IB Mathematics HL

Internal Assessment

How does Air Quality Affect the Stock Market?

Exam Session: May 2021

Candidate number: 000277-JJR446

1.0 Introduction

Modernization improves many aspects of human lives, yet it leads to various negative changes simultaneously. These changes include: over-consumption of energy and natural resources, air pollution, and climate change. In particular, it is known that air pollution leads to poor air quality which affects our health (Bruenkreef et. al, 2002). To a greater extent, there is actually a correlation between health and decision making (Grossman, 1972). An extensive body of literature on behavioral finance has shown a connection between weather and stock returns. The rationale is that air quality lowers the risk aversion of stock traders, which in turn affects the performance of the stock market. For the US market, Heyes et al. (2016) and Levy & Yagil (2011) saw a negative correlation between air pollution and stock returns. In China, Li & Peng (2016) saw a positive correlation. Wu et. al (2018) used firm-level data and confirmed the negative impacts of air pollution on the general market performance.

Building on the previous research by Heyes et al. (2016) and my mathematical knowledge gained from the IB programme, I aim to answer the following question: **How does air quality affect the stock market?**

1.1 Objectives

I aim to answer my research question through the following objectives:

1. Compute descriptive statistics to observe individual variables.
2. Attempt to reduce omitted variable bias to isolate the effect of the explanatory variable.
3. Construct a statistical model that describes the patterns and characteristics of stock market prices explained by air pollution.

1.2 Definitions

PM 2.5

PM 2.5, or fine particulate matter, are particles with diameters of around 2.5 microns (EPA, 2018). Since these particles can be inhaled, they are able to penetrate into the lungs and even the bloodstream, causing serious health problems. PM 2.5 is measured in micrograms per cubic meter ($\mu g/m^3$). According to the EPA, 0-40 is considered safe, while 65 and above is considered unhealthy. PM 2.5 levels were used in this investigation as the indicator of air quality.

S&P 500

There are a few ways of measuring the performance of the New York stock exchange. Of which, the Standard & Poor's 500 (S&P 500) index is a weighted index of the 500 largest companies in the US. The S&P 500 is widely used because of its breadth and depth (Kenton, 2020).

Time Series

A time series is simply a set of data points ordered in time (Peixeiro, 2019). Data for PM 2.5 and S&P 500 fall into this category since they are collected over time. However, time series data often comes with some internal structure (such as autocorrelation, stationarity, seasonality, which will be defined later) that must be accounted for.

1.3 Methodology

We try to build a model to explain the relationship between PM 2.5 levels in New York and the performance of the S&P 500 index (Heyes et. al, 2016). However, single variable regression with time series data is unreliable because of bias; specifically, omitted variable bias and stationarity. Omitted variable bias may lead us into mistakenly thinking that the effect we observe is more significant than it actually is, and non-stationarity causes random effects that will obscure the true relationship between the dependent and independent variables. We address these issues through statistical tests, data manipulation, and constructing a multivariate linear model estimated by the Ordinary Least Squares method (OLS).

Finally, we can conduct statistical tests to determine the significance of the target relationship.

Section 2 will present the descriptive statistics, section 3 will focus on reducing the effects of omitted variable bias and the construction of the model, and section 4 will be on regression analysis.

2.0 Descriptive Statistics

In this section, we first graph the two sets of data to get a general picture of what they look like. We obtained historical daily PM 2.5 data in New York city from The Air Quality Project database and S&P 500 index data in csv format from January 1st, 2014, to October 7th, 2020. The data is then cleaned up using Python, by only keeping the days where data is available for both PM 2.5 and S&P 500.

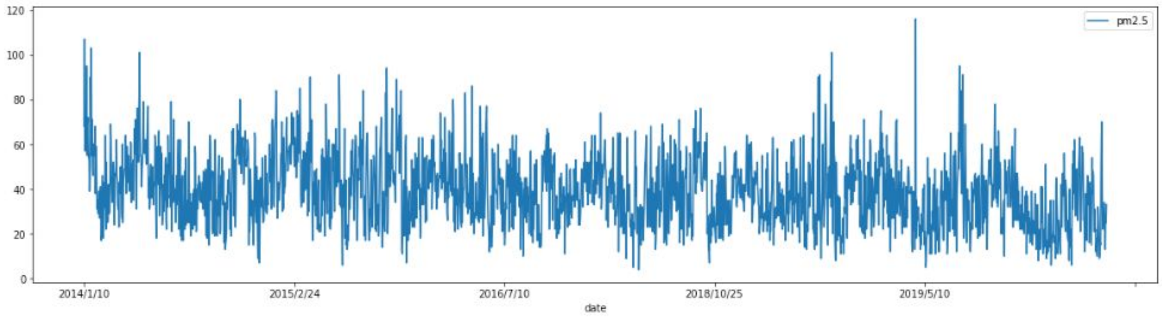


Figure 1: PM 2.5 levels in New York

We observe a great number of fluctuations in PM 2.5 levels over time. A general periodic fluctuation can be observed, indicating possible seasonal trends, which will be later controlled in the model.

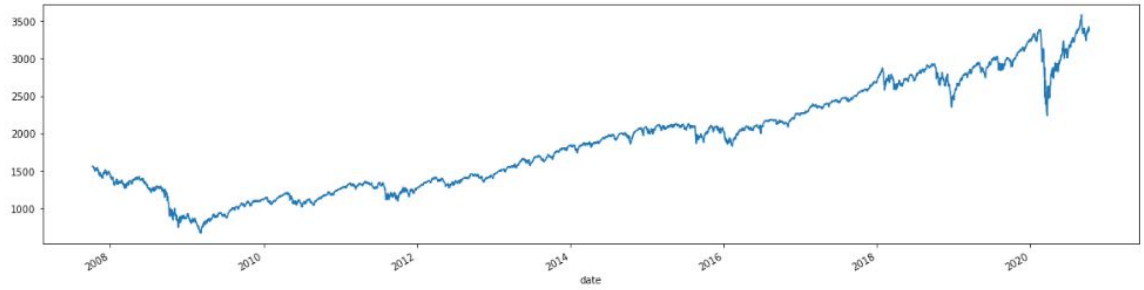


Figure 2: S&P 500 index

The S&P 500 index has a clear upward overall trend. Statisticians also commonly look at the returns of the stock, which is shown in Figure 3.

