MA-541 Financial Project Report

Sarah Thuman

Abstract

Predicting how stocks will perform is a field of study that has been heavily researched, and a wide range of techniques have been employed and developed in this pursuit. Because the stock is complex and can fluctuate widely from day to day, there is no one method that works perfectly, which is why research continues. In the paper, regression analysis is used to model the behaviors of four stocks, ETF, gold, crude oil, and J.P. Morgan Chase & Co., with the goal of predicting the daily relative change in price of ETF stock based on the other three. The data used in the study comes from a 2.75-year period and was pre-processed before use.

Introduction

The goal of this analysis is to determine if the relative daily change in price of one stock can be predicted using the relative daily change in price of other stocks. The daily relative change in price of four stocks, ETF, crude oil, gold, and J.P. Morgan Chase & Co. stock (JPM), were analyzed to check the hypothesis that the relative daily change in price of ETF could be predicted using the data from crude oil and gold. Along with the focus of this research, general statistics and hypothesis tests relating to how the means and standard deviations of the stocks compared to one another were performed to assess whether the stocks used are a good choice for the prediction model. The data contained 1000 data points for each variable, which is about 2.75 years of data that was studied. Since the data was provided and not collected, the actual dates the data was collected from is unknown. Previous analyses like this have used representative factors [1] instead of different stock for regression analysis and others have used different modeling methods such as time-series forecast models [2].

Analysis

Basic Statistics

The first step in analyzing the data was normalizing the ETF data. The data collected for ETF was the daily returns. To normalize the data, it was converted from daily returns to relative daily change in price. Due to not having the data for the day preceding the first entry, the first entry for the relative daily change in ETF price was assumed to be zero. Next the mean and standard deviation of each variable was calculated, and a correlation matrix was created which are summarized in Table 1 and Table 2 respectively. As can be seen in Table 1, the means of all the variables are close to zero with small standard deviations. These observations were used in hypothesis testing which will be discussed later in this report. A quick examination of the correlation matrix in Table 2 shows that the only variables that are strongly correlated are ETF and JPM.

	Mean	Standard Deviation
ETF	0.00047	0.00692
Crude Oil	0.00103	0.02109
Gold	0.00066	0.01129
JPM	0.00053	0.01102

Table 2: Summary Statistics

	ETF	Crude Oil	Gold	JPM
ETF	1	-0.0712	0.0897	0.7028
Crude Oil		1	0.2357	-0.1208
Gold			1	0.1002
JPM				1

Table 1: Correlation Matrix

To further develop an understanding of the data, histograms and time series plots were made. An example of each is provided in Figure 1 and Figure 2 respectively. The remaining histograms and time series plots can be found in Appendix A. A cursory examination of the histograms implies that all the variables are normally distributed. This assumption is tested and will be discussed later in this report. The time series plots show the maximum and minimum values of each variable. Since the variables were normalized all the time series plots are centered at about zero.

