Project Report

Ander B. Ibarrondo

Fall 2025

TABLE OF CONTENTS

1 Introduction

Semiconductor companies are critical to the functioning of the modern global economy, from personal computers to the "cloud" and advanced manufacturing; they are embedded in the fabric of global supply chains and therefore have been shown to be very responsive to changes in the global economy, geopolitics and the cost of producing goods. Semiconductor companies therefore need to understand how their stocks will respond to changes in the global economy and the associated risks and opportunities that will arise as a result of those changes.

The SOXX ETF is specifically designed to track the performance of the semiconductor industry, providing exposure to 9 semiconductor stock portfolios; however, due to its limited scope compared to more broad-based indices (such as the NASDAQ), it has been less researched, allowing for further understanding of the relationship between the macroeconomic environment and the semiconductor sector.

We can therefore use SOXX to study the behavior of the semiconductor sector in response to varying macroeconomic conditions. The sudden and widespread impact of the COVID-19 pandemic on supply chains and financial markets creates a unique opportunity to assess whether there were structural changes in the behavior of SOXX returns. Due to the rapid nature of the shock and its widespread reach, it is reasonable to assume that either volatility or sensitivity to economic factors changed during and subsequent to the pandemic.

Therefore, our research questions are organized into two primary areas:

1) Did the COVID-19 shock cause a statistically significant increase in the volatility of weekly SOXX returns?

To answer this question, we will calculate the variance of SOXX returns before and after the pandemic using robust methods that are appropriate for financial time series that exhibit heavy tails.

2) Which macroeconomic and sector-specific variables had a statistically significant effect on the performance of SOXX in the post-pandemic era?

Our goal is not to develop a predictive model of future performance, but to identify the key factors influencing the behavior of semiconductor equities currently.

2 Data Description

This study uses a collection of financial and macroeconomic time series obtained from two primary sources: Yahoo Finance and the Federal Reserve Economic Data (FRED). Throughout this section, we denote by n the number of daily observations and by v the number of variables in each dataset. For each series, we also report the number of missing values (NA count). Table 1 summarizes all variables, their tickers, a brief description, data provenance, and structural characteristics.[1]

[1]Note that the VIX variable will be used as a categorical variable in the future; see Section 5

Ticker	Description	n	v	NA	Source
SOXX	iShares Semiconductor ETF	2514	6	0	Yahoo Finance
^VXN	CBOE Nasdaq Volatility Index	2514	6	0	Yahoo Finance
REMX	Rare Earth/Strategic Metals ETF	2514	6	0	Yahoo Finance
^NYICDX	ICE U.S. Dollar Index	2514	6	0	Yahoo Finance
CL=F	Crude Oil WTI Futures	2519	6	48	Yahoo Finance
CNY=X	Chinese Yuan / USD Exchange Rate	2610	6	42	Yahoo Finance
HACK	Cybersecurity ETF	2514	6	0	Yahoo Finance
DGS10	10-Year U.S. Treasury Constant Maturity Rate	2609	1	110	FRED
T10YIE	10-Year Breakeven Inflation Rate	2609	1	110	FRED

Table 1: Summary of all datasets used in the analysis, including number of observations (n), number of variables (v), missing values, and data sources.

For the final models, we rely on adjusted prices whenever they are available, as they incorporate the effects of dividends, stock splits, and other corporate actions. This ensures that the series reflects true economic returns and avoids artificial jumps or distortions in the data. For variables such as DGS10 and T10YIE, adjusted prices are neither necessary nor applicable, since they are single, non-corporate time-series without adjustment events. The dataset used in this analysis for the second part of the proyect is summarized in the following image 1

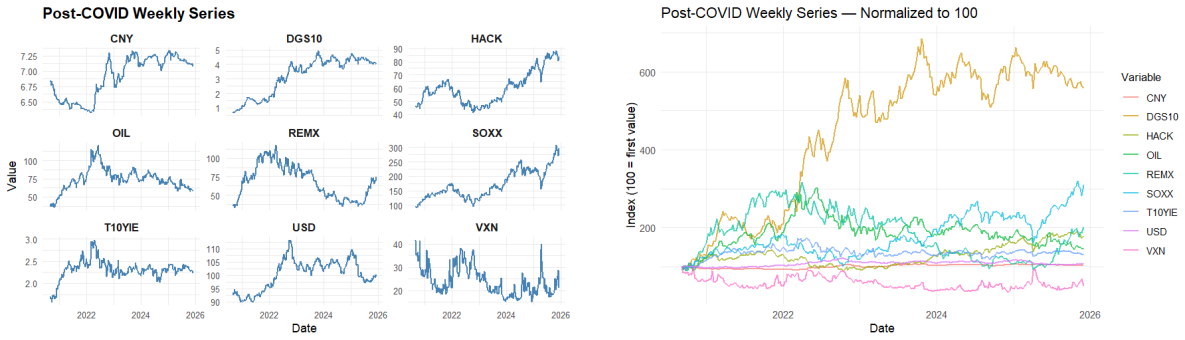

Figure 1: Data used

For the first part we take a look to the SOXX series (adjusted prize) and its log-returns, which is going to be what we finally use:

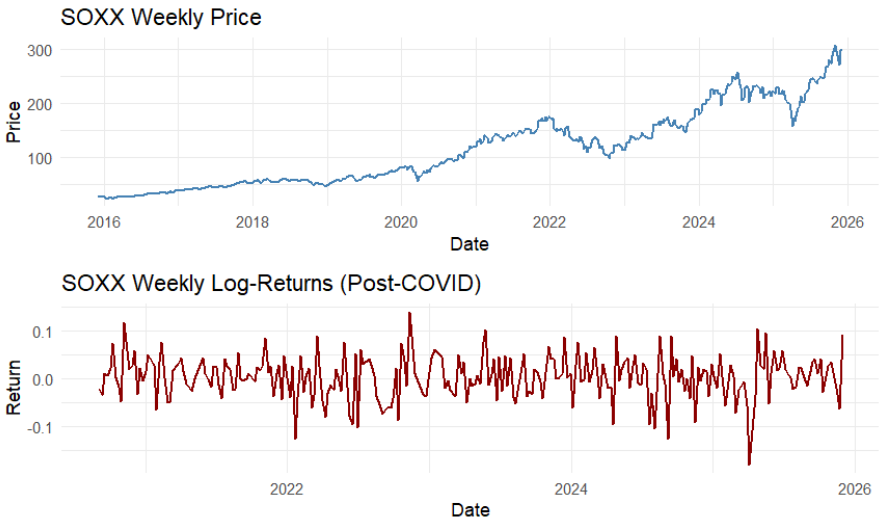

Figure 2: SOXX adjusted prize plus the log-returns

3 Exploratory data analysis

Financial instruments like SOXX have typical statistical characteristics that deviate from classical assumptions (normality, symmetry, dependence, homoskedasticity). To illustrate these violations, researchers use tests (Shapiro-Wilk statistic) and plots (histograms, Q-Q plots); many find that the prices of a financial instrument like SOXX can be described as having "fat tails," "volatility clustering" and time varying variance.