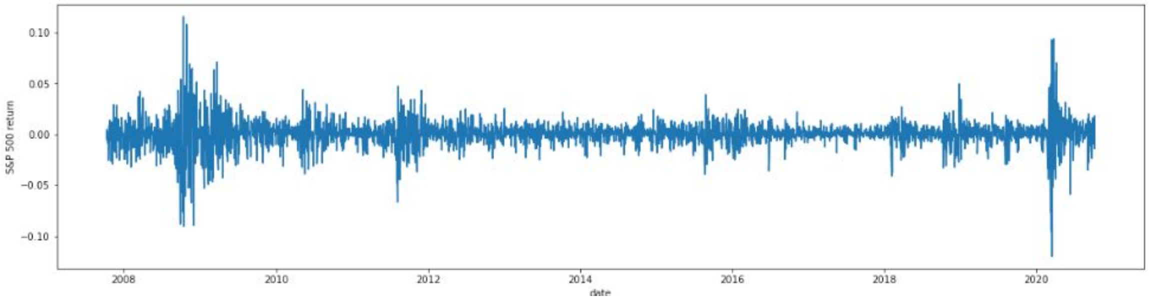


Figure 3: S&P 500 returns

Today’s return can be calculated by the following formula:

$$r_t = \frac{(x_t - x_{t-1})}{x_{t-1}} \text{ (Fool, 2016)}$$

The returns graph does not seem to have any upward or downward trends. Sharp spikes can be observed near the end of 2008, 2011, and early 2020. These seem to align with the stock market crashes associated with the 2008 recession, 2011 Black Monday, and the 2020 coronavirus pandemic.

Table 1: Summary statistics

	Mean	Standard deviation
S&P 500 daily return	0.00033	0.01334
PM 2.5	39.08	15.72

Here are some summary statistics for the datasets. The average stock return is essentially 0, while the average PM 2.5 level is 39, which is within a moderate range.

3.0 Model Construction

Before we construct a model, it is important to understand the motivation behind doing so.

As mentioned in subsection 1.3, hidden variables, such as the inherent performance of the market, may obscure or exaggerate the relationship between PM 2.5 and S&P 500. A multivariate model is therefore employed so that we can effectively control for other variables that can influence the value of S&P 500.

We construct the equation below through the following workflow:[1]

- Determine stationarity of data through the augmented Dickey Fuller Test
- Determine autocorrelation and thus the appropriate lag length through the F-statistic approach
- Determine seasonal indicators through boxplot
- Add interaction terms

$$Y_t = \text{ß}_0 + \text{ß}_1 X_t + \text{ß}_2 Y_{t-1} + \sum_{k=1}^{5} \beta_{k+2} X_{t-k} + \text{ß}_8 X_t Y_{t-1} + \text{ß}_{9,10,11} Season_t + \text{ß}_{12,13,14} X_t Season_t + \varepsilon_t$$

Where:

Y_t is the return of S&P 500 at time t,

X_t is the value of PM 2.5 at time t,

[1] Discussed in greater detail in subsequent sections.

$Season_t$ is spring, summer, winter (seasonal indicators) at time t,

$X_tSeason_t$ is the interaction term multiplying PM 2.5 and the seasonal indicators,

ε_t is the error term which captures the non-observable factors,

$ß_j$ are constants/ parameters to be estimated,

The model will be estimated using the **Ordinary Least Squares method (OLS)**.

The reasons we use the return of S&P 500 instead of the price series are two-fold. Firstly, measuring the daily return of the index would be more meaningful to capture its performance rather than its absolute price. Secondly, and arguably more importantly, the S&P 500 price index is not stationary while the return series is. In its most intuitive sense, a time series is stationary if its statistical properties (mean, variance) remain constant over time (Studenmund, 2017). The reason we want stationarity is because if one dataset had a linear trend (such as the S&P 500 index) and the other dataset (PM 2.5) does not, it is clear that the linear trend is not caused by the other dataset. Specifically, the overall increase in S&P 500 is due to the inherent performance of the stock market and does not relate to PM 2.5 in any way. This causes PM 2.5 to appear to have no causal relationship with S&P 500 because the linear trend outweighs the smaller, localized fluctuations in the two time series. More details regarding the stationarity test can be found in subsection 3.3.

Y_{t-1} is used as one of the predictors for Y_t is because S&P 500 returns are correlated with its value the day before (i.e. autocorrelation). Therefore, omitting Y_{t-1} would bias our estimations. More detail regarding the autocorrelation test and lag length selection can be found in subsection 3.4.

X_t, the daily values of PM2.5, is the main variable of interest. Hence, $ß_1$ is the main interest parameter, which describes the *ceteris paribus*[2] relationship between PM2.5 and S&P 500 performance. Therefore, **obtaining a good estimation of $ß_1$ is the main goal of this paper**.

Lagged values of PM 2.5 up to 5 days are also included as it is reasonable to think that PM 2.5 takes time to enter the stock traders' lungs before it can affect their behaviors.[3]

Seasonal (spring, summer, and autumn) indicators are employed to control for any seasonality effects that may influence S&P 500, discussed further in subsection 3.5.

[2] Fancy word for 'true'

[3] Results for testing each lag individually can be found in Appendix B

OLS selects coefficients $\widehat{ß}_j$ such that the sum of squared residuals over the sample of data points are minimized (Studenmund, 2017). More will be discussed in the next section.

3.1 Ordinary Least Squares

This section attempts to explain the workings of the Ordinary Least Squares (OLS) method. OLS attempts to model a theoretical equation like:

$$Y_k = \beta_0 + \beta_1 X_k + \varepsilon_k$$

with estimated coefficients using a set of data to create an equation

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k$$

By minimizing the function

$$SE_{LINE} = \sum_i \left(Y_k - \hat{Y}_k\right)^2$$

(Studenmund, 2017).

Substituting for $\widehat{Y}_k$, we have

$$= \sum_k \left(Y_k - (\hat{\beta}_0 + \hat{\beta}_1 X_k)\right)^2$$

And by expanding, we get

$$= \sum\nolimits_k Y_k^2 - 2Y_k\hat{\beta}_1 X_k - 2Y_k\hat{\beta}_0 + \hat{\beta}_1^2 X_k^2 + 2\hat{\beta}_0\hat{\beta}_1 X_k + \hat{\beta}_0^2$$

Grouping each term separately,[4]

$$= \sum Y_k^2 - 2\hat{\beta}_1 \sum X_k Y_k - 2\hat{\beta}_0 \sum Y_k + \hat{\beta}_1^2 \sum X_k^2 + 2\hat{\beta}_0\hat{\beta}_1 \sum X_k + k\hat{\beta}_0^2$$

Recall that the expected value is

$$E[Y] = \frac{\Sigma Y_k}{k}$$

Therefore, the equation can be simplified to

$$= kE[Y^2] - 2k\hat{\beta}_1 E[XY] - 2k\hat{\beta}_0 E[Y] + k\hat{\beta}_1^2 E[X^2] + 2k\hat{\beta}_0\hat{\beta}_1 E[X] + k\hat{\beta}_0^2$$

From here, we take the partial derivatives and equate them to 0.