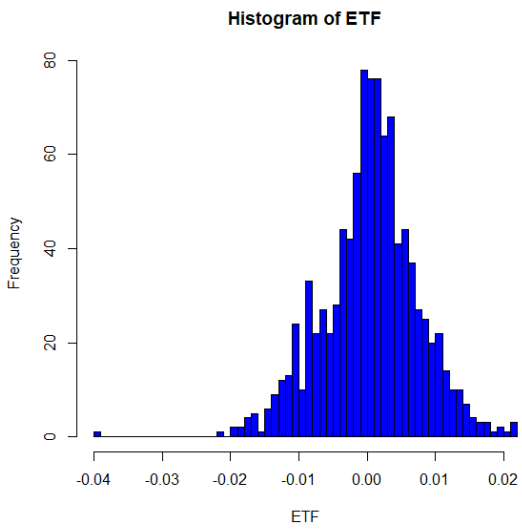

Figure 1: Histogram of ETF

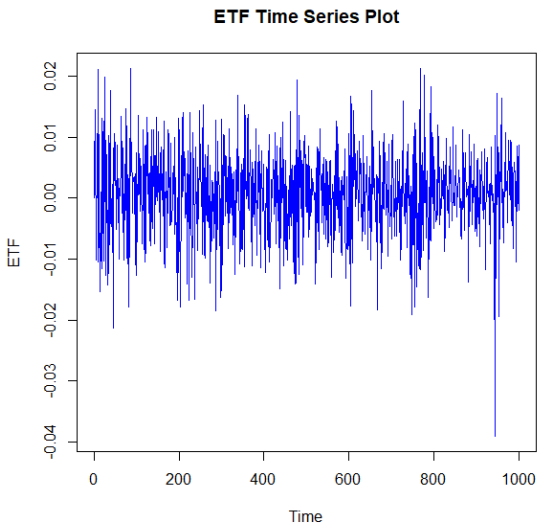

Figure 2: Time Series Plot of ETF

Scatter plots were also created and show the correlation between ETF and the other variables. The relationship between ETF and the remaining variables was the only one looked at graphically with a scatter plot since later it will be used at the dependent variable in modeling. All scatter plots can be found in Appendix A.

Hypothesis Tests

Several hypothesis tests were conducted on the data. The first one was to test if the variables were normally distributed. Hypothesis 1 was H_0: $variable\ is\ normally\ distributed$ and H_A: $variable\ is\ not\ normally\ distributed$ were applied to each variable. The Shapiro-Wilk Normality test was used to assess the hypotheses. All the tests had p-values that indicated statistical significance and therefore the null hypotheses were rejected. Because the hypotheses that the variables are normally distributed were rejected, a central limit theorem test was performed to check if it could be assumed that the distribution of the sample means approximates the normal distribution given a large enough sample size. To check this, four sampling methods were used: 1) split data into 50 groups of 20 entries, 2) split data into 10 groups of 100 entries, 3)

randomly sample 50 groups of 20 entries with replacement, and 4) randomly sample 10 groups of 100 entries with replacement. After doing this, histograms were created for each new group in the four categories. A 95% confidence interval was constructed from a random sample from groups 3 and from group 4 to test how well the random sampling was at capturing the population mean.

For the next few hypotheses, the data from groups 3 and was split 4 80/20 into training and test sets. Hypotheses 2 and 3 were $H_0: \mu = \mu_0$ and $H_A: \mu \neq \mu_0$. Hypothesis 2 used a random sample from method 3 and Hypothesis 3 used a random sample from method 4. Because there were greater than 30 values, a two-tailed z-test was used to test both training sets. Hypothesis 4 was $H_0: \sigma = \sigma_0$ and $H_A: \sigma \neq \sigma_0$. The F-test was used to check the hypothesis and used the Method 3 training set. The null hypothesis for Hypothesis 5 was the same as Hypothesis 4 but instead the alternative was $H_A: \sigma < \sigma_0$.

For the next group of hypotheses, the gold and crude oil columns are treated as random samples from two separate populations. Based on this assumption, two sample hypothesis tests were conducted. Hypothesis 6 was $H_0: \mu_{gold} - \mu_{oil} = 0$ and $H_A: \mu_{gold} - \mu_{oil} \neq 0$. A two-sided z-test was used based on the sample size being greater than 30. To obtain the sample populations a sample size of 100 was taken from both populations. Hypothesis 7 was $H_0: \mu_{gold-oil} = 0$ and $H_A: \mu_{gold-oil} \neq 0$. Again, a two-sided z-test was used. To obtain the sample data, first the oil column was subtracted from the gold column and then a random sample of 100 was taken from the new population. The last hypothesis, Hypothesis 8, tested was $H_0: \sigma_{gold} - \sigma_{oil} = 0$ and $H_A: \sigma_{gold} - \sigma_{oil} \neq 0$. A F-test was used to test the hypothesis and a sample size of 100 was taken from both populations to use as a sample population.

For all hypothesis tests, a 0.05 significance level was used to evaluate the tests.

