Analysis on Association between Natural Gas Consumption and Greenhouse Gas Emission

YUNFEI YU - 1005976918

Introduction

Since the pre-industrialization period in the mid-1800s and early 1900s, gradual heating in the Earth's climate system had been observed (NASA, 2021). Increasingly with human activities, especially the production and consumption of various types of energy such as fuel, coal and natural gas, global warming is intensified by the emission of greenhouse gas which traps heat within the atmosphere (NASA, 2021). Studies have shown that since the pre-industrialization period, the global temperature has risen by an average of 1 degree Celsius, and is continuing to rise by 0.2 degrees Celsius per decade (NASA, 2021). Global warming, and climate change resulting from it, leads to natural disasters and lack of resources and are posing threat to human beings as a whole. At the center of the issue with global warming is greenhouse gas emission due to the production and consumption of energy. Thus, reducing greenhouse gas is at the core of mitigating global warming. However, despite efforts in developing renewable energy, "traditional", non-renewable types of energy are still essential to daily life and production. Although controlling and hopefully reducing damage to the environment is urgent, the consumption and production of energy are also inevitable to satisfy basic human needs. Hence, the problem regarding finding the balance between the consumption of energy and mitigating global warming rises. Emerging naturally from this problem is the question of **how is our daily consumption of energy related to greenhouse gas emission**. What types of renewable or non-renewable energy contributes to greenhouse gas emissions and how significant is the contribution? In addition, one may also be interested in what types of households or operations emit more greenhouse gas than others and how is that influenced by their usage of energy? Intuitively, one may expect that the more energy consumed, regardless of the type of energy, the more greenhouse gas emitted; and operations that use more energy are the ones that operate for a longer duration.

Aiming at providing insights into these questions and assumptions, this research performs an analysis on the *Annual Energy Consumption* dataset (Environment and Energy, Open data dataset) provided by the Open Toronto Data Portal and published by Environment & Energy. The dataset is updated annually with the last update in 2018. The most recent version (2018 version) is the one used in this analysis. The dataset contains information on energy consumption and greenhouse gas emission of 1483 operations of various types and areas in the city of Toronto. The types, addresses, total floor area and weekly operating hours of each operation are also reported by the dataset. Among the energies consumed and reported in the dataset, **natural gas** is one of the most widely used and of the most important types. Natural gas, essential to both production and domestic activities, is used in a variety of ways from electricity generation and agriculture to domestic heating and cooking fuel (Uses for natural gas: Canada's oil and natural gas producers 2021). Furthermore, research done by Statistics Canada (Canada, 2020) has shown that natural gas production and consumption is a large contributor to the overall emission of greenhouse gas in Canada between the years 2008 and 2018. Natural gas consumption also accounts for 30% of total secondary energy consumption, which is the energy used by the final consumer in the economy, in Canada (Canada, 2020). Thus, it is meaningful to perform an analysis of natural gas consumption, aiming to unveil the association between natural gas consumption and greenhouse gas emission.

Finally, before ending this section, a brief overview of the analytic methods and statistical terminologies will be provided. To investigate the association between natural gas consumption and greenhouse gas emission, one or more *linear regression* models will be proposed and checked for appropriateness and validity. A detailed description of linear regression will be provided in the **Methods** section. However, prior to performing any analysis, it is necessary to ensure the data is in an analyzable format. Since the raw data retrieved from the Open Toronto Data Portal appears to be messy, it is cleaned and reduced to include only the valid *observations* or rows and the *variables* or columns of the primary interest. The detailed cleaning process will be introduced in the **Data** section of the report. Towards the end of the report are the **Results** and **Conclusion** sections, where the results section contains numerical results and their meaning and the conclusions comment on the overall takeaway of the analysis, limitations of the methods and suggestions for further study on the topic.

Data

The *Annual Energy Consumption* data used for this analysis is published by the City of Toronto's Environment and Energy division. Included in the dataset is information on energy usage and greenhouse gas emission of operations within the city of Toronto. The original data are in the form of an Excel sheet, containing basic information about the operations and data of their energy usage from January 2018 to December 2018. Certain columns are indicated in red and are mandatory for the operations to fill in, in addition to providing energy usage data (Environment and Energy, *Open data dataset*). Such columns include general information about the operation such as name, type, address, floor area and weekly operating hours. A total of 1483 operations provided data or information for 36 columns and are included in the original dataset (Environment and Energy, *Open data dataset*). However, due to the self-reporting nature of the data collection process, it is likely that there may be missing and invalid values of the variables, or that the values reported may be of inconsistent units.

When reading the original Excel sheet into RStudio, the data appears messy due to its complex layout. To be clear, the first few rows of the Excel sheet contain the title, the name of the organization that is collecting the data (i.e. City of Toronto), the twelve-month period and requirements for operations filling out the form. Besides, the column *energy type and amount purchased and consumed in natural units* is divided into sub-columns based on energy type, which are further divided to indicate the amount and corresponding units separately. Such layout leads to chaos when imported into RStudio, thus it is necessary to clean the data for interpretability. The first step of the data cleaning process is to delete redundant header rows such that the first row corresponds to the first operation (first observation). Since the observations start on the 8th row, the first 7 rows are removed. The removal of redundant rows is done by creating a new dataset by indexing `-c[1:7]` on the original dataset. Next, new column names are given to each column, corresponding to the column names in the original dataset. Sub-columns are given new variable names and are thus no longer affiliated to an overarching category. For example, in the original dataset, the column for *Electricity* is divided into *Quantity* and *Unit*; but in the cleaned dataset, this becomes a column for *Electricity_quantity* and another for *Electricity_unit*; similar is done for other energy types. However, the order of the columns is left unchanged. In addition, the column corresponding to the city-generated category of the operations is named "Comments" in the original dataset. This name is quite vague and thus is renamed in the new dataset as *Category.* The second step is done using the *colnames* function in R, with the input being a vector containing all 36 column names, while the renaming of columns is done through the *rename* function from the tidyverse package, setting the new variable name *Category* equal to the original name "Comments".

