12/6/2022

DATA603 Final Report (Group 4)

Canadian Energy Situation and Greenhouse Gas Investigation

Aung Khant Min (30160164)
Hao Su (30191442)
Sijing (Sarah) Li (10174020)
Yi-Chi Chang (30156294)
Zell Albusairi (30193671)

Table of Contents

Table of Figures

Table of Equations

Summary of Tables

1.0 Introduction

1.1 Motivation

1.1.1 Context

The topic of initial interest for investigation was energy efficiency, as this was the analyses completed as part of DATA604. The initial motivation behind energy efficiency was from environmental considerations. The Paris Agreement establishes the goal of limiting global warming to 1.5°C (degrees Celsius), or at least 2°C within this century (United Nations, n.d.) – reflecting the direness of our current environmental situation. According to IPCC (Intergovernmental Panel on Climate Change) AR5 report, if no further action is taken, the global temperature in 2100 will rise around 2.7°C to 4.5°C (Pachauri et al., 2015). Upon further research, energy efficiency is also crucial to cost and infrastructure (Environmental Protection Agency, 2022). As individuals, we can witness energy costs firsthand. In addition, when the strain on energy resources lessens, the longevity of the infrastructure used for energy generation is preserved as well.

When diving into environmental considerations – which again, was the main motivator for this topic of exploration, the root of why energy efficiency is emphasized for environmental concerns is because it is one of the main contributors to Greenhouse Gas (GHG) emissions (estimated to be responsible for over two-thirds of the global emissions). Therefore, when adapting the energy efficiency motivation to the DATA603 framework of model building, the logical conclusion was to choose GHG emissions as the response variable for this investigation (IEA, 2021).

1.1.2 Problem

In United States, the source of greenhouse gas emission is mainly split in three categories: transportation (29%), electricity production (28%) and industrial activity (22%) (MacMillan & Turrentine, 2021). This led to the consideration of finding what are the contributing factors to the GHG emission for Canada. Thus, this investigation hopes to identify these contributing factors and to what extent do they impact the GHG emissions in Canada.

As mentioned, energy efficiency is the main motivator for exploring this topic; thus, the investigation scope will focus around energy related predictors and determining which of the proposed energy generation or production predictors, as well as other categorical predictors is related to GHG emissions.

1.2 Objectives

1.2.1 Overview

The overall intent of this project is to give the audience a rudimentary understanding on Canada's energy situation and how it affects Canada's GHG emission modeled based on the available data from Open Canada.

1.2.2 Goals & Research Question

The expectation and goal from this investigation is to identify which predictors contribute to GHG emissions within Canada. It also seeks to quantify these effects.

By the end of the investigation, a multiple linear regression model will be generated to model GHG emission.

The importance of this research topic has been briefly discussed in section 1.1.1 Context. In addition to what has been mentioned, it is important to note that the global temperature is rising gradually every year and the environmental situation has not been improving at the rate required for sustainable life on Earth. If it is possible to pinpoint which factors influence the GHG the most, as a nation, it is possible to try and improve or reduce the effect of the identified factors and positively contribute to sustainability efforts.

2.0 Methodology

2.1 Data

Three separate data sets were retrieved for this analysis:

1) Greenhouse Gas (GHG) Emissions dataset (Statistics Canada, 2021b)
    a. This provided information on the response variable: GHG emissions (numerical predictor)
2) Household energy consumption dataset (Open Government, 2022a)
    a. This provided information on type of dwelling (categorical predictor) and yearly energy consumption based on the type of dwelling (numerical predictor)
3) Energy production dataset, broken down by resource type (Open Government, 2022b)
    a. This provided information on production of the following energy resources: crude oil, electricity, hydrogen, and natural gas (all numerical predictors)
    b. Note that each energy resource came in a separate table so four tables total were retrieved from this data source.

All datasets contained information on year and geographical location (i.e. province specification and overall 'Canada' specification). These columns were crucial to the data wrangling required prior to the start of model building. Since six separate tables were collected, they were processed using Excel and SQL (Structured Query Language). All datasets had more columns than necessary for the analysis, thus they were dropped and to merge the appropriate information, data needed to be transposed and joined using common columns.

For the final datafile used in the investigation, the following predictors and variables were included:

1) Quantitative (numerical):
    a. Yearly crude oil production (Thousand Cubic Metres per day)
    b. Yearly electricity generation (GWh)
    c. Yearly hydrogen production (Mega tonnes)
    d. Yearly natural gas production (Million Cubic Metres per day)
    e. Yearly energy consumption based on type of dwelling (this is for the total energy consumption across all energy types, in Gigajoules)
2) Qualitative (categorical):
    a. Region (i.e. province/territory), along with Canada
    b. Type of dwelling for the household energy consumption
3) Response/dependent variable: GHG emissions (Kilotonnes); this is the value for total industries and households

The data sets were all acquired from an open data source. Since they all come from the Government of Canada, they are licensed by "Open Government Data - Canada" (Open Government, 2019).

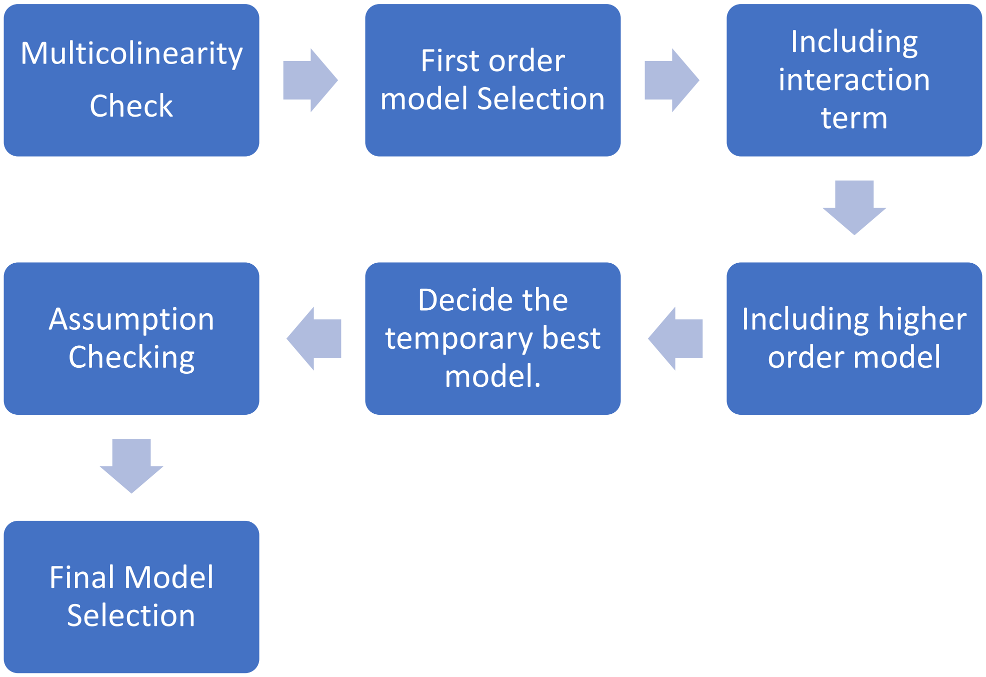
The goal with choosing the specified predictors is to ensure that again, the topic stayed within the scope of energy efficiency related predictors but also due to the fact that the description for the GHG emissions dataset specified industry and household contributions, it was desired to have both energy production representing industry contributions and household consumption data to represent household contributions.

2.2 Approach

In this project, multiple linear regression will be used to address the topic, using all methods taught in DATA 603. We will begin with multicollinearity testing. After determining the appropriate variables to be used in model building, we will use stepwise regression, backward regression, and individual t-test to find the best first order model. Next, appropriate interaction terms will be added to represent the effect of interaction between each independent variable. The potential for higher order relationship between GHG emissions and numerical variables will be determined as well. Based on the results from these steps, we will be able to identify the temporary best model for our project.

To ensure the validity and robustness of our temporary final model, several assumption checking procedures will be carried out, including Linearity assumption, independence assumption, equal variance assumption, normality assumption, and outlier review. Extra modifications and adjustments will be needed if some of the assumptions are not held for the temporary final model from the previous step. After the assumption checking, we will be able to select the final model for the project.

2.3 Workflow



Multicolinearity Check

First order model Selection

Including interaction term

Including higher order model

Decide the temporary best model.

Assumption Checking

Final Model Selection

From the summarized steps seen above, the higher order model determination and assumption checking would likely be the most difficult. This due to the fact that it requires a more iterative approach and careful consideration of calculated criterion to conclude which model should be proposed. For the assumption checking, this is foreseen to be more challenging as there are limited steps that can be taken to address the model if the assumptions fail. Thus, this step also introduces a more iterative approach to reviewing the model and multiple steps to try and fit the model to meet the assumptions.

Unfortunately, some of the assumptions were not met during this revision of the report, the proposed next steps and solutions have been discussed in section 4.4 Future Work.

2.4 Workload Distribution

The workload distribution can be seen in Table 1.

Table 1: Summary of workload distribution among team members

Aung	**R tasks:** - Further Improvement: Tested the model with "Canada" excluded from Region **Report tasks:** - Writing sections 1.0 and 3.4
Hao	**Report tasks:** - Writing section 3.2 and 3.3
Sarah (Sijing)	**R tasks:** - Model building (Workflow steps 1 – 4) - Improvements in revision 2 of this report: Predictions and AIC values - R file clean up and merging **Report tasks:** - Writing sections 2.1, 2.4, 3.0, 3.1, 3.2.5.1, 3.6, 4.0, references (section 5.0), Zotero references - Review and edit report
Yi-Chi	**R tasks:** - Assumptions checking (Workflow step 5) **Report tasks:** - Writing sections 2.2, 2.3, 3.3, 3.4, 3.5
Zell	**Report tasks:** - Review and edit report, as necessary

3.0 Main Results of the Analysis

This part of the report will be organized and reported in the same order as summarized in the 2.3 Workflow section above (i.e. presented in the same order as model development). Please note the Rmd file (R notebook file) is submitted separately and contains the appropriate code to generate the results seen throughout this section. For presentation considerations, the code will not be included within this report, only code outputs deemed crucial for understanding of this analysis is included in section 6.2 Important Code Outputs.

Please note that the hydrogen production column often results in an output of 'NA'. This is due to the fact that the input in this column is all zero. This indicates that hydrogen production is zero for the range of data chosen for the model. This column is still kept within the analysis as it was one of the columns originally considered for model building as it is a numerical variable relevant to the topic being investigated.

Finally, where possible, all outputs are rounded to four decimal places for presentation purposes and from a scientific standpoint, four decimal places has been continuously used as common convention to sufficiently represent a particular value.

3.1 Multicollinearity

Multicollinearity was checked first although it falls under the category of 'Assumptions' (see 3.5 Assumptions Review) as it is crucial to understand if there are any predictors which are highly correlated. This is a pre-screening step to determine if there are redundant information present within the dataset and if so, address it appropriately through dropping any of the highly correlated predictors from the model or creating a singular predictor from the highly correlated predictors.

This step needs to be performed prior to determining which base predictors should be present within the model as it will affect which predictors are being considered for model adoption.

To check for multicollinearity, the VIF (Variance Inflation Factors) are calculated in R. The detailed outcome of these calculations is seen in section 6.1 Table Outputs in Table Outputs

Table 15. The summarized outcome for if each predictor (i.e. variable) showed multicollinearity can be seen in Table 2.

Table 2: Multicollinearity Detection Summary

Variable	Multicollinearity Detected?	Comments
Region	Yes	Critical levels of multicollinearity detected.
Crude_Yearly_Average	Yes	
Elec_Yearly_Average	Yes	
NG_Yearly_Average	Yes	
Hydrogen_Yearly_Average	NA	Too many zeros within the dataset.
Dwelling_Type	No	Moderate collinearity detected but does not warrant further action.
Energy_Consumption	No	

To investigate this potential multicollinearity observance, a pairs plot was generated (which is essentially a grid view of multiple bivariate plots) to help visualize and identify which variables have multicollinearity with each other. Note that the pairs function only takes in numerical variables so the yearly production variables (Crude_Yearly_Average, Elec_Yearly_Average, and NG_Yearly_Average) are plotted.

This is also helpful as it is hypothesized that multicollinearity is only observed due to the fact that the categorical variable, Region, has more than three levels (Allison, 2012). Thus, if no correlation is observed between these numerical variables, it would support the hypothesis.