[4] Sigma notation is simplified for readability.

$$\frac{\partial SE}{\partial \hat{\beta}_1} = -2kE[XY] + 2k\hat{\beta}_1 E[X^2] + 2k\hat{\beta}_0 E[X] = 0$$

$$\frac{\partial SE}{\partial \hat{\beta}_0} = -2kE[Y] + 2k\hat{\beta}_1 E[X] + 2k\hat{\beta}_0 = 0$$

By isolating $\widehat{ß}_1$ and $\widehat{ß}_0$ we obtain

$$\hat{\beta}_1 = \frac{E[XY] - E[X]E[Y]}{E[X]^2 - E[X^2]}$$

$\widehat{ß}_0$ is simply modelled by $\widehat{ß}_1$:

$$\hat{\beta}_0 = E[Y] - \hat{\beta}_1 E[X]$$

For equations with more explanatory variables like the one used in this paper, one can use a similar procedure to derive for the constants $\widehat{ß}_j$. However, it would be extremely tedious to do it by hand and would require the assistance of a computer.

The OLS equation relies on a few key assumptions (Stock and Watson, 2010):

- $E(\varepsilon) = 0$ (The error term is 0 on average)
- $E(\varepsilon|X) = E(\varepsilon)$ (Error term is independent from explanatory variables)
- Variance of the error term is known
- There are no outliers

With these assumptions in mind, we can derive some desirable properties of $\widehat{ß}_1$ (Stock and Watson, 2010):

Property 3.1.1 (aka Law of Large Numbers): $\widehat{ß}_j$ converges to β_j at large sample sizes. This is important because it means that $\widehat{ß}_j$ is a good estimate of β_j .

Property 3.1.2: $E(\widehat{ß}_j) = \beta_j$ ($\widehat{ß}_j$ is unbiased). The average of OLS estimators for different samples equals exactly the true/population parameter.

Property 3.1.3: The estimated parameters $\widehat{ß}_j$ follow a normal distribution with the mean being the true value of $ß_j$.

According to the Gauss-Markov Theorem, these properties make OLS the best estimation of β_j in the class of linear models (Rothman, 2020).

To demonstrate how OLS works, a simple case containing only 1 explanatory variable and 7 samples were used, as it is highly impractical to perform hand calculations for anything larger. Values from consecutive entries in the dataset were used. The values are as follows:

Table 2: Randomly sampled 7 consecutive entries in the dataset.

Date	S&P 500 index	S&P 500 return	PM 2.5 level
2015-01-09	2044.8100	-0.008404	57
2015-01-12	2028.2600	-0.008094	60
2015-01-13	2023.0300	-0.002579	69
2015-01-14	2011.2700	-0.005813	55
2015-01-15	1992.6700	-0.009248	64
2015-01-16	2019.4200	0.013424	66
2015-01-20	2022.5500	0.001550	49

We start by rearranging the equation for $\widehat{\beta}_1$:

$$\hat{\beta}_1 = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2}$$

$$= \frac{\frac{\sum X_k Y_k}{N} - \frac{\sum X_k}{N} \times \frac{\sum Y_k}{N}}{\frac{\sum X_k^2}{N} - \left(\frac{\sum X_k}{N}\right)^2} \times \frac{N^2}{N^2}$$

$$= \frac{N \sum X_k Y_k - \sum X_k \sum Y_k}{N \sum X_k^2 - \sum X_k \sum X_k}$$

N=7 since there are 7 samples. The respective components are calculated.

$$\sum_k X_k = 57 + 60 + 69 + 55 + 64 + 66 + 49 = 420$$

$$\sum_k Y_k = -0.008404 - 0.008094 - 0.002579 - 0.005813 - 0.009248 + 0.013424 + 0.001550$$
$$= -0.019164$$

$$\sum_k X_k Y_k = 57 \times -0.008404 + 60 \times -0.008094 + 69 \times -0.002579 + 55 \times -0.005813$$
$$+ 64 \times -0.009248 + 66 \times 0.013424 + 49 \times 0.001550 = -1.092272$$
$$\sum_k {X_k}^2 = 57^2 + 60^2 + 69^2 + 55^2 + 64^2 + 66^2 + 49^2 = 25488$$

Therefore,

$$\hat{\beta}_1 \quad = \frac{7(-1.092272) - (420)(-0.019164)}{7(25488) - 420^2} \quad = 0.0001999$$

$$\hat{\beta}_0 = E[Y] - \hat{\beta}_1 E[X]$$

$$= \frac{(-0.019164)}{7} - 0.0002(60)$$

$$= -0.014731$$

3.2 Hypothesis testing

Hypothesis testing tests a hypothesis by using sample data (Majaski, 2020). We construct a null hypothesis and an alternate hypothesis, and test if the data is significant enough to reject the null hypothesis and support the alternative. In the context of this investigation, we can construct the following hypotheses:

H_0 (null hypothesis): There is no correlation between PM 2.5 and market returns ($\widehat{\beta}_1 = 0$).
H_A (alternative hypothesis): There is a correlation between PM 2.5 and market returns ($\widehat{\beta}_1 \neq 0$)

We want to find out the probability (p-value) of getting a result at least as extreme as the value compiled through the data, given that the null hypothesis is true. If this p-value is below a certain threshold, then getting the observed result is very rare given that the null hypothesis is true. Therefore, we say that we have enough evidence to claim that the null hypothesis is not true. Computing the p-value can be done through significance tests such as the t-test.

3.2.1 T-test

The t-statistic is described by:

$$t \ = \frac{estimator - hypothesized\ value}{standard\ error\ of\ estimator} = \frac{\widehat{\beta}_j - <\beta_j = 0>}{SE(\widehat{\beta}_j)}$$

Where $SE(\widehat{\beta}_j)$ is estimated by

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{MSE}{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}}$$ (Wooldridge, 2016)

and MSE denotes the mean squared error.

The standard error of the mean (in this case, estimator) measures how far the sample mean ($\widehat{\beta}_j$) of the data is likely to be from the true population mean (β_j), and is used to estimate the standard deviation.

The critical values of the t-statistic is described by a t-table[5].

There are some basic conditions the data must meet before being able to perform the z-test.

1. Samples are **random**.
2. Observations follow a **normal distribution.**
3. Observations are **independent** of each other (Khan Academy, n.d.).

Firstly, our data meets the conditions to perform the t-test.