Thus, research often converts financial variables to "log-return" variables because log-returns produce time series with relatively stable variances, remove scale effects, and provide an additive way to interpret changes in the values of the original variables. Log-returns are also useful when conducting statistical and econometric models; therefore, log-returns represent the most common method of measuring return on investment in time series data in finance.

Therefore, we plan to transform all our data to log-returns and in addition to that we will convert the variable VIX to a categorical variable, as explained in section 5:

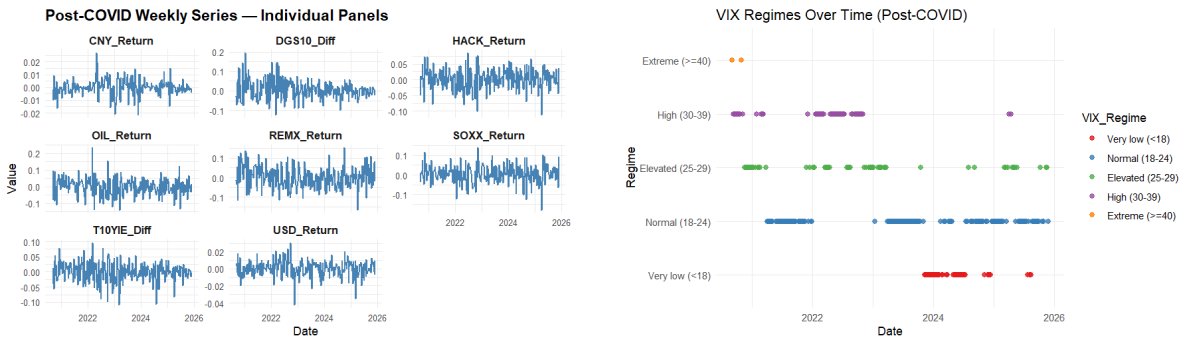

Figure 3: Final Data used

To examine whether the weekly data of SOXX follows a time-series process we ran two tests to determine its stationarity, the Augmented-Dickey Fuller (ADF) and the KPSS test. Both tests run under different hypotheses: the ADF test runs under the assumption that if the data is not stationary then it has at least one unit root, whereas the KPSS test runs under the assumption that if the data is not stationary then the data is locally stationary. Table 2 reports the findings.

Series	**ADF p-value**	**KPSS p-value**	**Conclusion**
SOXX Price Level	0.6444	< 0.01	Non-stationary
SOXX Log-Returns	< 0.01	0.10	Stationary

Table 2: ADF and KPSS test results for SOXX price levels and log-returns.

As expected in the literature, the price levels of SOXX failed to reject the null hypothesis of a unit root using the ADF test (thus indicating that the price levels of SOXX are non-stationary) and rejected the null hypothesis of stationarity using the KPSS test (thus supporting the fact that there is a stochastic trend in the price levels of SOXX). The weekly log-returns of SOXX, however, rejected the null hypothesis of a unit root using the ADF test (thus indicating that the log-returns of SOXX are stationary) and did not reject the null hypothesis of stationarity using the KPSS test (thus supporting the fact that the log-returns of SOXX are stationary).

This provides empirical support for the conversion of the data into log-returns and suggests that log-returns are suitable for the subsequent econometric modeling of the data.

We then examine how weekly log-returns of SOXX are distributed over two different time frames: before the COVID-19 pandemic (i.e., before Jan. 2020) and during the pandemic (i.e., from Sept. 2020 onward.). Figure 4 illustrates the variability in each time frame and highlights potential outliers.

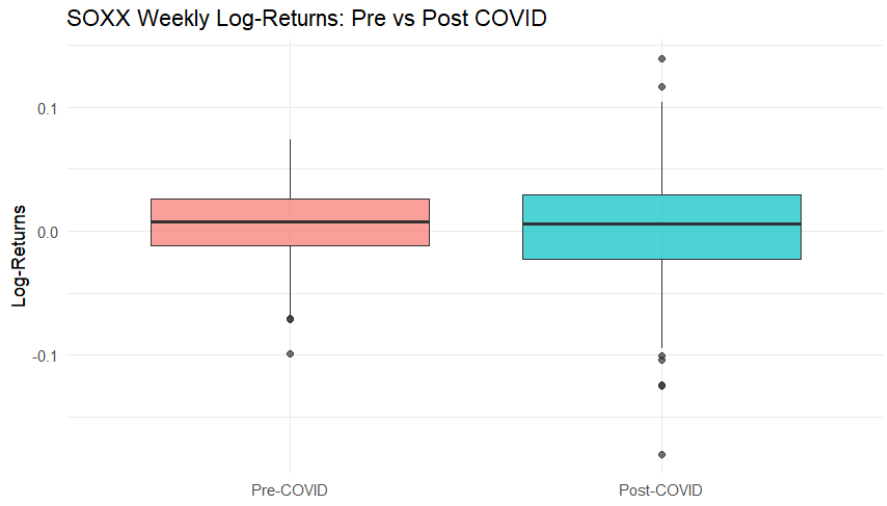

Figure 4: Boxplots Pre vs Post covid era

It can be observed that the median returns are relatively consistent across the two time

frames. However, the tail behavior clearly shows that large deviations — whether positive or negative — have become much more common since the COVID-19 shock. This provides some preliminary evidence related to the research question posed in the introduction, which will be formally evaluated in Section 4 by employing appropriate statistical methodologies.

With respect to further exploratory data analysis, we also observe in Figure 5 that for all the variables in the post-covid sample, we see histograms that show the data are generally not normally distributed (as is evident from the full time-series of SOXX log-returns; see Section 4 for tests and additional graphs on the full SOXX series).

However, all of the key methodological choices employed in our study were selected to be robust to violations of normality: the Brown–Forsythe test does not require that the data be normally distributed, and regression inference relies on large-sample properties rather than requiring strictly normal residuals. Thus, while the non-normal distributions of the returns are an important empirical property, they do not affect the validity of the statistical conclusions in this study.