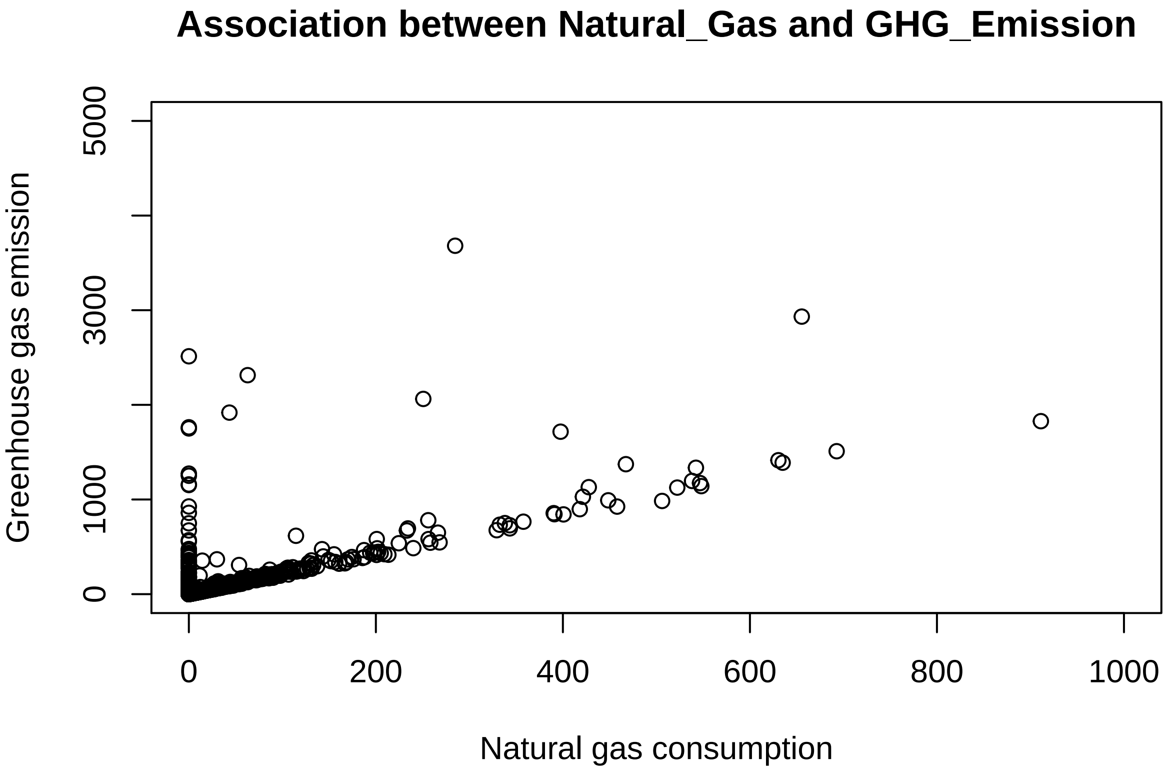
After removing redundant header rows and setting new column names, the data are in a readable and analyzable format. The next few steps of data cleaning regard shrinking the rather large dataset to include only variables of interest and removing invalid observations. Firstly, an *id* variable is created by piping the new dataset into the *mutate* function from the tidyverse package in RStudio with input being an equation setting the name of the variable (id) equal to an integer vector from 1 to 1483, which is the total number of observations, for referencing the observations in later analysis. Then, two additional variables (*Natural_Gas* and *GHG_Emission*) are created which contain numerical values corresponding to the variables *Natural_Gas_Quantity* and *GHG_Emission_kg.* This step may seem redundant, but it is necessary since RStudio reads in all values from Excel sheets as strings such that the values of variables for natural gas consumption (*Natural_Gas_Quantity*) and greenhouse gas emission (*GHG_Emission_kg*) are treated as "words" instead of numbers. However, since natural gas consumption or greenhouse gas emission are meant to be numerical, it is of necessity to convert their values from strings to true numbers to avoid any confusion in the analyzing process. Although values of variables for other types of energy consumption are also treated as strings instead of numbers, those variables will not be included in the final cleaned dataset since they are not of primary interest, thus their values are left unchanged at this step. The creation of *Natural_Gas* and *GHG_Emission* is also done using the *mutate* function in a similar fashion as when creating the *id* variable; the conversion from string to numerical values is done through the *as.numeric* function where the input is the corresponding variable name (*Natural_Gas_Quantity* or *GHG_Emission_kg*).

Since the variables of primary interests are greenhouse gas emission (*GHG_Emission*) and natural gas consumption (*Natural_Gas*), they are included in the final dataset whereas variables for other types of energy are excluded. The variable indicating the unit of natural gas consumption (*Natural_Gas_Unit*) is also included. In addition, the average weekly operation duration (*Avg_hrs_per_week*) is also a variable of interest and is included in the dataset. It is worth noting that, although *Avg_hrs_per_week* contains numerical values, such as 70, 100 and 168, it is in fact a categorical variable with each numerical value corresponding to a level of operation duration in hours with no values in between the categories. The selection of variables is done through the *select* function from the tidyverse package with inputs being the names of the variables that are being included. After the selection of variables is done, included in the dataset are a total of five variables (*id, GHG_Emission, Natural_Gas, Natural_Gas_Unit and Avg_hrs_per_week*). In addition, invalid observations are removed from the dataset. To be more specific, observations with inconsistent units and/or missing values are considered invalid. Since the units of greenhouse gas emission and average weekly operation hours are specified in the original Excel sheets, the only variable that may cause confusion in terms of units is natural gas consumption. By skimming through the entire dataset, it is noticed that the vast majority if not all observations reported natural gas consumption in terms of cubic meters. However, to ensure the units are actually consistent throughout the dataset, the command `filter(Natural_Gas_Unit == 'Cubic meter')` is added where the *filter* function selects observations that satisfy the condition in the input, that is, the unit is in cubic meter. The observations with missing values in any of the variables are detected by the *is.na* function which returns `TRUE` in the case with missing values and `FALSE` otherwise. The input to *is.na* function is the variable names. Similarly, the selection of variables without missing values is also done using the *filter* function with the condition being `is.na` returning `FALSE` (or `is.na` returning `TRUE`) for all variables. After the selection of observations and variables, the cleaned dataset contains **1481 observations and 5 variables**. The final step of the cleaning process modifies the values of *Natural_Gas* and *GHG_Emission.* Since the values of these two variables spread over a wide range from 0 to several hundred thousand, the scale is modified by dividing the values by 1000 and round to 5 decimal places to avoid complex numbers. Correspondingly, *Natural_Gas_Unit* is now in terms of thousand cubic meters as opposed to cubic meters, and greenhouse gas emission in thousand Kg. The change of variable values is done through the *mutate* function, with input being the corresponding variable with values either being original values divided by 1000 and then rounded to 5 decimal places using the *round* function (for *Natural_Gas* and *GHG_Emission*) or being 'Thousand cubic meters'.

After the entire cleaning process, the 5 variables included in the dataset are *id* of each observation, *GHG_Emission* or greenhouse gas emission in thousand Kg which is the dependent or response variable, *Natural_Gas* or natural gas consumption in thousand cubic meters which is a continuous predictor or independent variable, *Natural_Gas_Unit* representing the unit of natural gas consumption, and finally a categorical variable *Avg_hrs_per_week* for the average weekly operating hours of each observation.