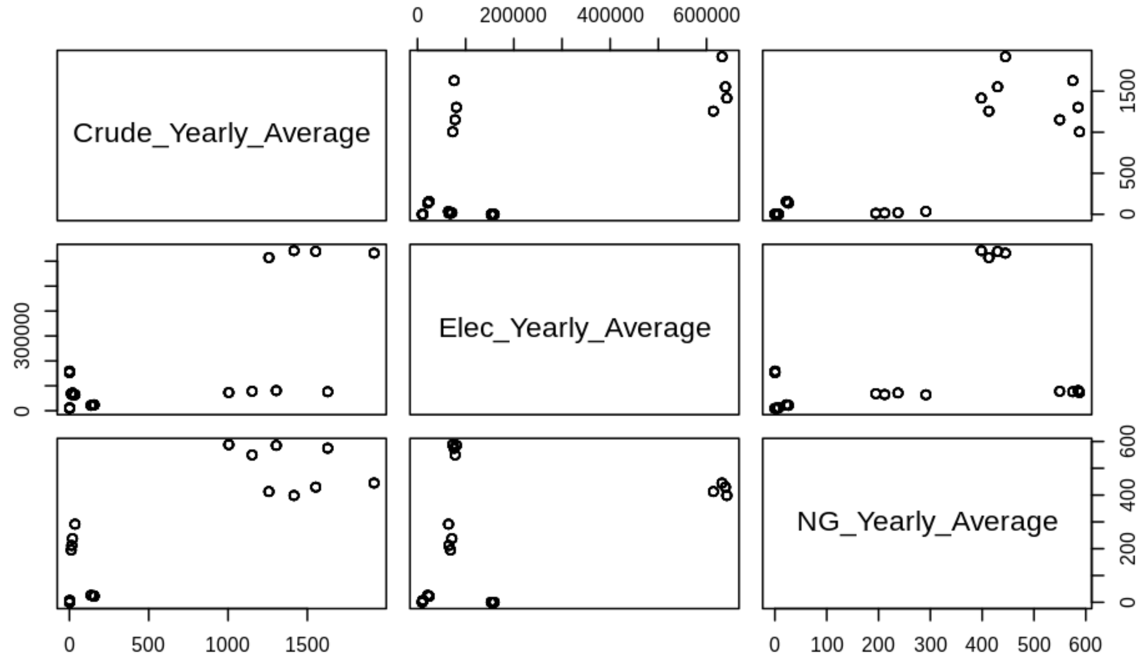


Figure 1: Pairs plot for multicollinearity detection

As seen in Figure 1, there does not appear to be a clear correlation between the numerical variables which had multicollinearity detected. Thus, it's likely the cause for the multicollinearity detection is as hypothesized – due to the categorical variable Region having six levels. Further supporting evidence is that all numerical variables flagged for multicollinearity originate from the same dataset as the one containing Region.

An updated set of VIF calculations was completed after removing the variable Region from the dataset, which showed no detection of multicollinearity, as seen in Table 3, with the detailed output found in Table 16.

Table 3: Multicollinearity Detection Summary (Region removed)

Variable	Multicollinearity Detected?	Comments
Crude_Yearly_Average	No	The VIF value is greater than 5 but the Diagnostics table indicate a detection of 0.
Elec_Yearly_Average	No	Moderate collinearity detected but does not warrant further action.
NG_Yearly_Average	No	
Hydrogen_Yearly_Average	NA	Too many zeros within the dataset.
Dwelling_Type	No	Moderate collinearity detected but does not warrant further action.
Energy_Consumption	No	

For the variable Crude_Yearly_Average, as mentioned in Table 3, the VIF value is greater than 5 but the R output indicates that detection is 0. As a cautionary measure, a visualization will be generated to confirm that Crude_Yearly_Average does not have multicollinearity with any other variable present within the model. Figure 1 already shows this variable has no correlation with Elec_Yearly_Average and

NG_Yearly_Average. Thus, it will be plotted against the remaining variables (Hydrogen_Yearly_Average and Energy_Consumption).

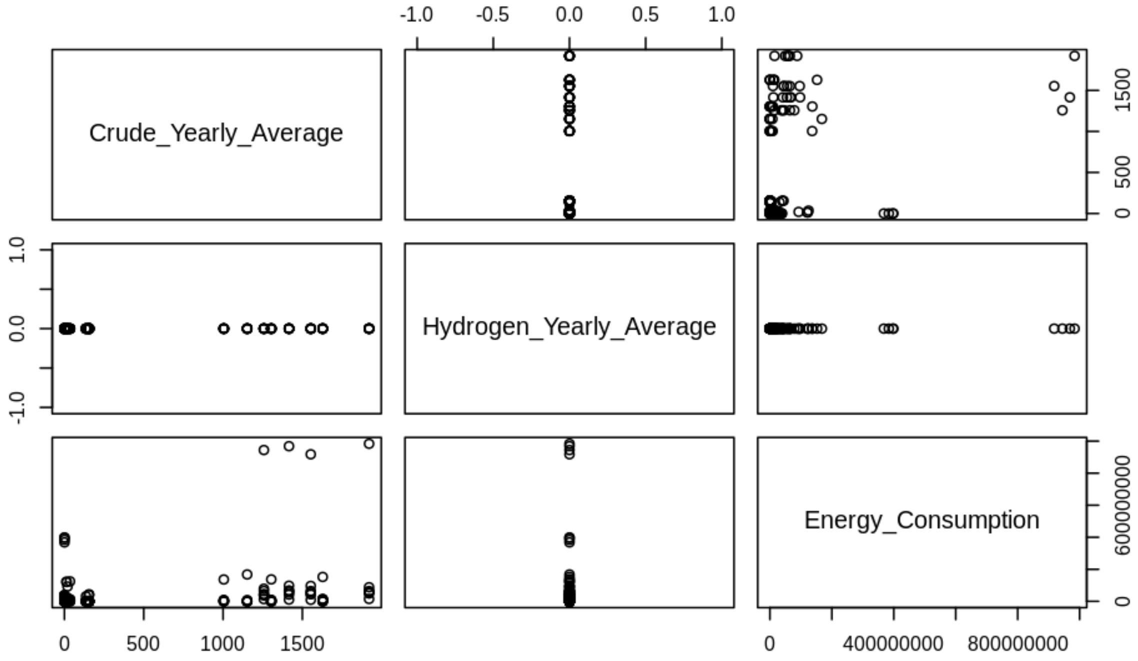


Figure 2: Additional pairs plot for multicollinearity detection specifically for Crude_Yearly_Average

As seen in Figure 2, there does not appear to be a clear correlation between the remaining numerical variables and the variable Crude_Yearly_Average. Note, the bivariate plots with Hydrogen_Yearly_Average show a singular line since it appears the only values for hydrogen production within the dataset is zero. Therefore, there does not appear to be multicollinearity between Crude_Yearly_Average and any of the other numerical variables within the dataset. It is possible that the remaining categorical variable Dwelling_Type is causing high VIF values. Similar to Region, Dwelling_Type also has more than three levels – there are seven levels.

As a last precautionary measure, the categorical variable Dwelling_Type is removed to evaluate the proposed hypothesis regarding Dwelling_Type causing the higher VIF values. The summary of results can be found in Table 4, while the detailed output can be found in Table 17.

Table 4: Multicollinearity Detection Summary (both categorical variables removed)

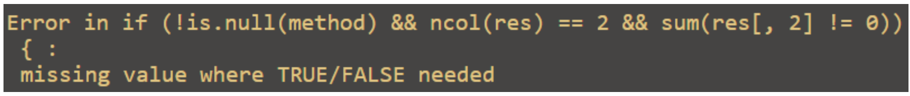
Variable	Multicollinearity Detected?	Comments
Crude_Yearly_Average	No	The VIF value is greater than 5 but the Diagnostics table indicate a detection of 0.
Elec_Yearly_Average	No	Moderate collinearity detected but does not warrant further action.
NG_Yearly_Average	No	
Dwelling_Type	No	
Hydrogen_Yearly_Average	NA	Too many zeros within the dataset.

Unfortunately, this does not alter the VIF value for Crude_Yearly_Average; however, as stated previously, based on the visualization generated, no clear correlation exists between the suspect variable and any other remaining numerical variables. Again, the categorical variables are not of concern as they both contain more than three levels (Allison, 2012).

Therefore, all variables present within the dataset will be considered for the base predictor evaluation.

3.1.1 Minor Commentary on R

Please note that to calculate the VIF values, the function 'imcdiag' is used. However, after running the code, an error message appears, as seen below:



```
Error in if (!is.null(method) && ncol(res) == 2 && sum(res[, 2] != 0))
 { :
 missing value where TRUE/FALSE needed
```

[1]

A search on this code reveals that this error appears when the input variable in dependent statements or operations are missing a value (ProgrammingR, 2022). However, there are no missing values within the imported data table – this was checked in the native Excel file and the beginning/end of the data file within R was checked with the 'head'/'tail' functions. Thus, the hypothesized issue is that there are many rows where multiple columns have a value of zero – which may be causing issues in VIF calculations. Since no missing values are in the datafile read into R and the VIF values are being generated as expected, no further investigation on this error was completed.

3.2 Base Predictor Determination

As the first step in model building, the predictors which can be kept within the model needs to be determined.

The first order base model with all the potential predictors can be represented as the equation seen in Equation 1.

Equation 1: First Order Base Model with all potential predictors

$$
\begin{aligned}
Yearly\_\widehat{GHG}\_Emission & \\
= {} & \beta_0 + \beta_1 Region_{i1} + \beta_2 Region_{i2} + \beta_3 Region_{i3} + \beta_4 Region_{i4} + \beta_5 Region_{i5} \\
& + \beta_6 \text{Crude\_Yearly\_Average} + \beta_7 \text{Hydrogen\_Yearly\_Average} + \beta_8 \text{NG\_Yearly\_Average} \\
& + \beta_9 \text{Elec\_Yearly\_Average} + \beta_{10}(\text{Dwelling\_Type}_{i1}) + \beta_{11}(\text{Dwelling\_Type}_{i2}) \\
& + \beta_{12}(\text{Dwelling\_Type}_{i3}) + \beta_{13}(\text{Dwelling\_Type}_{i4}) + \beta_{14}(\text{Dwelling\_Type}_{i5}) \\
& + \beta_{15}(\text{Dwelling\_Type}_{i6}) + \beta_{16}\text{Energy\_Consumption} + \epsilon
\end{aligned}
$$

3.2.1 Full Model Test

A Full Model Test (also known as the global F test) is conducted to determine if any of the potential predictors should be used in the model. For this scenario, it will be testing if any of the predictors included within the dataset is related to the response variable, GHG emissions.

[1] Note no figure title or number is assigned to this visual as this is a deviation to comment on the R code output and not part of the main report. Thus, this figure should not be included in the Table of Figures – hence no title is assigned.

To carry out this test, the hypotheses being tested are:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{16} = 0$$

$$H_a: \text{at least one } \beta_i \; is\ not\ zero\ (i = 1{,}2{,}3, \cdots, 16)$$

This test is being conducted using the F-statistic and can be generated in R using the 'anova' function. The resulting ANOVA table can be seen in Table 5.

Table 5: ANOVA table output for the Full Model Test

Source of Variation	DF	Sum of Squares	Mean Square	F-Statistic	p-Value
Regression	15	10680558451971	712037230131.4	57786	< 0.00000000000000022
Residual	152	1872947346	12322022.01		
Total	167	10682431399317	63966655086		

From the output of R code, the ANOVA table indicates that F_{cal}=57786, which results in a p-value less than 2.2e-16. Since the test is being performed at the 95% confidence level, indicating α=0.05, the p-value is less than the specified α and therefore, the null hypothesis should be rejected.

It provides compelling evidence against the null hypothesis H_0, and we can conclude at least one β_i is not zero (where i=1,2,3,…,16), meaning at least one of the independent variables is related to the GHG emissions.

3.2.2 Individual Coefficient Tests

Now that it is confirmed at least one of the predictors included in the dataset can be used to model GHG emissions, the next step is to determine which of the predictors is suitable and should remain in the final proposed model. To do this, individual t-tests will be performed for each predictor.

With each predictor, the hypotheses being tested are:

$$H_0: \beta_i = 0 \quad (\text{i}=1{,}2{,}3{,}\cdots,\ 16)$$

$$H_a: \beta_i \neq 0 \quad (\text{i}=1{,}2{,}3{,}\cdots,\ 16)$$

The detailed t-test output can be seen in Figure 12, in section 6.2 Important Code Outputs. Table 6 shows the summarized outcome of the R output, with only the significant predictors (for ease of presentation).

Table 6: Summary of Individual Coefficient Tests Outputs for Significant Predictors

Variable	Test Statistic[2]	p-value	Significance Level
Region – British Columbia	-27.887	< 0.0000000000000002	***
Region – Canada	9.888	< 0.0000000000000002	***
Region – Nova Scotia	-19.084	< 0.0000000000000002	***
Region – Ontario	-8.510	0.0000000000000155	***
Region – Saskatchewan	-15.441	< 0.0000000000000002	***

[2] Note that the outputs are given to 3 decimal places and thus, 4 decimal places have not been shown as a result.