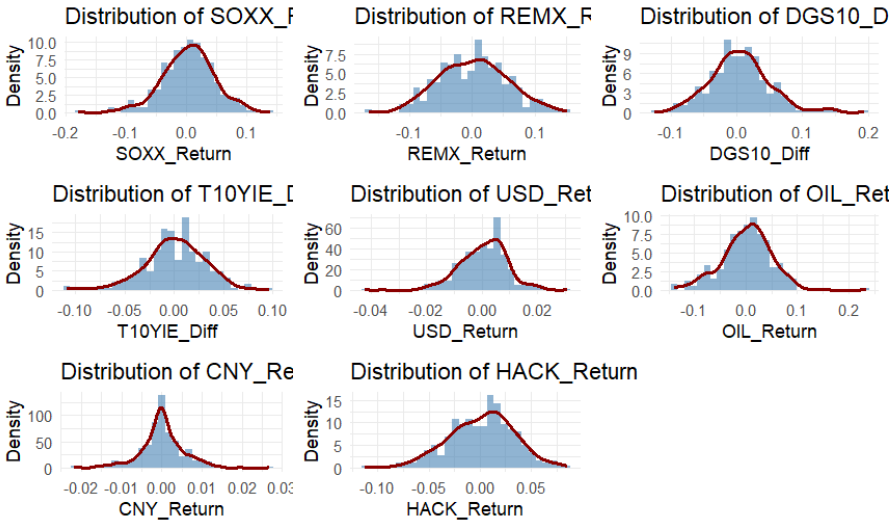

Figure 5: Distribution of the data (post covid era)

Regarding the remaining analysis involving autocorrelation, heteroscedasticity, normality, etc. for the SOXX log-return time series, please see Section 4.2, where all relevant assumptions are discussed in detail.

Before proceeding to Section 4 we analyze the short-term, weekly log-return volatility using a 15-week rolling window as shown in Figure 6. A rolling window gives us a "smoother" picture of the short-term changes that have occurred in the overall level of market risk. The COVID-19 crisis has clearly been an extreme event with respect to its effect on the stock price volatility and therefore is a natural point of departure for examining its impact upon the future behavior of the time series.

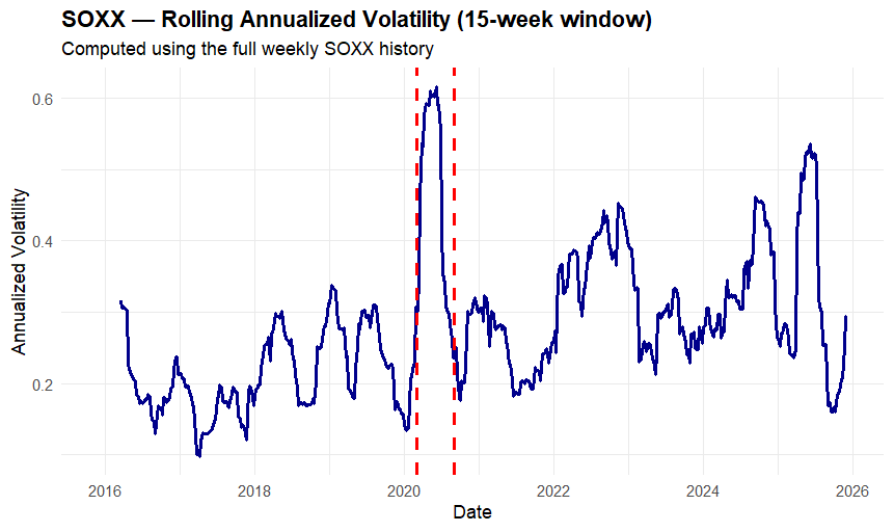

Figure 6: Rolling volatility of log-returns(15-week rolling window), with red lines indicating the covid period that separate pre vs post covid era

4 Hypothesis 1: Inference for Two Population Variances

The purpose of this area of study is to find out if the volatility of weekly SOXX returns was affected by the disruption caused by Covid-19 to the markets. Understanding the way that volatility moves is important for many areas including the valuation of assets, portfolio optimization, and managing risk. The changes in volatility are typically indicative of structural breaks or periods of high levels of market stress which can have a significant impact on the way you forecast risk and make investment decisions.

In order to determine how much of an effect the disruptions caused by Covid-19 had on the volatility of the SOXX returns the weekly log-return data was separated into two distinct time frames: a pre-event period that ended on 31-January-2020 and a post-event period that started on 01-September-2020. The time frame in between these two was excluded because it included a number of weeks where the markets were very volatile and distorted by the early stages of the pandemic.

Let R_t denote the weekly log-return, with $\{R_t^{\text{pre}}\}$ and $\{R_t^{\text{post}}\}$ representing the pre- and post-event samples. The preliminary assessment of how the events may have impacted the volatility will be based on the empirical variance of the data. However, the Brown-Forsythe test for equal variances will be used to formally assess the difference in volatility between the two time frames. It is preferable to use the Brown-Forsythe test rather than classical tests (e.g. F-tests) due to its ability to remain reliable when working with data that has heavy tailed distributions, is not normally distributed, and/or exhibits heterogeneous variance; characteristics that are common in return series generated from financial transactions. The goal is to determine if there are any differences in the level of volatility exhibited in the post-Covid timeframe compared to the level

of volatility exhibited in the pre-Covid timeframe.

4.1 Methodological Framework

Comparing the variance between two groups using the traditional F-test for testing equality of variance is generally based upon the fact that both samples are normally distributed. Since financial returns do not follow a normal distribution due to the presence of skewed distributions with fat tails among others, this assumption is not satisfied in most cases of financial returns.

For example, in order to verify this formally, we have run the Shapiro-Wilk normality test for the weekly returns of the SOXX index as it relates to the log-returns. This test provided a p-value = 1.306e-06, which strongly indicated rejection of the null hypothesis that the data is normally distributed. Furthermore, graphical diagnostic techniques were used to support these findings; a QQ-plot of the weekly returns (See Figure 7) showed considerable deviation from the theoretical quantiles of the normal distribution in the tails. As such, the nature of the empirical return distribution is clearly different than normal and therefore, the results of the F-test would be highly sensitive to the degree of skewness and tail behavior in the data, and thus inappropriate.