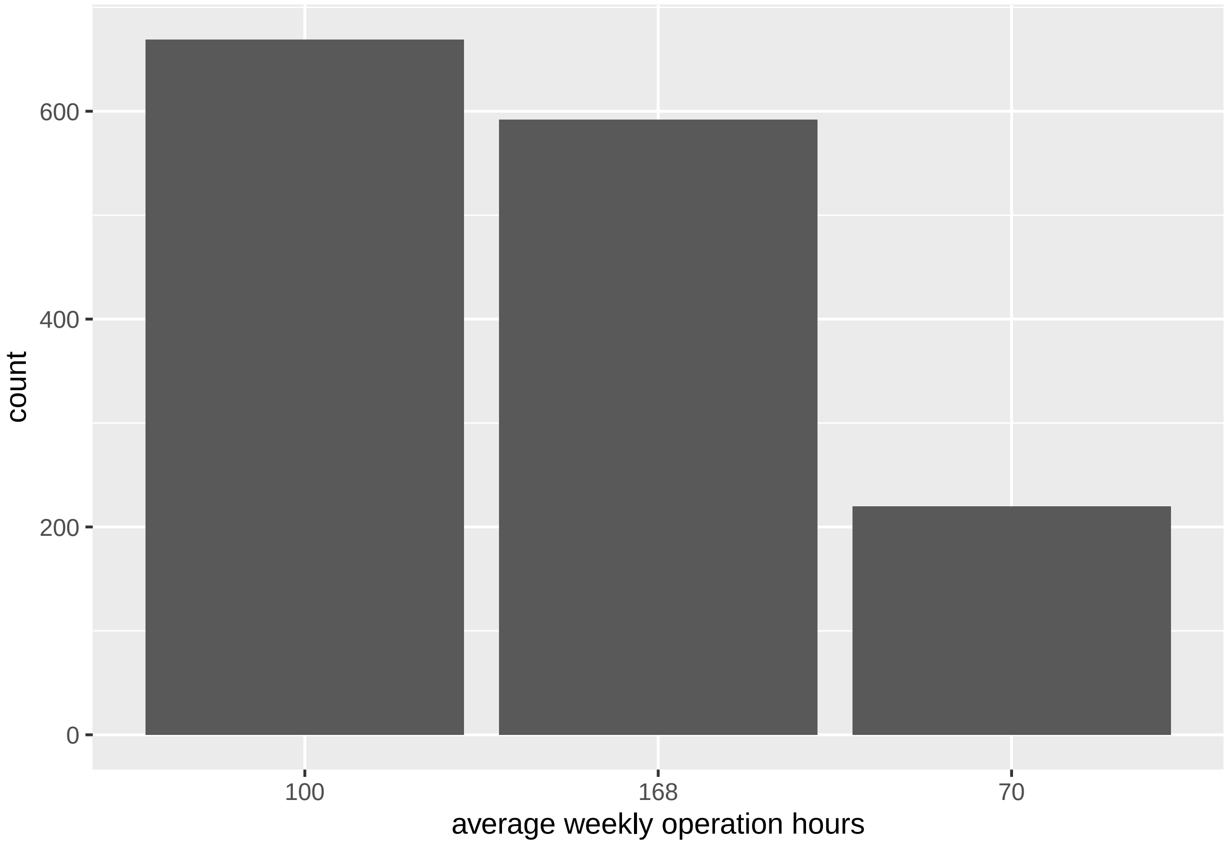
Finally, some numerical and graphical summaries of the dataset will be presented.

The scatterplot below shows the relationship between natural gas consumption and greenhouse gas emission. However, observations with values on the extreme end of the scale are not included in this graph for a better representation. This does not mean that those observations are removed from the dataset; instead, they are still included in the dataset and in further analysis but are excluded only in this plot so that the scatterplot better showcases the overall pattern. As shown in the scatterplot, *Natural_Gas* and *GHG_Emission* are mostly linearly related except for observations clustered around 0. This indicates that it may be appropriate to use linear regression to model the relation between these two variables. However, the model may explain only part of the relationship as the value of *GHG_Emission* seem to vary a lot despite the same level of natural gas consumption.



Graph 1: Association between natural gas consumption and greenhouse gas emission

Next is a bar plot showing the distribution of observations over different levels of average operation hours per week (*Avg_hrs_per_week*). The three categories of *Avg_hrs_per_week* are 70, 100 and 168. The majority of the observations operate at the 100-hour and 168-hour levels, with over 650 observations (over 43%) operating 100 hours per week on average and approximately 600 (about 40%) operating 168 hours. Only a little over 200 observations operate at the 70-hour level per week, which is about 13.5% of the total observations.



Graph 2: Distribution over average weekly operation hours

All analysis for this report was programmed using `R version 4.1.1`.

Methods

To perform the analysis on the association between greenhouse gas emission (*GHG_Emission*) and natural gas consumption (*Natural_Gas*), along with average weekly operating hours (*Avg_hrs_per_week*), both **simple linear regression** (SLR) and **multiple linear regression** (MLR) will be performed. The choice of linear regression is based on several facts and summaries of the data. Since the response variable (*GHG_Emission*) is continuous, it fits for linear regression as opposed to other types of regression, such as logistic regression which regress on categorical response variables. Besides, as demonstrated by the scatterplot (Graph1) in the **Data** section, *GHG_Emission* and *Natrual_Gas* seem to show a linear relation. Hence, fitting a linear regression model is reasonable in this case.

Theoretically, the model for SLR is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

, where Y is the response or dependent variable and X is the predictor or independent variable. β_0 and β_1 are the intercept and slope parameters, respectively. To be more specific, the value of β_1 can be interpreted as the average amount of the change in Y for a one-unit change in X; while the value of β_0 represents the value of the response when X is 0. The interpretation of β_0 is based on a theoretical setup, however, the value of β_0 may not always make practical sense. ϵ is the random error parameter accounting for the discrepancies between true values and predicted values of the response (Sheather, 2009).

Similarly, the theoretical model for MLR takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X2 + \ldots + \beta_p X_p + \epsilon$$

. While the interpretations for β_0 and ϵ are consistent with SLR, the interpretation for β_i is instead the average amount of change in Y for a one-unit change in the corresponding X_i holding all other X's constant. In this analysis, the response variable Y will be *GHG_Emission* and predictors X will be *Natural_Gas*, which is continuous, and *Avg_hrs_per_week*, which is categorical with three categories.

In addition, for correlated predictors, an interaction term may also be added to the multiple linear regression model. In the case with a continuous predictor X (*Natural_Gas*) and a categorical predictor D (*Avg_hrs_per_week*), the theoretical MLR model takes the form

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_{12} XD + \epsilon$$

(Sheather, 2009), where β_{12} is the interaction effect, representing the change in slope β_1 by the value of D (Sheather, 2009).

