Final Report
Statistical Methods

Nicolas Jorquera

Introduction

1 – Meeting the Data

In this project we analyze four random variables: the daily ETF return; the daily relative change in the price of the crude oil; the daily relative change in the gold price; and the daily return of the JPMorgan Chase & Co stock. Below we analyze the sample mean and sample standard deviation, with each variable having a sample size of 1000.

	Close ETF	Oil	Gold	JPM
Standard Deviation	12.570	0.0211	0.0113	0.00110
Mean	121.153	0.00103	0.000663	0.000530

Table 1 – Standard Deviation / Mean

We can also check the correlation between different random variables, and can see that there is no significant linear correlation (definition of the Pearson Correlation Coefficient); because most values are close to 0. This can also be visualized using a heat map, seen below.

	Close ETF	Oil	Gold	JPM
Close ETF	1.000	-0.00905	0.02299	0.0368
Oil		1.000	0.2356	-0.1208
Gold			1.000	0.1000
JPM				1.000

Table 2 – Correlation between Random Variables

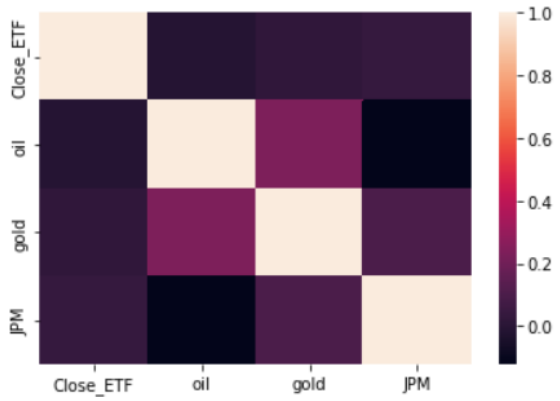


Figure 1 – Heatmap of Random Variables.

Methodology

2 – Describing the Data

In this project we will be using Python to analyze this data, and try to find patterns and results. We will do this by using statistical tests and methods to find patterns in the data, and see if they have any significance. To assist us we will use the statistical packages included in Python. For an in depth look at the code please refer to the Appendix.

However we will begin exploring the distribution of this data by utilizing a histogram, as well as exploring any patterns using a Time Series Plot:

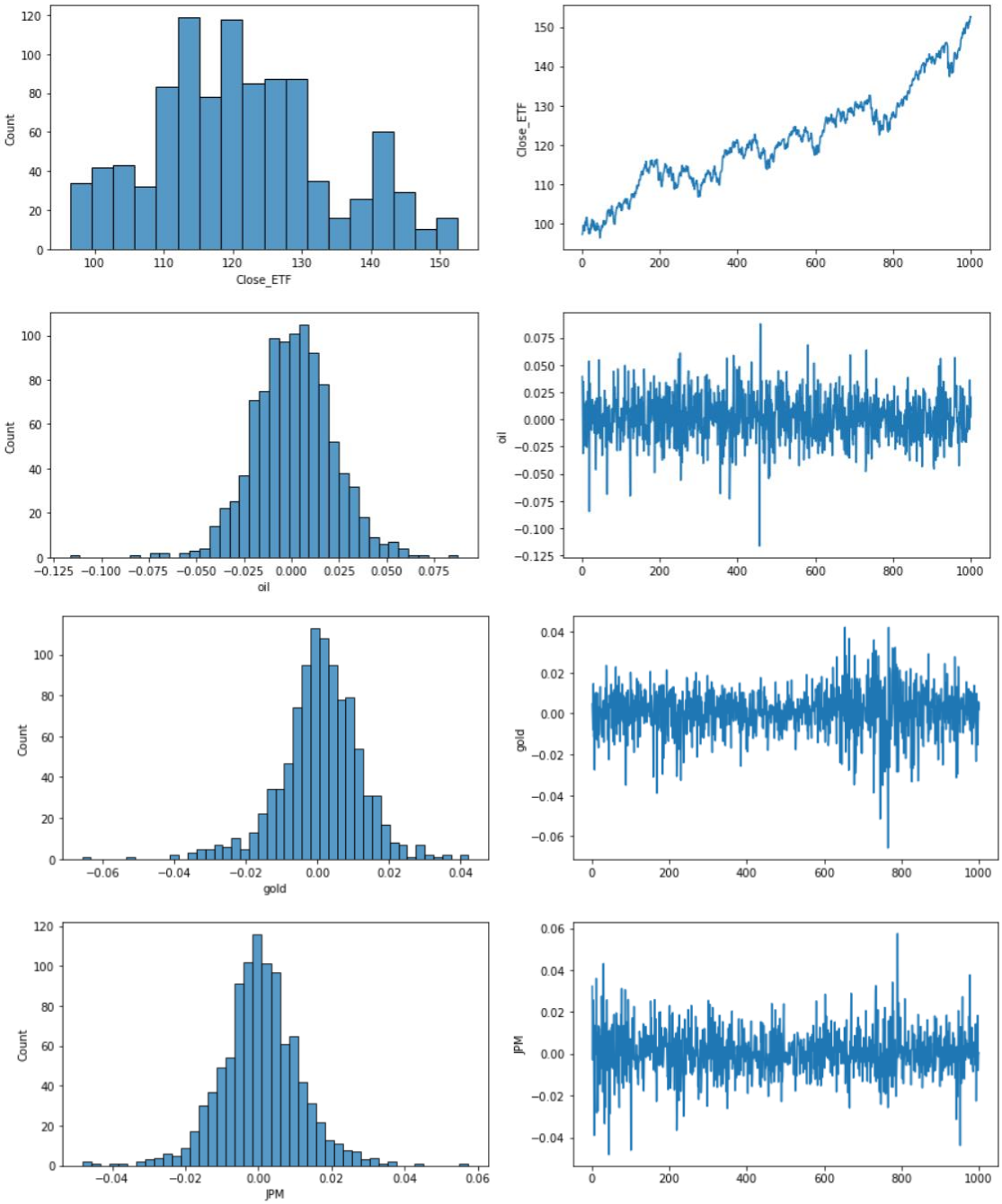


Figure 2 – Histogram Plot / Time Series Plot for each Random Variable

In Figure 2 we can see that it looks like most of the data is normally distributed, just skewed either left or right.

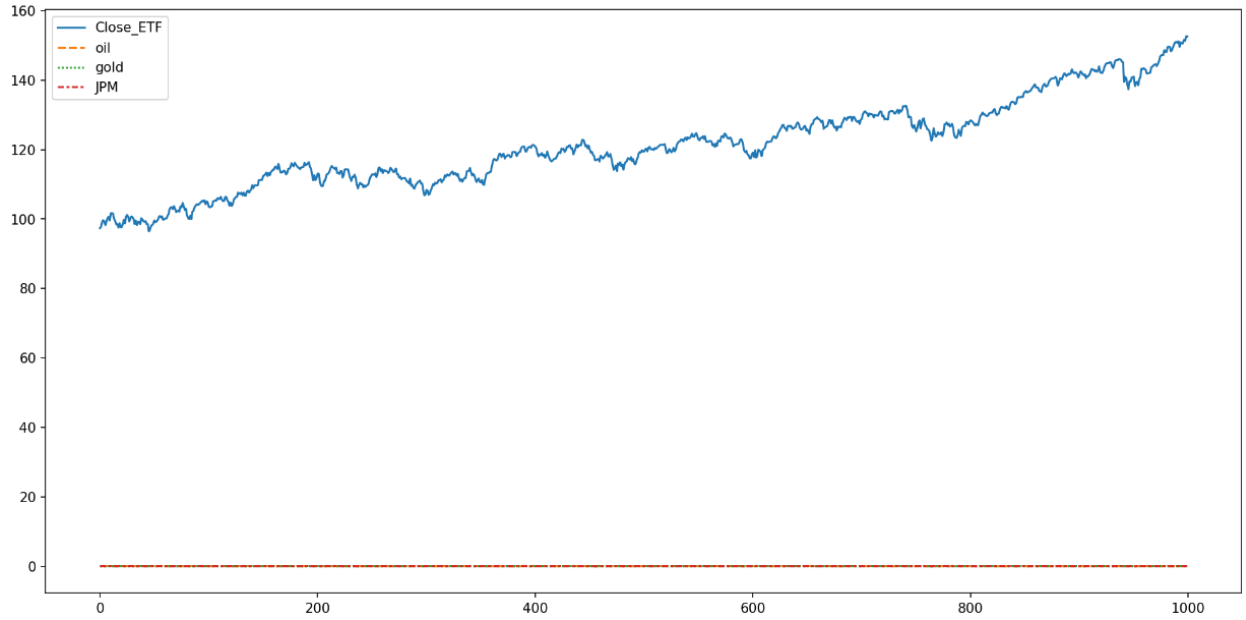


Figure 3 – Time Series Plot of all Random Variables.

