The Impact of Various Economic Sectors on the Price of the S&P/TSX Composite Index[*]

The Analysis of the Canadian Stock Market from 1998-2022

Rayhan Walia

28 April 2022

Abstract

In this report, we focus on the primary and secondary sectors of the economy and analyze their influence on the S&P/TSX Composite Index. This is important to be able to predict the reaction of the market (the composite index) to potential movements in individual sectors. The results of this report could help governments with maneuvering money within sectors to maintain the market. It was found that a combination of the energy, utility, materials, and industrial indices were significant in explaining the response. The model was validated, and all errors in model statistics were found to be less than 10%, indicating a *roughly* validated model.

Contents

[*] Code and data are available at: https://github.com/RayhanWalia/stock_market_influence

1 Introduction

The stock market is arguably the best indicator of the health of an economy and is used as a barometer of what cycle the economy is in (Tan 2020). With 48 countries having stock exchanges (Vodenska et al. 2016), Canada is amongst them. The study of various factors in the economy is essential for federal/central banks, as well as the regulation of monetary policy (Team 2021). This allows the growth of any economy. To further study the economy, it is broken up into four major sectors; primary, secondary, tertiary, and quaternary. The primary and secondary sectors account for raw materials and finished goods respectively, while the tertiary and quaternary sectors represent the service and public sectors (Pettinger et al. 2020).

While all companies are individually important, only a few have a real impact on the market. The S&P/TSX Composite is an equity index that tracks the performance of the largest 230-250 stocks (at any given time) on the Toronto Stock Exchange (Fernando 2021) (largest by market capitalization[1]). It is often used as an indicator of the strength of the Canadian economy (Fernando 2021). All of the sectors we will be analyzing today are publicly traded indices on the Toronto Stock Exchange, that comprise the largest companies in their respective sectors. Each will be detailed in Section 2.

A note; a market index is simply a method of measuring the performance of various securities at once; i.e. singular security representing numerous others (Chen 2022). This is how we are easily able to measure the health of individual sectors, without observing each subsequent security under them.

In this report; we focus on the primary and secondary sectors of the economy, and view their influence on the S&P/TSX Composite Index; from 1998 until 2022. With data from Statistics Canada (2022); this comprises the energy, industrial, materials, and utility sectors, provided with their respective indices (the Standard and Poor's/Toronto Stock Exchange Canadian Energy Index, ' ' ' ' Canadian Industrial Index, ' ' ' ' Canadian Materials Index, ' ' ' ' Canadian Utilities Index). Data was provided for past dates (from 1958), however, all the indices were not present since then, making it hard to quantitatively measure the value of sectors. This data, therefore, begins at the birth of these sectors (all in 1998). This analysis hopes to provide key insights into the workings of major sectors of the Canadian economy and their overall impact on it.

By applying a linear model to our data (with the S&P/TSX Composite as our response variable), we can attempt to analyze the factors influencing the response. Via first checking, then correcting assumptions to the linear model, we can perform statistical tests; such as t-tests (Student 1908) and ANOVA-F tests (Philippas 2014), that allow us to deduce quantitative measures of the data. Note, we use the model to, "help us explore and understand the data that we have" (Alexander 2022), meaning they are not absolute, or even *really* present in the data. We are simply aiming to make sense of what is provided, i.e. further understanding the data set, and a model helps us achieve that (Simsion et al. 2015).

The goal of this report is to then use this model to predict the future value of the TSX Composite Index, knowing trends in specific economic sectors. For example; if we knew the demand for raw materials is high, the Materials Index is likely to increase in price. The effect of this increase on the TSX Composite will be important to know; to prevent potential downfalls in the market. This report will assist the individuals setting monetary policy and those concerning themselves with macroeconomic impacts on the stock market. Note; predicting the stock market is an incredibly difficult task (Shiller 2019), with all 'predictions' being a mix of data and luck. The 'data' is where the prediction becomes difficult since all previous crashes in the stock market have had different reasons for their crashes (Williams 2022). Our goal is to therefore do this to the best of our ability, using the data we have at hand.

[1] market capitalization: total value of all the outstanding shares issued by a company. outstanding shares: shares held by the shareholders

2 Data

This entire report would not be possible without the R programming language (R Core Team 2020), which helps us analyze the data. Tidyverse (Wickham et al. 2019) and dplyr (Wickham et al. 2021) help us clean the data, while ggplot (Wickham 2016), knitr (Xie 2021) and kableExtra (Zhu 2021) help us visualize the data.

2.1 Toronto Stock Exchange (TSX) Statistics

The data has been provided from Statistics Canada (2022), with prices of each 'index' in CAD, from 1956 to 2022. The advantage of the chosen data is that it is publicly available through the Toronto Stock Exchange, for anyone to freely access. Note, 'index' is in quotations due to not all 25 variables being market indices, but all representing some individual part in the market. This is due to the price of the indices only being recorded after 1998 (technically, the recording began in December of 1997, so the beginning of 1998 was taken as a starting point). Before this, the prices were measured very vaguely; with variable names such as, "TSX, gold and silver, closing quotations," which doesn't relate to a formal measurement, but just a relative one to the prior measurements. Due to the informality of the measurement, there were many missing observations, which would have further hindered our analysis. Since these values are not very helpful, only the formal indices were taken as variables-of-interest, and therefore this report begins to analyze only after 1998.

In total, the prices of 12 different market indices are recorded on the first of every month, totaling 277 observations of each index; covering every possible sector of the Canadian economy. As described in Section 1, we are only interested in the primary and secondary sectors of the economy for this report and can reduce our sample to 5 market indices (including the response variable; the TSX Composite Index).

Due to this data being acquired from the stock exchange directly, there exists no bias; which significantly increases the strength of any predictions we make with its analysis. Also, due to strict regulations in the stock market via the CSE; the Canadian Securities Exchange (CSE, n.d.), manipulating/fudging the value of a stock; especially an entire index, is nearly impossible. We can therefore be confident in the data's validity.

2.2 Data Cleaning

The very first step, as with any time-series data, is to convert the date into a readable format; i.e., from a string as 'YYYY-MM-DD' to a decimal (for example; 1998-02-01 is converted to 1998.083).

The original, 'raw' data acquired from Statistics Canada (2022), had all 25 variables (12 indices) under a single column, with various other columns for querying the data (unnecessary columns for our analysis); totaling 11,162 observations of 16 variables (columns). After removing the unwanted 'indices' and the dates (starting from 1998), the data was 'transposed,' with each column now representing a different index, and each row representing the price of the index on a new month.

To better understand summaries, another data set was created, with each column containing *relative* prices; which will be termed 'relative data' from here on out. These are computed as the value divided by the maximum price of the index (across 1998-2022). This data set is created to better visualize relative differences in the data. Due to various prices set for different indices, there is no way to compare them on an absolute scale (for example, the gold index could be at a price of $200 today, and the communications index at $500. This does not imply that communications are more important in our economy). A relative scale would help here, since it would return the price of a sector relative to itself, making it easier for comparisons (for example, an average relative price of 0.5 vs 0.7 states the latter index performs much 'stronger' (higher average returns), with its average price being closer to its maximum). Another advantage of this scale is when visualizing the plots. This can now be done on a singular plot, where we can focus on each index's *pattern*, instead of absolute price. This greatly helps us visualize the influence of these indices on the TSX Composite Index. Note, there are various specific situations where the advantages of the relative price diminish (for example, the index reaches a maximum and stays there, with not much variability. This will have a very high average relative price). We will monitor if these situations arise with visualizations (plots and tables in the following sections).

2.3 Summary Statistics

We first visualize the general summary statistics for the data, which includes the mean, standard deviation, its 25 and 75 percentile, and minimum and maximum values. The data used in this case, just to better understand the relative differences; is the relative data.

Table 1: Summary statistics of market indices

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
SP/TSX Composite Index	290	0.571	0.17	0.261	0.416	0.694	1
SP/TSX Canadian Energy Index	290	0.455	0.207	0.116	0.283	0.617	1
SP/TSX Canadian Industrial Index	290	0.366	0.212	0.139	0.211	0.493	1
SP/TSX Canadian Materials Index	290	0.517	0.203	0.19	0.329	0.669	1
SP/TSX Canadian Utilities Index	290	0.58	0.175	0.221	0.424	0.667	1

With Table 1, we can observe the utility index is closest in mean price to the composite index, which we can roughly interpret as them having similar performance. With the Industrial index having the lowest mean relative price, we can roughly conclude the performance of this index is the worst as compared to the rest[2].