1. **Randomness**: the PM 2.5 and S&P 500 data is not picked specifically by dates.
2. **Normality**: $\widehat{\beta}_i$ follows a normal distribution (property 3.1.3).
3. **Independence**: The current data is not independent due to autocorrelation and other factors. We attempt to alleviate this in subsection 3.4.

According to the Central Limit Theorem, the T-distribution of a large sample size converges to the normal distribution (Stock and Watson, 2011). Since our sample consists of over a thousand data points, the t-test is a good approximation of the z-test and is therefore used in the analysis.

To demonstrate the use of the t-test in OLS, we build off the sample calculations done in section 3.1. The standard error can be rearranged into more familiar terms:

[5] See Appendix A

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{MSE}{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2}} \quad = \sqrt{\frac{MSE}{E[X^2] - E[X]^2}}$$

MSE is the mean squared error, defined as

$$MSE = \frac{SSE}{n-2}$$ (Wooldridge, 2016)

SSE is the sum of squared errors.

$$SSE = \sum\left(Y_k - \hat{Y}_k\right)^2$$

$$\begin{aligned} = &\left(-0.008404 - (-0.0033367)\right)^2 + (-0.0027370 - (-0.008094))^2 \\ &+ (-0.0009379 - (-0.002579))^2 + (-0.0037365 - (-0.005813))^2 \\ &+ (-0.0019374 - (-0.009248))^2 + (-0.0015376 - 0.013424)^2 \\ &+ (-0.0049359 - 0.001550)^2 \\ &= 0.00038074 \end{aligned}$$

Working backwards, we can solve for the standard error.

$$MSE = \frac{0.00038074}{7-2} = 0.00007615$$

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{0.00007615}{\frac{25488}{7} - \frac{420^2}{7^2}}} = 0.001360$$

The t-score can now be determined.

$$t = \frac{estimator - hypothesized\ value}{standard\ error\ of\ estimator}$$

$$= \frac{\hat{\beta}_1 - < \beta_1 = 0 >}{\hat{\sigma}_{\hat{\beta}_j}}$$

$$= \frac{0.0001999}{0.001360}$$

$$= 0.15$$

= NO STATISTICAL SIGNIFICANCE

3.3 Stationarity and the Augmented Dickey Fuller Test

A time series variable X_t is stationary if:

1. The **mean** of X_t remains constant over time.
2. The **variance** of X_t remains constant over time (Studenmun, 2017).

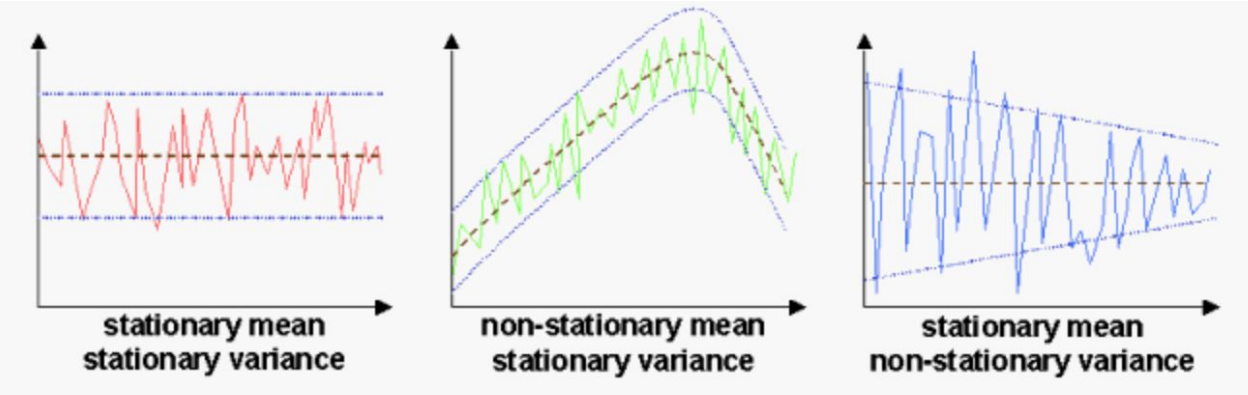


Figure 4: The mean and variance of a time series have to be constant. Taken from Palachy (2019). Figure 2 (S&P 500 index) is non-stationary because its mean is not constant.

To reiterate the importance of stationarity mentioned in sections 1.3 and 3.0, non-stationarity can cause coincidences that appear to be correlation but actually are not (called spurious correlation). Therefore, achieving stationarity is important before proceeding with the rest of the analysis. One way to determine the stationarity of time series data is the Dickey-Fuller Test. Before performing the Dickey-Fuller Test (DF), we explain how it works.

Many time series data, despite not having a time trend, behaves like a **random walk** (Studenmund, 2017)**.** This means that it can wander up and down without settling at an equilibrium mean. Consider the following function describing the change in Y over time .

$$Y_t = \gamma Y_{t-1} + v_t$$

Where:

Y_t is the value at time t,

Y_{t-1} is the value at time t-1,

v_t is a stochastic error term (white noise), and

γ is the coefficient of proportionality.

If $|\gamma| = 1$ (known as a unit root), then by substituting $Y_{t-1} = \gamma Y_{t-2} + v_{t-1}$, we have

$$Y_t = Y_{t-2} + v_{t-1} + v_t$$

Continuing this series of substitutions, we end up with

$$Y_t = \sum_k v_{t-k} \text{ (we assume } Y_0 = 0\text{)}$$

Having the error term grow without limit implies that this time series is not stationary due to a non-constant variance (Fig. 4).

On the other hand, if $|\gamma| < 1$, then the error term converges to 0 (ie. it is stationary). By subtracting Y_{t-1} on both sides, we end up with:

$$Y_t - Y_{t-1} = (\gamma - 1)Y_{t-1} + v_t$$

And we substitute $Y_t - Y_{t-1}$ and $(\gamma - 1)$ for ΔY_t and α_1 respectively.

$$\Delta Y_t = \alpha_1 Y_{t-1} + v_t$$

From here, we can construct a *one-sided hypothesis test* with:
H_0 (null hypothesis): $\alpha_1 = 0$
H_A (alternative hypothesis): $\alpha_1 < 0$

We use the one-sided test favoring $\alpha_1 < 0$ because we are interested in testing whether the time series is stationary, demonstrated by the alternative hypothesis above. Because $\alpha_1 > 0$ is no different than $\alpha_1 = 0$ in the sense that they both indicate nonstationarity, observing $\alpha_1 > 0$ just means that we cannot reject the null hypothesis. We only care about finding out if there is evidence for $\alpha_1 < 0$ to reject the null hypothesis.