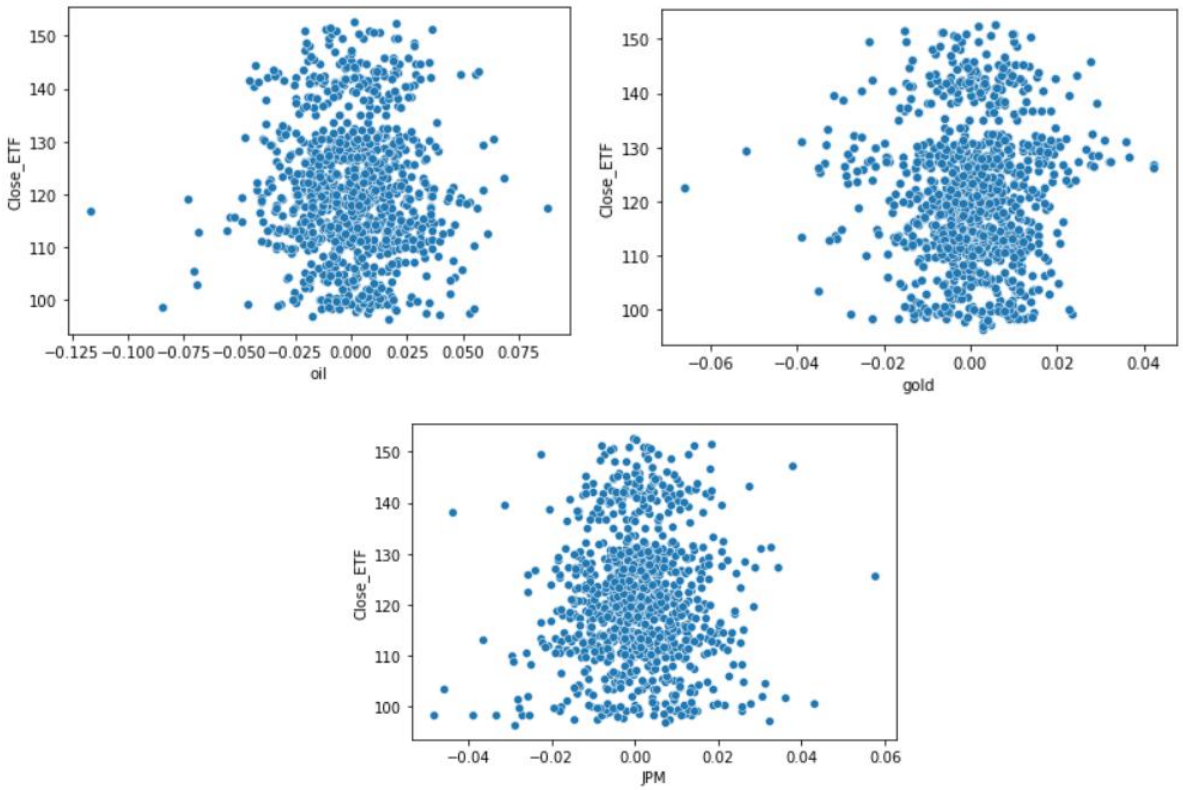


Figure 4 – Scatterplot of gold, oil and JPM in relation to Close ETF

3 – Distribution of Data

I will assume that each of these variables are distributed normally, as we can see in the graphs in Part 2. Although certain variables seem to be better alligned with a Gaussian distribution, we will determine this by using a normality test. In this case we will use a Shapiro-Wilk Test.

> "If the test is non-significant (p>. 05) it tells us that the distribution of the sample is not significantly different from a normal distribution. If, however, the test is significant (p < . 05) then the distribution in question is significantly different from a normal distribution."

Null Hypothesis: The data does not follow a normal distribution.

```
Close_ETF stat = 0.980, p = 0.000

Reject the null hypothesis, assume our distribution is normal
-------------------------------------------------------------------------
oil stat = 0.989, p = 0.000

Reject the null hypothesis, assume our distribution is normal
-------------------------------------------------------------------------
gold stat = 0.969, p = 0.000

Reject the null hypothesis, assume our distribution is normal
-------------------------------------------------------------------------
JPM stat = 0.980, p = 0.000

Reject the null hypothesis, assume our distribution is normal
-------------------------------------------------------------------------
```

Table 3 – Shapiro-Wilk Test Results

Therefore we can assume that all 4 random variable follow a normal distribution.

4 – Central Limit Theorem

In order to prove the Central Limit Theorem, we will explore the distribution of data for ETF. The population mean and standard deviations were found previously and can be found in Table 1. In order to explore the Central Limit Theorem, we will break the data into groups, calculate the mean; and explore the distribution of the mean. The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed.

Here we have 4 samples, 2 that are random samples and 2 that are sequentially grouped. We also varied the sample size in each group, creating one group has 50 samples while the other has 20.

1 – Sequential with 50 samples.