Apart from these initial conclusions, we can also witness the 'spread' of these indices, with their standard deviation and quantile values. We observe the Industrial index with the lowest 75th percentile, while all other indices have similar values. This tells us the Industrial index might have had the largest growth (in percentage terms) of all the sectors in the recent years. We can infer this since a low 75th percentile value indicates at the 75% mark it is the farthest from its maximum, which; provided the index is increasing over time, implies there is more growth to be observed in this sector.

While each sector index's minimum values all vary, we can observe the composite index with a minimum of nearly 30% of its maximum, much higher than other index minimums. This shows us the advantage of an index such as the TSX Composite Index, which is meant to weigh all sectors of the economy such that any one sector cannot single-handedly move the market. This also shows us the advantage of investing in an index such as the composite index over an individual index representing a sector; since you would not have unrealized losses[3] as large in the composite index. We can therefore think of the composite index as a hedge against potential drops in individual economic sectors.

2.4 Time Series Representation

We must note, that all variables are functions of time, and vary throughout the data. To better understand the data at hand, we can visualize it. Once again, we use the relative data; since we only care about the patterns in the indices, and not the values themselves.

[2] Note, the term 'roughly' is used in the previous sentences, since we must remember the mean relative price doesn't reflect anything directly on the performance on the underlying index but provides us simply with an *indication* of how the index is performing

[3] The decrease in the value of an investment that is still ongoing

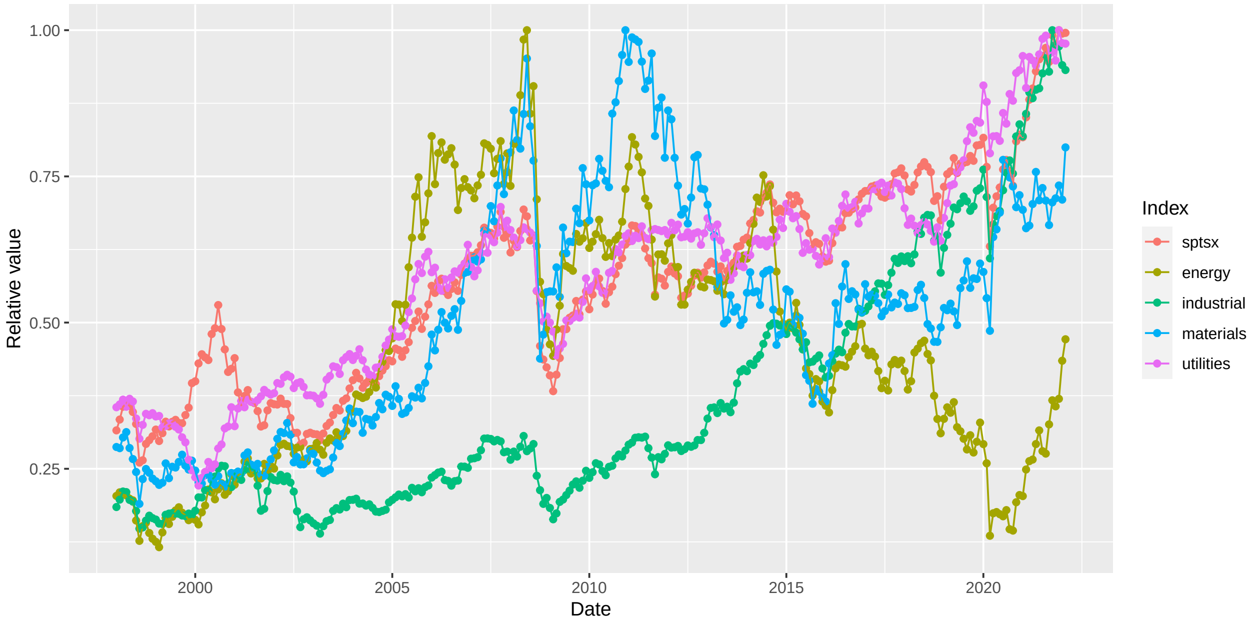

Figure 1: Relative prices of indices-of-interest over 1998-2022. "sptsx" = composite index

Here, we can observe very many qualities of the individual sector indices, and how they compare to the composite index. We can firstly verify our initial conclusion with the summary statistics; that the utility index matches in performance to the composite index. We observe that these indices closely follow each other, with rises and falls in the composite index matching those in the utilities. We also observe why the Industrial index had the lowest 75th percentile since its maximum is much higher than its average values as compared to the other indices; pointing to large growth in the sector in a short timespan (percentage growth in recent years (last 10-12) would be the greatest for the industrial sector, due to explosive growth seen roughly post-2017).

We can also observe various other qualities, such as the drastic increase of indices such as Materials and Energy, followed by the then-drastic decrease of the Energy index. There are various events we can speak of during these twenty-four years; that led to the rise and fall of different indices, and even the composite index itself; such as the 'dotcom' bubble burst in late 2000 (Hayes 2021), the housing market/financial crisis of 2008 (Singh 2022) as well as the recent crash related to the COVID-19 pandemic in 2020 (Frazier 2022). While further analysis into the indices themselves and the factors behind their rise and fall is out of the scope of this report, it is always good to know more about these indices that shape our economy.

3 Model

3.1 Multiple Linear Regression

To understand the relationship between a set of variables (called, 'predictors') and a singular variable (the response), a multiple regression model is used. With the given data, the intuitive prediction would be to use a *linear* model[4].

The results of a multiple-linear model convey the relationship (specifically, it returns the coefficient β) between the response and the predictor variables; which we can then use for further analysis. Primarily; holding the other predictors constant, the model allows to predict (hence the name, 'predictor' variable) a future value for the response, knowing the variation in a singular variable. Also, knowing some variation in multiple variables allows us the same prediction, due to the nature of the *multiple* linear model. What this translates to, for our

[4] due to the data representing price fluctuations over a linear time interval. Also, the assets being measured (the indices) represent large parts of the economy, and would not fluctuate heavily (i.e. linear approximation would work)

data; is to, theoretically, be able to predict the price of the S&P/TSX Composite Index, knowing a potential rise or fall in a certain sector(s) of the economy.

With there being 4 ‘predictor’ indices (energy, industrials, materials and utilities), the general form of the multiple linear regression model would be:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \tag{1}$$

where $\boldsymbol{\beta}$ represents the coefficients and ϵ represents the error (deviation) from the model. $\boldsymbol{\beta}$ contains β_0; representing the intercept and $\beta_{1...4}$; representing the slopes of each predictor ($X_{1...4}$), holding all others constant. Note, that this equation assumes all of the predictor variables are independent. If we wish to add any possible interaction between the terms, we can add a coefficient of a simple 2nd-order term, for example; $\beta_5 X_n X_m$. Note, this doesn’t disturb the *linearity* of the model, since the model must be linear in its coefficients; $\boldsymbol{\beta}$.

An important step that must be considered before the application of the model, is to check that the data satisfies the assumptions of a linear model. These are explicit assumptions, further detailed in Section 3.3; that we must verify before applying the model; since not doing so may render the model insignificant (if the data doesn’t satisfy the assumptions, the results of the model are meaningless). If assumptions are not satisfied, corrections (transformations) can be made to the data to rectify this. In essence, if these assumptions are satisfied, we can be confident that our estimated coefficients (what the model returns) are unbiased and that statistical tests such as t-tests and ANOVA F-tests are valid (otherwise the tests would be meaningless).

Creating a model is one thing while having it accurately describe the data is another. The final step would be to compute statistical tests on the model to verify this accuracy. This is detailed further in Section 3.5, and explains what properties of the model to verify, to be satisfied with the results of the model (the coefficients).

3.2 Validation

Due to the goal to predict a future price, we would need a method to validate the model. For this, before any modeling is done whatsoever; we create a 75-25 split in the data, such that we can apply and refine our model with the larger set of data (the **training** data set), and test if its predictions work on the smaller set (the **test** data set). This is a very powerful tool that allows us to create confidence levels such that we can apply our model to (theoretically) any economy in the world!

To be certain our data sets aren’t too different, we compute error percentages in both the mean and standard deviation of each variable in both data sets. Note, we wish for these values to be under 10%, but also not so low as to be almost identical to the training data.