Variable	Test Statistic[2]	p-value	Significance Level
Crude_Yearly_Average	11.161	< 0.0000000000000002	***
Elec_Yearly_Average	5.155	0.0000007798433600	***

The test is being performed at a 95% confidence level, indicating α=0.05; the p-values for the predictors Region, Crude_Yearly_Average and Elec_Yearly_Average are less than the specified α and therefore, the null hypothesis should be rejected for these predictors. This indicates that the estimated parameters for these predictors are not zero, thus these predictors are related to the GHG emissions and should be kept within the model.

Conversely, the remaining variables have a p-value greater than 0.05 and therefore, the null hypothesis cannot be rejected, indicating the estimated parameters are zero – thus, these predictors should not be kept within the model and is not related to GHG emissions.

Another set of individual t-tests was performed to confirm the proposed variables are significant and should remain in the model. Please reference section 3.2.6 Conclusion for this analysis.

3.2.3 Stepwise Regression

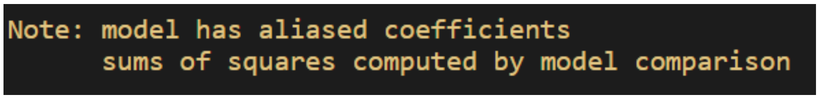
Stepwise regression is completed to supplement the individual coefficient test. This allows multiple different methods of determining which base predictors should be chosen and therefore, maximizes the likelihood of achieving the best possible model. Typically, the p-enter value (i.e. p-value threshold for a predictor to be adopted in the model) is 0.1 while the p-remove value (i.e. p-value threshold for a predictor to be removed from the model) is 0.3. However, to get a better model, more strict threshold values were chosen, where p-enter = 0.05 while p-remove = 0.1. The p-enter value was chosen with the consideration that the model is meant to be consistently produced at the 95% confidence level. The p-remove value was chosen such that it is higher than the p-enter value but still allows an objective evaluation of the variables (i.e. not too high of a threshold).

The output of this analysis indicated to keep Region, Crude_Yearly_Average and Elec_Yearly_Average – which aligns with the outcome of the individual coefficients tests as well. The summary output of this analysis can be found in Figure 13.

Please note that the Forward Regression Procedure was omitted as it carries out the same steps as the Stepwise Regression Procedure, without the potential of eliminating added variables. Since we are applying a stricter criteria to the model, only the outcome of Stepwise Regression would be referenced in this case.

3.2.3.1 Minor Commentary on R

Please note that to calculate the VIF values, the function 'ols_step_both_p' is used. However, after running the code, an error message appears, as seen below:



```
Note: model has aliased coefficients
      sums of squares computed by model comparison
```

[3]

A search on this code reveals that this error appears when there is a high correlation between predictors (Zach, 2021). We already knew this was true from the multicollinearity evaluation seen in section 3.1 Multicollinearity. However, as detailed in the review of the presence of multicollinearity, it is acceptable, and no predictors will be manipulated or altered. Thus, the impact of the R code error has already been evaluated and again, deemed acceptable.

3.2.4 Backward Regression Procedure

Backward regression is completed with the same motivation as the Stepwise regression – to provide another source of reference for which predictors to keep. And again, similarly to Stepwise regression, a more strict p-remove was chosen, 0.05, to better reflect the 95% confidence level at which all tests within this report are consistently following.

The output of this analysis indicated to remove Dwelling_Type, Energy_Consumption, and NG_Yearly_Average. Therefore, the proposed model keeps Region, Crude_Yearly_Average, Hydrogen_Yearly_Average, and Elec_Yearly_Average – which is similar to the outcome of the individual coefficients tests and Stepwise regression as well. The additional predictor Hydrogen_Yearly_Average only takes on values of zero within the dataset, as mentioned. Thus, this likely influenced the results of the Backward Regression Procedure and regardless of this inclusion, this predictor would not have an appropriate estimated parameter with only zeros as values. The summary output of this analysis (R code) can be found in Figure 14.

Please refer to section 3.2.3.1 Minor Commentary on R for a commentary on the R Outputs for this procedure.

3.2.5 All-Possible-Regressions-Selection

For selecting the best predictors, the Mallows's Cp Criterion, BIC (Bayesian information criteria), AIC (Akaike Information Criterion)[4], Adjusted R^2 and RMSE were calculated. The outcome of these calculations can be represented with a figure, as seen in Figure 3; Table 7 represents the numerical outputs of the calculations. Six scenarios with up to 6 variables are included in the analysis as there are six potential predictors for this model after the removal of the hydrogen column (see section *3.2.5.2 AIC Updates* for further details).

[3] Note no figure title or number is assigned to this visual as this is a deviation to comment on the R code output and not part of the main report. Thus, this figure should not be included in the Table of Figures – hence no title is assigned.

[4] Unfortunately, the value AIC could not be calculated initially. As discussed with professor Ngamkham, there was an issue when using the "ols_step_best_subset" function. However, in revision #2 of this report, a hypothesis was tested which resolved the issue – see section 3.2.5.2 AIC Updates for further details.

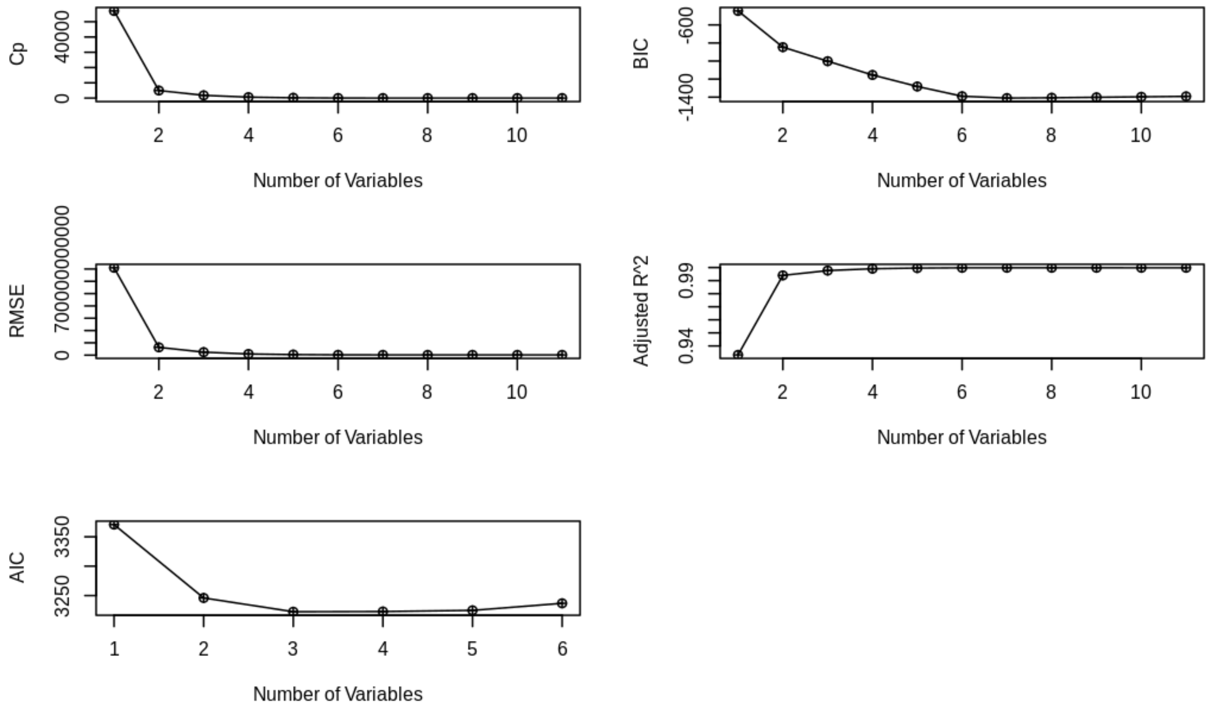


Figure 3: Visual representation of the calculations completed for Cp, BIC, RMSE, AIC, and R_{adj}^2

Table 7: Numerical outputs for the calculations completed for Cp, BIC, RMSE, AIC, and R_{adj}^2

Number of Variables	Cp	BIC	RMSE	Adjusted-R^2[5]	AIC[6]
1	57068.2393	-445.2420	709887223841	0.9331460	3370.725
2	4937.5885	-846.3349	63253382482	0.9940070	3245.973
3	1799.9623	-1001.8584	24310636822	0.9976826	3222.508
4	607.6321	-1154.6508	9496613579	0.9990892	3222.869
5	192.1223	-1281.9357	4317977905	0.9995833	3224.868
6	23.4378	-1390.0327	2200871380	0.9997863	3236.868

From the outcome of these calculations, it is proposed that the model with three predictors is adopted. It was difficult to choose between the model with three or four predictors since the improvements in R_{adj}^2 was relatively minute with each additional variable. The Cp value, BIC value, R_{adj}^2 value, and RMSE show improvement up to the addition of the sixth variable. The AIC value shows improvement only up until the third variable. The main predictor used to evaluate how many variables to include is the R_{adj}^2 value as it provides insight on overfitting as well; after the fourth variable addition, the increase in value indicates that any further addition would indicate overfitting.

The improvement from the second and third variable addition is around 0.3676% improvement while the improvement from the third and fourth addition is around 0.1407%. Since the model already has a

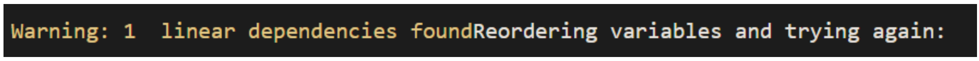
[5] The R_{adj}^2 were not rounded since the values were already very high, this way each level can be distinguished from one another.

[6] Note that the outputs are given to 3 decimal places and thus, 4 decimal places have not been shown as a result.

high R_{adj}^2 value, the mode with three variables is chosen to avoid over-fitting. In addition, this would be supported by what is observed for the AIC value – which indicates the model has already been optimized at three variables.

3.2.5.1 Minor Commentary on R

Please note the following error was observed when running the code for this section, as seen below:



```
Warning: 1  linear dependencies foundReordering variables and trying again:
```

[7]

In attempt to eliminate this warning, the nvmax was reduced (Lumley, 2005). Unfortunately, this did not eliminate the error. Another potential solution is to use a Random Forest approach (Placidia, 2012); however, this is out of the scope of this course – thus, this operation is kept within the analysis. It appears the operation does produce results even with linear dependencies – thus, it is important to note that there may be linear dependencies; however, this has already been discussed in the multicollinearity review (Gundersen, 2021).

3.2.5.2 AIC Updates

During the initial version of this report, the function "ols_step_best_subset" would not work and thus AIC values were not included. In the final revision of this report, the hypothesis of removing the hydrogen production column (since this column only contained zeros), was tested. This was the cause of the issue seen in the operation and AIC values could be calculated. Figure 3 and Table 7 were updated accordingly. However, please note that the number of predictors within the dataset was reduced to six from seven as a result of this update.

3.2.6 Conclusion

It appears there are 3 variables in common from all of the evaluations: Region, yearly production of crude oil and yearly production of electricity. Therefore, the updated proposed model will take the format:

Equation 2: First Order Model format after the model selection process

$$\begin{aligned}\widehat{Yearly\_GHG\_Emission} &= \beta_0 + \beta_1 Region_{i1} + \beta_2 Region_{i2} + \beta_3 Region_{i3} + \beta_4 Region_{i4} + \beta_5 Region_{i5} \\ &+ \beta_6 \text{Crude\_Yearly\_Average} + \beta_7 \text{Elec\_Yearly\_Average}\end{aligned}$$

The individual t-test will be applied to the proposed updated model, which has the following hypothesis:

$$H_0 : \beta_i = 0 \quad (\text{i}=1,2,3,\cdots,7)$$

$$H_a : \beta_i \neq 0 \quad (\text{i}=1,2,3,\cdots,7)$$

The summary output of this analysis (R code) can be found in Figure 15, while Table 8 shows the relevant R outputs in a table format.

[7] Note no figure title or number is assigned to this visual as this is a deviation to comment on the R code output and not part of the main report. Thus, this figure should not be included in the Table of Figures – hence no title is assigned.