To perform a thorough analysis on the association between response and predictor variables, an SLR model is fit on *GHG_Emission* versus *Natural_Gas* and an MLR model will be fit on the *GHG_Emission* and both predictors (*Natural_Gas* and *Avg_hrs_per_week*). In addition, the interaction of the predictors may also be included in the MLR model since, intuitively, different operation duration may have an impact on the pattern of energy consumption. Thus, it makes sense to check the interaction between *Natural_Gas* and *Avg_hrs_per_week.* However, before performing the actual regression, several assumptions must be met to ensure that the model is appropriate and valid. The first assumption regards the linearity of the relationship between predictors and the response, which forms the basis of a linear regression model (Sheather, 2009). If the linearity assumption is violated, the linear regression model would be invalid since the relation is not linear. The second and third assumptions regard the distribution of the error term. To be more specific, for a linear regression model to be valid, the errors are supposed to be independently and normally distributed with constant variance (Sheather, 2009). Thus, to check these assumptions, residual plots will be constructed. Residuals are estimated values of the errors, defined as the difference between the fitted values of y and the actual values of Y. Residual plot, which results from plotting the residuals against fitted values of Y, provides insight into whether the assumptions are met. Specifically, if there is no salient pattern of the residuals so that they are evenly spread across the entire space, it is an indicator that all assumptions hold and the model is valid (Sheather, 2009). On the other hand, any pattern in the residual plot indicates that at least one of the assumptions is not met and transformations of the model may be required.

To check whether the assumptions are met, a residual plot (see Appendix) is created for each of the models by plotting the residuals against the fitted values response variable. The residual plots show no significant linear or fanning pattern which provides evidence that the linearity and constant variance assumptions are not violated. Although on the graph the residuals appear to be clustered at the lower end of the scale, it is due to the existence of several extreme values of *GHG_Emission* on the higher end which stretches out the scale, while the majority of data points are on the lower end. When "zooming in" to focus on only the lower range of values, the residuals show no salient pattern. Additionally, a normal Q-Q plot for each model is constructed to check the normality assumption (see Appendix). A normal Q-Q plot plots the quantiles of standardized residuals of the model against the theoretical quantiles of a standard normal distribution. If the residuals appear to be a straight line (i.e the relation is one-to-one), then there is evidence that the normality assumption is met. The Q-Q plots for both the SLR and the MLR models show an overall one-to-one relationship between the quantiles of the standardized residuals and of standard normal. Although the points at both ends seem to wiggle, they are not too off of the line considering the large values of the response. Hence, there is evidence that the normality assumption is also met. Thus, based on the results of the residual plots and normal Q-Q plots, both SLR and MLR models satisfy all assumptions and thus are valid and appropriate for analyzing the association between *GHG_Emission*, *Natural_Gas* and *Avg_hrs_per_week.*

A total of three linear regression models are fit to the response variable, with one SLR model being *GHG_Emission* regressed solely on *Natural_Gas*, an MLR model fit on *GHG_Emission* and both predictors without an interaction term, and another MLR model with an interaction term. The linear regression takes a Frequentist approach, which assumes all parameters of interest are fixed real numbers. The linear regression models are created in r by the `lm()` function, where the results of the models are obtained by the `summary(model)` command. To select the model that best describes the association between the response and predictors, the model with the highest adjusted R^2 value is chosen. R^2 is the coefficient of determination indicating the proportion of total sample variability in the response variable that is explained by the regression model (Sheather, 2009), thus for a fixed number of predictors in the model, larger values of R^2 indicate a better fit. Typically, as more predictors are included in the model, R^2 increases, but in this case, larger values of R^2 do not guarantee a better model (Sheather, 2009). Thus, when comparing models with different numbers of predictors, for instance when comparing an SLR model to MLR models, adjusted R^2 is preferred over R^2 because it adjusts for the rise in R^2 due to the increase in the number of predictors (Sheather, 2009). The third model, which is the multiple linear regression model including both predictors and an interaction term of the two predictors, has the highest value of adjusted R^2 and thus is selected as the final model.

Results

The SLR model fit on *GHG_Emission* and *Natural_Gas* results in a slope parameter β_1 of approximately 2.397 and an intercept parameter β_0 of approximately 28.217. Thus, based on the SLR model, for a one thousand cubic meter increase in natural gas consumption, greenhouse gas emission increases by 2.397 thousand Kg on average; at 0 cubic meters natural gas consumption, greenhouse gas emission is about 28.217 thousand Kg. Compared to the scatterplot (Graph 1) of *GHG_Emission* versus *Natural_Gas* provided in the **Data** section, the regression parameters seem reasonable that natural gas consumption is positively linearly related to greenhouse gas emission, but does not determine greenhouse gas emission on its own since even without consumption of natural gas, greenhouse gas is still emitted. The adjusted R^2 of the SLR model is 0.9161, which means that 91.61% of the sample variability in GHG_Emission is explained by the model. This also aligns with the scatter plot which presents a strong positive linear relationship between the two variables except at 0 natural gas consumption. The value of the adjusted coefficient of determination indicates that natural gas consumption is a strong predictor of greenhouse gas emission, but does not fully determine it.

The results from fitting a multiple linear regression model on *GHG_Emission* versus *Natural_Gas* and *Avg_hrs_per_week* (without accounting interaction between the two predictors) show an adjusted R^2 value of 0.9174 such that the MLR model explains 91.74% of the sample variability in *GHG_Emission.* Since the SLR model accounts for 91.61% of the variability in the response, this also implies that the average weekly operation hours is not a strong predictor of greenhouse gas emission as adding it to the model does not contribute much to the explanation of the total variability. However, *Avg_hrs_per_week* does pose some influence on the average change in *GHG_Emission.* To be more specific, when *Avg_hrs_per_week* is 100, that is, when an operation operates at the 100-hour level per week, greenhouse gas emission increase by 2.391 thousand Kg on average for a one thousand cubic meter increase in natural gas consumption; for operations operating at the 70-hour level, an average of 49 thousand Kg more greenhouse gas is emitted compared to those operating at the 100-hour level when natural gas consumption is held constant; and for operations that operate 168 hours per week emit 51.97 thousand Kg more greenhouse gas, on average, than those operating 100 hours at the sample level natural gas consumption. β_0 of the MLR model is 0.37 such that 0.37 thousand Kg of greenhouse gas is emitted when natural gas is not consumed at all.