Table 2: Error % of training data with respect to the test data

	Mean Error %	Std. Dev Error %
sptsx	3.58	0.46
energy	5.09	0.51
industrial	1.31	3.70
materials	4.01	2.89
utilities	3.01	5.87

Here we observe an average error of $\sim 2 - 3$ %, conveying our training data accurately represents our data.

Note, the test data will only become important once an optimal model has been found, such that we can *test* our model.

3.3 Model Assumptions

For a linear model to be applicable to the data, the data must roughly follow some assumptions. They can be mathematically written as:

$$\boldsymbol{Y}|\boldsymbol{X} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \tag{2}$$

Where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are the response and predictor variables respectively, n represents normality in all predictors, with a mean of $\boldsymbol{X}\boldsymbol{\beta}$ and variance of $\sigma^2\boldsymbol{I}$. This can be better understood with 4 explicit assumptions. They are:

1. Population conditional mean responses (conditioned on all other predictors being constant) are given by $\boldsymbol{X}\boldsymbol{\beta}$. Equivalently, the population errors have 0 mean.
2. Population responses (therefore errors) are uncorrelated with each other.
3. Population responses (therefore errors) have constant variance around the conditional mean (assumption 1).
4. Population responses (therefore errors) are normally distributed with a mean from assumption 1.

Since the residuals represent errors (which return similar conclusions to the response), we can formally analyze potential model violations with the residuals. Recall, that the residuals are given by the difference between the prediction (model) and the observation. By monitoring patterns in these residuals, we can observe if our model is incorrect. For this, any pattern in the residuals can be observed; from *clumped* residuals to ones showing fanning, to complete polynomial-like patterns in the residuals. All these imply an incorrect model has been fit. To be satisfied with our model, the residuals must be scattered randomly around the 0-line. Note, if a pattern is observed, we can further assess it using additional conditions (checking for additional assumptions) to be able to identify *what* in our model is incorrect.

We observe 3 major residual plots; versus predictors, versus predicted (fitted) values, and the normal QQ (quantile-quantile) plot. The normal QQ plot computes quantiles from the residuals of the model and compares them to standard normal quantiles. This returns the error distribution in our model (and therefore the distribution of the response). It must be noted that; the tails of these plots (high quantile, low probability observation) usually deviate from these lines, since highly unlikely observations could be scattered.

We can also view the distribution of errors (residuals), which provides us insight into if the model can be applied to the data.

With a combination of these 4 plots, we can formally assess our data for linear model violations. We, therefore, wish to see randomly scattered residuals for the first two plots, a normal QQ plot that resembles a 1:1 distribution (as $y = x$) and a normal error distribution with mean and standard deviation 0 and 1 respectively.

Finally, if we are unsatisfied with the results from the assumptions, assuming the predictor and response are normally distributed; we can apply the Box-Cox transformation (Box and Cox 1964) to the variables. It is a transformation used for 'variance-stabilization,' i.e., to correct normality and/or linearity. It returns optimal powers (hence it is also named a power transformation) to the variables for the linearity assumptions to be best satisfied so that a model used on this transformed data holds meaning.

3.4 Multiple Models

Various models were tested, with differing predictors and their combinations. There were many reasons for trying a new model, from high collinearity between variables to non-significant tests. With various models on our hand, we can compare them using statistical tests to select a singular, *best* model. Recall, that this model is only a way to understand the data, and may not truly be present (Alexander 2022).

8 different models were tested, varying the predictors in each. Table 3 describes all models. The following variables represent the TSX composite, materials, energy, industrial, and utility index respectively; Y, M, E, I, U.

Table 3: Various models used on the training data

Model #	Model
1	$Y_1 = \beta_1 M + \beta_2 I + \beta_3 E + \beta_4 U + \beta_0$
2	$Y_2 = \beta_1 M + \beta_2 E + \beta_0$
3	$Y_3 = \beta_1 MI + \beta_2 EU + \beta_0$
4	$Y_4 = \beta_1 M + \beta_2 I + \beta_3 E + \beta_4 U + \beta_5 MI + \beta_6 EU + \beta_0$
5	$Y_5 = \beta_1 M + \beta_2 I + \beta_3 U + \beta_4 MI + \beta_5 EU + \beta_0$
6	$Y_6 = \beta_1 M + \beta_2 I + \beta_3 U + \beta_4 ME + \beta_5 EU + \beta_0$
7	$Y_7 = \beta_1 M + \beta_2 I + \beta_3 U + \beta_4 M^2 + \beta_5 ME + \beta_6 EU + \beta_0$
8	$Y_8 = \beta_1 M + \beta_2 I + \beta_3 U + \beta_4 ME + \beta_5 EU + \beta_6 MI + \beta_0$

There are valid reasons for testing each model, which will be explained here. While Model 1 represents the general model, Model 2 removes the variables with high collinearity[5] (via measuring the variable's variance inflation factor (VIF)). Model 3 then adds interaction terms only. The variables to interact with were chosen carefully. The Global Industry Classification Standard (MSCI 1999) explicitly defines each sector. The materials and industrial sectors both deal in manufacturing (the materials sector deals in more raw goods), and the energy and utility sectors both deal closely with energy and its transportation (the energy sector deals with a more raw form, i.e., the direct production). Model 4 simply adds all other predictors, while Model 5 removes the energy sector, due to an insignificant t-test. Model 6 then aims to improve on the interaction terms via replacing the material interaction with the industrial sector with that with the energy sector. This is due to them both being involved in the production of raw materials, making them primary sectors of the economy. The industrial and utilities sectors were also made to interact and test the coefficient; however, this lead to an insignificant test. This is assumed to be due to the dependencies of these sectors on the primary sectors. Model 7 aims to *guess* a correction for possible issues seen in the residual plot with the materials sector, via adding a quadratic term (aiming to correct variance). Finally, Model 8 is a union of Model 5 and 6, to test if all may be significant together.

Testing these various models allows us to deduce a rough idea of a *best* model. Note, since this is only from our selection of models, a *better* model can exist. This is why, near the end, we must validate our data, for understanding how good this model would be with another data set. This will be done using the other half of the data set (the test data) to be able to directly test our model.

3.5 Model Tests

Provided the assumptions of the model hold, we can compute t-tests on each predictor (holding the others constant); under the null hypothesis that the predictor does not influence the response. A significant test (low p-value) therefore conveys the predictor is indeed related to the response and should not be ignored when aiming to predict the response. An ANOVA F-test can also be performed on each model, which returns the significance of the variation explained by the predictors in the model. The null hypothesis is that the best fit is given by the intercept-only model (no predictors). A significant result, therefore, conveys the predictors in the model explain the data better than the intercept-only model. Both of these tests are returned by the `summary()` function in R.

These tests should return a few models that have all significant statistics. To then determine the strongest model out of these, we must explore additional tests on the models. We use 3 major tests for these models;

1. Adjusted R-squared
2. (Corrected) Akaike Information Criterion (AICc) (Sugiura 1978)
3. Bayesian Information Criterion (BIC) (Schwarz 1978)

With a combination of these 3 tests, the goal is to be able to determine an optimal model; one that satisfies all of these tests.

[5] indicating the variables may be related to other variables in the data

The adjusted R-squared (adjusted coefficient of determination) is a measure of the amount of the response's variance explained by the predictors in the model. The higher this value, the more variation is being explained in the response. The 'adjusted' conveys that this value considers the number of predictors (since more predictors automatically imply more variation is explained, this should be corrected). We, therefore, accept the model with the greatest adjusted R-squared.

The AICc uses the log-likelihood method of measuring how well the model fits the data, with a 'penalty' (worsening its value) based on the number of predictors. While analyzing this statistic, the smaller the value, the stronger the model. The 'corrected' term comes from this statistic being derived from the AIC, with a stricter penalty to correct for over-fitting of the statistic.

The BIC is similar to the AICc, with an even stricter penalty for the number of predictors. Similarly, smaller values of BIC indicate stronger models.

Finally, to test if our model explains this data, and data similar to the training set (i.e., the test data), we can now validate our model using the optimal one found using the tests. The goal is to observe similar coefficients and significances in the variables when applying the model to both data sets. If the error between the coefficients is minimal, we can successfully validate our model.