Table 8: Summary of Individual Coefficient Tests Outputs

Variable	Test Statistic[8]	p-value	Significance Level
Region – British Columbia	-61.211	< 0.0000000000000002	***
Region – Canada	10.315	< 0.0000000000000002	***
Region – Nova Scotia	-50.617	< 0.0000000000000002	***
Region – Ontario	-15.595	< 0.0000000000000002	***
Region – Saskatchewan	-46.516	< 0.0000000000000002	***
Crude_Yearly_Average	12.385	< 0.0000000000000002	***
Elec_Yearly_Average	5.117	0.000000878	***
ADJ-R-squared			0.9998
RMSE			3438

The test is being performed at a 95% confidence level, indicating α=0.05; the p-values for the predictors Region, Crude_Yearly_Average and Elec_Yearly_Average are less than the specified α and therefore, the null hypothesis should be rejected for these predictors. This indicates that the estimated parameters for these predictors are not zero, thus these predictors are related to the GHG emissions and should be kept within the model.

Therefore, the updated model is:

Equation 2: First Order Model with estimated parameters after the model selection process

$$\widehat{Yearly\_GHG\_Emission} = 218648.2250 - 158454.5095 Region_{i1} + 323170.1674 Region_{i2} - 202531.7569 Region_{i3} - 90895.8567 Region_{i4} - 156503.3431 Region_{i5} + 24.9340 \text{Crude\_Yearly\_Average} + 0.2902 \text{Elec\_Yearly\_Average}$$

3.3 Interaction Terms

With the updated base model, interaction terms will be evaluated. The hypothesized model is:

Equation 3: Proposed model with interaction terms format

$$\widehat{Yearly\_GHG\_Emission} = \beta_0 + \beta_1 Region_{i1} + \beta_2 Region_{i2} + \beta_3 Region_{i3} + \beta_4 Region_{i4} + \beta_5 Region_{i5} + \beta_6 \text{Crude\_Yearly\_Average} + \beta_7 \text{Elec\_Yearly\_Average} + \beta_8 (Region_{i1})(\text{Crude\_Yearly\_Average}) + \beta_9 (Region_{i2})(\text{Crude\_Yearly\_Average}) + \beta_{10} (Region_{i3})(\text{Crude\_Yearly\_Average}) + \beta_{11} (Region_{i4})(\text{Crude\_Yearly\_Average}) + \beta_{12} (Region_{i5})(\text{Crude\_Yearly\_Average}) + \beta_{13} (Region_{i1})(\text{Elec\_Yearly\_Average}) + \beta_{14} (Region_{i2})(\text{Elec\_Yearly\_Average}) + \beta_{15} (Region_{i3})(\text{Elec\_Yearly\_Average}) + \beta_{16} (Region_{i4})(\text{Elec\_Yearly\_Average}) + \beta_{17} (Region_{i5})(\text{Elec\_Yearly\_Average}) + \beta_{18} (\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})$$

[8] Note that the outputs are given to 3 decimal places and thus, 4 decimal places have not been shown as a result.

The individual t-test will be applied to the proposed updated model, which has the following hypothesis:

$$H_0: \beta_i = 0 \quad (i=1,2,3,\cdots, 18)$$

$$H_a: \beta_i \neq 0 \quad (i=1,2,3,\cdots, 18)$$

The summary output of this analysis (R code) can be found in Figure 16, while Table 9 shows the relevant R outputs in a table format.

Table 9: Summary of Individual Coefficient Tests outputs, with Interaction terms

Variable	Test Statistic[9]	p-value	Significance Level
Region – British Columbia	2.323	0.02153	*
Region – Canada	-1.287	0.19994	
Region – Nova Scotia	0.233	0.81570	
Region – Ontario	7.889	0.000000000000603	***
Region – Saskatchewan	0.756	0.45055	
Crude_Yearly_Average	3.253	0.00141	**
Elec_Yearly_Average	6.205	0.000000005150568	***
(Region – British Columbia) *(Crude_Yearly_Average)	7.101	0.000000000047196	***
(Region – Canada) *(Crude_Yearly_Average)	2.639	0.00921	**
(Region – Nova Scotia) *(Crude_Yearly_Average)	0.335	0.73840	
(Region – Ontario) *(Crude_Yearly_Average)	18.626	< 0.0000000000000002	***
(Region – Saskatchewan) *(Crude_Yearly_Average)	4.352	0.000024921265676	***
(Region – British Columbia) *(Elec_Yearly_Average)	-6.286	0.000000003406823	***
(Region –Canada) *(Elec_Yearly_Average)	-14.632	< 0.0000000000000002	***
(Region –Nova Scotia) *(Elec_Yearly_Average)	-0.810	0.41915	
(Region –Ontario) *(Elec_Yearly_Average)	-8.038	0.000000000000260	***
(Region –Saskatchewan) *(Elec_Yearly_Average)	-4.902	0.000002454472246	***
(Crude_Yearly_Average) *(Elec_Yearly_Average)	-2.637	0.00926	**
ADJ-R-squared			0.9999
RMSE			1129

From the output of the R code, all interactions terms are significant, therefore they will all be kept in the model, as the null hypothesis is rejected.

[9] Note that the outputs are given to 3 decimal places and thus, 4 decimal places have not been shown as a result.

Therefore, the interaction model is:

Equation 4: Proposed model with interaction terms and their respective estimated parameters

$$
\begin{aligned}
&Yearly\_\widehat{GHG\_Emission} \\
&\quad = -19806.3345 + 97451.4007 Region_{i1} - 451413.0413 Region_{i2} \\
&\quad + 9977.0474 Region_{i3} + 346753.6210 Region_{i4} + 30915.8281 Region_{i5} \\
&\quad + 114.2416\mathrm{Crude\_Yearly\_Average} + 3.4741\mathrm{Elec\_Yearly\_Average} \\
&\quad + 192.1408(Region_{i1})(\mathrm{Crude\_Yearly\_Average}) \\
&\quad + 680.3384(Region_{i2})(\mathrm{Crude\_Yearly\_Average}) \\
&\quad + 288.4150(Region_{i3})(\mathrm{Crude\_Yearly\_Average}) \\
&\quad + 40259.1968(Region_{i4})(\mathrm{Crude\_Yearly\_Average}) \\
&\quad + 389.0909(Region_{i5})(\mathrm{Crude\_Yearly\_Average}) \\
&\quad - 3.5028(Region_{i1})(\mathrm{Elec\_Yearly\_Average}) \\
&\quad - 1.5671(Region_{i2})(\mathrm{Elec\_Yearly\_Average}) \\
&\quad - 0.7520(Region_{i3})(\mathrm{Elec\_Yearly\_Average}) \\
&\quad - 4.5478(Region_{i4})(\mathrm{Elec\_Yearly\_Average}) \\
&\quad - 3.9027(Region_{i5})(\mathrm{Elec\_Yearly\_Average}) \\
&\quad - 0.0012(\mathrm{Crude\_Yearly\_Average})(\mathrm{Elec\_Yearly\_Average})
\end{aligned}
$$

3.4 Higher Order Terms

After the interaction model, we would like to see if there exists a nonlinear relationship between greenhouse gas emissions and numerical variables such as yearly average of electricity production and yearly average of Crude oil production. First, we use ggpair to draw the correlation table and find out that Elec_yearly_average has the highest correlation with greenhouse gas emissions, with correlation equal 0.966. Based on the former result, we will start testing the nonlinear relationship between Elec_yearly_average and greenhouse gas emission by adding higher order of Elec_yearly_average into the first order model.

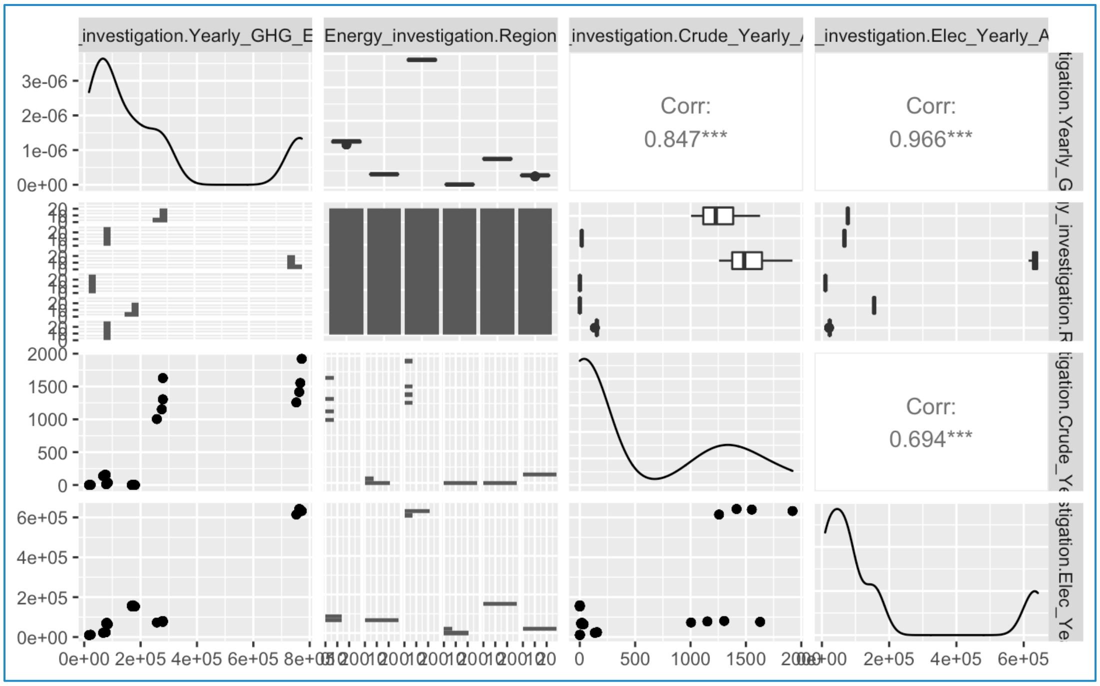


Figure 4: GGpair to check correlation

First, we put second order of Elec_Yearly_Average into the first order model. From table 7, we can see that the higher order of Elec_Yearly_Average is significant with P-value < 0.05. However, comparing with the interaction model, the adjusted R-squared drop from 0.99998 to 0.99982 and the RMSE increase from 1129.475 to 3385.032.

Table 10: Higher order model (Including Elec_Yearly_Average)

Coefficients	p-value	Significance
Intercept	< 0.0000000000000002	***
Region – British Columbia	< 0.0000000000000002	***
Region – Canada	0.00000000136	***
Region – Nova Scotia	< 0.0000000000000002	***
Region – Ontario	< 0.0000000000000002	***
Region – Saskatchewan	< 0.0000000000000002	***
Crude_Yearly_Average	< 0.0000000000000002	***
Elec_Yearly_Average	0.0000413	***
I(Elec_Yearly_Average^2)	0.0149	*
R_{adj}^2		0.99982
RMSE		3385.032

Next, we put second order of Crude_Yearly_Average to the model. From table 8, we can see that the higher order of Crude_Yearly_Average is significant with P-value < 0.05. However, comparing with the interaction model, the adjusted R-squared drop from 0.99998 to 0.99985 and the RMSE increase from 1129.475 to 3092.967.

Table 11: Higher order model (Including Crude_Yearly_Average)

Coefficients	p-value	Significance
Intercept	< 0.0000000000000002	***
Region – British Columbia	0.00000000000000196	***
Region – Canada	< 0.0000000000000002	***
Region – Nova Scotia	< 0.0000000000000002	***
Region – Ontario	0.0184	***
Region – Saskatchewan	< 0.0000000000000002	***
Crude_Yearly_Average	0.000000000000758	***
Elec_Yearly_Average	0.000107	***
I(Crude_Yearly_Average^2)	0.00000000419	***
R_{adj}^2		0.99985
RMSE		3092.967

Although the tables above show that the coefficient of the numerical variables' higher order term is significantly greater than 0, the Adjusted R-squared is lower than the original interaction model. Hence, as adding higher order term does not increase the performance of the model, we decided not to include any higher order variable into the model to reduce the complexity and avoid over fitting.

3.5 Assumptions Review

Multiple assumptions reviews are required to ensure the validity and robustness of the temporary final model retrieved from the previous step.

3.5.1 Linearity Assumption

By implementing multiple regression analysis, we assume that there exists a linear relationship between dependent and independent variables. In this project, we will be using residual plot to check the linearity assumptions.

From the residual graph below, as the line is aligned to the middle line, there is not any significant pattern showing the existence of non-linear associations in the data. In conclusion, based on visual inspection, there does not exist a significant non-linear transformation needed for the model and hence the relationship between the predictors and the response seems to be linear.