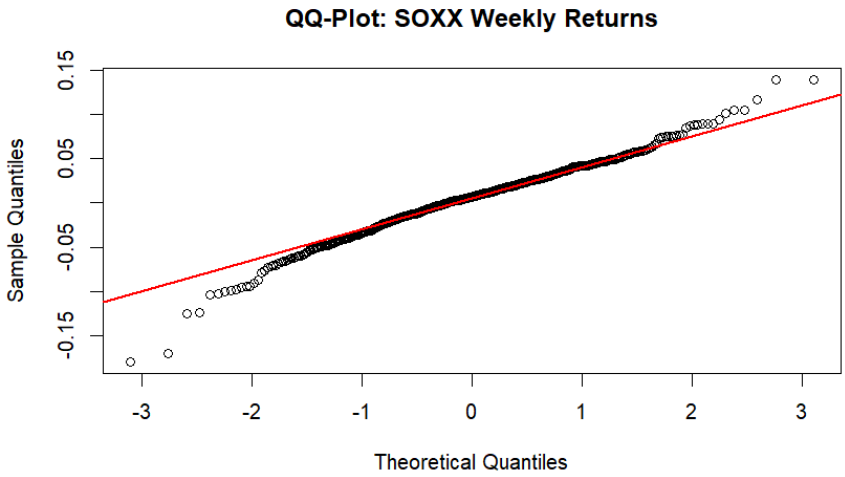

Figure 7: QQ-plot of weekly SOXX log-returns.

To overcome these limitations, the Brown–Forsythe test is employed. This method is a robust alternative designed to remain valid under heavy-tailed and asymmetric distributions. Let Y_{it} denote the return in group $i \in \{\text{pre}, \text{post}\}$ at time t, and let $\widetilde{Y}_i$ be the sample median of group i. The test constructs the transformed variables

$$Z_{it} = |Y_{it} - \widetilde{Y}_i|,$$

and performs a one-way ANOVA on $\{Z_{it}\}$. Under the null hypothesis of equal variances, the group means of Z_{it} should not differ significantly. Because it relies on median-based deviations, the

Brown–Forsythe procedure is substantially less sensitive to non-normality and heavy tails than the classical F-test, making it particularly suitable for financial return series.

In summary, the non-Gaussian empirical distribution of SOXX returns justifies the use of a robust variance comparison method, and the Brown–Forsythe test provides an appropriate and theoretically sound framework for evaluating whether volatility differs across the pre- and post-Covid regimes.

4.2 Assumptions

Before applying the Brown–Forsythe test to evaluate whether the volatility of weekly SOXX returns differs between the pre- and post-Covid periods, we conduct a series of diagnostic checks. These tests ensure that the statistical assumptions relevant to variance comparison procedures are reasonably satisfied and that the return series does not exhibit dependence structures that could invalidate the inference. All tests are performed separately for each subsample.

Test for Autocorrelation (Ljung–Box)

An important assumption to perform a comparison of variances is the *independence* of data points within each time period; we employ the Ljung-Box test with 10 lags to evaluate whether the first 10 autocorrelations of the SOXX return series are statistically significantly different than 0.

The p-value associated with the test is .36 for the pre-COVID sample and .95 for the post-COVID sample. For both samples, we fail to reject the null hypothesis of no serial correlation between successive data points.

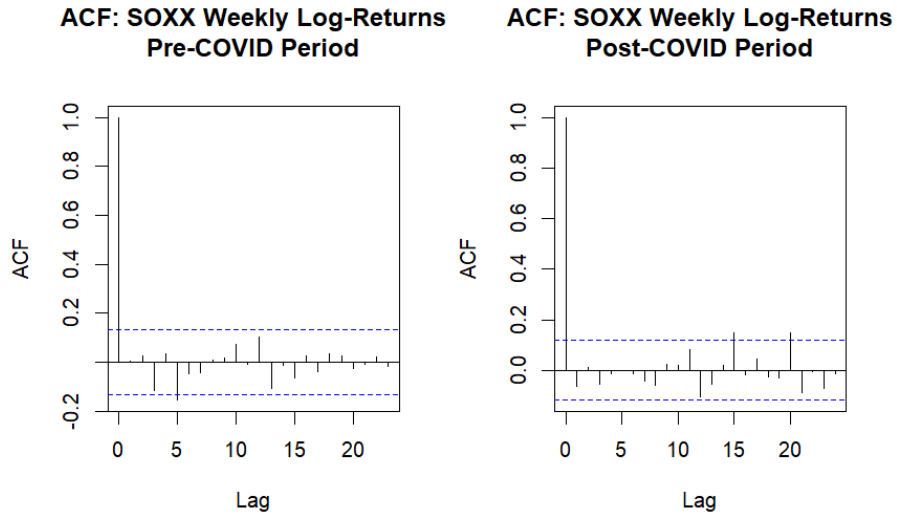

Figure 8: ACF of weekly returns of SOXX.

The correlograms seen in Figure 8 further confirm that all sample autocorrelations lie comfortably within the confidence bands, except for minor and economically negligible deviations

(there is a small peak at lag 5 for the pre-COVID sample and there are some mild oscillations around lags 15-20 for the post-COVID sample). Collectively, these results suggest that weekly SOXX returns can be viewed as independent processes over time in both time periods.

Test for Conditional Heteroskedasticity (ARCH LM Test)

While the Brown-Forsythe test relies on the serial independence of the returns, we can also determine if there is evidence of conditional heteroscedasticity in our time series. If our time series has ARCH effects, then we may see evidence of volatility clustering, a common trait found in high-frequency financial data. To examine the existence of ARCH effects in the squared residuals, we will use the ARCH LM test by Engle (1982) as the test statistic. For the ARCH LM test, the null hypothesis is no ARCH effects; while the alternative hypothesis is some form of ARCH effects.

In our results, the p-value for the ARCH LM test was very large for both the pre-COVID-19 period (p = 0.309) and the COVID-19 period (p = 0.816). Therefore, the null hypothesis of no ARCH effects cannot be rejected. These results indicate that the returns do not exhibit statistically significant volatility clustering at a weekly interval, a relatively rare occurrence in many financial time series (weekly returns tend to smooth out such bursts of volatility).

Implications for the Brown–Forsythe Test

The Brown–Forsythe test is a robust procedure for comparing population variances in a way that it will be a good option even if the data has very long tails and is non-normal. The main assumption for the Brown-Forsythe test is independent observations are present for each group; this means the Brown-Forsythe test does not require normality or homoscedasticity of the data to function properly.

Diagnostic testing indicates that (i) no autocorrelation is evident in weekly SOXX returns, (ii) no conditional heteroskedasticity exists in the return data, and (iii) the return distribution does not contain structural features that would make the use of robust variance comparisons inappropriate. As such, the statistical diagnostic evidence supports the use of the Brown-Forsythe test on the weekly SOXX return series as being statistically appropriate.