4 Results

4.1 Checking Assumptions

We begin by checking the assumptions of a linear model, using the simplest model (Model 1; from Table 3). The following plots are the first two residual plots; including the residuals versus the fitted and versus the predictors.

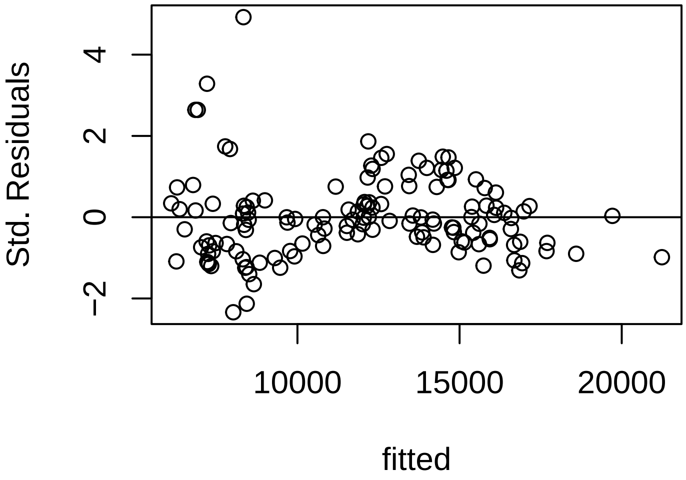

Figure 2: Standardized residual plots of fitted values

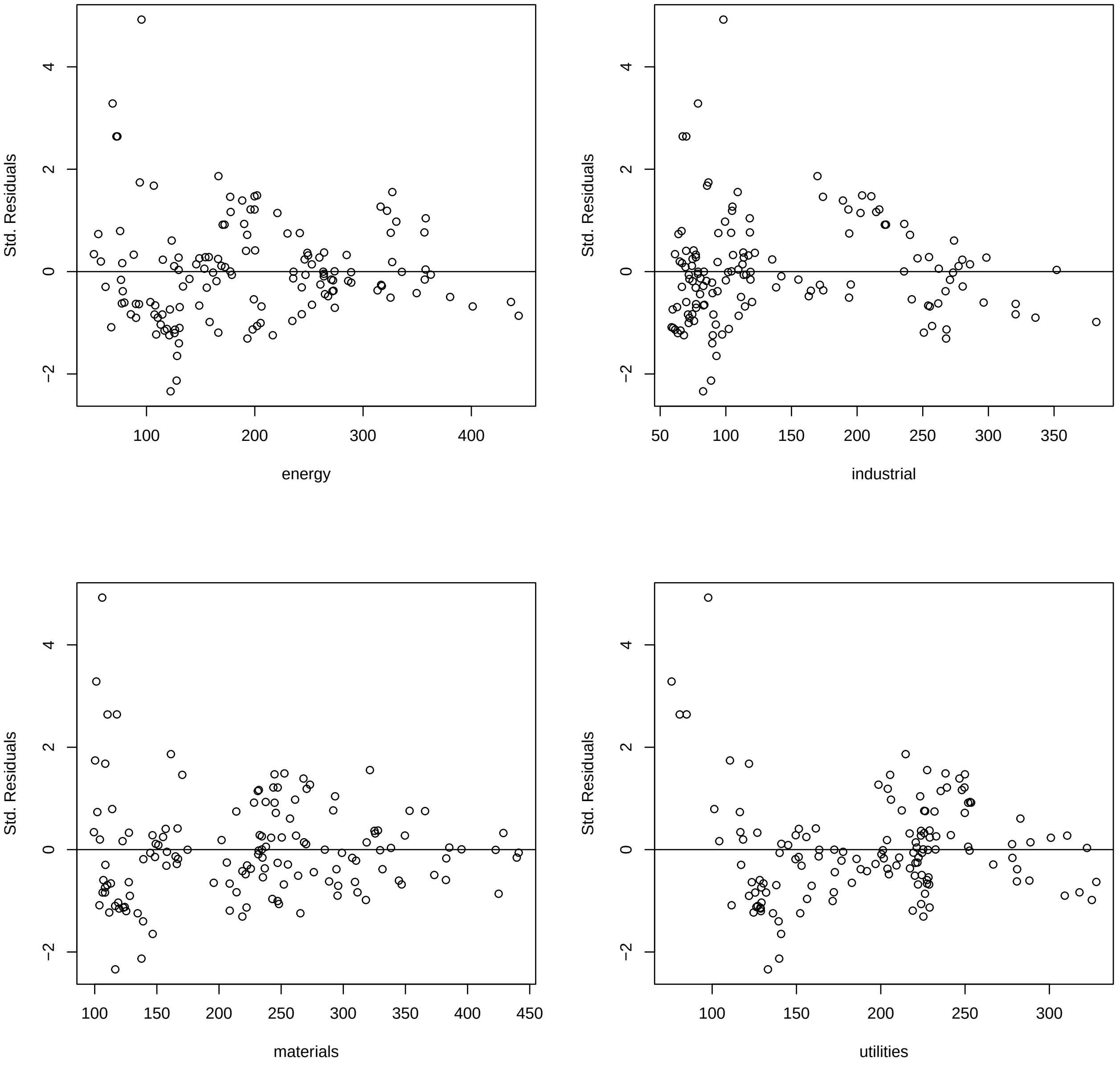

Figure 3: Standardized residual plots of all predictors

In Figure 2, we observe most of the residuals are randomly scattered. There do seem to be influential observations[6]. In Figure 3, we can observe the materials plot showing a *slight* amount of fanning, where the residuals gradually increase/decrease in a triangular pattern. This may imply an issue with the variance/normality, which is why Model 7 was created, via guessing a quadratic form (to potentially fix the normality). As we will later observe in Section 4.2, this produced insignificant results; implying that it was an incorrect guess.

We can now look at the normal QQ plot, shown in Figure 4. With this model, we expect a minimal deviation from the line shown in the figure. We can observe some deviation at the tails, however, there is no significant pattern in the plot, to which we can conclude the model is roughly accurate to be applied to the data; i.e., it satisfies the model assumptions.

[6] outliers, leverage points; points that affect our model and estimated coefficients

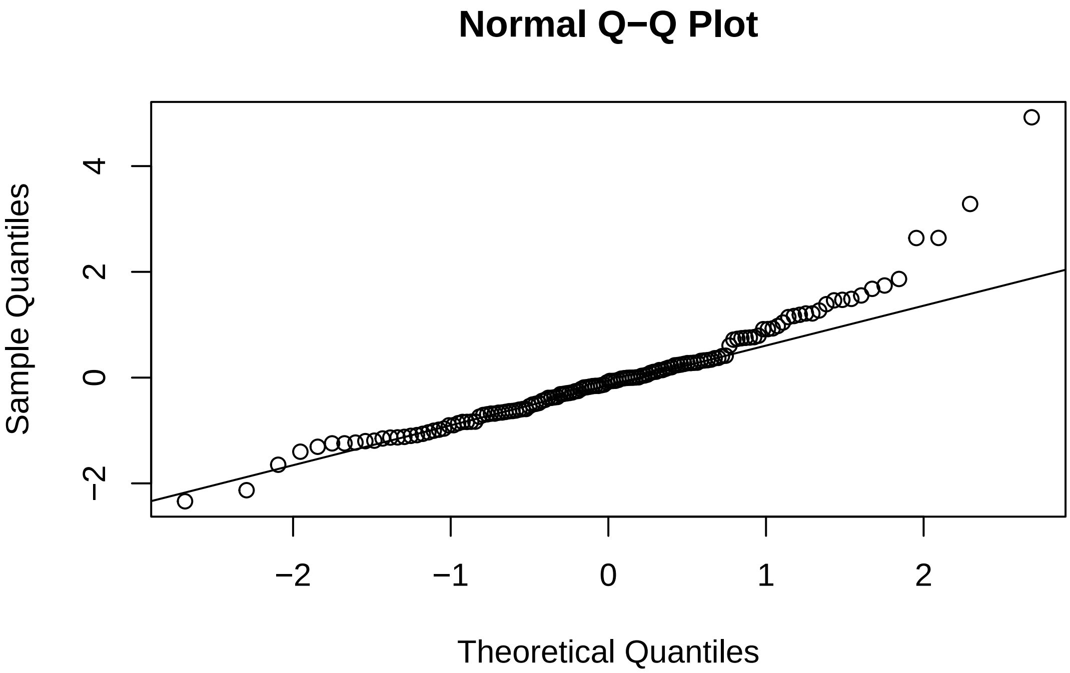

Figure 4: Plot showing normality of errors (and therefore response)

Finally, we can observe the distribution of errors (residuals) to be certain it follows a normal distribution (with mean 0).

