GROUP WORK PROJECT # 2
GROUP NUMBER: 3352

MScFE 610: FINANCIAL ECONOMETRICS

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Jefferson Bien-Aimé	USA	jefferson.bienaime@gmail.com	
Marin Stoyanov	Bulgaria	azonealerts@gmx.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).	
Team member 1	Jefferson Bien-Aimé
Team member 2	Marin Stoyanov
Team member 3	

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed. **Note:** You may be required to provide proof of your outreach to non-contributing members upon request.
N/A

1. Multicollinearity

1.1. Covariance and Correlation Definitions:

Covariance is a measure used to determine how much two variables change together. The formula for covariance is as follows:

$$Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

A positive covariance indicates that the two variables move in the same direction, while a negative covariance indicates that they move in opposite directions. A covariance of 0 means that the two variables are linearly uncorrelated. The strength of the relationship between the two variables is indicated by the absolute value of their covariance. However, since the value of covariance can change when the scales of the two variables change, correlation is often used instead in data analysis. [11]

Correlation measures the co-movement of two variables while eliminating the issue of scale by dividing the covariance by the square root of the product of the variances of the two variables. The formula for correlation is as follows:

$$Corr = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

In finance, it is common for variables to move together in systematic ways. When two variables are correlated, this is known as collinearity. When more than two variables are correlated, this is known as multicollinearity. To detect multicollinearity, one can use various statistical methods. Usually, a correlation matrix is a good starting point, where a table filled with the variables correlations is built. [12]

After doing some exploratory data analysis in the notebook we decided to do time series analysis. But first let's give some additional definitions.

1.2. Time Series Definition

A time series is a dataset where every cell is filled with a random variable (let's say X_t which does not need to be independent or identically distributed) and this dataset is ordered by an index that represents an ordered time stamp like date, week, month etc.

The term "time lag" is used in time series analysis and it simply means the fixed amount of passing time between two sets of observations. [13] The formula looks like this:

$Lag_1(X_2) = X_1 \rightarrow$ this is a fist lag example

$Lag_3(X_{10}) = X_7 \rightarrow$ this is a third lag example

1.3. Autocorrelation

Autocorrelation is when a time series is linearly related to a lagged version of itself or in simple words it measures the relationship between a variable's current value and its past values (depending on the time lag that we explained above).

1.4. Mean function of time series

This is a measure to describe the center of the time series and the formula looks like this:

$$\mu_{X_t} = E(X_t) = \int_a^b x f_t(x) dx$$

Where:

- a and b are the bounds
- f_t is the marginal density function for X_t

1.5. Autocovariance function

Autocovariance is the covariance of a variable with a time lagged state of itself and the formula looks like this

$$\gamma_X(s,t) = cov(X_s, X_t) = E[(X_s - \mu_{X_s})(X_t - \mu_{X_t})]$$

Where s and t are the different time stamps of the time series variable [14]

1.6. Autocorrelation function (ACF)

This is the mathematical equation that shows if the current observation has any significant correlation with observations in different time lags. The formula looks like this:

$$\rho_X(s,t) = \frac{\gamma_X(s,t)}{\sqrt{\gamma_X(s,s)\gamma_X(t,t)}}$$

It is important to say that the the values for $\rho_X(s,t)$ are bounded between -1 and 1 [15]

1.7. Partial autocorrelation

The partial autocorrelation function (PACF) helps us understand the correlation between a time series and its lagged versions, excluding the effects of intermediate time points. The formula look like this:

$$\phi_X(t, t-3 \mid t-1, t-2) = \frac{cov(X_t, X_{t-3} \mid X_{t-1}, X_{t-2})}{\sqrt{var(X_t \mid X_{t-1}, X_{t-2}) var(X_{t-3} \mid X_{t-1}, X_{t-2})}} \quad [15]$$

For simplification of the notation we can use just this:

$\phi_X(t, t-3)$

1.8. Diagnosis

- For proving that there is a multicollinearity issues with the independent variables we can use the Variance Inflation Factor (VIF) [16]
- Also a good practice is to do Exploratory Data Analysis and plot the correlation matrix
- Also a good choice for visualizations is a heatmap and a pairplot
- Then when we are more familiar with the dataset we can make time series analysis and test for autocorrelation with ACF plot and partial autocorrelation with PACF plot

1.9. Damage

Multicollinearity can lead to the following problems:

- The coefficient estimates of the model can fluctuate significantly based on which other predictor variables are included in the model. This means that small changes in the model specification can result in large changes in the estimated coefficients.
- It reduces the precision of the estimated coefficients. This means that the standard errors of the coefficients are larger than they would be if there were no multicollinearity. As a result, the p-values associated with the coefficients are less reliable, making it more difficult to determine which predictor variables are statistically significant.
- It can cause the coefficient estimates to swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model, making it difficult to interpret their effects.
- It also weakens the statistical power of our regression model by reducing the precision of the estimated coefficients. This can lead to wider confidence intervals for the coefficients, which produce less reliable probabilities in terms of the effect of independent variables in a model. As a result, the statistical inferences from a model with multicollinearity may not be dependable.