4.3 Hypothesis

Formally, the hypotheses are:

$$H_0 : \sigma^2_{\text{pre}} = \sigma^2_{\text{post}}, \qquad H_1 : \sigma^2_{\text{pre}} \neq \sigma^2_{\text{post}}.$$

The test is two-sided, as volatility could plausibly either rise or fall following the Covid-19 disruption.

```
Levene's Test for Homogeneity of Variance (center = median)
```

```
       Df  F value     Pr(>F)
group   1  21.479 4.594e-06 ***
      489
---
```

The Brown–Forsythe statistic strongly rejects the null hypothesis of equal variances. The test yields an F-value of 21.479 with a p-value of 4.594×10^{-6}, providing overwhelming evidence that the volatility of weekly SOXX returns differs across the two regimes. This result indicates a statistically significant shift in the variance structure between the pre- and post-Covid periods.

4.4 MCAR & MNAR

Missing data mechanisms can change the statistical analysis of the data. To evaluate how robust the Brown–Forsythe test is depending on the type of missing data that was added to the data set, we used two types of missing data mechanisms: missing completely at random (MCAR) and missing not at random (MNAR). These two types of missing data mechanisms represent different ways that the probability an observation will be removed from the data set may depend on the characteristics of the observations that were removed or retained in the data set.

MCAR: Random Removal of Observations

For the MCAR mechanism, the probability that an observation will be missing is completely unrelated to any characteristics of the observation (both observed and unobserved). For our simulation of MCAR, a random subset of the full data set was used to create missing data. The percentage of data points in the full data set that were randomly deleted from the subset ranged from 10 percent to 60 percent. Since MCAR removes the observations independently of both the value of the returned data point and whether the data point corresponds to a time period before or after the Covid pandemic, MCAR maintains the same fundamental structure as the original data and should remove no systematic bias when comparing variances across time periods.

In order to determine how the missingness under the MCAR mechanism would affect the statistical inference regarding Hypothesis 1 using the Brown-Forsythe test, the Brown-Forsythe test was recomputed for each of the levels of missingness represented in the previous paragraph. Table 3 displays the p-values obtained from each of these tests. As can be seen from the table, regardless of the amount of missingness that had been added to the data set, the test remained strongly statistically significant. Therefore, the conclusion that the variances of the returns for the stock prices are equal during the pre-pandemic period and less than equal during the post-pandemic period is very robust to missing data that has been lost due to random loss of data. This behavior is consistent with what would have been theoretically expected: since MCAR loses observations uniformly, it reduces the sample size; however, it does not distort the variance structure of the underlying data.

Table 3: Brown–Forsythe p-values under varying MCAR missingness levels

Percentage Missing	p**-value**
10%	5.83×10^{-5}
20%	3.73×10^{-4}
30%	3.30×10^{-4}
50%	2.47×10^{-3}
60%	2.00×10^{-3}

That the statistical significance of the test persisted even after the removal of over 50 percent of the sample demonstrates that the variance shift detected in Section 2.3 is not a result of sample size or sensitive to random omission of data. Thus, the results from the MCAR confirm the stability of the inference made in relation to this hypothesis.

MNAR: Selective Removal Based on Return Magnitude

Missing Not At Random (MNAR) occurs when the probability that an observation is missing is not randomly determined. While MNAR is theoretically possible, it is also economically plausible as financial returns are subject to market microstructure effects (e.g., trading interruption, data filters) that have a disproportionate impact on extreme returns; therefore, it is likely that unusual large positive and/or negative returns are omitted from financial databases.

In order to replicate MNAR for each return, a probability of removal was assigned based on the absolute value of the return. The "missingness strength" variable determines the extent to which the MNAR mechanism is able to target extreme values of returns. Low values of the missingness strength will result in a larger number of extreme returns being removed from the database. Conversely, high values of the missingness strength will result in fewer extreme returns being removed.

The results of the Brown–Forsythe test for MNAR are reported in Table 4. The results show that the Brown–Forsythe test is highly significant when the missingness is somewhat selective (i.e., strength = 0.3–1.5). Although the percentage of removed returns varied from approximately 54% to 8%, the test remained highly significant. This robustness is primarily caused by the fact that the moderately selective removal of extreme returns still provides sufficient data to determine the existence of a difference in volatility between the two time periods.

Conversely, when the mechanism is extremely selective and removes more than 70% of the original sample (i.e., strength = 0.2), the Brown-Forsythe test is no longer able to detect a statistically significant difference in variance. When the mechanism is extremely selective, it essentially eliminates almost all of the tail behavior in the data that causes a difference in volatility between the two time periods prior to and after Covid.

Table 4: Brown–Forsythe p-values and missing rates under MNAR mechanisms

Missingness Strength	**Missing Rate**	p**-value**
0.2	0.6904	1.95×10^{-1}
0.3	0.5458	6.52×10^{-3}
0.5	0.3585	5.02×10^{-3}
0.8	0.2037	4.42×10^{-3}
1.2	0.1466	1.85×10^{-3}
1.5	0.0855	1.77×10^{-4}

Therefore, only when the mechanism has removed an overwhelmingly large portion of the data (essentially eliminating almost all of the tail behavior) does the difference in variance between the two regimes become masked.

Implications

Overall, the results from both MCAR and MNAR demonstrate that the conclusion of Hypothesis 1 is highly robust. Random missingness has virtually no impact on the inference, and even under selective removal of extreme returns, the variance difference remains detectable in all but the most extreme and unrealistic MNAR scenarios. Therefore, the evidence of a post-Covid increase in volatility remains strong and reliable.

4.5 Conclusion

The results of the Brown–Forsythe test provide statistically compelling evidence that the variance of weekly SOXX returns differs between the period before and after the onset of the Covid-19 pandemic. In particular, the test statistic F = 21.479 with a p-value of 4.594Ö10^{-6} allows us to confidently reject the null hypothesis of equal variances. Moreover, the diagnostic analyses confirm independence and the absence of effects that would compromise the validity of the inference in either time window, ensuring the appropriateness of the applied test. The simulations under MCAR and MNAR missing data mechanisms show that statistical significance persists except under extremely selective removal of observations, which reinforces the robustness of the conclusion. Consequently, we conclude that the disruption caused by Covid-19 generated a significant structural shift in SOXX volatility, with direct implications for risk evaluation and portfolio management within the semiconductor sector.