The Augmented Dickey-Fuller Test (ADF) follows a similar logic as the regular Dickey-Fuller Test but accounts for serial correlation (error term for each period is correlated to the previous error terms) as well, which is commonly observed in time series data. Thus, we choose ADF over DF.

Table 3: Results of the ADF Test

		S&P 500 index	**S&P 500 returns**	**PM 2.5**
ADF Statistic		0.627	-14.141	-8.261
p-value		0.988	0.000	0.000
Critical Values[6]	**1%**	-3.432	-3.432	-3.433
	5%	-2.862	-2.862	-2.863
	10%	-2.567	-2.567	-2.567

According to the table, we fail to reject the null hypothesis for the S&P 500 index, which confirms that the test works correctly, seeing from Fig. 2 that there is a clear linear trend in the data. However, we can reject the null hypothesis for both the returns and PM 2.5 at the 99% confidence level. In other words, both returns and PM 2.5 are stationary and do not require any further transformations.

3.4 Autocorrelation and Lag length selection

In time series analysis, data is often correlated with past values, known as **autocorrelation** or **serial correlation** (Studenmund, 2017). Consequently, omitting lags would contribute to omitted variable bias, as correlation which is actually driven by the lagged value of S&P 500 could be falsely attributed to PM 2.5. Therefore, in this subsection, we aim to determine whether our time series data contains autocorrelation and at what lag length(s) this occurs. The lag length selection is presented below, where we first define correlation formally. Then, we use an F-statistic approach outlined in Stock and Watson (2010), aided with autocorrelation and partial autocorrelation plots, to find the correlation of S&P 500 returns with the lagged versions of itself.

[6] These are slightly different from those in a normal t-table. Referenced from Studenmund (2017)

The **simple correlation** between y_t and y_{t-k} is defined as:

$$r_k = \frac{\sum_{t=k+1}^{T}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

(Stock and Watson, 2011)[7]

Partial correlation is the direct correlation between Y_t and Y_{t-k}, while ignoring the correlation between them and the intermediate periods (Brownlee, 2017). For example, this looks at the correlation between Y_1 and Y_3 while removing the effects of the correlation between $Y_1 - Y_2$ and $Y_2 - Y_3$. This is defined as:

$$r_k^* = \frac{\sum_{t=k+1}^{T} Y_k^* Y_{t-k}^*}{\sum_{t=k+1}^{T}(Y_{t-k}^*)^2}$$

Where:

${Y^*}_k$ are the residuals from the regressing Y_k on Y_{t-1} up to Y_{t-k}

${Y^*}_{t-k}$ are the residuals from regressing Y_{t-k} on Y_{t-1} up to Y_{t-k}

Each vertical line in Fig. 5 and Fig. 6 below indicate how strongly y_t and y_{t-k} are correlated, ranging from -1 to 1. If the correlation = -1, the two variables are perfectly and negatively correlated. If correlation = 0, they are not correlated. If correlation = 1, they are perfectly positively correlated (Ganti, 2020).

For instance, in Fig. 5, the first line tells us that the correlation r_0 between the price of S&P 500 at time t and time $t - 0$ (i.e. correlation at time t to itself) is, unsurprisingly, 1. However, the second line tells us that r_1 (the correlation between returns at time t and time $t - 1$) is nearly -0.2.

The shaded blue region shows the 95% confidence interval (2 standard deviations away from zero). Hence, if a line ends outside of the region, the correlation is significantly different from zero (ie. there is a correlation). For example, the first two lines (r_0 to r_k) in Fig. 5 are significantly different from zero. If this happens, then there is autocorrelation between y_t and y_{t-k}, k being the number of lags from t.

[7] Note that this is also Covariance over Variance

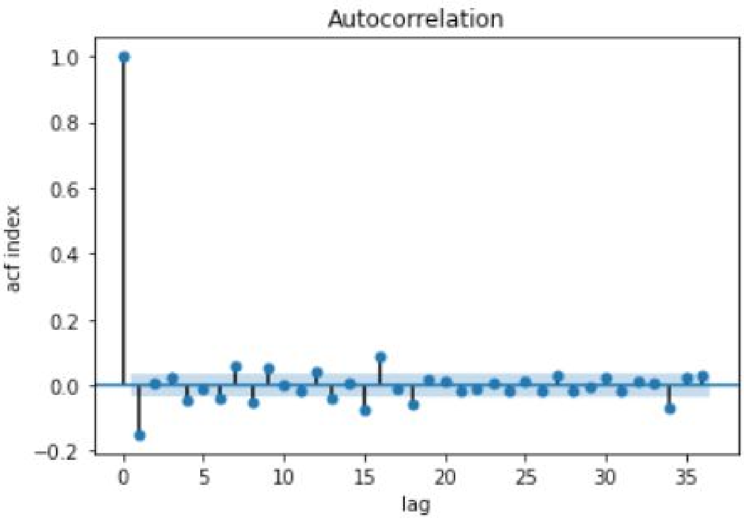


Figure 5: ACF plot for S&P 500 returns.

Only 1 value does not lie within the 95% confidence interval at lag k=1 (k=0 means it is correlated with itself so we can ignore it). This is evidence to say that lag length is 1, meaning it is correlated only with 1 previous period. We include a 1-lagged term of the return of S&P 500 as an explanatory variable to account for this autocorrelation.

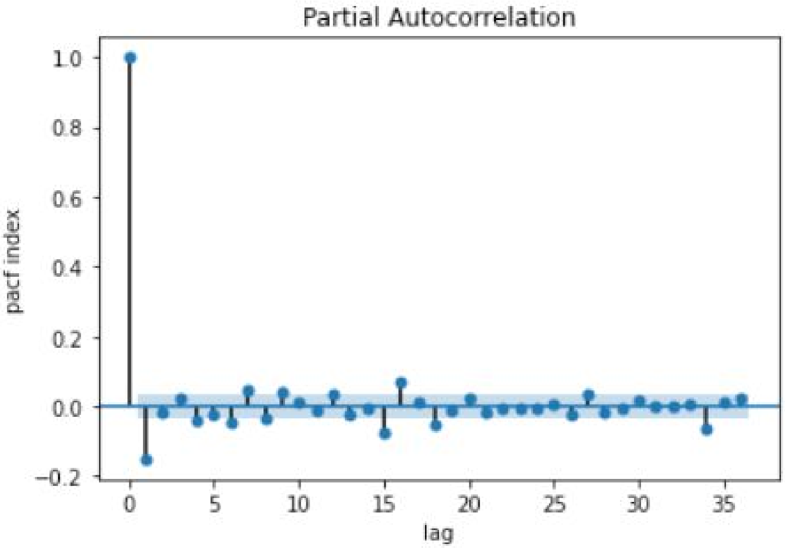


Figure 6: Partial Autocorrelation plot of SP 500 returns

A spike is observed at k=1, so there is only correlation between Y_t and Y_{t-1}, which further confirms the results of the ACF.