Finally, an MLR model was fitted on GHG_Emission versus Natural_Gas and Avg_hrs_per_week with an additional interaction term, which is also selected as the **final model** based on the value of adjust R^2. The final model indicates that when no natural gas is consumed, an average of approximately 4.94 thousand Kg of greenhouse gas is emitted. For operations that operate 100 hours per week on average, the average greenhouse gas emission rise by 2.15 thousand Kg for a one thousand cubic meter increase in natural gas consumption. For operations operating at the 168-hour level, an additional 46.69 thousand Kg of greenhouse gas is emitted on average at the same level of natural gas consumption and the average effect of natural gas consumption on greenhouse gas emission increase by 0.26. That is, operations that operate 168 hours per week emits approximately 2.41 thousand Kg greenhouse gas for a unit increase in natural gas consumption, compared to 2.15 thousand Kg for operations that operates 100 hours. In addition, for operations operating at the 70-hour level, an average of 65.23 thousand Kg more greenhouse gas is emitted compared to those operating at the 100-hour level when natural gas consumption is held constant and the average effect of natural gas consumption on greenhouse gas emission decrease by 0.17. Thus, for operations with a weekly operation duration of 70 hours, 1.97 thousand Kg more greenhouse gas is emitted for a one thousand cubic meter increase in natural gas consumption. The adjusted R^2 value for the second MLR model is 0.9181 such that this model explains 91.81% of the total variability in *GHG_Emission*, which is also the **highest** of all 3 models fitted to the data.

The three linear regression models have similar adjusted R^2 values and all models explain over 91% of the sample variability in the response variable, *GHG_Emission.* The linear regression model with the largest value of adjusted *R ˆ2*, and thus best explains the sample variability in the response, is the multiple linear regression model that include both predictors (*Natural_Gas* and *Avg_hrs_per_week*) and the interaction between the predictors. Hence, this multiple linear regression model (the last model) is selected as the final model and the results of this model, as described above, is included in the following table (Table 1)

Table 1: Results of multiple linear regression model (with interaction term)

Parameter Estimate	Value
$\hat{\beta}_0$	4.9440
$\hat{\beta}_1$	2.1456
$\hat{\beta}_2$, *average weekly operating hours* = 168	46.6855
$\hat{\beta}_2$, *average weekly operating hours* = 70	65.2257
$\hat{\beta}_{12}$, *average weekly operating hours* = 168	0.2597
$\hat{\beta}_{12}$, *average weekly operating hours* = 70	0.1736

All analysis for this report was programmed using `R version 4.1.1`. I used the `lm()` function in base `R` to derive the estimates of a frquentist linear regression in this section [4].

Conclusions

Natural gas, among a variety of types of energy, is both an essential resource for production and a major source of consumption (Canada, 2020). It is expected that the annual natural gas consumption for operations serves as a significant predictor of greenhouse gas emission. Intuitively, higher levels of natural gas consumption would correspond to a higher level of greenhouse emission and the opposite would also be true. In addition, household or operations that operates for a longer duration per week are expected to consume more energy and thus emit more greenhouse gas. A scatterplot (Graph 1) of natural gas consumption versus greenhouse gas emission based on the *Annual Energy Consumption* data collected by the city of Toronto in 2018 provided insight into the association between the two variables, showing a seemingly linear, positive and quite strong trend which seems to support the hypothesis. To verify the association shown by the scatterplot as well as the hypothesis, linear regression techniques are performed where a simple linear regression model is fit to the cleaned annual energy consumption dataset on these two variables. In addition, considering other factors, such as the average weekly operating hours of operations, may as well be associated with greenhouse gas emission and may or may not influence the effect of natural gas consumption on greenhouse gas emission, two multiple linear regression models are also fit in addition to the simple linear regression model. The predictors of both MLR models are natural gas consumption and weekly operating hours, but only the second MLR model included an interaction term that accounts for the correlation between the two predictors.
The results of both simple and multiple linear regression are in line with the expectation. Both SLR and MLR models indicate that natural gas consumption and greenhouse gas emission are positively and linearly correlated and all three models account for approximately 91% of the total variability in the response variable, namely, *GHG_Emission.* In addition, the MLR models convey several important aspects of the effect of natural gas consumption on greenhouse gas emission over different levels of operation durations. First, operations that consume a certain amount of natural gas and operate 70 hours or 168 hours per week emit more greenhouse gas compared to those operating at the 100-hour with the same level of natural gas consumption. This result may seem counter-intuitive as one would probably expect that the longer an operation operates, the more energy would be consumed and thus more greenhouse gas emitted. However, if other variables such as operation type and floor area are taken into consideration, the results are reasonable. For example, an administrative office may have a different pattern of energy consumption than an outdoor

recreation facility along with different operating hours; at the same level of natural gas consumption, an outdoor recreation facility may consume more of other types of energy than an administrative office. Thus operating hours alone does not serve as a significant indicator of greenhouse gas emission. Second, operation hours also influence the effect of natural gas consumption on greenhouse gas emissions. Specifically, the amount of natural gas consumed has a larger effect on the amount of greenhouse gas emitted for operations that operate 168 hours per week; and a smaller effect for operations operating at the 70-hour level. The difference in the effect of natural gas consumption on greenhouse gas emission is due to the interaction between the two predictors, and thus average weekly operation hours can be viewed as a moderator such that the effect of *Natural_Gas* on *GHG_Emission* depends on it. To conclude, natural gas consumption explains a large proportion (over 90%) of the sample variability in greenhouse gas emissions and is a strong predictor of it. On average, the more natural gas consumed, the more greenhouse gas emitted. However, the size of the effect of the amount of natural gas consumed on the amount of greenhouse gas emitted, although always positive, depends on the level of average weekly operating hours of the operations. Although a significant indicator, natural gas alone does not determine greenhouse gas emissions. A variety of other types of energy that are vastly produced and consumed, either included or not in the *Annual Energy Consumption* dataset, are not included in the simple and multiple linear regression models. Thus, the models created in this analysis provide a rather simplistic explanation of the total variability in greenhouse gas emissions. Large adjusted R^2 values of 0.91 only mean that the models account for most variability in greenhouse gas emission in this specific dataset, but not the total variability. In addition, the operations included in the dataset used for this analysis vary a lot in terms of operation type and total floor area. The type of operations ranges from indoor to outdoor recreational facilities, from administrative office to streetlights and from community centers to fire stations. The total floor area of various operations ranges from less than one square foot to several hundred thousand square feet. Different types of operations may have different patterns and different amounts of energy consumption. Yet, the linear regression models do not account for the variability in the operations. To address these limitations and to conduct a more comprehensive analysis of patterns of greenhouse gas emission, a different approach may be used in further analyses. For instance, instead of specifying certain variables of interest before building regression models, researchers could start with the full model that includes all potential predictors and from there select the most significant ones based on p-values or other criteria, such as AIC and BIC. In addition, **energy intensity**, which is also a variable from the *Annual Energy Consumption* dataset, may be a better predictor than any specific type of energy. Energy intensity is defined as the ratio of the energy consumed to activity units such as the floor area (Canada, 2020), thus reflecting both overall energy consumption and the variability inherent in the observations. Finally, the dataset itself also has limitations and weaknesses. The *Annual Energy Consumption* dataset was last updated in 2018, which is the one used in this analysis. Thus, the data regarding energy consumption are not the most recent ones and may or may not be significantly different from the current energy consumption pattern (i.e. in 2021). Events such as the COVID-19 pandemic and change in political and economic atmosphere since 2018 may lead to a change in the energy consumption pattern. Hence, although analysis of past data provides insight into the association between energy consumption and greenhouse gas emission, the results may not be as representative as if the analysis were performed on the most recent data. In further analysis, a more insightful result may be produced if the most up-to-date is used. Finally, the *Annual Energy Consumption* dataset only includes data of operations within the City of Toronto, thus may not be representative of the energy consumption and GHG emission in Canada or over the World. Thus, the results would be more meaningful if a dataset including, for example, data from Ontario or Canada was analyzed. To conclude, the analysis on the *Annual Energy Consumption* dataset shows that, in the City of Toronto in 2018, natural gas consumption contributes largely to the greenhouse gas emission in the city. Based on results of frequentist multiple linear regression, over 91% of the sample variability in greenhouse gas emission is expalined by the variability in natural gas, yet the effect of natural gas consumption on greenhouse gas emission, although always positive, is affected by the average weekly operating hours of operations. The results provides some insight into the association between natural gas consumption and greenhouse gas emission. However, due to limitations inherent from the dataset, the results of this analysis may not be generalizable to the entire country or the world. More meaningful and insightful researches on the topic can be done using larger and more comprehensive dataset.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: October 12, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: October 12, 2021)

4. Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*

5. Canada, N. R. (2020, October 6). *Energy and Greenhouse Gas Emissions (GHGs).* Statistics Canada. Retrieved October 24, 2021,from https://www.nrcan.gc.ca/science-and-data/data-and-analysis/energy-data-and-analysis/energy-facts/energy-and-greenhouse-gas-emissions-ghgs/20063.

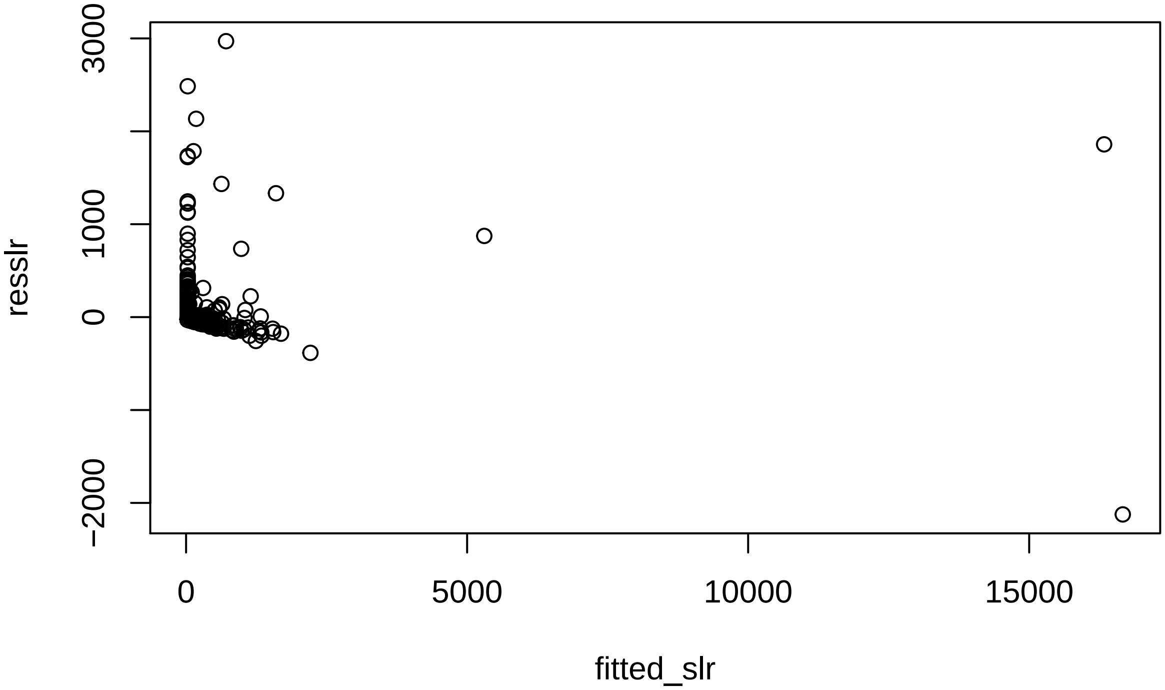
6. Environment & Energy. (n.d.). *Open data dataset.* City of Toronto Open Data Portal. Retrieved October 24, 2021, from https://open.toronto.ca/dataset/annual-energy-consumption/.

7. NASA. (2021, August 24). *Overview: Weather, Global Warming and climate change.* NASA. Retrieved October 24, 2021, from https://climate.nasa.gov/resources/global-warming-vs-climate-change/.

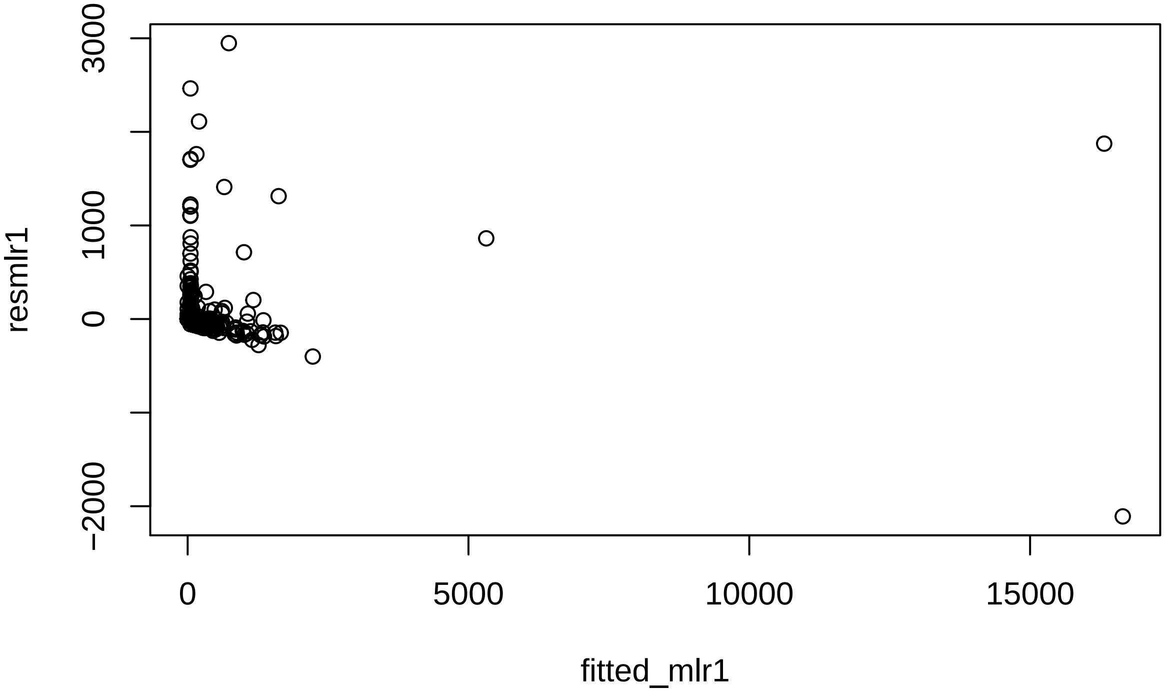
8. Sheather, S. J. (2009).*A modern approach to regression with R.* Springer.