Considering all the above, multicollinearity can make it difficult to accurately estimate and interpret the effects of predictor variables in a multiple regression model. [17]

1.10. Directions.

For reducing multicollinearity there are a few suitable approaches:

- Feature generation: some kind of combination, interaction or a simple ratio between some of the independent variables is a good way to save information and lower the number of the variables. (usually after this transformation the old variable should be dropped) [18]
- Principal Component Analysis (PCA) can be used to lower the dimensionality of the dataset and relatively saving the most of the information (after such kind of transformation a little part of the information is lost)
- Support Vector Machine is another approach for lowering dimensionality
- Lasso regression is suitable for measuring the feature importance and thus dropping the not so important variables. [18]

2. Unit Root Testing

2.1 Definition

Unit root tests are statistical procedures used to determine if a time series data set is non-stationary, exhibiting a trend or random walk, and possesses a unit root. According to Robert Nau at the Fuqua school of Business at Duke University, The time series variable has a constant mean and variance over time, but the trend is not constant. This is called stationarity. [1][2][7]

For Unit Root Testing, you will likely detect a pattern in the time series. The equation for a unit root process can be represented as the following

$$Y_t = Y_{t-1} + u_t$$

where Y_t is the value of the series at time t, Y_{t-1} is the value at the previous time period, and u_t is a random error term.

2.2 Description

Unit root tests are used to test the null hypothesis that a time series variable has a unit root. The alternative hypothesis is that the variable is stationary. If the null hypothesis is rejected, then the variable is non-stationary.

There are different types of tests for stationary in a time series. They all followed the same logic in terms of saying what is the null hypothesis or the alternative hypothesis. [3][5][7]

Null Hypothesis (H0): The time series has a unit root (i.e., it is non-stationary).

Alternative Hypothesis (H1): The time series does not have a unit root (i.e., it is stationary).

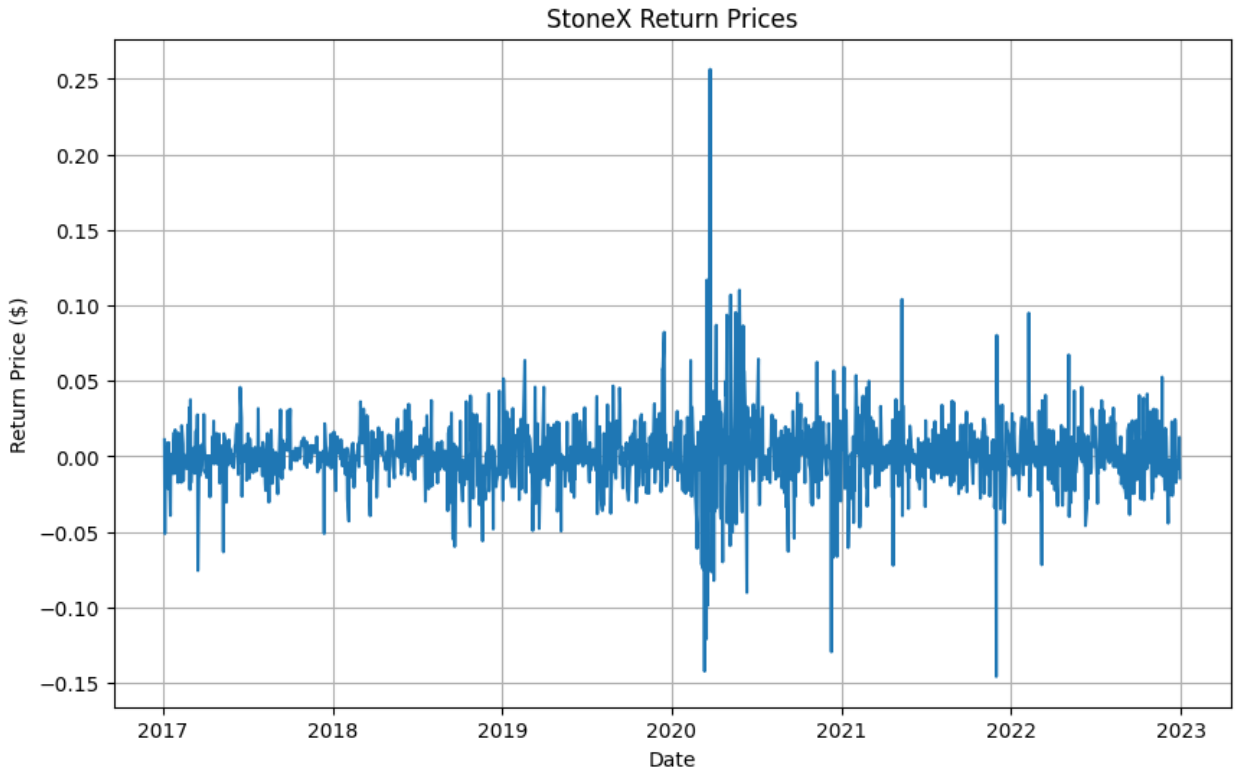
The following tests can be used:

- The Dickey-Fuller Test (sometimes called a Dickey Pantula test), which is based on linear regression.
- Augmented Dickey-Fuller (ADF) test can be used to handle bigger, more complex models and serial correlation.
- The ADF-GLS Elliott–Rothenberg–Stock Test or ERS. This test removes the trend in the time series data.
- The Schmidt Phillips which is based on regression.
- The statistical test Phillips Perron is a modified version of the Dickey-Fuller test, accounts both heteroscedasticity and autocorrelation.
- The Zivot-Andrews test excludes exogenous structural change in its unit root process[4].

2.3 Demonstration

For the demonstration, please refer to the notebook for more information. I have taken a specific stock price, the close price of the public company known as StoneX Group, Symbol is SNEX. I have looked at the prices from January 2017 to December 2022. After getting the data from yahoo finance, using the yfinance library, we will plot the Close Price for each day. The next thing we do is to use the ADF test to check whether or not the series is stationary or not. We obtain a p-value = 0.92 which says the series is indeed non-stationary. See Diagrams section for more diagrams

2.4 Diagrams:



2.5 Diagnosis:

The need for unit root testing typically arises when working with time series data, as many statistical models require the input data to be stationary. Non-stationary data often results from trends, cycles, random walks, or other structural changes in your data. To test this, we use one of the test mentioned in the demonstration section. If the p-value is less than 0.05, we reject the null hypothesis. We deduct the series is stationary, data doesn't have a unit root. That means if the p-value is greater than 0.05, the series is non-stationary and has a unit root. See the directions section on how we solve this.

2.6 Damage:

Non-stationarity in time series data can lead to spurious regression results, leading to incorrect inferences about the relationships and predictions based on the data. For example, If two series each have a unit root, standard regression techniques may indicate a statistically significant relationship between the two, even if the two series are entirely unrelated. This can result in misleading analytics and poor decision-making[6].

2.7 Directions:

If your data has a unit root (i.e., it is non-stationary), you could consider differencing the data, applying a transformation like logging or deflating, or using statistical models like ARIMA that can handle non-stationarity.

1. Differencing: This is a method that helps make the data stationary by subtracting the previous observation from the current observation.
2. Transformation: This involves methods like logarithms, square roots, etc., to stabilize the non-constant variance of a series.
3. Using models like ARIMA that handle non-stationarity in the data.
4. Use non-linear models or models that take non-stationarity into account like ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) Models

In our case, we have used a differencing method by calculating the return of the data and redoing the ADF test which led to a p-value less than 0.05. To solve the issue, we have taken the difference of the close price and drawn both the graph and computed the ADF. The p-value is 3e-18 which is less than 0.05. We can conclude the time series is now stationary.

3. Feature Extraction

3.1. Data normalization

Before we start applying Lasso regression we have to normalize the data. There is a very useful method for normalization called: StandartScaler() and it is part of the scikit-learn module. It can also be called: the z-score transformation and the formula behind this procedure is this:

$$Z = \frac{x_i - \mu}{\sigma}$$

Where:

- x_i is the value of the current variable at moment i
- μ is the mean of the given time series
- σ is the standard deviation of the given time series

3.2. Lasso Regression

For the procedure of feature extraction we are going to use Lasso regression which is part of the Penalized Regressions family. Lasso regression is based on the OLS regression with the addition of a penalty function. In an OLS regression model, we try to minimize the following objective function:

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \ldots - \beta_p X_{pi})^2$$

Here is no restriction about how many coefficients we can have and thus this can lead to overfitting.The objective function of a Penalized regression looks like this:

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \ldots - \beta_p X_{pi})^2 + \lambda \sum_{j=1}^{p} f(\beta_j)$$

For Lasso regression (the full name is **least absolute shrinkage and selection operator**) has the following penalty function:

$$f() = \sum_{j=1}^{p} |\beta_j| = ||\beta_1||$$

And this penalty function is called L1 penalty function

3.3. Damage

- Without feature extraction procedure and if we let all the independent variables stay and be used in the model, we are letting the possibility of overfitting materialize.
- It is a good practice to drop the not important feature and make the model lighter for easier and faster calculations.
- Especially after some feature engineering, then the new variables that contain more information like ratios and other transformations should stay in the model and the rest may become redundant.