Modeling

Two models were used to fit the data, looking to determine if the relative daily change in ETF price could be predicted based on the relative daily change in price of the other three stocks. The first model was a linear regression model where the relative daily change in ETF price was treated as the response variable and the relative daily change in gold price was treated as the explanatory variable. The coefficient of correlation between the ETF and gold variables was 0.090. The linear regression equation produced by the model was: $\hat{y} = 0.00055 + 0.040x_{gold}$. A couple of hypothesis tests were conducted based on the model. One to test $H_0: \beta_1 = 0$ and $H_0: \beta_1 \neq 0$. A two-tailed test was used to test the hypothesis. The other hypothesis test conduct on the linear regression model was $H_0: \mu_{gold} - \mu_{ETF} = 0.005127$ and $H_A: \mu_{gold} - \mu_{ETF} \neq 0.005127$, where it was assumed that the daily relative change in gold price was 0.005127. A 99% confidence interval was calculated to be (-0.000884, 0.001587). A 995 Prediction interval was also calculated and was (-0.0170844, 0.0185215).

The other model used to fit the data was a multiple linear regression model where ETF was the response variable and gold, and crude oil were the explanatory variables. The data was split 80/20 into a training set and test set. The training set was used to compute the multiple linear regression model and the equation was $\hat{y} = 0.00055 + 0.0576x_{gold} - 0.0388x_{oil}$. The multiple linear regression model was then used to predict the daily relative change in ETF price using the test set. The residuals were then calculated and were tested to check that they were normally distributed, have a mean equal to zero, homoscedastic, and independent. A Shapiro-Wilk Normality test, z-test, Breusch-Pagan Test, and Durbin Watson Test were used to test the residual assumptions.

Results

Hypothesis Tests

As mentioned in the analysis section, histograms were created to determine the distribution of the four variables. While the histograms all appeared to be normally distributed, a Shapiro-Wilk Normality test was conducted on every variable to confirm. The Shapiro-Wilk Normality test was conducted assuming that H_0: $variable\ is\ normally\ distributed$ (Hypothesis 1). Based on the p-values, which are summarized in Table 3, the null hypothesis was rejected for all variables, indicating that none of the variables are normally distributed. The results from the Central Limit Theorem tests (sampling the population data by either splitting the population into 10 or 50 groups or sampling with replacement into 10 or 50 groups), showed that sampling the total population multiple times results in a normal distribution. A 95% confidence interval was calculated for a random sample from each of the two random samplings with replacement methods and showed that the sampled data was successful in capturing the true mean of the population. As expected, the method that had larger group sizes was better at predicting the true mean of the population. The population mean for ETF was 0.00047. The 95% confidence interval for the simple random sample with ten groups was (-0.00102, 0.00153) and the 95% confidence interval for the simple random sample with 50 groups was (-0.00346, 0.00178). As can be seen, both intervals capture the population mean.

	p-value
ETF	2.153 e-07
Crude Oil	5.487 e-07
Gold	1.02 e-13
JPM	1.538 e-10

Table 3: Shapiro-Wilk test results

The results from hypotheses 2-5 will be discussed next. For this group of hypotheses, both the simple random sample with 10 and 50 groups was split 80/20 into a training and test set. All tests were conducted on the ETF column and used a significance level of 0.05. For

Hypothesis 2 (H_0: $\mu = \mu_0$), a two-tailed z-test was performed on a random sample from the training set for the simple random sample with ten groups. The resulting p-value was 0.84 which meant the null hypothesis was not rejected and that the population mean is equal to the sample mean. For Hypothesis 3 (H_0: $\mu = \mu_0$), a two-tailed z-test was performed on a random sample from the training set for the simple random sample with 50 groups. The p-value was 0.65 which meant the null hypothesis was not rejected and that the population mean is equal to the sample mean. Hypothesis 4 (H_0: $\sigma = \sigma_0$) was tested using a two-sided f-test on the simple random sample with 50 groups training set. The p-value was 0.83, meaning there is not enough evidence to reject the null hypothesis that the population standard deviation is less than the sample standard deviation. The last hypothesis in this group, Hypothesis 5 (H_0: $\sigma = \sigma_0$; H_A: $\sigma < \sigma_0$), was tested using a one-sided f-test on the simple random sample with 50 groups training set. The p-value was 0.41, meaning there is not enough evidence to reject the null hypothesis that the population standard deviation is less than the sample standard deviation.