9. Uses for natural gas: *Canada's oil and natural gas producers.* CAPP. (2021, January 8). Retrieved October 24, 2021, from https://www.capp.ca/natural-gas/uses-for-gas/.

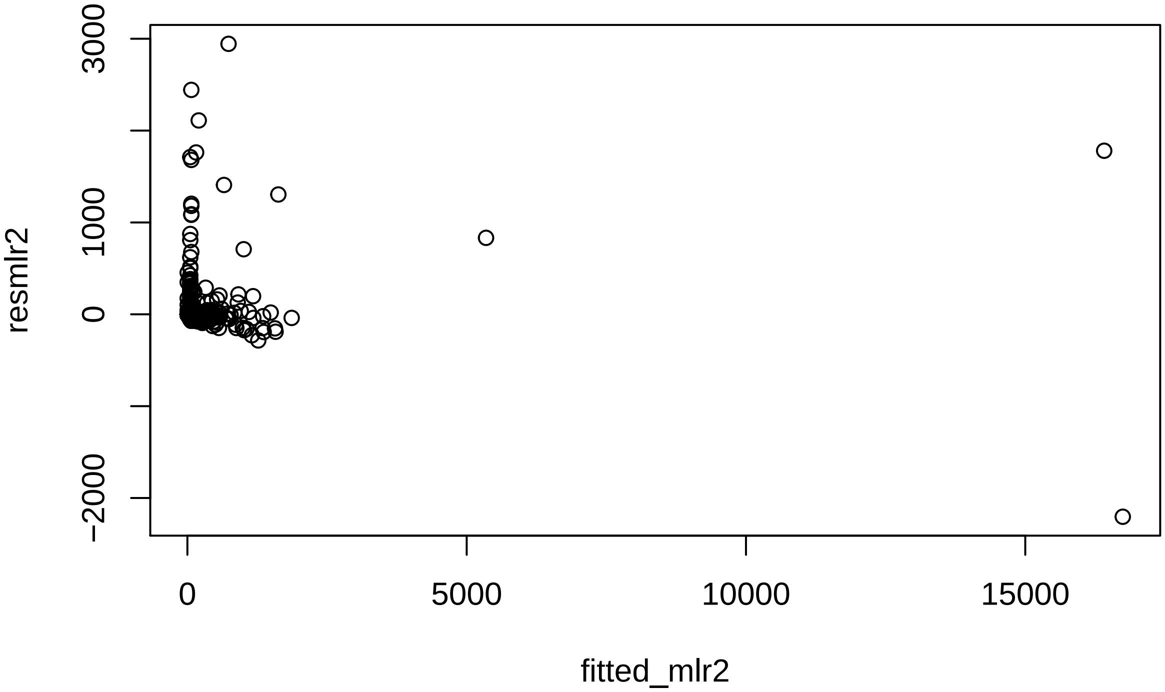
Appendix



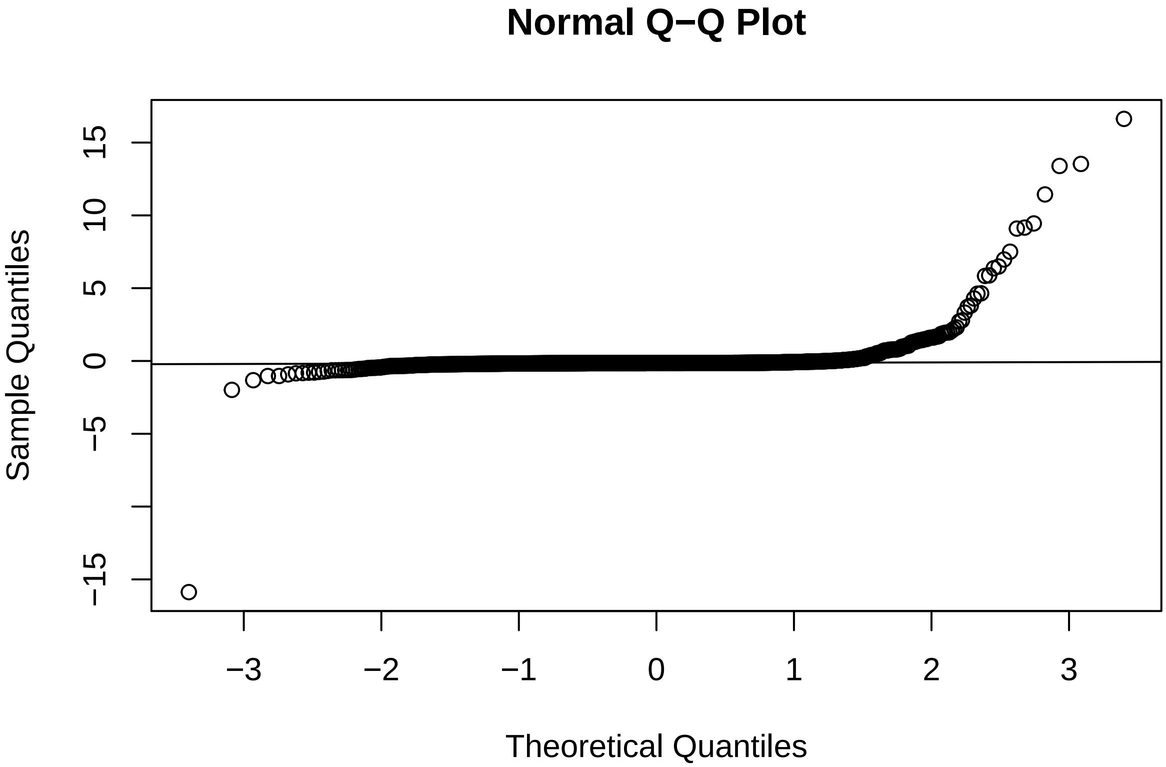
Graph 3: Residual plot of simple linear regression model



