossil fuel commercial – 0.002 * renewable energy residential – 0.005* renewable energy commercial + 0.000003* electric vehicle sales*

F-test:

Ho: B0=B1=B2=B3=B4=B5=B6=B7

Ha: At least one B is not 0

Level of significance (alpha) = 5% or 0.05

If F-calculated > F-table score = reject Ho

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	49.54546	7.07792	22.22	<.0001
Error	54	17.20065	0.31853		
Corrected Total	61	66.74610			

Explanatory variables = 7

Sample size minus 2 = 60

Alpha = 0.05

f-table output = 2.1665

As F value of 22.22 > f calculated of 2.1665, we have enough evidence to reject Ho and conclude that the model is a good fit.

Forward Selection:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	58.65411	1.72699	393.76153	1153.49	<.0001
CO2 emission	-0.10803	0.03457	3.33335	9.76	0.0028
Transportation_Emission	0.00365	0.00070308	9.18755	26.91	<.0001
total_fossil_fuel_residential	-0.00080397	0.00014601	10.35046	30.32	<.0001
renewable_energy_residential	-0.00179	0.00052461	3.97953	11.66	0.0012

Backward Elimination:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	60.50131	1.93467	311.50711	977.95	<.0001
CO2 emission	-0.07475	0.04183	1.01726	3.19	0.0795
Transportation_Emission	0.00278	0.00110	2.04176	6.41	0.0143
total_fossil_fuel_residential	-0.00146	0.00030672	7.25649	22.78	<.0001
total_fossil_fuel_commercial	0.00083837	0.00039828	1.41140	4.43	0.0400
renewable_energy_residential	-0.00290	0.00066901	5.96976	18.74	<.0001
renewable_energy_commercial	-0.00577	0.00279	1.36217	4.28	0.0434
electric_vehicule_sales	0.00000313	0.00000162	1.18931	3.73	0.0586

Bounds on condition number: 18.793, 434.21

Stepwise Selection:

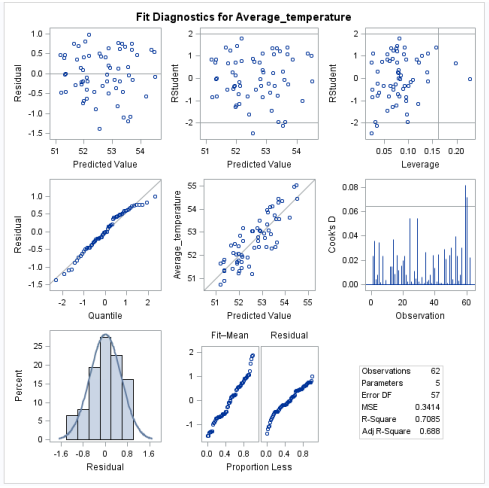
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	58.65411	1.72699	393.76153	1153.49	<.0001
CO2 emission	-0.10803	0.03457	3.33335	9.76	0.0028
Transportation_Emission	0.00365	0.00070308	9.18755	26.91	<.0001
total_fossil_fuel_residential	-0.00080397	0.00014601	10.35046	30.32	<.0001
renewable_energy_residential	-0.00179	0.00052461	3.97953	11.66	0.0012

We do find commonalities in the forward and stepwise selection, where CO2 emission, transportation emission, total fossil fuel residential and renewable energy residential were kept in the regression at 10% alpha. Backward elimination kept all 7 explanatory variables in the model.

Regression equation using forward and stepwise with the best set of variables:

Average Temperature = 58.65 – 0.108 CO2 emission + 0.003* transportation emission – 0.008* total fossil fuel residential – 0.001*renewable energy residential*

Residual analysis:



Residual analysis attached with 20% prediction excel

Fossil fuel Commercial hypothesis:

Ho: B=0 Ha B=/= 0 T-score (0.025,60) = 0.679 < t-value 1.48.

We do not have enough evidence to reject Ho and conclude that Fossil fuel commercial is a contributing factor in global warming. But it is not the only factor as transportation, CO2 emission are also important variables

Sampling and predicting

Performing an 80-20 random sampling using SAS survey select with sample size of 50, and performing a regression analysis we obtain the output:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	60.71571	2.93512	20.69	<.0001
CO2 emission	1	-0.07236	0.05416	-1.34	0.1899
Transportation_Emission	1	0.00245	0.00125	1.96	0.0573
total_fossil_fuel_residential	1	-0.00142	0.00050987	-2.79	0.0083
total_fossil_fuel_commercial	1	0.00077973	0.00052861	1.48	0.1489
renewable_energy_residential	1	-0.00285	0.00113	-2.53	0.0159
renewable_energy_commercial	1	-0.00425	0.00536	-0.79	0.4323
electric_vehicule_sales	1	0.00000306	0.00000423	0.72	0.4734

Regression equation:

Average temperature = 60.71571 – 0.07236*CO2 emission +0.00245* Transportation emission – 0.00142* fossil fuel residential + 0.00077973*fossil fuel commercial – 0.00285*renewable energy residential – 0.00425* renewable energy commercial + 0.00000306* electric vehicle sale

Applying the linear regression to the remaining 20% sample, we can conclude that we are able to predict the average temperate using both sample and full dataset and stepwise/forward aproach

Regression prediction (80%)	forward&stepwise prediction	regression prediction (100% dataset)	Average Temperature from dataset
53.72517474	53.45086506	53.73524609	54.13
53.46847186	53.55240717	53.4761052	52.36
53.81577843	53.59236012	53.57406669	53.55
52.9211689	52.90829146	52.96667901	53.05
52.11565831	52.37129096	52.06876373	51.91
53.25899595	53.35369772	53.31912197	54.1
51.3055243	51.30416192	51.30978735	51.7
52.90384515	53.12161731	52.82099095	52.75
53.37831232	53.26169766	53.44348588	53.07
51.89388875	51.96459906	51.85445979	51.23
54.54516607	54.50919229	54.44774595	54.43
51.40689205	51.88907023	51.35646553	52.04
54.58493394	53.69983041	54.38417922	54.29
52.38750569	52.54752182	52.33884806	51.17
52.70955778	52.60631649	52.76898614	53.03
51.2688017	51.1977752	51.29380744	50.72
53.24446541	53.39377643	53.18395409	52.29
51.43233886	51.33570472	51.42717171	51.8
53.52523972	53.54850094	53.50467203	53.72
51.86602958	51.91929335	51.81546635	51.7

Conclusion

The original hypothesis that claimed pollution is mostly due to commercial pollution was proven to be false. Although commercial pollution plays a role in global warming, other variables were also shown to be larger factors in polluting, such as CO2 emission and transportation emission. Hence, the hypothesis test to confirm if our data model was a good fit was proven to be true. The analyses align with various expert opinions that transportation, electricity production, and commercial pollution are among the three largest causes.

References

Nordell, B. (2003). Thermal pollution causes global warming. Global and planetary change, 38(3-4), 305312.

Ritchie, H., & Roser, M. (2020). CO_2 and greenhouse gas emissions. Our world in data.