For hypotheses 6-8, the gold and crude oil columns are treated as random samples from two separate populations. A significance level of 0.05 was used to evaluate all hypotheses. Hypothesis 6 (H_0: $\mu_{gold} - \mu_{oil} = 0$) was tested using a two-sided z-test. A sample of 100 entries was taken from the gold and crude oil columns to test the hypothesis. The p-value was 0.48, meaning there is not enough evidence to reject the null hypothesis that the difference between the mean of the gold column and the mean of crude oil column was equal to 0. For hypothesis 7 (H_0: $\mu_{gold-oil} = 0$), first the entries in the crude oil column were subtracted from the corresponding entries in the gold column. Next, a sample of 100 entries with replacement was taken from the difference of the gold and crude oil columns that was calculated. This sample was used in a two-sided z-test to check the hypothesis. The p-value was 0.41, meaning the null

hypothesis that the mean of the difference between the gold and crude oil columns was not rejected. The last hypothesis tested was Hypothesis 8 (H_0: $\sigma_{gold} - \sigma_{oil} = 0$). Samples of 100 entries were taken from both the gold column and crude oil column. A two-sided f-test was used to evaluate the hypothesis. The p-value was 1.41 e-09 which is smaller than the significance value. Therefor the null hypothesis that the difference between the standard deviation of the gold sample and the standard deviation from the crude oil sample are not equal to 0. A summary of all hypotheses can be found in Table 4.

Hypothesis	H_0	H_A	Test Used	p-value	Reject?
2	$H_0: \mu = \mu_0$	$H_A: \mu \neq \mu_0$	z-test	0.84	No
3	$H_0: \mu = \mu_0$	$H_A: \mu \neq \mu_0$	z-test	0.65	No
4	$H_0: \sigma = \sigma_0$	$H_A: \sigma \neq \sigma_0$	f-test	0.83	No
5	$H_0: \sigma = \sigma_0$	$H_A: \sigma < \sigma_0$	f-test	0.41	No
6	$H_0: \mu_{gold} - \mu_{oil} = 0$	$H_A: \mu_{gold} - \mu_{oil} \neq 0$	z-test	0.48	No
7	$H_0: \mu_{gold-oil} = 0$	$H_A: \mu_{gold-oil} \neq 0$	z-test	0.41	No
8	$H_0: \sigma_{gold} - \sigma_{oil} = 0$	$H_A: \sigma_{gold} - \sigma_{oil} \neq 0$	f-test	1.41e-09	Yes

Table 4: Summary of Hypothesis Tests

Modeling

As mentioned in the analysis section, two models were created using linear regression and multiple linear regression to find a best fit line through the data. The linear regression model used a training set of the ETF and gold columns and was used to see how well the daily relative change in ETF price could be predicted from the daily relative change in gold price. The equation of the linear regression line was $\hat{y} = 0.00055 + 0.040x_{gold}$. A scatter plot of the variables with the equation line is shown in Figure 3. The slope of the equation shows that there is a slight positive correlation between the daily relative price change in ETF and gold stock. A two-tailed t-test was used to test the hypothesis that $\beta_1 = 0$ to determine if the positive

correlation was significant or not. A significance level of 0.01 was used to determine if the null hypothesis should be rejected. The resulting p-value of the t-test was 0.0542 and was greater than the significance level of 0.01. This means that there is not enough evidence to reject the null hypothesis that the slope of the linear regression line is zero. This indicates that there is no relationship between the daily relative change in price of ETF and gold. Further delving into how good a fit this model is for the data, the R^2 value of the regression model was assessed. With $R^2 = 0.0047$, the conclusion is that this model is not a good fit for the data. Given that it was determined there is no relationship between the daily relative change in price of ETF and gold, it makes sense that this model would not work. A 99% confidence interval was calculated. The predicted y value was 0.00075 assuming that the mean of the daily relative change in gold price was 0.005127. The confidence interval was (0.00013, 0.00138). A 99% prediction interval was also calculated assuming that the mean of the daily relative change in gold price was 0.005127. The 99% prediction interval was (0.0007513178, 0.0007515335). Some of the assumptions that were made when testing this model were that there was a linear relationship between the two variables, there were no outliers in the data set, and that the variables were normally distributed.