5 Hypothesis 2: Multiple linear regression

This portion of the project will not be developing a prediction model based on the classic forecasting definition; instead, it seeks to identify and measure the economic influences that contribute to weekly SOXX returns.

Because the first part of the study has confirmed by statistical means that there has been a structural shift in the volatility of SOXX returns after the COVID-19 market disruption, the post-COVID era is investigated independently.

Therefore, the estimated relationships of interest are representative of current market conditions that are economically relevant to the semiconductor industry today, and as such prevent the combination of structurally heterogenous regimes.

Formally, let R_t^{SOXX} denote the weekly log-return of SOXX, modeled as:

$$R_t^{SOXX} = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \varepsilon_t, \tag{1}$$

where $X_{i,t}$ denotes macroeconomic or sector-related explanatory variables motivated by financial theory, and ε_t represents the regression disturbance term.

The core inference question is:

> *Which macroeconomic and sector-specific variables exert a statistically significant impact on weekly SOXX returns?*

5.1 Predictor Selection and Economic Motivation

The selection of variables is guided by economic intuition regarding factors that can influence semiconductor performance:

- **REMX (Rare Earths/Metals ETF):** captures input-cost exposure. Rare earth materials are essential for semiconductor production. Positive dependence may reflect industrial demand cycles.
- **DGS10 (10-year U.S. Treasury yield):** proxy for interest rate expectations. Higher rates typically compress equity valuations, particularly in tech.
- **T10YIE (Breakeven inflation rate):** represents long-term inflation expectations. Elevated inflation may raise production costs and discount rates.
- **ICE U.S. Dollar Index:** a stronger dollar increases costs for multinational semiconductor exporters and tightens global liquidity.
- **WTI Crude Oil (CL=F):** energy is a major cost input in the semiconductor supply chain. Oil shocks can impair margins and industrial activity.
- **CNY/USD Exchange Rate:** proxies China-related economic pressures, supply chain dynamics, and reshoring trends.
- **HACK (Cybersecurity ETF):** represents the performance of cybersecurity companies and technology infrastructure.

- **VIX Regime (from $\hat{}VXN$):** a measure of implied volatility for Nasdaq firms and a widely used indicator of market risk sentiment in the tech sector. To avoid imposing a linear relationship between volatility and returns, VXN is incorporated as a categorical variable representing different risk environments:
  - Very low: $VXN < 18$
  - Normal: $18 \leq VXN \leq 24$
  - Elevated: $25 \leq VXN \leq 29$
  - High: $30 \leq VXN \leq 39$
  - Extreme: $VXN \geq 40$

  This classification enables the model to capture changes in SOXX sensitivity under distinct volatility regimes.

Quadratic terms are included to allow nonlinear sensitivity to large movements in the explanatory factors.

Why market benchmarks (NASDAQ or S&P 500) are excluded:

- SOXX is a component of major U.S. equity benchmarks and shares many constituents with NASDAQ.
- Including SPY or QQQ would lead to severe multicollinearity and make inference economically trivial.
- The question of interest is not whether SOXX moves with the broad market, which is obvious, but how exposure to specific external risk drivers explains semiconductor dynamics.

Thus, our specification is designed to avoid capturing only the general equity beta and instead isolate sector-relevant economic forces.

5.2 Hypothesis

This area's aim will be to formally assess those macro-economic and sector-specific factors that significantly affect weekly SOXX returns in the post-Covid time frame.

Our objective with the study is to be inferential (i.e., to explain how economic forces impact semiconductor equities) as opposed to being predictive; therefore, we are concerned with evaluating the statistical significance of each of the individual regression coefficients, and identifying if the various economic forces have a meaningful effect upon the performance of semiconductor equities.

Let

$$R_t^{SOXX} = \beta_0 + \sum_{i=1}^{k} \beta_i X_{i,t} + \varepsilon_t$$

denote the multiple linear regression model estimated on weekly post-Covid data, where $X_{i,t}$ includes the economically motivated predictors described in Section 3.1 together with their quadratic terms and the categorical VIX regime.

For each explanatory variable X_i, the inferential question is:

Does X_i have a statistically significant effect on weekly SOXX returns?

Formally, for every regression coefficient we test:

$$H_0 : \beta_i = 0 \qquad \text{vs.} \qquad H_1 : \beta_i \neq 0.$$

Under H_0, the predictor has no explanatory power for SOXX returns, whereas under H_1 the variable contributes significantly after controlling for all other factors.

This comprehensive model serves as an exploratory benchmark: it allows us to observe how each factor behaves when all potential sources of variation are included simultaneously.

```
Call:
lm(formula = SOXX_Return ~ REMX_Return + DGS10_Diff + T10YIE_Diff +
    USD_Return + OIL_Return + CNY_Return + HACK_Return + REMX_Return_sq
        +
    DGS10_Diff_sq + T10YIE_Diff_sq + USD_Return_sq + OIL_Return_sq +
    CNY_Return_sq + HACK_Return_sq + VIX_Regime, data = df_merged)

Residuals:
      Min         1Q     Median         3Q        Max
-0.068162 -0.020602 -0.000486  0.017097  0.085284

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0017889  0.0057769   0.310 0.757091
REMX_Return     0.1760029  0.0460504   3.822 0.000169 ***
DGS10_Diff     -0.0185325  0.0541824  -0.342 0.732625
T10YIE_Diff     0.0140303  0.0789477   0.178 0.859097
USD_Return     -0.0881902  0.2763656  -0.319 0.749927
OIL_Return     -0.0333564  0.0435333  -0.766 0.444304
CNY_Return     -0.1363115  0.3706271  -0.368 0.713361
HACK_Return     0.8123984  0.0720873  11.270  < 2e-16 ***
REMX_Return_sq -0.0004774  0.5429322  -0.001 0.999299
DGS10_Diff_sq   0.9964947  0.5823095   1.711 0.088338 .
```

```
T10YIE_Diff_sq -2.4051477  1.3101257  -1.836 0.067638 .
USD_Return_sq   1.8621716 12.2424397   0.152 0.879231
OIL_Return_sq  -0.4507441  0.4278753  -1.053 0.293209
CNY_Return_sq  46.3844571 25.9038086   1.791 0.074627 .
HACK_Return_sq  2.2021601  1.4307261   1.539 0.125092
VIX_Regime.L    0.0010219  0.0152974   0.067 0.946796
VIX_Regime.Q    0.0142536  0.0126427   1.127 0.260706
VIX_Regime.C    0.0064894  0.0080982   0.801 0.423743
VIX_Regime^4    0.0018893  0.0049106   0.385 0.700774
---

Residual standard error: 0.0302 on 237 degrees of freedom
Multiple R-squared:  0.5806,    Adjusted R-squared:  0.5487
F-statistic: 18.23 on 18 and 237 DF,  p-value: < 2.2e-16
```