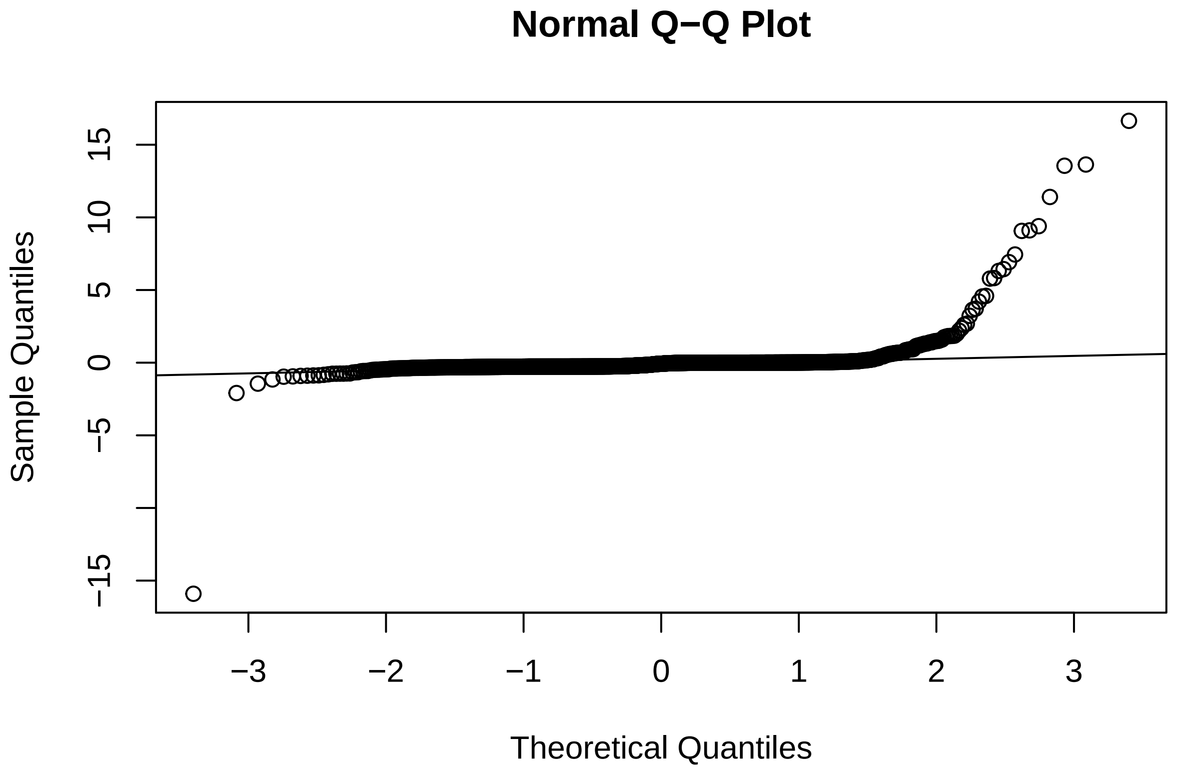
Graph 4: Residual plot of simple linear regression model



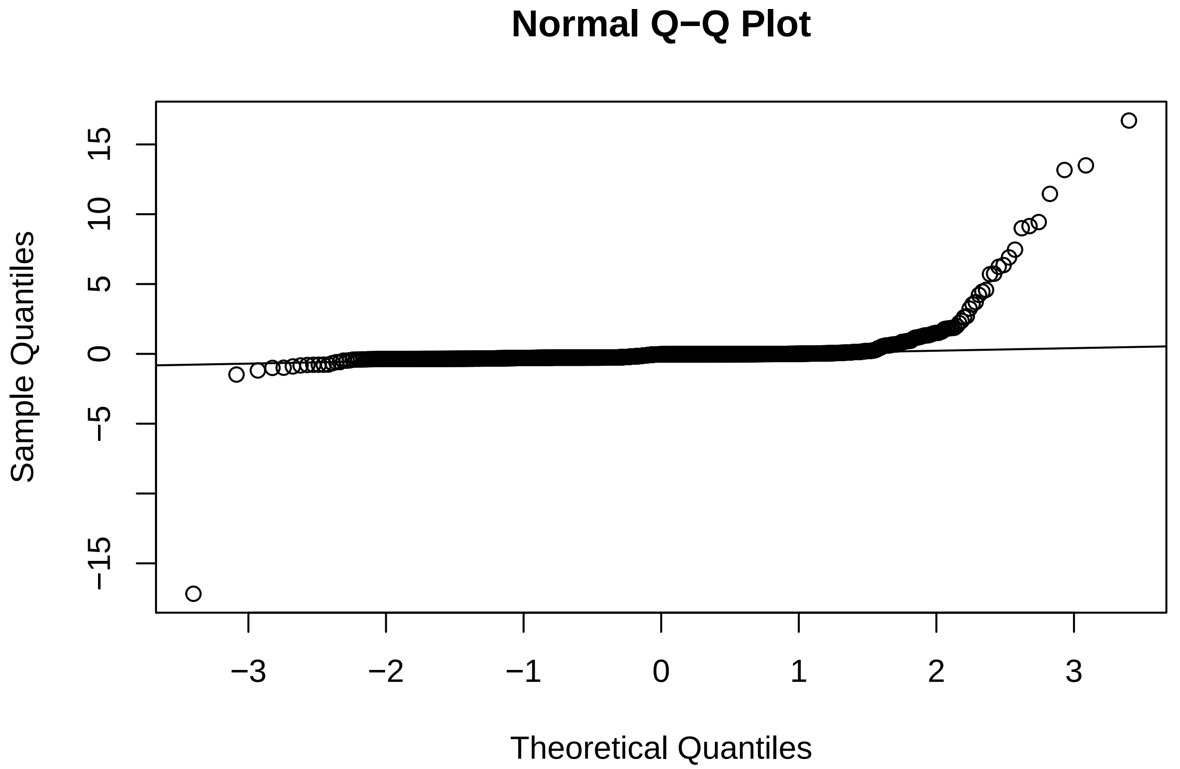
Graph 5: Residual plot of simple linear regression model



Graph 6: Normal Q-Q plot for simple linear regression model



Graph 7: Normal Q-Q plot for multiple linear regression model (without interaction)



Graph 8: Normal Q-Q plot for simple linear regression model (with interaction)