Figure 3: Scatter plot with linear regression line.

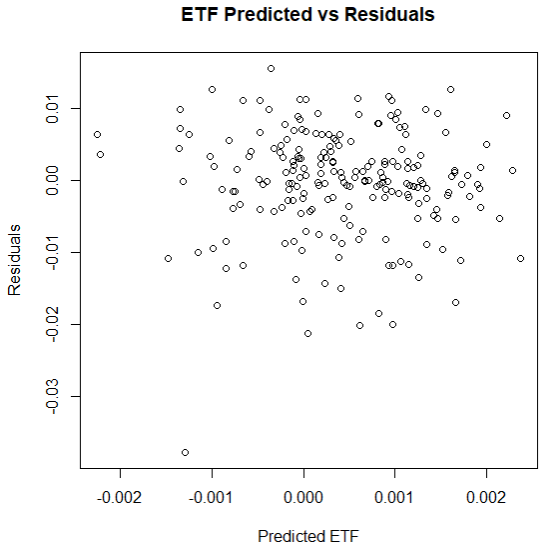

Figure 4: Scatter plot of residuals

The second model was created using multiple linear regression. This time in addition to gold, the daily relative change in crude oil price was used to build the model. The multiple linear regression equation was $\hat{y} = 0.00055 + 0.0576x_{gold} - 0.0388x_{oil}$. The R^2 value for this model was 0.01871 was an improvement from the linear regression model. The R^2 value of the multiple linear regression model still indicated that the model was not a good fit for the data, however because it was an improvement from the linear regression model, the multiple linear regression model was used as the prediction model. The test set was used with the multiple linear regression model to predict the daily relative change in ETF price. Residuals were then calculated to determine how accurate the predictions were. A plot of the residuals is shown in Figure 4. Four assumptions on the residuals were checked. The $\mathrm{E}(residuals) = 0$ was tested using a two-sided z-test and a significance level of 0.05. The p-value was 0.49 meaning there was not enough evidence to reject the null hypothesis that the expectation of the residuals = 0. The next assumption that was checked was that the residuals have constant variance. This was tested using the Breusch-Pagan Test. The p-value from the Breusch-Pagan Test was 0.48, so the null hypothesis that the variance of the residuals is constant was accepted. The next assumption checked was that the residuals were normally distributed. The Shapiro-Wilk Normality test was used and resulted in a p-value of 3.613 e-07 meaning the null hypothesis was rejected and that the residuals are not normally distributed. The last assumption checked was that the residuals were statistically independent. A Durbin Watson

Test was used. The p-value was 0.528 meaning the null hypothesis that the residuals were statistically independent was accepted.

Conclusion

In this experiment, two models were built to predict the daily relative change in ETF price based on three other stocks. While neither of the models were a good fit for the data, the multiple linear regression model was chosen since it outperformed the linear regression model. Multiple hypotheses were tested to determine the relationship between the four stocks. One conclusion that can be drawn from the hypothesis tests and the predictive power of the models is that these four stocks are not good predictors for one another. This could be because the stocks are from different sectors and have little impact on one another. For future experiments, stocks from the same or similar sectors could analyzed to determine if there is a relationship between the daily relative change in price of the stock and if they can be used to predict how one of the stocks will perform based on the others. Adding more stocks to the analysis could also help mitigate some of the issues and produce a better model. It is also possible that selecting different model methods could produce better results.