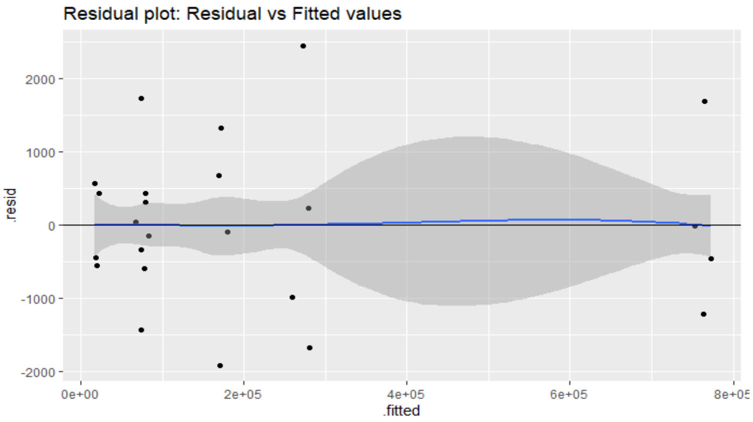


Figure 5: Residual plot for linearity assumption check

3.5.2 Independence Assumption

To implement MLR, the error terms should be uncorrelated to each other. We will use residual scatter plot to test whether our model fulfill the independence assumption. From the figure below, we can find out that there is no significant relationship between each residual, meaning that the model fulfill the independence assumption.

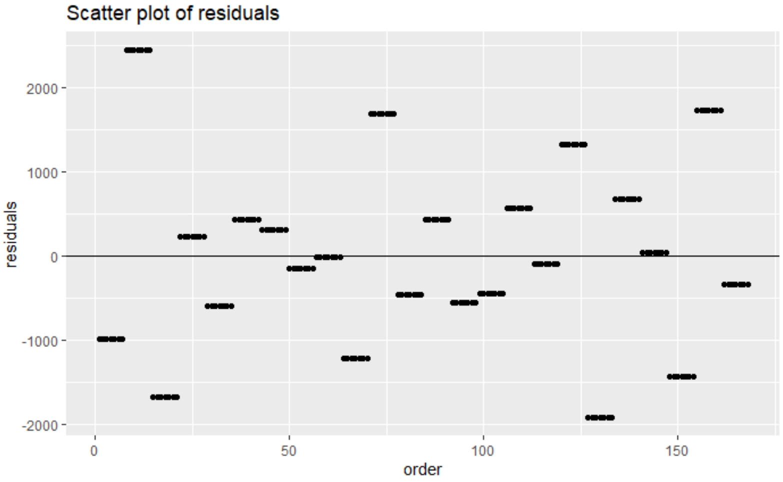


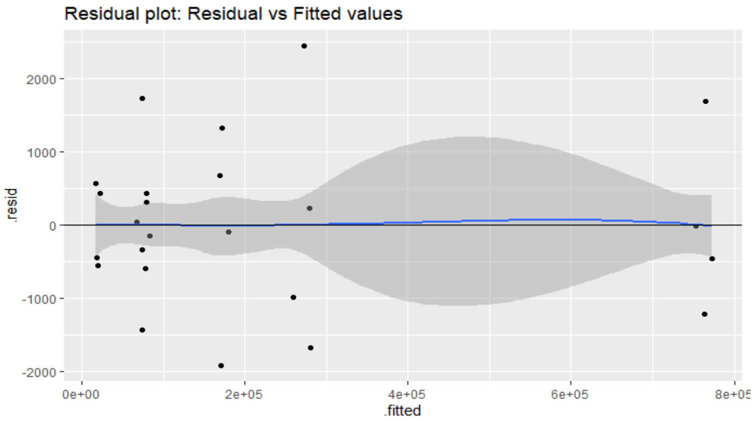
Figure 6: Scatter plot of residuals

3.5.3 Equal Variance Assumption

The homoscedasticity should be held to ensure a robust MLR model. In this project, we will be using residual plot and Breusch-Pagan test to check whether the homoscedasticity assumption holds for our final model. The hypothesis of the test listed below:

$$H_0:\ Heteroscedasticity\ is\ not\ present$$

$$H_a: Heteroscedasticity\ present$$



:

Figure 7: Residual plot for homoscedasticity assumption check

From the residual plot above, we found that there exists a wedge shape within the graph, meaning that the homoscedasticity condition might not hold for our model. The p-value of the Breusch-Pagan test on our final model is 2.2e-16, meaning that we should reject the null hypothesis and conclude that heteroscedasticity present. The Equal Variance Assumption does not hold.

3.5.4 Normality Assumption

Another assumption for MLR is that the residuals should be normally distributed. We will be checking the assumption through histogram, normal probability plot and the Shapiro-Wilk test. From the graph below, we can see from the histogram that the shape of residuals' distribution does not look like a normal distribution. Besides, the QQ-plot also show that the points are not perfectly align to the line. We suspect that our model might not satisfied the normality assumption. The result of Shapiro-Wilk test shows that, with p-value = 0.0002583, we should reject the null hypothesis and conclude that the sample data are not significantly normally distributed.

$$H_0: The\ sample\ data\ are\ significantly\ normally\ distributed$$

$$H_a: The\ sample\ data\ are\ \boldsymbol{not}\ significantly\ normally\ distributed$$

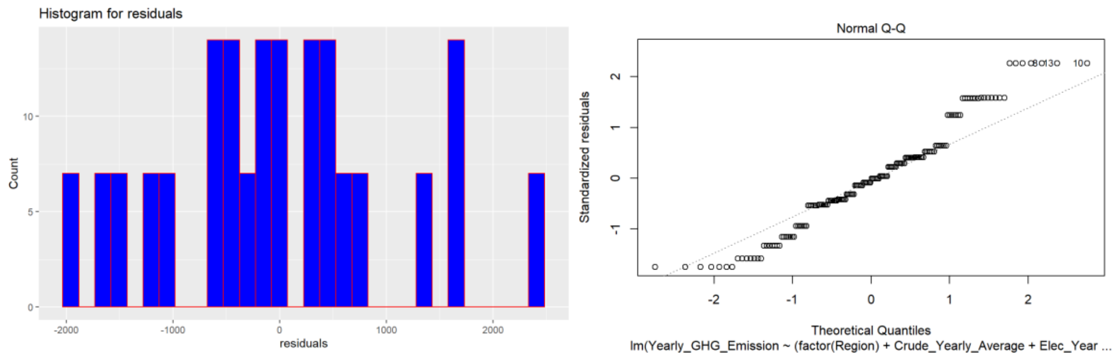


Figure 8: Histogram and QQ-plot for normality assumption check.

3.5.5 Box-Cox Transformation for unequal variances and nonnormality.

According to 3.5.3 and 3.5.4, our final model does not fulfill the homoscedasticity and normality assumption. To address the issue, we try to implement Box-Cox transformation on the response variable to see whether the method will help improve the model to solve the issue of unequal variances and nonnormality.

The method of maximum likelihood had shown that the lambda should be around 0.7778. Hence, we will be using $\frac{(Y^{\lambda}-1)}{\lambda}$ $(\lambda = 0.7778)$ transformation to rerun the model.

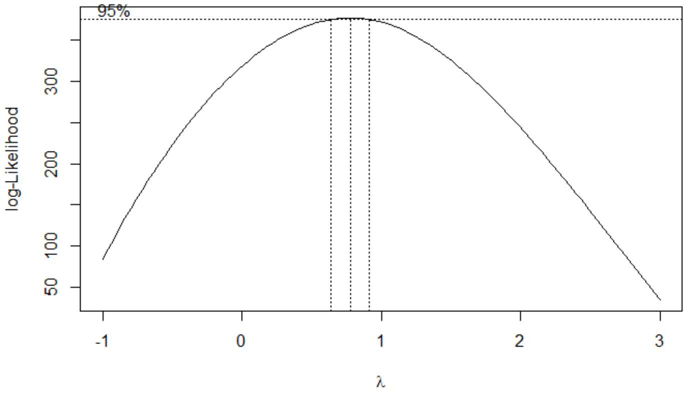


Figure 9:Maximun log-likelyhood for Box-Cox transformation

After re-running the model, the p-value for BP test is still 2.2e-16, meaning the transformed model still hasn't met the homoscedasticity assumption. The p-value for Shapiro-Wilk on the rerun model is 0.0006291, the p-value is higher than the normal model but still under 0.05, meaning the nonnormality problem has not been solved through the Box-cox transformation. Hence, we decided to use the original final model to reduce the complexity of the model.

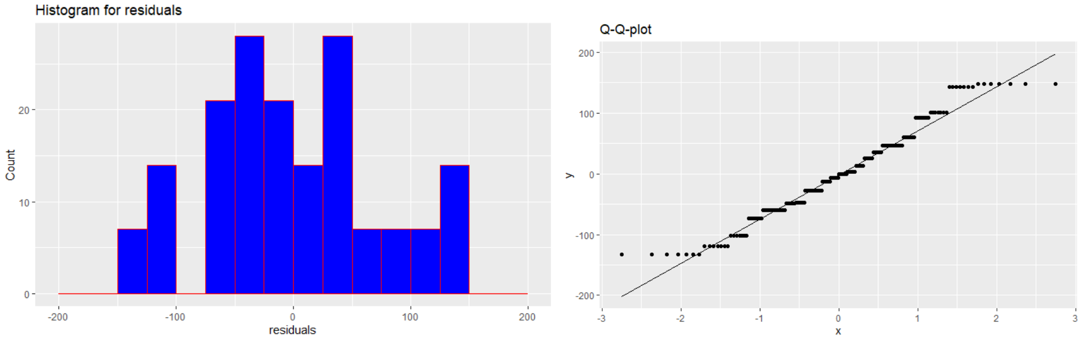


Figure 10 Histogram and QQ-plot after transformation

3.5.6 Outlier Review

Outlier might have a huge effect on the model construction. In this project, we will be using leverage points method to check whether there are outliers in our dataset. The number of predictors in our final model is 19 and the number of the sample size is 168. Hence, the threshold will be 38/168 = 0.2262. The figure below has shown that all hat values are below the threshold, meaning that there is no outlier in our dataset.

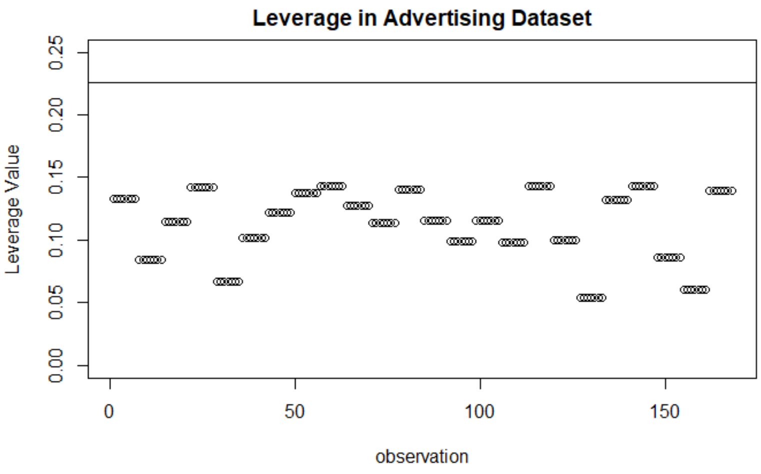


Figure 11: Leverage points method for outliers checking.

3.5.7 Final model selection

After the model selections process and assumption reviews, the final model is summarized in Equation 5.

Equation 5: Final proposed model with estimated parameters

$$
\begin{aligned}
&\widehat{Yearly\_GHG\_Emission} \\
&\quad = -19806.3345 + 97451.4007 Region_{i1} - 451413.0413 Region_{i2} \\
&\quad + 9977.0474 Region_{i3} + 346753.6210 Region_{i4} + 30915.8281 Region_{i5} \\
&\quad + 114.2416 \text{Crude\_Yearly\_Average} + 3.4741 \text{Elec\_Yearly\_Average} \\
&\quad + 192.1408 (Region_{i1})(\text{Crude\_Yearly\_Average}) \\
&\quad + 680.3384 (Region_{i2})(\text{Crude\_Yearly\_Average}) \\
&\quad + 288.4150 (Region_{i3})(\text{Crude\_Yearly\_Average}) \\
&\quad + 40259.1968 (Region_{i4})(\text{Crude\_Yearly\_Average}) \\
&\quad + 389.0909 (Region_{i5})(\text{Crude\_Yearly\_Average}) \\
&\quad - 3.5028 (Region_{i1})(\text{Elec\_Yearly\_Average}) \\
&\quad - 1.5671 (Region_{i2})(\text{Elec\_Yearly\_Average}) \\
&\quad - 0.7520 (Region_{i3})(\text{Elec\_Yearly\_Average}) \\
&\quad - 4.5478 (Region_{i4})(\text{Elec\_Yearly\_Average}) \\
&\quad - 3.9027 (Region_{i5})(\text{Elec\_Yearly\_Average}) \\
&\quad - 0.0012 (\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})
\end{aligned}
$$

3.6 Prediction

Data has been acquired which shows the projected energy usage for 2050 (under scenarios where energy policies have not been updated for environmental conservation considerations), which are summarized in Table 12 (Open Government, 2022b). Thus, the proposed model was used to predict the quantity of GHG emissions based on the energy production projections.