Now that we have determined where the autocorrelation is, we have justified our use of the lagged term S&P 500 as an explanatory variable in our model.

3.5 Seasonal indicators

PM 2.5 levels change with the time of year (Russell et. al, 2004). This seasonal trend cannot be bundled together with PM 2.5 to explain S&P 500 returns, because we want to observe the irregular fluctuations of PM 2.5's effect on S&P 500 returns, not seasonal changes. We group PM 2.5 levels based on the month and display it using a boxplot, which makes it easy to observe the seasonal fluctuations of PM 2.5.

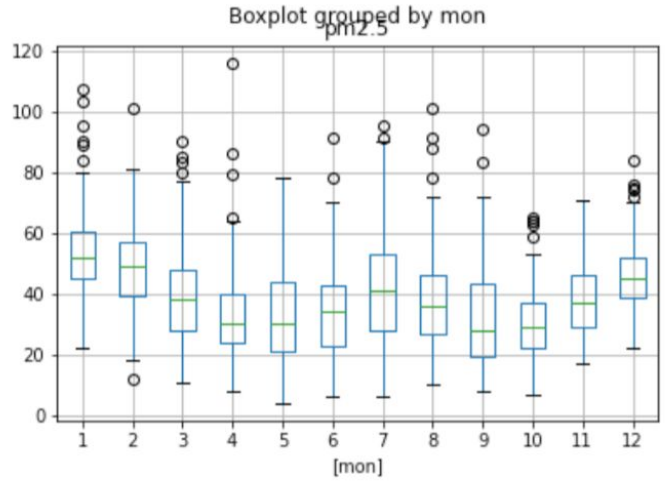


Figure 7: Monthly boxplot for PM 2.5

PM 2.5 levels are high in summer and winter and low in spring and autumn, a clear seasonality trend. Based on the boxplot, the seasons can be grouped into:

- Winter: 12-2
- Spring: 3-5
- Summer: 6-8
- Autumn: 9-11

We use seasons instead of the more specific monthly effects because we want to minimize time-fixed effects, as too many can cause overfitting. Only 3 seasonal indicators are needed because including all 4 will cause the problem of multicollinearity (Frost, 2017).

3.6 Interaction Terms

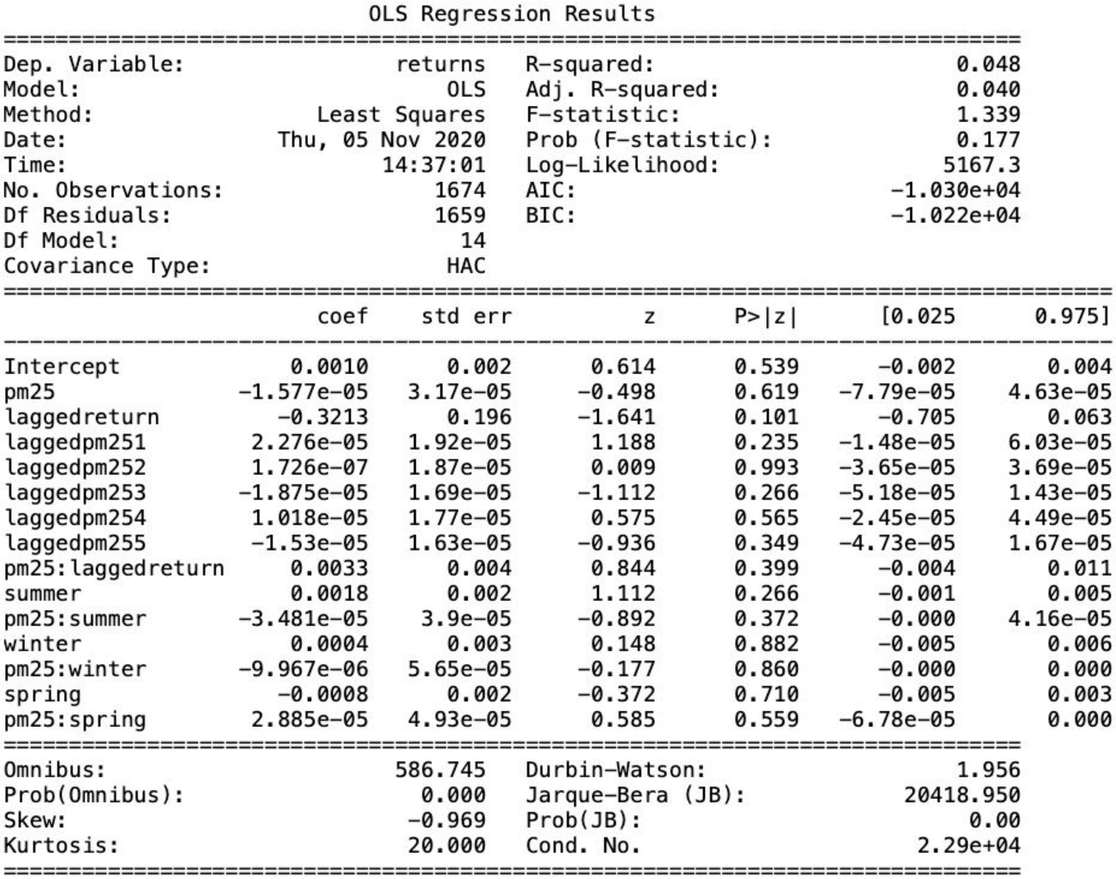
Sometimes, variables may have an influence on something collectively, but have no effect individually. For example, PM 2.5 levels and winter together may have an effect on the return of S&P 500, but PM 2.5 and winter alone may not. These are called *interaction terms*, and are created by multiplying two explanatory variables together (Stock and Watson, 2011). We include the interaction terms of PM 2.5 with each season, and with the lagged return, and see if they are of any statistical significance.