- **REMX_Return** and **HACK_Return** are strongly significant ($p < 0.001$), suggesting that both industrial-input dynamics and the broader technology ecosystem influence semiconductor performance.
- Several nonlinear terms (e.g., $DGS10\_Diff^2$, $T10YIE\_Diff^2$, CNY_Return^2) exhibit borderline significance ($0.05 < p < 0.10$), indicating possible nonlinearities, although the evidence is weak.
- Most macroeconomic linear terms (Treasury yields, inflation expectations, USD index, oil prices, CNY/USD exchange rate) are individually insignificant.
- The VIX regime dummies are not statistically significant, implying that once economic and sector-specific fundamentals are included, shifts in implied volatility do not add explanatory power for weekly SOXX returns.

Thus, the full specification serves primarily as a diagnostic benchmark: it highlights potential drivers while simultaneously revealing substantial redundancy among the predictors.

Model Selection via Exhaustive Best Subset

Because the number of variables is modest, an exhaustive search over all possible subsets is performed, as is indicated in [1] (pp. 205).

And we select the model that minimizes the Bayesian Information Criterion (BIC).

This approach penalizes unnecessary complexity and ensures that the final model retains only predictors with meaningful explanatory power.

The best-subset procedure identifies a remarkably parsimonious model:

$$R_t^{SOXX} = \beta_0 + \beta_1 \,\mathrm{REMX}_t + \beta_2 \,\mathrm{HACK}_t + \varepsilon_t.$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001866   0.001899   0.983    0.327
REMX_Return 0.196056   0.038988   5.029 9.36e-07 ***
HACK_Return 0.827730   0.064895  12.755  < 2e-16 ***
---