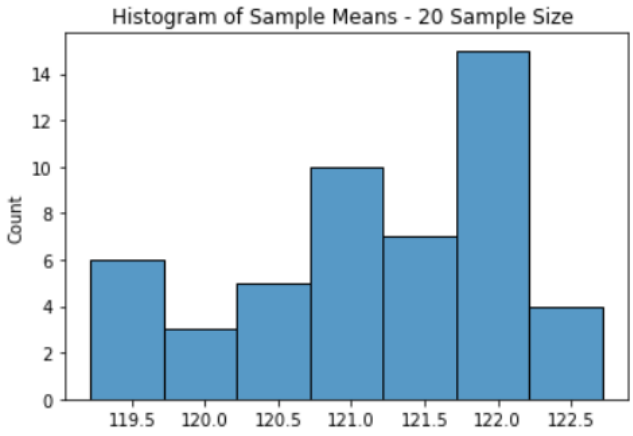
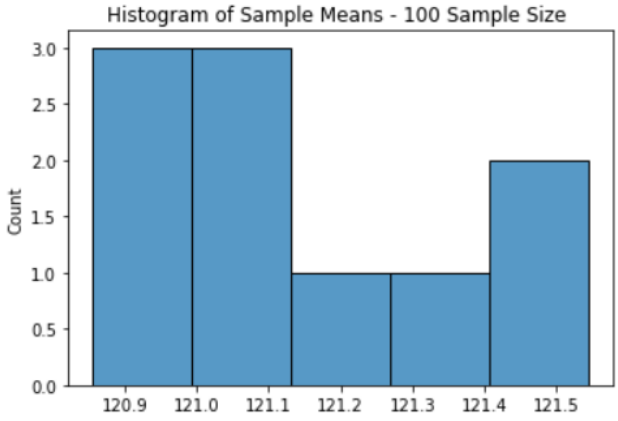


Figure 5 – Histogram of Sample Means – 20 sample size

By calculating the mean and standard distribution we see that the mean has stayed the same (the sample has not changed); and that there is a difference between the sample standard deviation and the standard deviation of the data divided by the square root of the number of sample means. However because we did not take any random samples, our results were not normally distributed; therefore it is not consistent with the central limit theorem.

2 – Sequential with 10 samples.

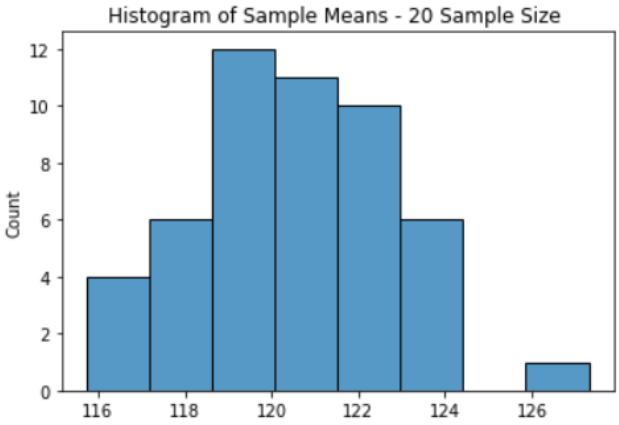



```
Comparison of Means
121.1529600120001
121.15296001199998
-----------------------
Comparison of Standard Deviation
0.2493916967824131
3.9749167100151
```

Figure 6 – Histogram of Sample Means – 10 sample size

Here we can see that the means is still the same, however the standard deviation has decreased for both the population and the sample. This result is still not consistent with the Central Limit Theorem for the same reasons mentioned above: we did not take any random samples therefore our results were not normally distributed.

3 – Random with 50 samples.

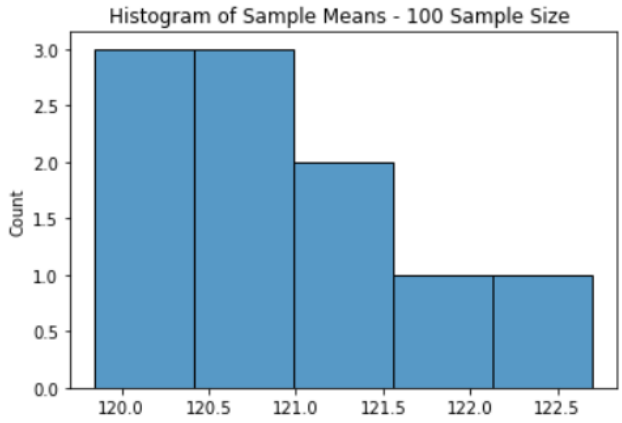



```
Comparison of Means
121.1529600120001
120.55296007199999
-----------------------
Comparison of Standard Deviation
2.36920455369242
1.7776367936987165
```

Figure 7 – Histogram of Sample Means – 50 sample size

Here we can see that the mean is slightly different (less than 1%), however this is expected as there is now replacements. The standard deviation follows the same trend as above, which is expected as we have a larger number of sample sizes (50 vs 10). This result is now consistent with the Central Limit Theorem: we took random samples therefore our results are more normally distributed around the mean.