Specifically, we add the following terms to our model:

- 1-lagged return × PM 2.5
- PM 2.5 × spring
- PM 2.5 × summer
- PM 2.5 × winter

4.0 Results and Analysis

Now that we have all the variables we need, the model can be run.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                returns   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.040
Method:                 Least Squares   F-statistic:                     1.339
Date:                Thu, 05 Nov 2020   Prob (F-statistic):              0.177
Time:                        14:37:01   Log-Likelihood:                 5167.3
No. Observations:                1674   AIC:                        -1.030e+04
Df Residuals:                    1659   BIC:                        -1.022e+04
Df Model:                          14
Covariance Type:                  HAC
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               0.0010      0.002      0.614      0.539      -0.002       0.004
pm25                -1.577e-05   3.17e-05     -0.498      0.619   -7.79e-05    4.63e-05
laggedreturn           -0.3213      0.196     -1.641      0.101      -0.705       0.063
laggedpm251          2.276e-05   1.92e-05      1.188      0.235   -1.48e-05    6.03e-05
laggedpm252          1.726e-07   1.87e-05      0.009      0.993   -3.65e-05    3.69e-05
laggedpm253         -1.875e-05   1.69e-05     -1.112      0.266   -5.18e-05    1.43e-05
laggedpm254          1.018e-05   1.77e-05      0.575      0.565   -2.45e-05    4.49e-05
laggedpm255          -1.53e-05   1.63e-05     -0.936      0.349   -4.73e-05    1.67e-05
pm25:laggedreturn       0.0033      0.004      0.844      0.399      -0.004       0.011
summer                  0.0018      0.002      1.112      0.266      -0.001       0.005
pm25:summer         -3.481e-05    3.9e-05     -0.892      0.372      -0.000    4.16e-05
winter                  0.0004      0.003      0.148      0.882      -0.005       0.006
pm25:winter         -9.967e-06   5.65e-05     -0.177      0.860      -0.000       0.000
spring                 -0.0008      0.002     -0.372      0.710      -0.005       0.003
pm25:spring          2.885e-05   4.93e-05      0.585      0.559   -6.78e-05       0.000
==============================================================================
Omnibus:                      586.745   Durbin-Watson:                   1.956
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            20418.950
Skew:                          -0.969   Prob(JB):                         0.00
Kurtosis:                      20.000   Cond. No.                     2.29e+04
==============================================================================
```

Figure 8: Final model

The Python package I used for this OLS regression outputted a z-value instead of the expected t-value. This is perhaps due to the fact that a T-test is indeed being run, but the critical values of the z-table[8] are referenced instead.

[8] The z-table can be found at http://www.z-table.com.

The interaction term of PM 2.5*spring has a positive coefficient, while its components, PM 2.5 and spring, have negative coefficients, showing that interaction terms change the way that the individual variables affect the returns of S&P 500.

$|z| > 2$ is significant (or p-value < 0.05). Thus, the effect of PM2.5, with a z-value of -0.5, is not statistically significant. None of the other variables are statistically significant as well. Consequently, S&P 500 is not strongly correlated with any of the explanatory variables, and we are unable to reject the null hypothesis. The R^2 value also appears quite low, meaning that a low percentage of variation in data that can be explained by the model.

5.0 Conclusion

In this paper, I have failed to replicate the results of Heyes et. al (2016), who claimed to have found evidence that PM 2.5 is correlated with the performance of the stock market. This may have been due to several reasons.

Firstly, the data on PM 2.5 was limited; I was only able to obtain daily data since 2014. Heyes et. al (2016) collected hourly data directly from monitoring stations directly outside of the New York Stock Exchange building, something I was not able to obtain given limited resources. As the effects of PM 2.5 happen on a rather local scale (might affect stock traders within hours), the low frequency of the data I used may have been a major reason why I did not obtain the desired results.

Secondly, risk-taking may only be triggered when there is widespread panic behaviour. Therefore, it may be valuable to look at the specific periods where the stock market experiences significant fluctuations, such as during the 2008 Great Recession or the recent 2020 coronavirus pandemic.

Thirdly, Heyes et. al (2016) used data for the weather and other pollutants as explanatory variables as well. As I did not use this sort of data in my investigation, this may have caused nontrivial omitted variable bias and obscured the correlation between PM 2.5 and S&P 500 returns.