Table 12: Projected Crude Oil Production and Electricity Generation Data for 2050, for Energy Policies Without Environmental Conservation Considerations

Region	Crude Oil Production Prediction	Electricity Generation Prediction
Alberta	2058.5961*	110144.9376

Region	Crude Oil Production Prediction	Electricity Generation Prediction
British Columbia	120.7866	104601.4391
Nova Scotia	0	10027.2334
Ontario	0.1725	187101.9445
Saskatchewan	278.5973	29765.1645
Canada	2516.4795*	777986.0470*

Crude oil production is provided in units of Thousand Cubic Metres per day while electricity generation is provided in units of GWh. Note that rounded values are presented as part of the report but unrounded values were used as part of the calculations.

The values with an asterisk (*) indication shows that the provided quantity is out of range of the data that was used to initially generate the model. Thus, any predictions made would be forecasting (not predicting) and be an inaccurate use of the model. Therefore, only predictions with the provinces of British Columbia, Nova Scotia, Ontario, and Saskatchewan are made; the results can be seen in Table 13.

Table 13: 2050 GHG Emission Predictions

Region	GHG Emission Predictions (Kilotonnes)
British Columbia	96488.56
Nova Scotia	17465.84
Ontario	132981.56
Saskatchewan	128628.20

As expected, Ontario being one of the most populous provinces has the highest quantity of GHG emissions. Although surprisingly this is followed by Saskatchewan, British Columbia, and Nova Scotia.

4.0 Conclusion and Discussion

4.1 Conclusion

The final model proposed to approximate GHG emissions involves the base predictors Region, yearly production of crude oil, and yearly generation of electricity. The model includes interaction terms between all three predictors but no higher order terms.

The linearity and independence assumptions were met but unfortunately the equal variance and normality assumptions were not met. The Box-Cox transformation was applied but it did not help the model meet the assumptions. Lastly, there are no outliers within our dataset.

Multicollinearity was initially observed but that was attributed to the categorical variable Region which had greater than three levels.

<u>The GHG emission for the province of British Columbia</u>:

$$Yearly\_\widehat{GHG}\_Emission = 77645.0662 + 306.3824(\mathrm{Crude\_Yearly\_Average}) - 0.0287(\mathrm{Elec\_Yearly\_Average}) - 0.0012(\mathrm{Crude\_Yearly\_Average})(\mathrm{Elec\_Yearly\_Average})$$

<u>The GHG emission for the province of Nova Scotia</u>:

$$\begin{aligned}Yearly\_\widehat{GHG}\_Emission &= -9829.2871 + 402.6566(\text{Crude\_Yearly\_Average}) + 2.7221(\text{Elec\_Yearly\_Average}) \\ &\quad - 0.0012(\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})\end{aligned}$$

<u>The GHG emission for the province of Ontario:</u>

$$\begin{aligned}Yearly\_\widehat{GHG}\_Emission &= 326947.2865 + 40373.4384(\text{Crude\_Yearly\_Average}) \\ &\quad - 1.0737(\text{Elec\_Yearly\_Average}) \\ &\quad - 0.0012(\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})\end{aligned}$$

<u>The GHG emission for the province of Saskatchewan</u>:

$$\begin{aligned}Yearly\_\widehat{GHG}\_Emission &= 11109.4936 + 503.3325(\text{Crude\_Yearly\_Average}) - 0.4286(\text{Elec\_Yearly\_Average}) \\ &\quad - 0.0012(\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})\end{aligned}$$

<u>The GHG emission for the province of Alberta</u>:

$$\begin{aligned}Yearly\_\widehat{GHG}\_Emission &= -19806.3345 + 114.2416(\text{Crude\_Yearly\_Average}) \\ &\quad + 3.4741(\text{Elec\_Yearly\_Average}) \\ &\quad - 0.0012(\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})\end{aligned}$$

<u>The GHG emission for Canada, as a whole</u>:

$$\begin{aligned}Yearly\_\widehat{GHG}\_Emission &= -471219.3758 + 794.5800(\text{Crude\_Yearly\_Average}) \\ &\quad + 1.9070(\text{Elec\_Yearly\_Average}) \\ &\quad - 0.0012(\text{Crude\_Yearly\_Average})(\text{Elec\_Yearly\_Average})\end{aligned}$$

The summary of the parameter estimates can be found in Table 14.

Table 14: Summary of parameter estimates for each region from the finalized model

	Intercept	Crude_Yearly_Average	Elec_Yearly_Average	Interaction term: (Crude_Yearly_Average)* (Elec_Yearly_Average)
British Columbia	77,645.0662	306.3824	-0.0287	-0.0012
Nova Scotia	-9,829.2871	402.6566	2.7221	-0.0012
Ontario	326,947.2865	40,373.4384	-1.0737	-0.0012
Saskatchewan	11,109.4936	503.3325	-0.4286	-0.0012
Alberta	-19,806.3345	114.2416	3.4741	-0.0012
Canada	-471,219.3758	794.5800	1.9070	-0.0012

4.2 Discussion

The result of this analysis is surprising in that natural gas production was not kept as a significant predictor for GHG emissions while crude oil and electricity was. Upon further research, the significance of crude oil and electricity production in contributing to GHG emissions is supported. As per a report published by the Government of Canada on GHG emissions, when summarizing economic sectors contributions, four sectors are mentioned: oil and gas, transport, agriculture, and electricity (Government of Canada, 2022). Another report published in 2021, titled for the year 2019 confirms that oil and gas remains "Canada's top industrial energy user" (Statistics Canada, 2021a).

It was also surprising that the residential energy consumption was not kept as a predictor as the GHG emission data is the total from industries and households. However, it is possible that consumption drives energy production, which is a more significant and insightful predictor and therefore explains the fact that only production predictors have been kept in the final model.

There are various ways the model can be continually explored, reference sections 4.3 Approach and 4.4 Future Work.

There are drastic differences between each region in terms of contribution for GHG emissions. Surprisingly, Canada has the lowest starting point for GHG emissions. Since Canada is an amalgamation of all the different provinces and territories, it indicates that the other provinces and territories not included as part of this analysis must have low quantities of GHG emissions. Note that the provinces which have been represented as part of this analysis are some of the most populous provinces within Canada and thus, it would be reasonable to propose that the other provinces or territories have lower GHG emissions.

The province which has the highest quantity of GHG emissions without the contributions of yearly crude oil production and electricity generation is Ontario. This is not surprising as Ontario is the most populous province (also with a high density of population). What was surprising is that crude oil production has the largest effect in Ontario while the expected response would have been for Alberta since it is the province known for oil and gas production. Ontario's production contributes to less than 0.1% of Canada's total production (Canada Energy Regulator, 2022). However, it is likely that crude oil production has a larger effect on GHG emissions in Ontario – regardless of the quantity of crude oil produced. In addition, since Alberta produces a large quantity of crude oil, for each thousand cubic metres per day, although the effect on GHG emissions is less, the overall GHG produced could be at a larger quantity. The province most heavily impacted by yearly electricity generation is Alberta when looking at the amount of GHG emissions. This can be due to the fact a large quantity of electricity generated in Alberta comes from fossil fuels (around 89%), which is a non-renewable energy source and actively contributes to GHG emissions (Canada Energy Regulator, 2022).

4.3 Approach

The steps taken to generate the final model is promising as the appropriate pre-screening steps were taken (i.e. determining multicollinearity) prior to model building. In addition, when determining the base predictors which should be kept within the model, various methods were referenced prior to concluding Region, yearly production of crude oil and yearly production of electricity should be kept. However, as seen in section 4.4 Future Work, there are improvements that can be made for building an even more robust model.

The aspects the approach could have been improved has been briefly discussed in the Future Work section as well. First, further research on this topic could have been completed such that the developers of the model were more familiar with this topic and in turn, understand the importance of each predictor and potentially introduce more predictors into the model. The importance of understanding each predictor would have aided in the multicollinearity analysis. Perhaps with further background information, it would have been clear which variables would have had correlation and why it is significant or insignificant. In addition, the approach could have been approved if it would be possible to model using time series modelling techniques and comparing both models to determine the best model to be proposed.

A minor update that would be made if there was a chance to redo the model from the beginning is to report the GHG emission in a higher unit, so the coefficients are easier to manage (i.e. not such large values). For example, in megatonnes instead of kilotonnes.

4.4 Future Work

The following aspects are aspects that can be researched or completed to aid with further understanding of the model:

1) Review of the independence assumption with the 'Year' included as this is the original form of the dataset.
2) A more iterative approach will be taken for inclusion of higher order terms to determine if the addition of these will help satisfy more assumptions. However, note that higher order terms were not initially included in the model to avoid over-fitting. Thus, research needs to be conducted to determine which should be prioritized: avoidance of over-fitting or satisfaction of assumptions? In addition, there could be more appropriate ways of transforming the data (which we have yet to learn as part of DATA603) to meet the assumptions instead of adding higher order terms which may contribute to over-fitting. All these factors need to be considered.
3) Evaluate the multicollinearity observed in the predictor Region. Although it has been deemed that this observance of multicollinearity can be ignored, based on initial understanding of this exclusion, the reasoning behind this is due to unequal proportions (Allison, 2012). However, there are no obvious signs of unequal proportions within this dataset so this concept will be future explored to determine if ignoring this multicollinearity is appropriate.

Beyond these conceptual understanding improvements, additional future work includes:

1) Further research on this topic to continuously add predictors as required.

Originally, these datasets were selected for the estimation of GHG emissions due to initial understanding of this topic from background knowledge and initial research (as well as focussing the scope of analysis to the motivation of energy efficiency). However, since the findings of the model is different from what was expected, additional research can be completed for this topic to find additional predictors to be included in the model for continuous improvement as the understanding of the topic continuously improves.

2) Include new data as necessary.

The data selection for this model pertains to data from the years 2011, 2013, 2015, and 2019 due to dataset limitations. Therefore, it would be ideal to continuously collect additional data from a greater time range to allow for a more robust model as it takes more data points into consideration. GHG emissions data will continue to be collected by the Government of Canada. Thus, with additional time, additional data will become available.

3) Modelling as time series data.

This is out of the scope of this course, however, since the data is originally represented as time series data, it may be appropriate to treat it as such and develop the model accordingly. Although the development of this multiple linear regression model is still accurate, as it takes into account the various values from all the selected predictors and generates a model based on the predictors. It would still be of interest to compare the appropriate model that would be generated when treating the data as time series and this multiple linear regression model.

4) Methods to improve the assumption and multicollinearity observed.

The methods learned as part of the DATA603 course was applied to attempt to meet the assumptions that were failed initially within the evaluation. Unfortunately, the assumptions still are not met after the transformations. Thus, in the future, additional transformation methods can be learned and adopted to improve the model to ensure it meets the assumptions.