3 – Random with 10 samples.




```
Comparison of Means
121.1529600120001
120.93559020599997
-----------------------
Comparison of Standard Deviation
0.8868944880482097
3.9749167100151
```

Figure 8 – Histogram of Sample Means – 10 sample size

Here we can see that the mean is slightly different (less than 1%), however this is expected as there is now replacements. The standard deviation also differs more than the example above, which is expected as we have a smaller number of random samples (50 vs 10). This result is also consistent with the Central Limit Theorem; however not as evident as the example above: we took random samples therefore our results are more normally distributed around the mean, however the normal distribution is a lot more evident with higher random samples.

It's important to note that the central limit theorem says that the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal. Therefore the distribution of the population does not have an impact on the distribution of the sample mean(s).

Statistical Results

In this section we will explore various statistical tests, and hypothesis tests in an attempt to find meaningful results from our data. This is also a good way to show the importance of these statistical tests.

5 – Confidence Intervals – for ETF

10 simple random samples with a 95% confidence interval of the mean.

120.2301	125.1681

50 simple random samples with a 95% confidence interval of the mean.

116.9826	128.4156

The mean of the population is 121.1530. In both confidence intervals above, we can see that the intervals include the mean of the population (entire ETF column). Although both confidence intervals are nearly identical, the one that is more accurate is the one with 50 simple random samples generated. This is because there is a larger amount of random sample, meaning that it is not as bias as the 10 simple random samples.

6 – Hypothesis Tests – same dataset

50 sample Random Samples - test H_0: $\mu=100$ vs. H_a: $\mu\neq100$ at the significance level 0.05.

Using a two-sided z-test to determine whether the mean is equal 100. Our sample size is above 30, therefore all conditions of the z-test are met. Below we can see that the p-value is less than 0.05, therefore we can reject the null hypothesis and say that the mean is not 100.

Z Test Findings and P-Value:

5.67713	1.36974e-08

10 sample Random Samples - test H_0: μ=100 vs. H_a: μ≠100 at the significance level 0.05.

Here we will use a t-test to determine if the mean is 100. Our sample size is less than 30, therefore the conditions of the z-test are not met. Below we can see that the p-value is less than 0.05, therefore we can reject the null hypothesis and say that the mean is not 100.

t Test Findings and P-Value:

18.01932	5.0614e-03

7 – Hypothesis Tests – different datasets

Considering the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, we will form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

We will use a t-test because we know that we have two independent variables and we want to compare their mean. The t-test tests the null hypothesis that 2 independent samples have identical average (expected) values.

t Test Findings and P-Value:

-0.4854	0.6275

The p-value is not less than 0.05, therefore we cannot reject the null hypothesis that the 2 independent samples have equal means.

We will then Subtract the entire Gold column from the entire Oil column and generate a sample of differences. Consider this sample as a random sample from the target population of differences between Gold and Oil. Form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

8 – Linear Regression Model

We will now consider the data in the ETF column and gold column to create a linear regression model. We explore this linear relationship by drawing a scatter plot as well as finding the Pearson coefficient – 0.022996. This is not the best score; which proves that the linear relationship is not statistically significant. We can further explore this relationship in Figure 9. The slope indicates the steepness of a line and the intercept indicates the location where it intersects an axis.

Slope – `25.6043`	**Intercept -** `121.1360`

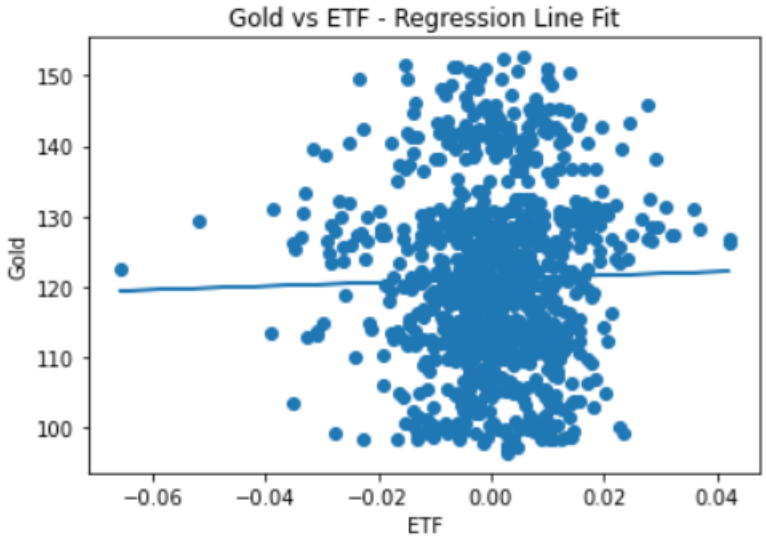


Figure 9 – Scatter Plot of ETF vs Gold

In this case study the coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0. This is not a good model because although it may show a linear relationship, it fails to show that the data is not homoscedastic.