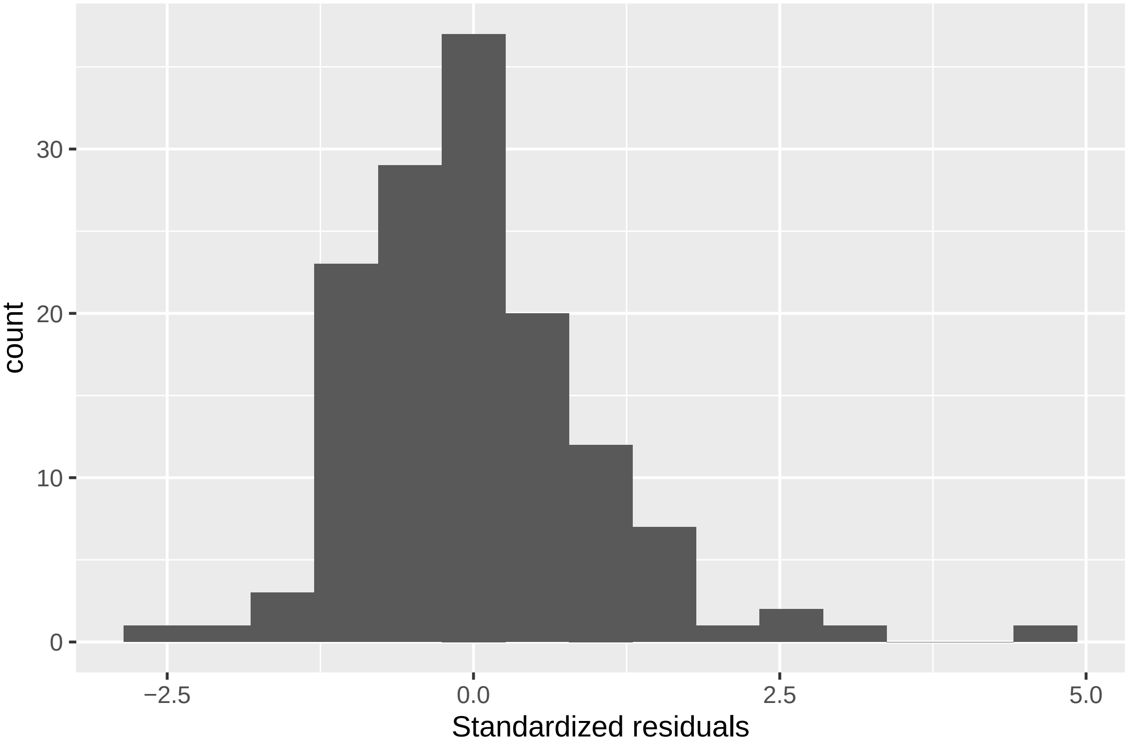

Figure 5: Histogram of errors showing its distribution

Here, we can observe influential observations (outliers; extreme residuals); that with the right set of reasoning, could be removed from the data. Mainly, we observe a mean centered at 0.

The mean and standard deviation of the standardized residuals were both computed, which returned a value of ~ 0 and 1.01 (~ 1) respectively.

Using a combination of these residual plots, we have seen that the data indeed follows the assumptions of a linear model.

4.2 Results of Models

We now attempt to use the tests described in Section 3.4 to attempt to find the *best* model.

It was found that all but Model 5 and 6 returned at least one insignificant coefficient (insignificant t-test; representing the model does not depend on said coefficient/predictor). ANOVA F-tests were performed on each model, where all returned significant statistics; implying all models significantly explained variation in the data.

The 3 major tests were then performed on these models (adjusted r-squared, AICc, aend BIC); the results of which we can see in Table 4.

Table 4: Results of model tests to determine superior model

Model #	# Predictors	Adjusted R-squared	AICc	BIC
Model 5	5	0.9754	2149	2169
Model 6	5	0.9765	2143	2162

From the results of the tests, we can observe Model 6 is the strongest model, with the highest adjusted R-squared and lowest AICc and BIC. We can therefore be confident with this model and aim to use it to validate our prediction.

We can observe the results of the model, given by the following equation:

$$Y_6(\$) = 13M + 39I - 18U - 0.03ME + 0.10EU + 4200 \tag{3}$$

Note, all coefficients have been rounded to two significant figures, to preserve continuity.

Here, we can observe the interaction terms (ME and EU) as not affecting the value as much as the other predictors, yet are highly significant (both p-values are under 0.01). We can therefore not ignore them.

With this model, we can now observe individual plots with the predictor and the response variable, giving us an indication of the difference between the individual linear model coefficients and the multiple linear model (i.e., testing how much the other predictors affect each coefficient). Note, in this plot; we also view the relationship with the date. Even though this is not part of the model, the relationship of price with-respect-to date is an important indication of the future of the stock market (while with other variables, the future represents varying values in each sector, the future here literally represents the future of an observation).

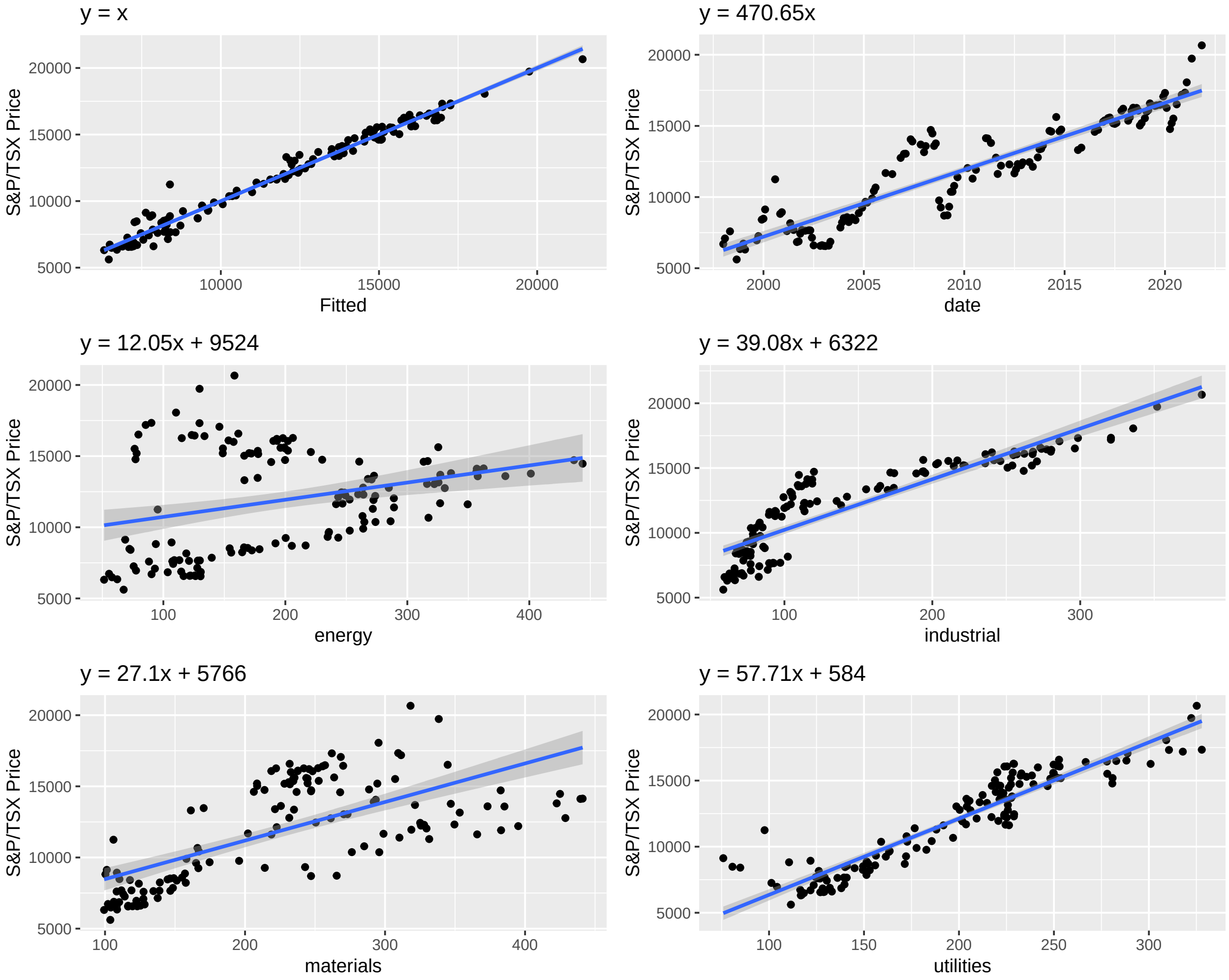

Figure 6: Individual linear model applied to each variable in the data, along with the fitted values (top left); from Model 6. Line of best-fit from individual models overlaid on each predictor, with standard errors. The equation of each line is provided atop the plot

Comparing the individual coefficients from each model (to the multiple linear model), we can view some interesting relationships, i.e., how the presence of other predictors affects the individual relationships. They will be spoken about in detail in Section 5.