5.0 References

1. Allison, P. (2012, September 10). *When Can You Safely Ignore Multicollinearity?* Statistical Horizons. https://statisticalhorizons.com/multicollinearity/
2. Canada Energy Regulator. (2022, July 28). *Provincial and Territorial Energy Profiles – Alberta*. Canada Energy Regulator. https://www.cer-rec.gc.ca/en/data-analysis/energy-markets/provincial-territorial-energy-profiles/provincial-territorial-energy-profiles-alberta.html#:~:text=About%2089%25%20of%20electricity%20in,and%2054%25%20from%20natural%20gas.
3. Environmental Protection Agency. (2022). *Local Energy Efficiency Benefits and Opportunities*. United States Environmental Protection Agency. https://www.epa.gov/statelocalenergy/local-energy-efficiency-benefits-and-opportunities#:~:text=and%20Other%20Sponsors-,Benefits%20of%20Energy%20Efficiency,and%20meet%20growing%20energy%20demand.
4. Government of Canada. (2022). *Greenhouse gas emissions*. Government of Canada. https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html
5. Gundersen, G. (2021, July 12). *Multicollinearity*. https://gregorygundersen.com/blog/2021/07/12/multicollinearity/#:~:text=If%20the%20predictors%20of%20X,predictors%20are%20effectively%20the%20same.
6. IEA. (2021, November 10). *Greenhouse Gas Emissions from Energy Data Explorer*. https://www.iea.org/data-and-statistics/data-tools/greenhouse-gas-emissions-from-energy-data-explorer
7. Lumley, T. (2005, May 11). *[R] Regsubsets()*. https://stat.ethz.ch/pipermail/r-help/2005-May/071305.html
8. MacMillan, A., & Turrentine. (2021, April 7). *Global Warming 101*. https://www.nrdc.org/stories/global-warming-101
9. Ngamkham, T. (2022a). *MLRModelling Part4 Class Notes*. Univeristy of Calgary. https://d2l.ucalgary.ca/d2l/le/content/472042/viewContent/5610393/View
10. Ngamkham, T. (2022b). *MLRModellingPart 3-ClassNotes*. Univeristy of Calgary. https://d2l.ucalgary.ca/d2l/le/content/472042/viewContent/5601101/View
11. Ngamkham, T. (2022c). *MLRModellingPart1-ClassNotes*. Univeristy of Calgary. https://d2l.ucalgary.ca/d2l/le/content/472042/viewContent/5569429/View

12. Ngamkham, T. (2022d). *MLRModellingPart2-ClassNotes*. Univeristy of Calgary. https://d2l.ucalgary.ca/d2l/le/content/472042/viewContent/5572726/View
13. Open Government. (2019). *Open Government Licence—Canada*. Government of Canada. https://open.canada.ca/en/open-government-licence-canada
14. Open Government. (2022a). *Household energy consumption, Canada and provinces* (No. 45fc5597-4ea3-4903-8f9b-405a5468bb0d; Table). Government of Canada. https://open.canada.ca/data/en/dataset/45fc5597-4ea3-4903-8f9b-405a5468bb0d
15. Open Government. (2022b). *Canada's Energy Future 2021: Energy Supply and Demand Projections to 2050*. Government of Canada. https://open.canada.ca/data/en/dataset/5a6abd9d-d343-41ef-a525-7a1efb686300
16. Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., Dubash, N. K., Edenhofer, O., Elgizouli, I., Field, C. B., Forster, P., Friedlingstein, P., Fuglestvedt, J., Gomez-Echeverri, L., Hallegatte, S., … Ypersele, J.-P. van. (2015). *Climate Change 2014 Synthesis Report* (Fifth Assessment Report of the Intergovernmental Panel on Climate Change). https://www.ipcc.ch/site/assets/uploads/2018/02/SYR_AR5_FINAL_full.pdf
17. Placidia. (2012, November 25). *Regsubsets with leaps fails*. StackExchange. https://stats.stackexchange.com/questions/44358/regsubsets-with-leaps-fails
18. ProgrammingR. (2022). *How to Fix the R Error: Missing value where true/false needed*. ProgrammingR. https://www.programmingr.com/r-error-messages/r-error-missing-value-where-true-false-needed/
19. Statistics Canada. (2021a). *Canadian System of Environmental–Economic Accounts: Energy use and greenhouse gas emissions, 2019*. Government of Canada. https://www150.statcan.gc.ca/n1/daily-quotidien/211213/dq211213c-eng.htm
20. Statistics Canada. (2021b). *Physical flow account for greenhouse gas emissions* (https://doi.org/10.25318/3810009701-eng). Government of Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3810009701&cubeTimeFrame.startYear=2009&cubeTimeFrame.endYear=2019&referencePeriods=20090101%2C20190101
21. United Nations. (n.d.). *The Paris Agreement*. United Nations. Retrieved November 7, 2022, from https://www.un.org/en/climatechange/paris-agreement
22. Zach. (2021, October 21). *How to Fix in R: there are aliased coefficients in the model*. Statology. https://www.statology.org/r-aliased-coefficients-in-the-model/

6.0 Appendix

Various detailed outputs are included within this section to ensure the flow and continuity of the main body of the report is not disrupted.

6.1 Table Outputs

Table 15: Detailed VIF Outputs for Multicollinearity Detection, from R notebook

Variable	VIF	Detection
Region – British Columbia	56.7059	1
Region – Canada	1965.2535	1
Region – Nova Scotia	190.3813	1
Region – Ontario	174.5282	1
Region – Saskatchewan	168.7672	1
Crude_Yearly_Average	27.9732	1
Elec_Yearly_Average	2214.2639	1
Hydrogen_Yearly_Average	NaN	NA
NG_Yearly_Average	179.2641	1
Dwelling_Type – Duplex	1.7143	0
Dwelling_Type – High-rise apartment (5 stories or more)	1.7143	0
Dwelling_Type – Low-rise apartment (fewer than 5 stories)	1.7154	0
Dwelling_Type – Mobile home	1.7154	0
Dwelling_Type – Row or terrace	1.7147	0
Dwelling_Type – Single-detached	2.4580	0
Energy_Consumption	2.0760	0

Table 16: Detailed VIF Outputs for Multicollinearity Detection, after Region was removed, from R notebook

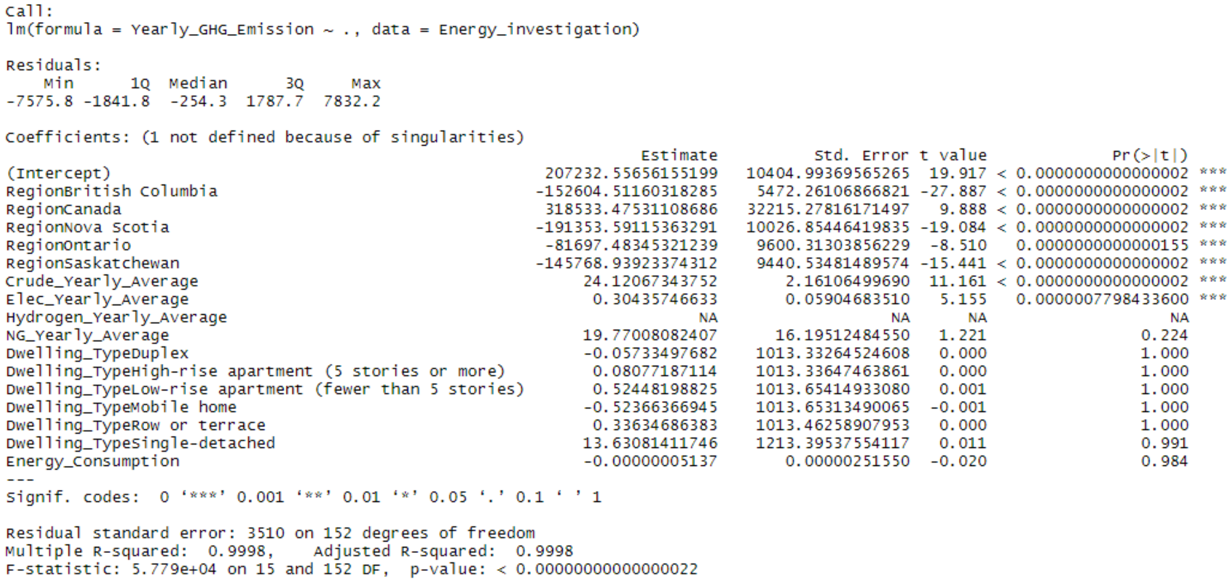
Variable	VIF	Detection
Crude_Yearly_Average	7.5992	0
Elec_Yearly_Average	2.7858	0
Hydrogen_Yearly_Average	NaN	NA
NG_Yearly_Average	4.8318	0
Dwelling_Type – Duplex	1.7143	0
Dwelling_Type – High-rise apartment (5 stories or more)	1.7143	0
Dwelling_Type – Low-rise apartment (fewer than 5 stories)	1.7154	0
Dwelling_Type – Mobile home	1.7154	0
Dwelling_Type – Row or terrace	1.7147	0
Dwelling_Type – Single-detached	2.4527	0
Energy_Consumption	2.0612	0

Table 17: Detailed VIF Outputs for Multicollinearity Detection, after both categorical variables were removed, from R notebook

Variable	VIF	Detection
Crude_Yearly_Average	7.5990	0
Elec_Yearly_Average	2.6352	0
Hydrogen_Yearly_Average	NaN	NA
NG_Yearly_Average	4.8316	0
Energy_Consumption	1.1880	0

6.2 Important Code Outputs

Note, not all code outputs will be included in the report. Only the ones that are deemed suitable and important for communicating crucial findings will be included. For detailed code outputs, please refer to the Rmd file.



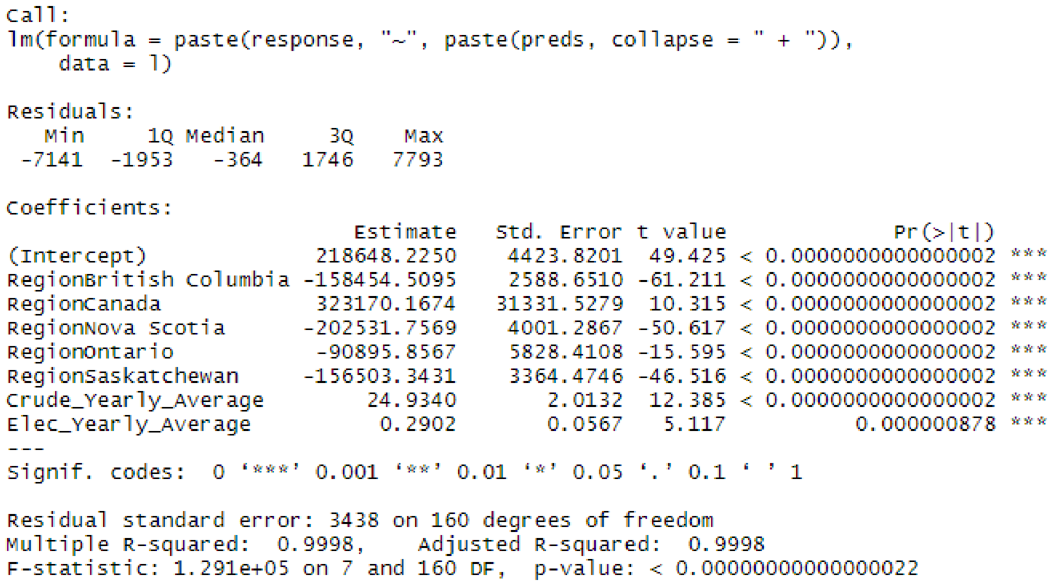
```
Call:
lm(formula = Yearly_GHG_Emission ~ ., data = Energy_investigation)

Residuals:
    Min      1Q  Median      3Q     Max 
-7575.8 -1841.8  -254.3  1787.7  7832.2 

Coefficients: (1 not defined because of singularities)
                                                                  Estimate         Std. Error t value             Pr(>|t|)    
(Intercept)                                             207232.55656155199  10404.99369565265  19.917 < 0.0000000000000002 ***
RegionBritish Columbia                                 -152604.51160318285   5472.26106866821 -27.887 < 0.0000000000000002 ***
RegionCanada                                            318533.47531108686  32215.27816171497   9.888 < 0.0000000000000002 ***
RegionNova Scotia                                      -191353.59115363291  10026.85446419835 -19.084 < 0.0000000000000002 ***
RegionOntario                                           -81697.48345321239   9600.31303856229  -8.510   0.0000000000000155 ***
RegionSaskatchewan                                     -145768.93923374312   9440.53481489574 -15.441 < 0.0000000000000002 ***
Crude_Yearly_Average                                        24.12067343752      2.16106499690  11.161 < 0.0000000000000002 ***
Elec_Yearly_Average                                          0.30435746633      0.05904683510   5.155   0.0000007798433600 ***
Hydrogen_Yearly_Average                                                 NA                 NA      NA                   NA    
NG_Yearly_Average                                           19.77008082407     16.19512484550   1.221                0.224    
Dwelling_TypeDuplex                                         -0.05733497682   1013.33264524608   0.000                1.000    
Dwelling_TypeHigh-rise apartment (5 stories or more)         0.08077187114   1013.33647463861   0.000                1.000    
Dwelling_TypeLow-rise apartment (fewer than 5 stories)       0.52448198825   1013.65414933080   0.001                1.000    
Dwelling_TypeMobile home                                    -0.52366366945   1013.65313490065  -0.001                1.000    
Dwelling_TypeRow or terrace                                  0.33634686383   1013.46258907953   0.000                1.000    
Dwelling_TypeSingle-detached                                13.63081411746   1213.39537554117   0.011                0.991    
Energy_Consumption                                          -0.00000005137      0.00000251550  -0.020                0.984    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3510 on 152 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998 
F-statistic: 5.779e+04 on 15 and 152 DF,  p-value: < 0.00000000000000022
```