3.4. Directions

- Choosing which model for data extraction to work with
- Data normalization (In our case, we chose Lasso regression so this step is mandatory) by the StandartScaler() method from the sklearn library. (a z-score transformation would do the job as well)
- Doing some additional feature engineering like ratios or other combinations so that some new and more informative variables can be created
- Running the Lasso regression model and fighting it to the data
- Visualizing the feature importance graph by plotting the Lasso regression coefficients
- Dropping the not important independent features and the data extraction is ready

4. **Challenges relations**.

- Skewness, outliers and multicollinearity do relate because they all can lead to distorted results. Outliers can skew the data distribution away from the normal distribution. At the same time

multicollinearity issues can cause a failure in outlier detection which can lead to a skewed data distribution, so all in all they are all connected. [8]

- Unit root testing is used to determine whether a time series variable is non-stationary.. The presence of a unit root indicates that the time series is non-stationary, meaning that its statistical properties such as mean and variance are not constant over time[3]. Unit root tests can be used to determine if trending data should be first differenced or regressed on deterministic functions of time to render the data stationary[4].
- Unit root testing and further time series modeling do relate in such a way that the procedure of unit root test should be done before the modeling because thanks to it we can find if the data is stationary or not. The presence of a unit root indicates that the time series is non-stationary, meaning that its statistical properties such as mean and variance are not constant over time. Thanks to unit root testing we can determine if trending data should be first differenced or regressed on deterministic functions of time to render the data stationary. After that we can continue with the modeling by looking at the ACF and PACF functions so that we can choose the suitable settings for the model. [9]
- Feature extraction and multicollinearity are related because thanks to feature extraction methods like Lasso, the issue with multicollinearity can be dealt with. Thanks to the penalty function of the Lasso regression the independent variables can be evaluated and specific feature importance metrics can be given to them. This makes some of them not as important as others and so they can be dropped which can remove the previous multicollinearity issues. [10]

REFERENCES

1. Hamilton, J. D. (1994). Time series analysis (Vol. 2, No. 1994). Princeton: Princeton University Press. This book provides a comprehensive introduction to the analysis of time series data, including unit root testing.
2. Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 71(3), 599-607. This paper introduced the Augmented Dickey-Fuller test for unit roots.
3. Kotz, S.; et al., eds. (2006), Encyclopedia of Statistical Sciences, Wiley.
4. Zivot, E., & Andrews, D. W. K. (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. Journal of Business & Economic Statistics, 20(1), 25-44.
5. Andrew Ozbun (2021). Unit Root Testing
6. Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. Journal of Econometrics, 2(2), 111-120.

7. https://medium.com/codex/unit-root-in-time-series-38d451d742ce
8. Jurczyk, T. (2011). Outlier detection under multicollinearity. Journal of Statistical Computation and Simulation
9. Dickey, D. and W. Fuller (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," Journal of the American Statistical Association
10. Hastie, T.J., Tibshirani, R.J., and Wainright, M. (2015). Statistical Learning with Sparsity. CRC Press, Boca Raton, FL.
11. Oxford Dictionary of Statistics, Oxford University Press, 2002
12. "Correlation (in statistics)", Encyclopedia of Mathematics, EMS Press, 2001
13. Keogh, Eamonn J. (2003). "On the need for time series data mining benchmarks". Data Mining and Knowledge Discovery
14. Gubner, John A. (2006). Probability and Random Processes for Electrical and Computer Engineers. Cambridge University Press
15. Shumway, Robert H.; Stoffer, David S. (2017). Time Series Analysis and Its Applications: With R Examples. Springer Texts in Statistics. Cham: Springer International Publishing
16. O'Brien, R. (2007) A Caution Regarding Rules of Thumb for Variance Inflation Factors, Quality & Quantity
17. Greene, W.H. (2002) Econometric analysis, 5th edition. Prentice Hall.
18. Taboga, Marco (2021). "Multicollinearity", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix. https://www.statlect.com/fundamentals-of-statistics/multicollinearity.