References

[1] Chen, S. (2020). Forecasting Daily Stock Market Return with Multiple Linear Regression. https://digitalcommons.latech.edu/cgi/viewcontent.cgi?article=1015&context=mathematics-senior-capstone-papers

[2] Tsai, M.-C., Cheng, C.-H., Tsai, M.-I., & Shiu, H.-Y. (2018). Forecasting leading industry stock prices based on a hybrid time-series forecast model. PLOS ONE, 13(12), e0209922. https://doi.org/10.1371/journal.pone.0209922

Appendices

Appendix A

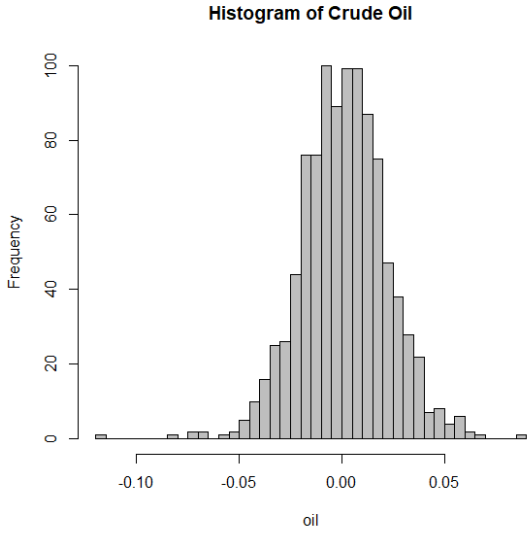

Figure 5: Crude Oil Histogram

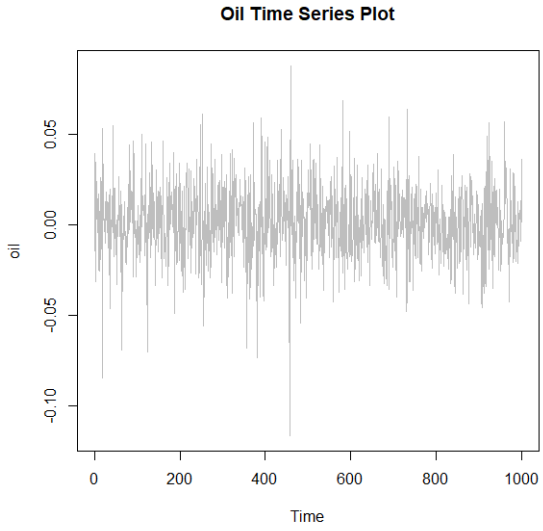

Figure 6: Crude Oil Time Series Plot

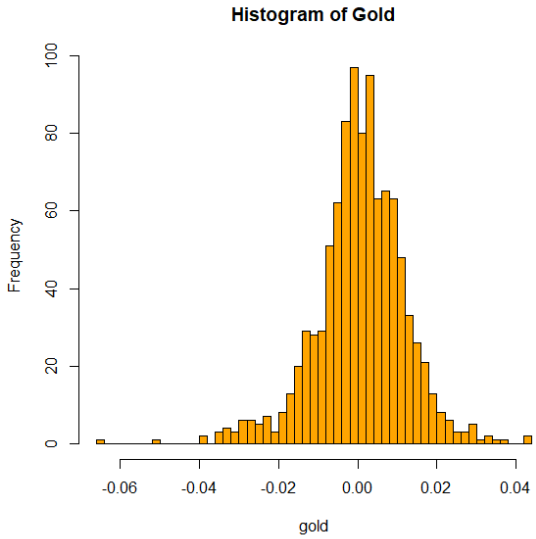

Figure 7: Gold Histogram

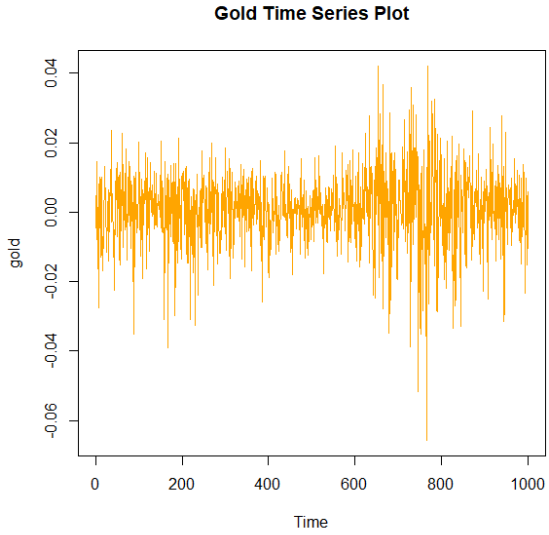

Figure 8: Gold Time Series Plot

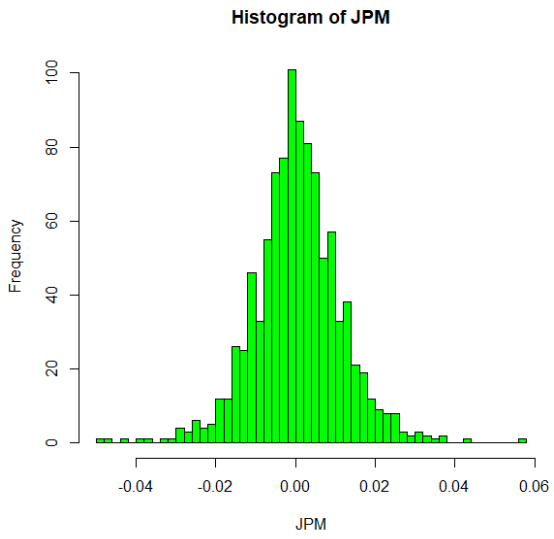