Figure 12: Summary output from individual coefficient tests



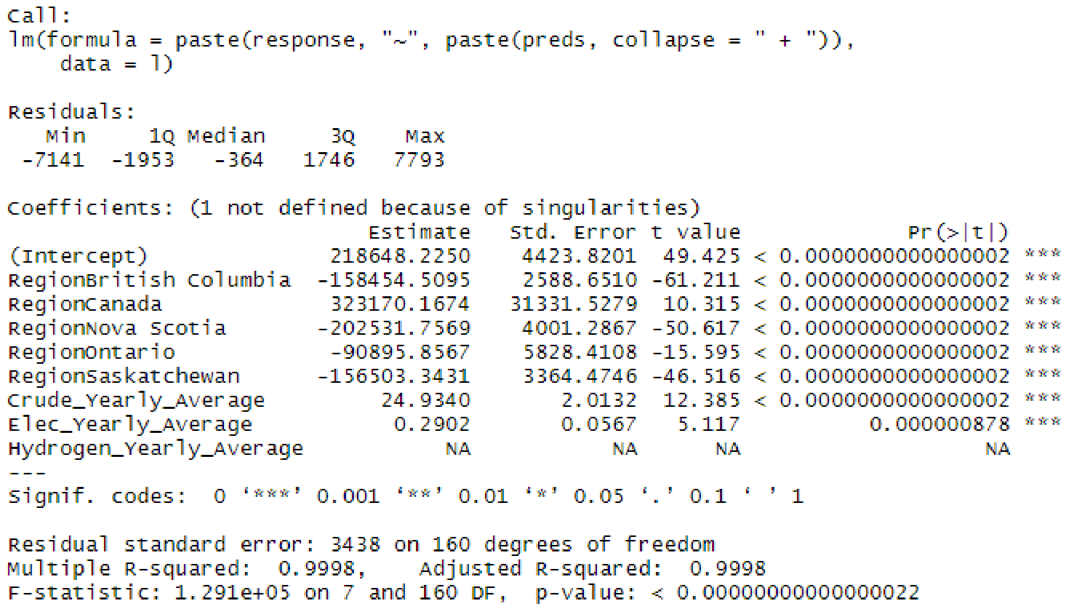
```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")), 
    data = l)

Residuals:
   Min     1Q Median     3Q    Max 
 -7141  -1953   -364   1746   7793 

Coefficients:
                           Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)             218648.2250    4423.8201  49.425 < 0.0000000000000002 ***
RegionBritish Columbia -158454.5095    2588.6510 -61.211 < 0.0000000000000002 ***
RegionCanada            323170.1674   31331.5279  10.315 < 0.0000000000000002 ***
RegionNova Scotia      -202531.7569    4001.2867 -50.617 < 0.0000000000000002 ***
RegionOntario           -90895.8567    5828.4108 -15.595 < 0.0000000000000002 ***
RegionSaskatchewan     -156503.3431    3364.4746 -46.516 < 0.0000000000000002 ***
Crude_Yearly_Average        24.9340       2.0132  12.385 < 0.0000000000000002 ***
Elec_Yearly_Average          0.2902       0.0567   5.117          0.000000878 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3438 on 160 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998 
F-statistic: 1.291e+05 on 7 and 160 DF,  p-value: < 0.00000000000000022
```

Figure 13: Summary output from Stepwise Selection Procedure



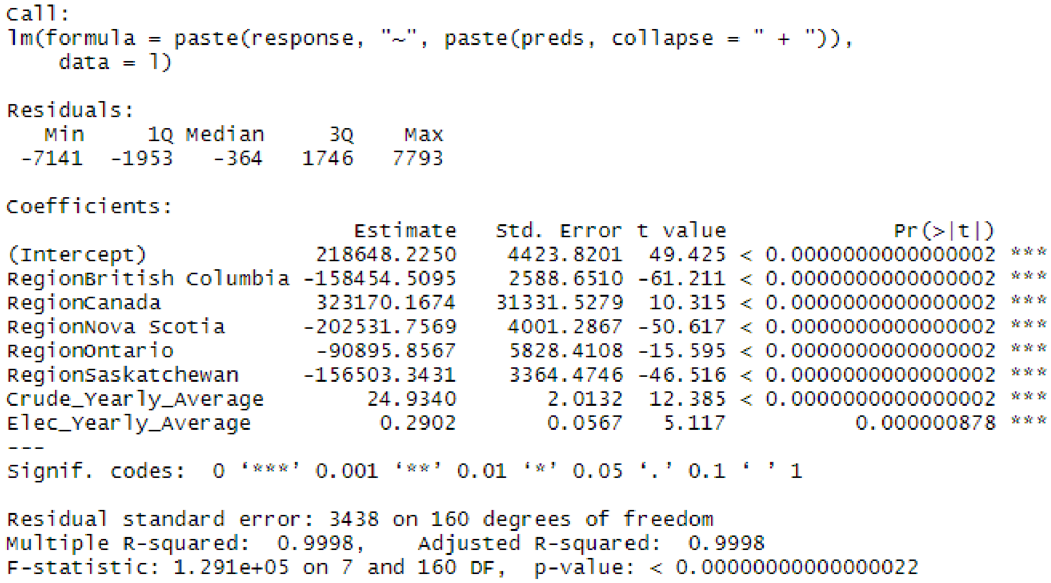
```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")), 
    data = l)

Residuals:
   Min     1Q Median     3Q    Max 
 -7141  -1953   -364   1746   7793 

Coefficients: (1 not defined because of singularities)
                            Estimate   Std. Error t value             Pr(>|t|)    
(Intercept)              218648.2250    4423.8201  49.425 < 0.0000000000000002 ***
RegionBritish Columbia  -158454.5095    2588.6510 -61.211 < 0.0000000000000002 ***
RegionCanada             323170.1674   31331.5279  10.315 < 0.0000000000000002 ***
RegionNova Scotia       -202531.7569    4001.2867 -50.617 < 0.0000000000000002 ***
RegionOntario            -90895.8567    5828.4108 -15.595 < 0.0000000000000002 ***
RegionSaskatchewan      -156503.3431    3364.4746 -46.516 < 0.0000000000000002 ***
Crude_Yearly_Average         24.9340       2.0132  12.385 < 0.0000000000000002 ***
Elec_Yearly_Average           0.2902       0.0567   5.117          0.000000878 ***
Hydrogen_Yearly_Average           NA           NA      NA                   NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3438 on 160 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998 
F-statistic: 1.291e+05 on 7 and 160 DF,  p-value: < 0.00000000000000022
```

Figure 14: Summary output from Backward Regression Procedure



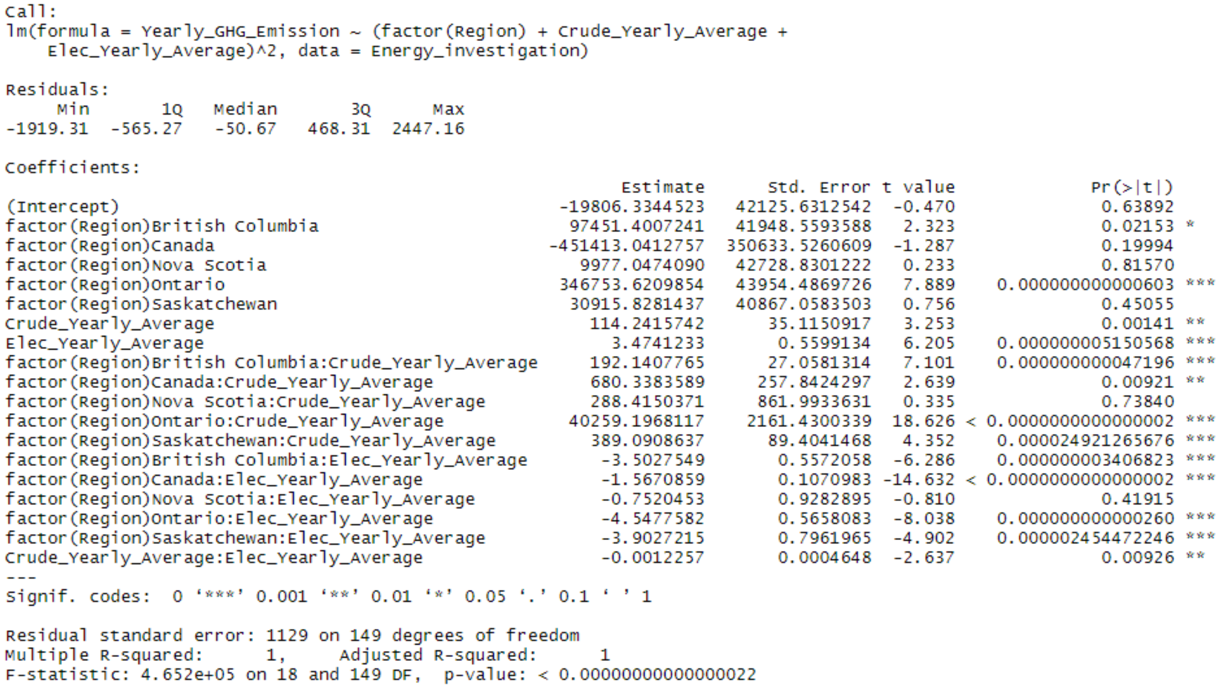
```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")), 
    data = l)

Residuals:
   Min     1Q Median     3Q    Max 
 -7141  -1953   -364   1746   7793 

Coefficients:
                          Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)            218648.2250   4423.8201  49.425 < 0.0000000000000002 ***
RegionBritish Columbia -158454.5095   2588.6510 -61.211 < 0.0000000000000002 ***
RegionCanada           323170.1674  31331.5279  10.315 < 0.0000000000000002 ***
RegionNova Scotia     -202531.7569   4001.2867 -50.617 < 0.0000000000000002 ***
RegionOntario          -90895.8567   5828.4108 -15.595 < 0.0000000000000002 ***
RegionSaskatchewan    -156503.3431   3364.4746 -46.516 < 0.0000000000000002 ***
Crude_Yearly_Average       24.9340      2.0132  12.385 < 0.0000000000000002 ***
Elec_Yearly_Average         0.2902      0.0567   5.117          0.000000878 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3438 on 160 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998 
F-statistic: 1.291e+05 on 7 and 160 DF,  p-value: < 0.00000000000000022
```

Figure 15: Summary output from the final proposed model with Region, Crude_Yearly_Average, and Elec_Yearly_Average



```
Call:
lm(formula = Yearly_GHG_Emission ~ (factor(Region) + Crude_Yearly_Average + 
    Elec_Yearly_Average)^2, data = Energy_investigation)

Residuals:
     Min       1Q   Median       3Q      Max 
-1919.31  -565.27   -50.67   468.31  2447.16 

Coefficients:
                                                 Estimate     Std. Error t value             Pr(>|t|)    
(Intercept)                                -19806.3344523  42125.6312542  -0.470              0.63892    
factor(Region)British Columbia              97451.4007241  41948.5593588   2.323              0.02153 *  
factor(Region)Canada                      -451413.0412757 350633.5260609  -1.287              0.19994    
factor(Region)Nova Scotia                    9977.0474090  42728.8301222   0.233              0.81570    
factor(Region)Ontario                      346753.6209854  43954.4869726   7.889    0.000000000000603 ***
factor(Region)Saskatchewan                  30915.8281437  40867.0583503   0.756              0.45055    
Crude_Yearly_Average                          114.2415742     35.1150917   3.253              0.00141 ** 
Elec_Yearly_Average                             3.4741233      0.5599134   6.205    0.000000005150568 ***
factor(Region)British Columbia:Crude_Yearly_Average   192.1407765     27.0581314   7.101    0.000000000047196 ***
factor(Region)Canada:Crude_Yearly_Average     680.3383589    257.8424297   2.639              0.00921 ** 
factor(Region)Nova Scotia:Crude_Yearly_Average   288.4150371    861.9933631   0.335              0.73840    
factor(Region)Ontario:Crude_Yearly_Average  40259.1968117   2161.4300339  18.626 < 0.0000000000000002 ***
factor(Region)Saskatchewan:Crude_Yearly_Average   389.0908637     89.4041468   4.352    0.000024921265676 ***
factor(Region)British Columbia:Elec_Yearly_Average    -3.5027549      0.5572058  -6.286    0.000000003406823 ***
factor(Region)Canada:Elec_Yearly_Average       -1.5670859      0.1070983 -14.632 < 0.0000000000000002 ***
factor(Region)Nova Scotia:Elec_Yearly_Average    -0.7520453      0.9282895  -0.810              0.41915    
factor(Region)Ontario:Elec_Yearly_Average      -4.5477582      0.5658083  -8.038    0.000000000000260 ***
factor(Region)Saskatchewan:Elec_Yearly_Average    -3.9027215      0.7961965  -4.902    0.000002454472246 ***
Crude_Yearly_Average:Elec_Yearly_Average       -0.0012257      0.0004648  -2.637              0.00926 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1129 on 149 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 4.652e+05 on 18 and 149 DF,  p-value: < 0.00000000000000022
```

Figure 16: Summary output from the final proposed model with Region, Crude_Yearly_Average, Elec_Yearly_Average and the appropriate interaction terms