The assumptions for this model working are there is There is a linear regression relation between Y and X, the error terms (residuals) are normally distributed, the variance of the error terms is constant over all X values (homoscedasticity), the error terms are independent.

Analysis

9 – Linear Regression Prediction

We will now consider data including the ETF, Gold and Oil column. Here we fit a multiple linear regression model to the data with the ETF variable as the response. We used our independent variables to train the data, and split the data into a testing and training set. The training data was used to create the regression model, with the testing dataset used to make the prediction.

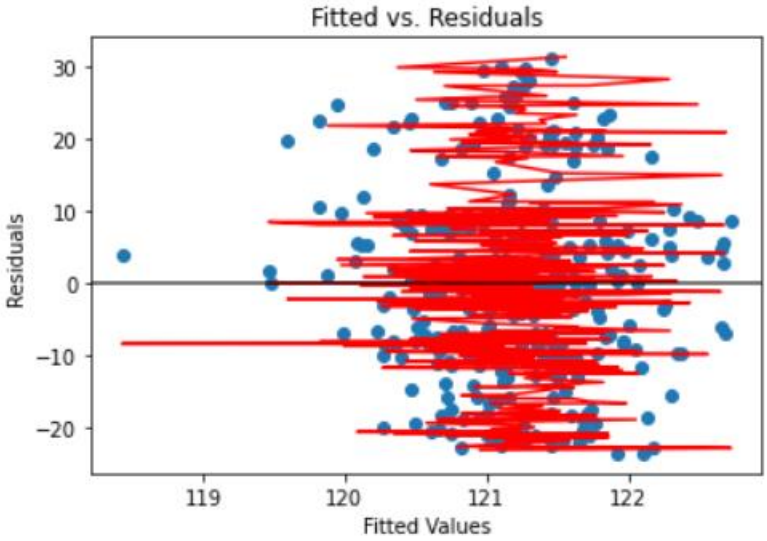
The adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input. Therefore we can see that the ETF is not a good indicator of the variance in the other variables.

Adjusted R2 score – `-0.0020718`

10 – Analyzing Model Performance

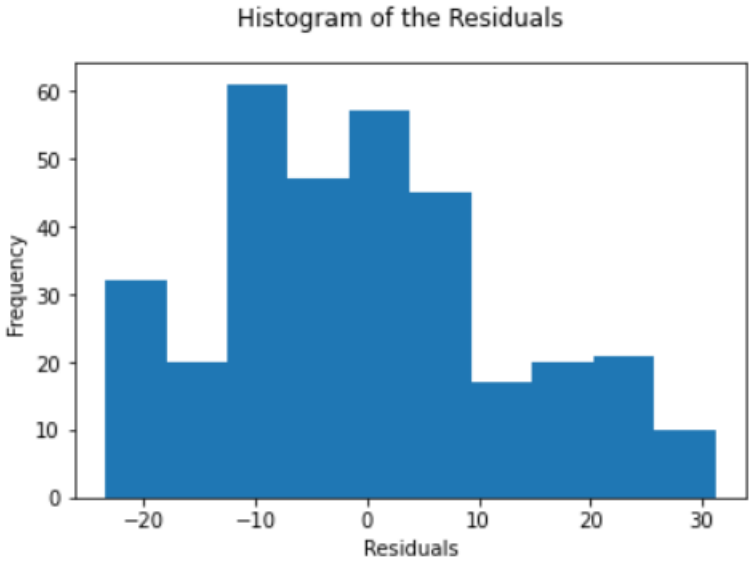
We will explore whether a linear regression model was a good fit. We will do this by exploring the conditions required for a Linear Regression Model. These conclusion can be drawn by exploring the figure below.