The issue with observing the results of the multiple linear model is that it is difficult to view a plot in 5 dimensions (1 response and 4 unique variables). We can however partially represent the model, via plotting individual predictors holding others constant. Using the car package (Fox and Weisberg 2019), we can do this. These plots, termed ‘added variable plots,’ show us the results of our model in a visualizable way. This is shown in Figure 7.

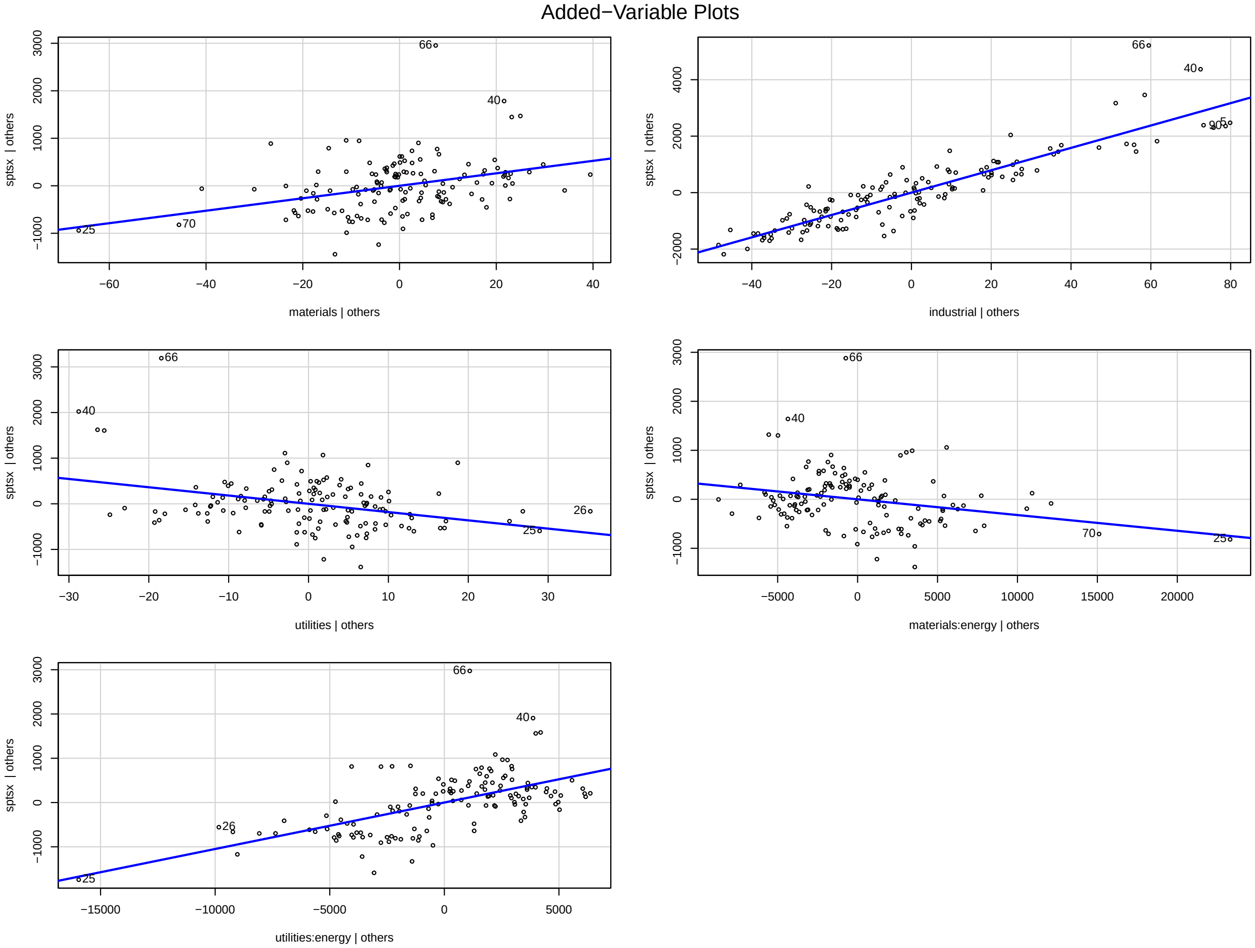

Figure 7: Plots showcasing individual predictor relationships conditioned on all others (being constant); using Model 6. Labelled points represent influential observations

Here, we can now observe the results of the multiple linear model (as opposed to the individual relationships), where we can observe the effect of conditioning on the predictors. For example, we can observe a negative relationship with the utility index with the multiple linear model, but a positive relationship when plotted individually. This will be spoken about further in Section 5.

4.3 Model Validation

Finally, we can apply Model 6 to our test data, to observe the coefficients and their significance. Interestingly, it was found that the interaction between materials and energy was not significant in the test data (it was quite insignificant; with a p-value of 0.33). The coefficients returned are shown in the following equation

$$Y_{\text{test}} = 7.0M + 43I - 19U - 0.01ME + 0.09EU + 5081 \tag{4}$$

With this, we can compare the coefficients to those returned from the training data set; and compute the error. Note, due to the insignificance of materials:energy, we expect the error in its coefficient to be high. Also, note, we cannot remove the interaction term from the test model due to its insignificance. Recall, we have already concluded Model 6 was the strongest from the training data; and we therefore must apply only this model to the test data. Therefore, the model should not be altered due to observations from the test data.

Table 5: Error % of training and test coefficients from Model 6

Predictor	Error in Coefficient (%)
Materials	89
Industrial	9
Utilities	7
Materials:Energy	131
Energy:Utilities	14
Intercept	17

Table 6: Error % of training and test statistics from Model 6

Statistic	Error in Statistic (%)
Adjusted R-squared,	0.28
AICc	9.84
BIC	9.8

In Table 5, we observe a large error in the materials index. This likely has something to do with the insignificant interaction term between materials and energy. A large error can also be (therefore) observed in the interaction term; between materials and energy. The rest of the errors are under 20%, roughly representing validated coefficients.

Then, in Table 6, we also observe an under 10% error in all statistics related to computing the 'best' model, namely; the adjusted R-squared, the AICc and BIC. This gives us hope in validating our model.

5 Discussion

5.1 Summary

Via utilizing the prices of primary and secondary economic sector indices on the Toronto Stock Exchange, we were able to analyze their effect on the S&P/TSX Composite Index; which is widely regarded as an indicator of the strength of the Canadian economy (Tan 2020). With this analysis, the goal is to be able to predict the value of the composite index, knowing a potential rise or fall in certain sectors (sector indices) of the economy.

Satisfied with the data following the assumptions of a linear model, we were able to apply and test various models, to which we concluded Model 6 (Equation 3) best described our data. Now, knowing a combination of values in the materials, industrials, energy, and utility index; one can predict values for the TSX Composite Index. The plots for each predictor (Figure 7) show us these predictions when a single predictor is varied.

Finally, to test the results of the model; it was validated via applying the same model to the test data. Here, it was (interestingly) found that the materials and energy interaction term was insignificant, with a p-value of 0.33. This indicates our model derived from the training data does not fully encompass the 'population' (the entire sample). Since this has been applied to the test data, we can only comment on this relationship and not alter any part of the model. This is due to the nature of the test data. If we re-do the analysis of our model due to observations from our test data, we will be directly adding bias to our model, as we will no longer be able to state that our model is valid in an external data set since we've used it to build/change our model. We can therefore only discuss the potential impacts of the insignificant coefficient on our data. Due to small errors observed in all but the materials index (Table 5), as well as small errors in the model statistics (Table 6); we can conclude that (apart from the insignificance of the materials index), the model has been *roughly* validated.