Lastly, this may have boiled down to the problem of utilizing a model that was too simple. Further research directions could include utilizing more complex models, such as the General Linear Mixed-effects model, or other non-linear models. Some examples may be found in Wu (2009).

6.0 References

Air Quality Historical Data Platform. (n.d.). World Air Quality Index Project. Retrieved November 9, 2020, from https://aqicn.org/data-platform/

Air Quality Index (AQI) Air Quality Communication Workshop. (2012). In *Environmental Protection Agency*. https://www.epa.gov/sites/production/files/2014-05/documents/zell-aqi.pdf

Brownlee, J. (2017). *A Gentle Introduction to Autocorrelation and Partial Autocorrelation*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. The lancet, 360(9341), 1233-1242.

Chiou, L., Liou, L., & Arron Kau, P. C. (n.d.). *Covariance | Brilliant Math & Science Wiki*. Brilliant.org. Retrieved November 9, 2020, from https://brilliant.org/wiki/covariance/

EPA,OAR. (2018, November 14). *Particulate Matter (PM) Basics | US EPA*. US EPA. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

Fool, M. (2015, December 13). *How to Calculate Return on Indices in a Stock Market*. The Motley Fool. https://www.fool.com/knowledge-center/how-to-calculate-return-on-indices-in-a-stock-mark.aspx

Frost, J. (2017, April 3). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions - Statistics By Jim*. Statistics By Jim. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. Journal of Political Economy 80(2): 223–255.

Heyes, A., Neidell, M., & Saberian, S. (2016). The effect of air pollution on investor behavior: evidence from the S&P 500 (No. w22753). National Bureau of Economic Research.

Kenton, W. (2020). *Understanding S&P 500 Index – Standard & Poor's 500 Index*. Investopedia. https://www.investopedia.com/terms/s/SP500.asp

Khan Academy. (n.d.). *Conditions for inference on a proportion*. Khan Academy; Khan Academy. Retrieved November 9, 2020, from https://www.khanacademy.org/math/ap-statistics/estimating-confidence-ap/one-sample-z-interval-proportion/a/conditions-inference-one-proportion

Levy, T., & Yagil, J. (2011). Air pollution and stock returns in the US. Journal of Economic Psychology, 32(3), 374-383.

Li, Q. and C.H. Peng (2016). "The stock market effect of air pollution: evidence from China." Applied Economics 48(36): 3442-3461.

Majaski, C. (2020). How Hypothesis Testing Works. Investopedia. https://www.investopedia.com/terms/h/hypothesistesting.asp

Peixeiro, M. (2019, August 7). The Complete Guide to Time Series Analysis and Forecasting. Medium; Towards Data Science. https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

Purdue Writing Lab. (2018). *General Format // Purdue Writing Lab*. Purdue Writing Lab. https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/general_format.html

Rothman, A. (2020, June 8). *OLS Linear Regression, Gauss-Markov, BLUE, and understanding the math*. Medium. https://towardsdatascience.com/ols-linear-regression-gauss-markov-blue-and-understanding-the-math-453d7cc630a5

Russell, M., Allen, D. T., Collins, D. R., & Fraser, M. P. (2004). Daily, Seasonal, and Spatial Trends in PM2.5 Mass and Composition in Southeast Texas Special Issue of Aerosol Science and Technology on Findings from the Fine Particulate Matter Supersites Program. *Aerosol Science and Technology*, *38*(S1), 14–26.

S&P 500 Index - 90 Year Historical Chart. (2009). Macrotrends.net. https://www.macrotrends.net/2324/sp-500-historical-chart-data

Stock, J. H., & Watson, M. W. (2011). *Introduction to econometrics* (3rd ed.). Pearson.

Studenmund, A. H., & Bruce Kenneth Johnson. (2017). *A practical guide to using econometrics*. Harlow, England Pearson.

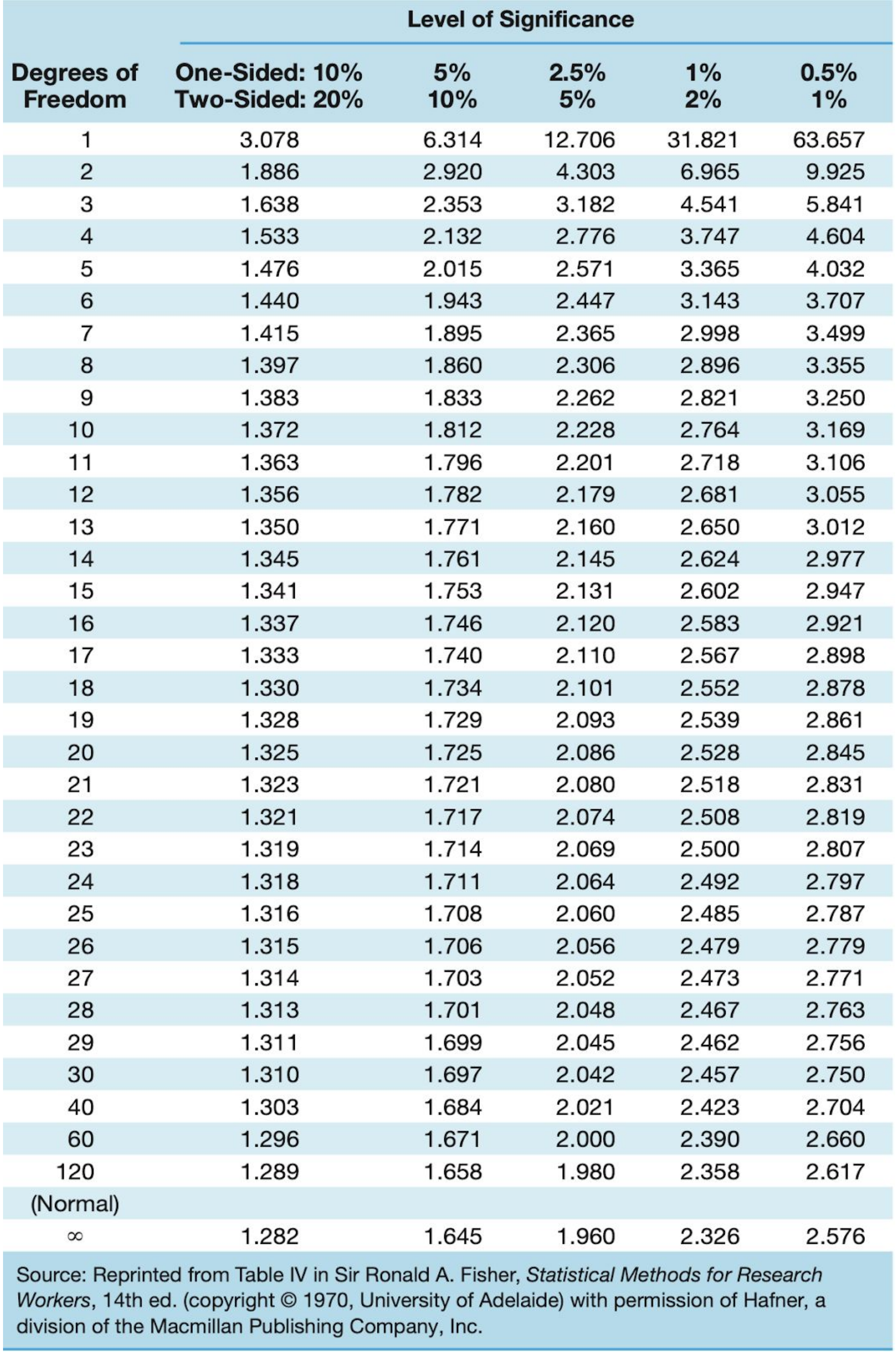
Wooldridge, J. M. (2016). *INTRODUCTORY ECONOMETRICS : a modern approach.* (6th ed., pp. 47–50). Cengage Learning.

Wu, L. (2009). "Mixed Effects Model for Complex Data," Chapman and Hall/CRC, pp. 31, 60-67. Isbn-10: 1420074024.

Wu, Q., Hao, Y., & Lu, J. (2018). Air pollution, stock returns, and trading activities in China. Pacific-Basin Finance Journal, 51, 342-365.

Appendix A

T-table (taken from Studenmund, 2017)



	Level of Significance				
Degrees of	One-Sided: 10%	5%	2.5%	1%	0.5%
Freedom	Two-Sided: 20%	10%	5%	2%	1%
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
(Normal)					
∞	1.282	1.645	1.960	2.326	2.576

Source: Reprinted from Table IV in Sir Ronald A. Fisher, *Statistical Methods for Research Workers*, 14th ed. (copyright © 1970, University of Adelaide) with permission of Hafner, a division of the Macmillan Publishing Company, Inc.

Appendix B

Lagged PM 2.5

As mentioned in section 3.0, PM 2.5 takes time for its effects to manifest. We test lagged terms of PM 2.5 up to 5 days.

Table 3: Adding lagged PM 2.5 to the base model

	coefficient	Std. error	z	P>|z|
1	7.914×10^{-6}	1.63×10^{-5}	0.485	0.628
2	-3.571×10^{-6}	1.73×10^{-5}	-0.206	0.837
3	-1.565×10^{-5}	1.6×10^{-5}	-0.976	0.329
4	-2.541×10^{-6}	1.74×10^{-5}	-0.146	0.884
5	-1.385×10^{-5}	1.65×10^{-5}	-0.838	0.402

It seems like the unlagged version of PM 2.5 has the lowest p-value (although still not statistically significant), and is therefore used as a constituent of the interaction terms.