Linear Relationship / Constant Variance

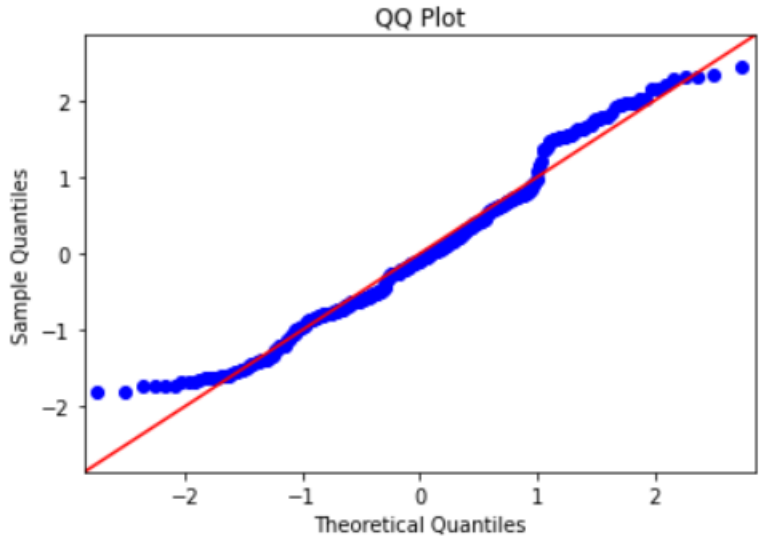


The linear relation is assumed to be satisfied if there are no apparent trends in the plot. The above plot satisfies this assumption. This plot also shows that the mean of the residuals is around 0. The constant variance assumption is assumed to be satisfied if the vertical spread of the residuals remains roughly consistent across all fitted values, and as we can see this is the case. The linear relation is assumed to be satisfied if there are no apparent trends in the plot.

Normality Test



Here we can see that the residuals are normally distributed.



In a Q-Q plot if the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x; which is the case above. Therefore we can assume that these the error is normally distributed.

Error Terms Are Independent

If there is a trend in the residual values over time, then this assumption is violated. In the plot above we can see that this assumption has not been violated.

In Conclusion, we can assume this linear regression model to be a good use, since all conditions for this test were met. However, this does not mean that the model would be a good predictor, as proven by the adjusted r square test. This is why it's important to explore various statistical tests to determine which is best; and see which assumptions are necessary to use it.

Appendix

Part 3 -

```
# Import Shapiro-Wilk test form statisitical package

from scipy.stats import shapiro
```

```
cols = ['Close_ETF', 'oil', 'gold', 'JPM']
```

```
for col in cols:
    stat, p = shapiro(df[col])
    print(col, 'stat = %.3f, p = %.3f\n' % (stat,p))
    if p > 0.05:
        print('Fail to reject the null hypothesis, assume our distribution is not normal')
        print('--------------------------------------------------------------------------')
    else:
        print('Reject the null hypothesis, assume our distribution is normal')
        print('--------------------------------------------------------------------------')
```

Part 4 –

10. Generate 10 simple random samples or groups (with replacement) from the population. The size of each sample is 100, i.e., each group includes 100 values.

```
choices2 = np.random.choice(etf, size=1000, replace=True)
```

```
x3 = np.array_split(choices2,100)
```

```
# Sample means and histogram of sample means

s_means3 = [sum(elem)/len(elem) for elem in zip(*x3)]
_ = sns.histplot(data=s_means3).set(title='Histogram of Sample Means - 100 Sample Size ')
```

```
etf_mn_2 = stat.mean(s_means3)
print('Comparison of Means')
print(etf_mn)
print(etf_mn_2)

print('-----------------------')

etf_std_2 = stat.stdev(s_means3)
print('Comparison of Standard Deviation')
print(etf_std_2)
print(etf_std / np.sqrt(10))
```

Part 5 –

```
import scipy.stats as st
```

1. Pick up one of the 10 simple random samples you generated in Step 10) of Part 4, construct an appropriate 95% confidence interval of the mean μ.

```
# Selecting the first simple random sample

data = [elem for elem in zip(*x3)][0]
```

```
# Finding the 95% confidence interval using the normal distribution

st.norm.interval(alpha=0.95, loc=np.mean(data), scale=st.sem(data))
```

```
(120.23011534341776, 125.16808463658222)
```

2. Pick up one of the 50 simple random samples you generated in Step 8) of Part 4, construct an appropriate 95% confidence interval of the mean μ.

```
# Selecting the first simple random sample

data1 = [elem for elem in zip(*x2)][0]
```

```
# Finding the 95% confidence interval using the normal distribution

st.norm.interval(alpha=0.95, loc=np.mean(data), scale=st.sem(data1))
```

```
(116.9826391352873, 128.4155608447127)
```

3. In Part 1, you have calculated the mean μ of the population (the entire ETF column) using Excel function. Do the two intervals from 1) and 2) above include (the true value of) the mean μ? Which one is more accurate? Why?

```
print('The mean of the population is: ', etf.mean())
```

```
The mean of the population is:  121.1529600120001
```

Part 6 –

```
from statsmodels.stats.weightstats import ztest
from scipy import stats
```

1. Use the same sample you picked up in Step 1) of Part 5 to test H_0: μ=100 vs. H_a: μ≠100 at the significance level 0.05. What's your conclusion?