5.2 Individual Models

To test the effect of individual predictors on the response, single-predictor models were created with each variable in the data; shown in Figure 6. Here, we can comment on the individual relationships between the predictors and the model. Note, all of these individual models had adjusted R-squared values less than 0.5 (compared to 0.97 for Model 6), indicating they do not explain all the variation in the data. These models are therefore not accurate with the given data; however, their relationships can be viewed and commented on. Due to all coefficients in the models being highly significant (the intercept in the 'date' model was insignificant (assumed to be due to strong dependencies on the other predictors), and hence not included in the individual model plots (Figure 6)), they are displayed atop the plots. Here, we can view some *interesting* relationships.

We can first view the relationship between the response and Model 6's fitted values. It can be observed that the data strongly follows the prediction. We compute a chi-squared test[7], from the Janitor package (Firke 2021), which returns a p-value of 0.24. Since this is greater than the 'significance' cutoff (0.05), i.e., it is an insignificant test, we can deem the model correctly predicts the data. We can now look at the individual models.

Firstly, the relationship with date; although not impacting our model, gives us an indication of the 'time-series' performance of the composite index. We observe a coefficient of 470.65, indicating; that on average, the S&P/TSX Composite Index increases by $470.65 every month. This data is taken over 24 years, representing 36% of the total period since the index's inception (incepted in 1956 (data begins); the total period is 66 years); a large percentage of the existence of the composite index. This can therefore give us a rough estimation of the market's performance in the future. This predictor also has the largest coefficient when modeled with the response; (roughly) indicating that time influences the market the most!

Due to the rapid increase and decrease of the energy index during the observation period, we can observe large non-injectivity[8] (large differences between the same x-value), where the standard errors of the prediction are the largest within all other predictors.

The Utility Index *looks* as it fits the response the best, as was observed with their similar time-series distributions in Figure 1. We can observe this due to most of the data residing *close* to the line-of-best-fit. Also, the standard errors of this model are the lowest compared to other predictors.

5.3 Multiple Linear Model

Being confident with Model 6, we can use its results to interpret the effect of these predictors; in the presence of all other predictors. This lets us analyze their relationship with the response in unison, instead of individually (performed with the individual linear models). We can view the results of this model in Figure 7.

Here, we observe all relationships but utilities and the interaction between materials and the energy index produce positive slopes (coefficients). We can observe in all plots that the points are roughly randomly scattered around the lines of best fit, indicating these models used are accurate. Again, we can also see this from a near-perfect 1:1 relationship between the fitted values and the observations; as seen in the (top left) plot in Figure 6.

An interesting comparison that can be made between the multiple linear model and the single-variable linear model is the utility index. We can observe (in Figure 6) that the utility index has a positive relationship with the composite index. In fact, when observing the relative patterns in the indices (Figure 1); the utility index was nearly identical to the composite index; which was also supported via the numerical summaries (Table 1) being similar. However, with the multiple-variable model used (Figure 7), we observe a negative relationship with utilities. This indicates the presence of the other predictors affects the relationship between the composite index and the utility index the most (since all other coefficients remained positive). This can also indicate that another predictor which is correlated to the utility index could be affecting its relationship with the composite index. The interaction of the index (with the Energy index) is likely what is causing this

[7] a statistical hypothesis test, with the null hypothesis, that the predictors have the same distribution. A significant test conveys the model does not describe the data accurately

[8] injective function: every x value has a unique y value. Note, due to this data being real-world observations, perfect injectivity is nearly impossible (unless it's time-series data)

issue. This is because when isolating for the utility index, we also must use the energy index (since they interact together); which then also includes the material index. A dependence somewhere in this chain of dependencies is likely what is causing the difference in the coefficients. Note, a different coefficient is not a *bad* thing, however, can be commented on.

While all coefficients were statistically significant, they do not equally contribute to the price of the composite index (interaction terms contribute the least). The largest contribution (via the absolute value of the coefficients) was observed from the industrial index, representing a $39 increase in the price of the composite index with every $1 increase in the industrial index, holding all other indices constant. The smallest, on the other hand, is observed in the interaction between the materials and energy indices.

With a combination of this model and the crucial information we obtained from the model validation (the insignificance of the interaction between the materials and energy indices); we can now use this model to *roughly* predict the price of the composite index; knowing variation in the predictor variables.

5.4 Weaknesses

Due to the data being directly sourced from the Toronto Stock Exchange, there are not many *changes* we can make to the data in the future; since it is dictated via real-time events, that can be *somewhat* random (most events can be traced back to a *starting-point*, however since most individuals aren't keeping track of every possible starting-point, the events that occur in the economy that shift economical sectors can be assumed to be random).

Therefore, only possible improvements can be made to the model itself. However, as we checked the 4 basic assumptions; none were found to violate linearity. The *slight* pattern in the standard residuals of the materials index (Figure 3) could potentially explain the issues we observe with the validation of the model. A power transformation (Box-Cox transformation) could be attempted; however, it should be performed on all predictors and the response, to not include bias in the transformation.

5.5 Next Steps

The future of reports such as this one is to analyze the economic sectors that affect different markets in the world; and compare them to each other. For example, if the USA has a greater dependence on the Energy Index on its stock exchange as compared to Canada, is it because they export more energy, have more factories, have more available resources for energies such as oil, etc? These are economic factors that influence the global economy, and further study into global markets may yield valuable information.

This particular report is aimed to assist those interested in monitoring macroeconomic factors that influence the Canadian economy; particularly those individuals with the power (such as policymakers) to prevent turmoil in the economy to do so aptly.

Appendix

.1 Datasheet: Describing the Data

Extract of the questions from Gebru et al. (2018)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to record changes in the prices of various indices on the Toronto Stock Exchange (TSX). We found a publicly available dataset from Statistics Canada (2022) that contained this information. This dataset specifically satisfies the requirements for the International Monetary Fund's Special Data Dissemination Standards (SDDS) Plus initiative, and used for various purposes; specifically economic, financial and wealth accounts. The data contains this variety of information in a structured format.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Statistics Canada obtained the data from the TSX; publicly available data, however, does not apply directly to Statistics Canada (available and used by numerous internal and external users)
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - No funding was required, since the data set is simply a compilation of publicly available data.
4. *Any other comments?*
   - Since all this data is based from a third party source (the TSX), there is not much that the data itself can be incorrect about. While we can correctly apply and use it to analyze the Canadian economy, models such as these may not work in other economies. This should be noted.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - All the variables in this dataset represent price; of various indices listed on the TSX. These prices represent the strength/condition of the sector the respective index represents. Relative to its past and future prices, we are able to judge the strength of each index.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There is simply one instance; representing the strength of a sector (via the price of the index).
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - This dataset contains all possible instances, from the birth of the TSX to date; when aiming to predict the Canadian economy. However, if we aim to predict other economies, due to various other indices and measuring techniques, we are not certain if it would be perfectly representative. Any other economy could have other primary and secondary economic sectors affect the price of their composite index. For example, in the model found using the Canadian economy, the Energy sector did not seem to directly correlate to the price; however in another country where their primary export (for example) is oil, their energy sector would be *predicted* to have a stronger relationship to the price of the composite index.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - All the instances convey the price of a certain security; be it a commodity index or an economic sector index. We are simply interested in the sector indices.
5. *Is any information missing from individual instances? If so, please provide a description, explaining why*

this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

- There are missing observations in the P.E. Ratio (price to earnings ratio); which is presumed to be due to this value also being directly supplied by the TSX; and earnings are not always recorded. Also, the CPI (Consumer Price Index) misses one observation; where the reason is unknown. Due to both of these variables not influencing our model or analysis, these observations do not affect our results.

6. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - No, since each sector is presumed to be independent. However, due to a few commonalities, there are securities that exist in multiple indices, making the variables dependent. We can omit this with interaction terms in the linear model.
7. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - A data split is performed to seperately train on one set of data (the training data) and then test the model on another (the test data). This is done after the data cleaning and exploratory data analysis (numerical summaries, etc.)
8. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are no errors in the data, since all have been obtained from the TSX, which is a strictly controlled exchange that doesn't allow manipulation. Also, because the data has been directly sourced from the TSX, there is no worry about error.
9. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
   - This data will always be available and free-of-charge, due to it being sourced from a publicly available database. Due to the availability, it is constantly updated monthly, to allow any new data to be available for analysis. -Every observation is important, and tells us something about the state of a part of the economy at one point in time. The data will therefore always exist, and continuously represent something that we can learn from.
10. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No. All data is available publicly.
11. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
12. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - No. One could think of individual sector indices as 'sub-populations' of the composite index, each which contain a specific part of the economy. Each have securities pertaining to its respective sector, which convey the strength of any individual sector in the economy.
13. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for*

example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

- The data was directly collected from the Toronto Stock Exchange, a freely publicly available database that has been structured in a reasonable and important format. These numbers represent large quantities of information, and are hard to be tampered with.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - This dataset simply collected the data from a 3rd party source, the TSX. Due to this data being available publicly, there was no complex mechanism to acquire the data.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - This data represents population level data for the Canadian economy.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - All publically traded companies listed on the TSX were involved in this data, since each of them contribute in some way to the price; which is then what we analyze.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - The data is recorded over 66 years; from 1956 to 2022, and we will be simply analyzing 24 years of data; from 1998-2022.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - None were needed, since the information was directly from the TSX.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - All the information was obtained from a third party source; the TSX.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - Not every single company responded due to not every single one listed on the TSX being a part of an index (i.e., there are companies not listed on any index, yet listed on the stock exchange, for example; penny stocks[9]). All other companies do not have an option to be not listed, since they are publicly traded.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - Since the data is available publicly, no consent was required.
10. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - The impact of the stock market and indices have been studied heavily; where everyone involved in the market wishes to predict its direction.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

[9] small companies in regard to ones listed on the exchange (less than $500M market capitalization)

- Due to an inefficient structure in the dataset, the data had to be restructured, which was the major processing step in the pre-model stage of the analysis.
- Variables that were unimportant to the analysis were removed (only primary and secondary economic sector indices were retained).

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - The raw data was stored directly from Statistics Canada, and is available in 'inputs/data/raw_stock.csv.'
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - The software used was `R` (R Core Team 2020) and `tidyverse` (Wickham et al. 2019).

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - This is not known; but it wouldn't be a surprise, due to the availability of the data. Also, the concise form of including only the sector indices and important securities must be beneficial to many individuals interested in macroeconomic indicators.
2. *What (other) tasks could the dataset be used for?*
   - Due to all commodity indices and sector indices being available in this dataset; from 1956, any other analysis can be performed. For example, an alysis of the Gold index would be one that may be of interest to many individuals looking to get into the 'gold-business.'
3. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - Due to the data being sourced directly from the TSX, there is no bias present. All indices that were listed on the exchange are included, and therefore there can exist no 'unfair treatment.'
4. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - This dataset cannot be used to aim to predict another economy. For example, one may assume due to the similar nature of stock exchanges, we can easily apply our data to the US Stock Exchange; however this is not true. Factors affecting only the country would affect the individual sector indices, and therefore this data should not be aimed to predict the world's economy.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - The data is distributed to any individual who seeks it.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The data set is available from Statistics Canada; which can be referenced in R. The doi of this dataset is: 'https://doi.org/10.25318/1010012501-eng.'
3. *When will the dataset be distributed?*
   - The dataset is updated monthly, and 'distributed' to the Statics Canada website.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - Dud to the 'open-source' nature of the data, there is no copyright to it.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If*

so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
   - No
7. *Any other comments?*
   - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The TSX
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - The email of Statistics Canada (infostats@statcan.gc.ca) can be used for contact; however, once again, due to the data being sourced from the TSX; this individual would have no say in the data.
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - This dataset is updated monthly; however past observations are not updated, only present ones.
5. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - Each older dataset is still valid if that is the time period one is interested in. Due to the price of a security on one day is unaffected by the price much later, all past data is relevant; if that is the interest.
6. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - No, since the data is directly sourced from the TSX.

References

Alexander, Rohan. 2022. "Telling Stories with Data." *Chapter 14 It's Just A Linear Model*. https://www.tellingstorieswithdata.com/ijalm.html.

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)* 26 (2): 211–52. http://www.jstor.org/stable/2984418.

Chen, James. 2022. "Index." *Investopedia*. Investopedia. https://www.investopedia.com/terms/i/index.asp.

CSE. n.d. "Trading Rules and Regulations." *CSE*. https://www.thecse.com/en/trading/trading-rules-and-links/trading-rules-and-regulations.

Fernando, Jason. 2021. "What Is the s&p/TSX Composite Index?" *Investopedia*. Investopedia. https://www.investopedia.com/terms/s/sp-tsx-composite-index.asp.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Frazier, Liz. 2022. "The Coronavirus Crash of 2020, and the Investing Lesson It Taught Us." *Forbes*. Forbes Magazine. https://www.forbes.com/sites/lizfrazierpeck/2021/02/11/the-coronavirus-crash-of-2020-and-the-investing-lesson-it-taught-us/?sh=5f22cfa846cf.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. 2018. "Datasheets for Datasets." arXiv. https://doi.org/10.48550/ARXIV.1803.09010.

Hayes, Adam. 2021. "Dotcom Bubble." *Investopedia*. Investopedia. https://www.investopedia.com/terms/d/dotcom-bubble.asp.

MSCI. 1999. "GICS - Global Industry Classification Standard." *MSCI*. https://www.msci.com/our-solutions/indexes/gics.

Pettinger, Tejvan, Asit Purohit, Suba, Vijay, and Nawrin Rahman Anty. 2020. "Sectors of the Economy." *Economics Help*. https://www.economicshelp.org/blog/12436/concepts/sectors-economy/.

Philippas, Dionisis. 2014. "Analysis of Covariance (ANCOVA)." In *Encyclopedia of Quality of Life and Well-Being Research*, edited by Alex C. Michalos, 157–61. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_82.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.

Shiller, Robert J. 2019. "Why Is It so Hard to Predict Stock Market Trends?" *World Economic Forum*. https://www.weforum.org/agenda/2019/04/why-is-it-so-hard-to-predict-stock-market-trends/.

Simsion, Author Graeme, Author John Giles, Author Robert S. Seiner, Author Sakthi Rangarajan, Author Ted Hills, and Author Charles Betz. 2015. "Data Modeling: Understanding and Being Understood." *TDAN.com*. https://tdan.com/data-modeling-understanding-and-being-understood/4589.

Singh, Manoj. 2022. "The 2007-2008 Financial Crisis in Review." *Investopedia*. Investopedia. https://www.investopedia.com/articles/economics/09/financial-crisis-review.asp.

Statistics Canada. 2022. "Table 10-10-0125-01 Toronto Stock Exchange Statistic [Data Table]." *Investopedia*. Investopedia. https://doi.org/10.25318/1010012501-eng.

Student. 1908. "The Probable Error of a Mean." *Biometrika*, 1–25.

Sugiura, Nariaki. 1978. "Further Analysts of the Data by Akaike' s Information Criterion and the Finite Corrections." *Communications in Statistics - Theory and Methods* 7 (1): 13–26. https://doi.org/10.1080/03610927808827599.

Tan, Han. 2020. "The Importance of Stock Markets." *Market News &Amp; Forecasts, Charts, Broker Reviews*. https://www.fxempire.com/forecasts/article/the-importance-of-stock-markets-663262.

Team, The Investopedia. 2021. "What Impact Does Economics Have on Government Policy?" Edited by Michael JEditor Boyle. *Investopedia*. Investopedia. https://www.investopedia.com/ask/answers/031615/what-impact-does-economics-have-government-policy.asp.

Vodenska, Irena, Hideaki Aoyama, Yoshi Fujiwara, Hiroshi Iyetomi, and Yuta Arai. 2016. "Interdependencies and Causalities in Coupled Financial Networks." *PloS One* 11 (March): e0150994. https://doi.org/10.1371/journal.pone.0150994.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Williams, Ward. 2022. "Timeline of US Stock Market Crashes." *Investopedia*. Investopedia. https://www.investopedia.com/timeline-of-stock-market-crashes-5217820.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.