Figure 9: JPM Histogram

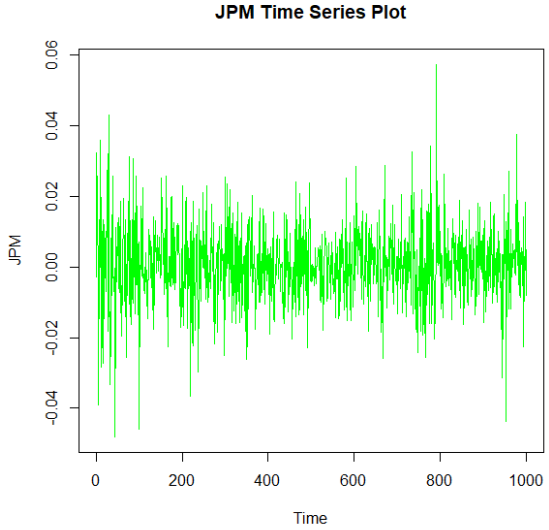

Figure 10: JPM Time Series Plot

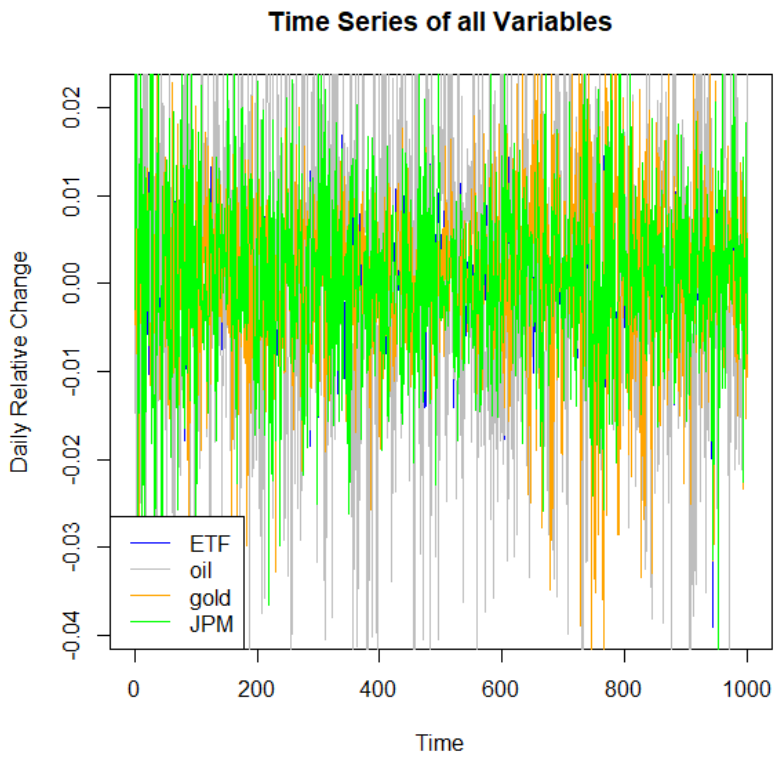

Figure 11: Time Series Plot All Stocks

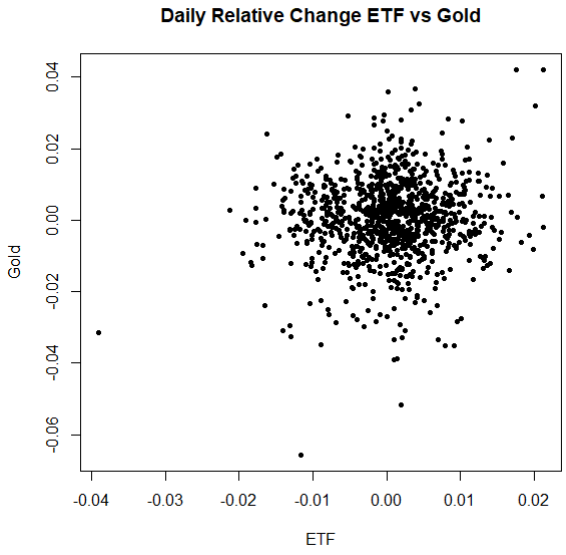

Figure 12: ETF vs Gold Scatter Plot

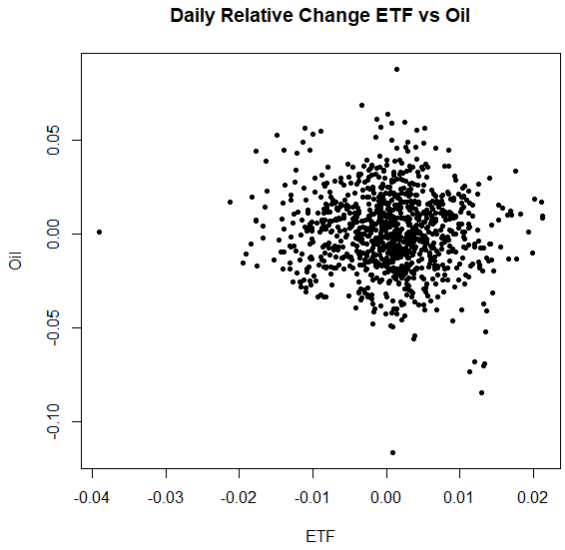

Figure 13: ETF vs Crude Oil Scatter Plot

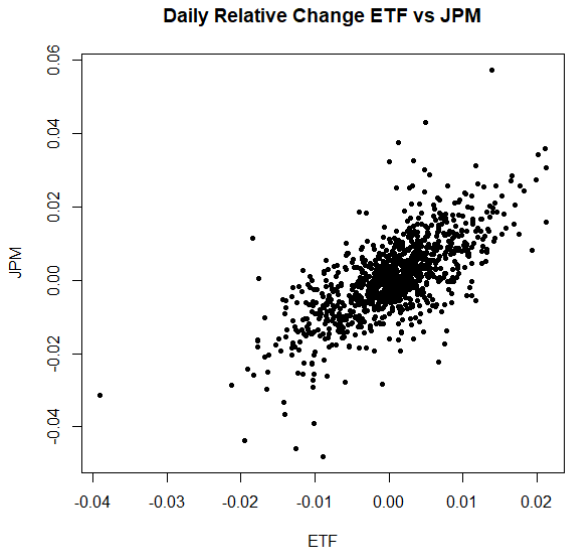

Figure 14: ETF vs JPM Scatter Plot