Using a two-sided z-test to determine whether the mean is equal 100. Our sample size is above 30, therefore all conditions of the z-test are met.

```
ztest(data1, value=100, alternative='two-sided')
```

```
(5.677128968918836, 1.369742029016157e-08)
```

- Above we can see that the p-value is less than 0.05, therefore we can reject the null hypothesis and say that the mean is not 100.

2. Use the same sample you picked up in Step 2) of Part 5 to test H_0: μ=100 vs. H_a: μ≠100 at the significance level 0.05. What's your conclusion?

Here we will use a t-test to determine if the mean is 100. Our sample size is less than 30, therefore the conditions of the z-test are not met.

```
stats.ttest_1samp(data, popmean=100)
```

```
Ttest_1sampResult(statistic=18.019317586060875, pvalue=5.0614382285231476e-33)
```

- Above we can seme that the p-value is less than 0.05, therefore we can reject the null hypothesis and say that the mean is not 100.

3. Use the same sample you picked up in Step 1) of Part 5 to test H_0: σ=15 vs. H_a: σ≠15 at the significance level 0.05. What's your conclusion?

```
stats.chisquare(data, 15)
```

```
Power_divergenceResult(statistic=78374.63997436938, pvalue=0.0)
```

Part 7-

```
gold = df.gold
oil = df.oil
```

1. Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

We will use a t-test because we know that we have two independant variables and we want to compare their mean. The t-test tests the null hypothesis that 2 independent samples have identical average (expected) values.

```
stats.ttest_ind(gold, oil)
```

```
Ttest_indResult(statistic=-0.48536661382360874, pvalue=0.6274695258306375)
```

- The p-value is not less than 0.05, therefore we cannot reject the null hypothesis that the 2 independant samples have equal means.

2. Subtract the entire Gold column from the entire Oil column and generate a sample of differences. Consider this sample as a random sample from the target population of differences between Gold and Oil. Form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

```
# Generating a sample of differences

dif = oil - gold
```

```
ztest(dif, value=0)
```

```
(0.5413309278514735, 0.5882795066915598)
```

- The p-value is not less than 0.05, therefore we cannot reject the null hypothesis that both oil and gold have a difference of 0

Part 8 –

```
#find line of best fit
a, b = np.polyfit(gold, etf, 1)

#add points to plot
plt.scatter(gold, etf)

#add line of best fit to plot
_ = plt.plot(gold, a*gold+b)
plt.xlabel('ETF')
plt.ylabel('Gold')
plt.title('Gold vs ETF - Regression Line Fit')
```

```
print('The slope of the line is - ', a)
print('The intercept of the line is - ', b)
```

Part 9 –

Multiple Linear Regression Model

```
# Labeling the X and y variables

y = df.Close_ETF
X = df.drop(['Close_ETF'], axis=1)
```

```
model = LinearRegression()
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
model.fit(X_train, y_train)
```

```
LinearRegression()
```

```
# Adjusted R-Squared

1 - (1-model.score(X_train, y_train))*(len(y_train)-1)/(len(y_train)-X_train.shape[1]-1)
```

```
-0.0020718454220578497
```

Part 10 –

```
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

```
# Setting up predicted Y value

y_pred = model.predict(X_test)
```

```
# Caluclulating the residuals

residuals = y_test - y_pred
```

1. Linear Relationship / Constant Variance

```
# Get the smoothed lowess line
lowess = sm.nonparametric.lowess
lowess_values = pd.Series(lowess(residuals, y_test)[:,1])
```

```
# Plot the fitted v residuals graph
plt.scatter(y_pred, residuals)
plt.plot(y_pred, lowess_values, c='r')
plt.axhline(y=0, c='black', alpha=.75)
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.title('Fitted vs. Residuals')
plt.show()
```

2. Normality Test

```
fig = plt.figure()
plt.hist(residuals)
fig.suptitle('Histogram of the Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()
```

```
sm.qqplot(residuals, line='45', fit=True)
plt.title('QQ Plot')
plt.show()
```

4. The Error Terms are Independent

```
l = list(range(1,331))
```

```
plt.scatter(l, residuals)
plt.axhline(y=0, c='black', alpha=.75)
plt.xlabel('Order')
plt.ylabel('Residuals')
plt.title('Order vs. Residuals - Tests for Independent Error Terms')
plt.show()
```