Residual standard error: 0.03035 on 253 degrees of freedom
Multiple R-squared:  0.5478,    Adjusted R-squared:  0.5442
F-statistic: 153.2 on 2 and 253 DF,  p-value: < 2.2e-16
```

Both coefficients are highly statistically significant ($p < 10^{-6}$), and the reduced model achieves an adjusted R^2 comparable to that of the full specification. Thus, after examining all candidate variables, only **REMX** and **HACK** consistently explain weekly SOXX returns in the post-Covid period.

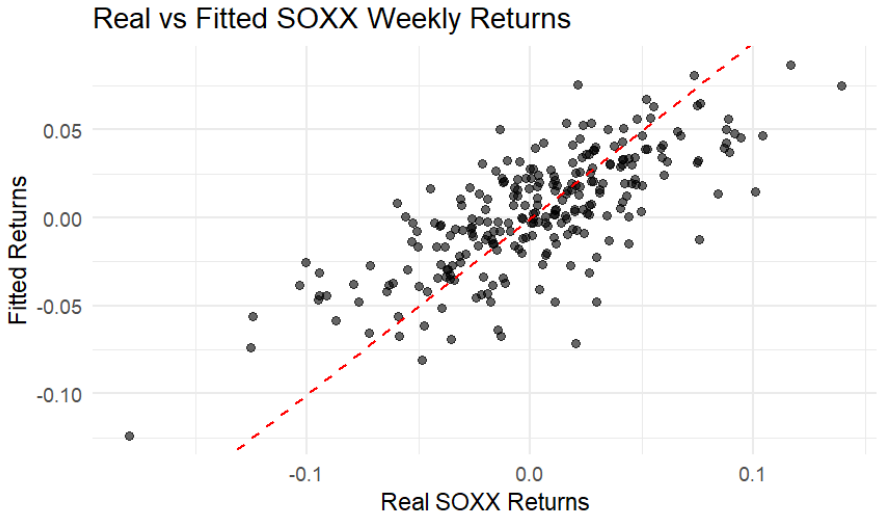

Figure 9: Real vs Fitted SOXX Weekly Returns

The comparison of real versus fitted values see Figure 9) confirms that this reduced specification captures the main dynamics of SOXX without notable structural distortions.

Overall we see the adjust in the next Figure 10:

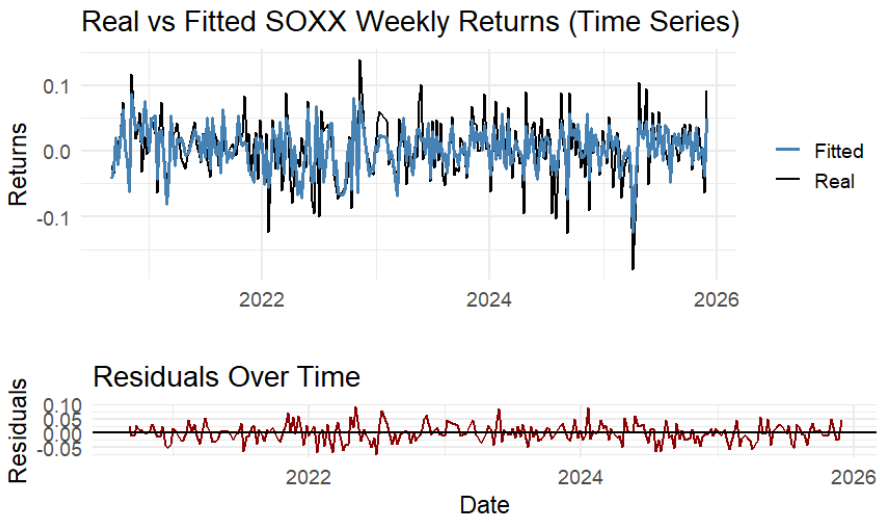

Figure 10: Fitted model

Interpretation of the Hypothesis Testing Outcome

The individual hypothesis tests yield the following conclusions:

- We **reject** H_0 for both REMX_Return and HACK_Return. These variables exert statistically significant and economically intuitive influences on SOXX.
- We fail to reject H_0 for the remaining macroeconomic predictors and for all VIX regime variables. Several quadratic terms (e.g., $DGS10\_Diff^2$, $T10YIE\_Diff^2$, CNY_Return^2) exhibit marginal significance at conventional levels ($0.05 < p < 0.10$), suggesting the possibility of nonlinear sensitivities when the full set of predictors is included. However, the evidence is weak and not robust: none of these terms are retained by the BIC-optimal model, indicating that their explanatory contribution is not sufficiently strong once model complexity is penalized.

5.3 Assumptions Checks

To confirm that the inferential results in the final regression are valid, the residuals of the final regression were subject to a comprehensive analysis to check against all of the key diagnostics (normality, homoscedasticity, independence, linearity). Overall, these evaluations, which can be both graphical as well as statistical, will provide an overall assessment as to whether the model's estimation is sufficient for drawing inferences about the parameters of interest.

Normality of the Residuals

Graphical and formal hypothesis testing were used to determine the distribution of residuals.

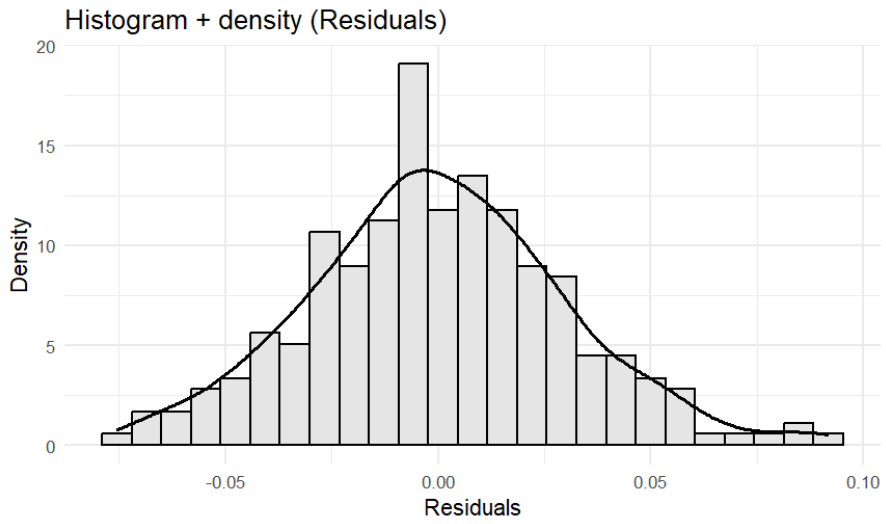

Figure 11: histogram with a kernel density estimation

The histogram with a kernel density estimation (see Figure 11) illustrates the residuals' distribution is generally symmetric and centered at zero. Although the residuals have slightly heavier tails than one would expect in other economic time-series data (which can be due to the volatility of the financial markets), the histogram does not exhibit obvious skewness. There is also little indication of extreme or unusual residual clusters in the data.

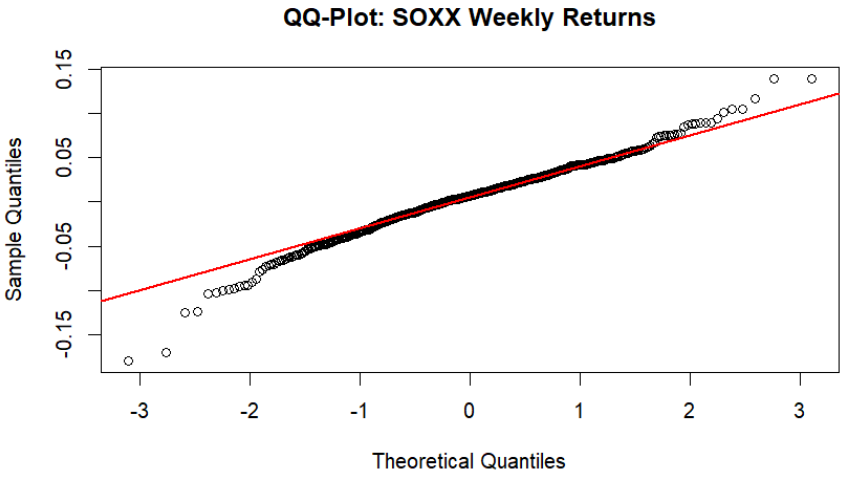

Figure 12: histogram with a kernel density estimation

A QQ plot (see Figure 12) provides additional support to the conclusions drawn from the histogram; the residuals match the theoretical normal quantiles well over most of the middle range of the residuals' distribution. The residuals deviate somewhat in the upper end of their distribution.

The Shapiro-Wilk test (p-value of 0.4029) further supports the graphical interpretations above as it fails to reject the null hypothesis of normal residuals. Therefore, both graphical and formal assessments support that the residuals can be appropriately modeled using a Gaussian distribution for inference.

Linearity

The residuals vs. fitted values plot (Figure 13) is used to assess whether the data follow a linear relationship. The data exhibit an unstructured and random cloud of points centered on the horizontal axis that represents zero residual, and do not show evidence of non-linear relationships such as curvature or trending; nor does the data demonstrate non-uniform variance over the fitted values range. In addition, the loess smoothed line indicates no change in trend which also supports the linearity assumption. The overall pattern is consistent with the assumed linear function. It may be worth noting that the lone anomaly in the regression can be linked back to a known event (trade policy announcement) but does not represent a systematic violation of the linearity assumption.

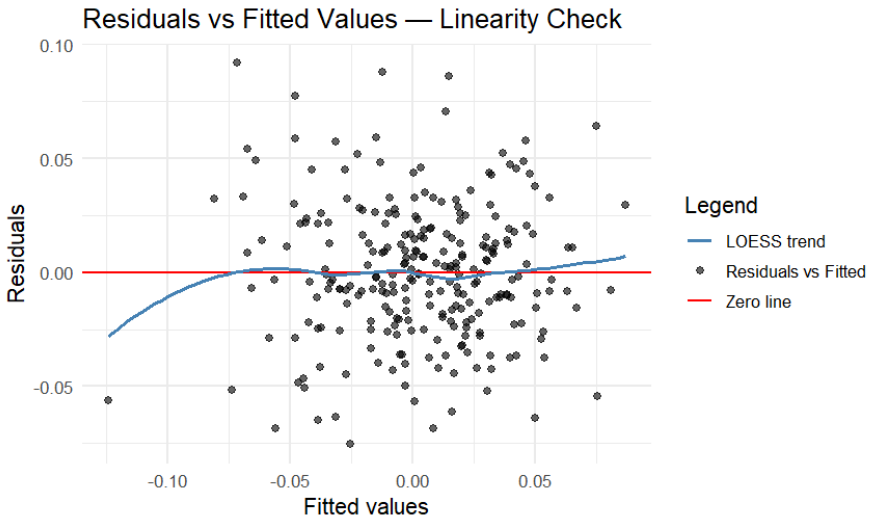

Figure 13: histogram with a kernel density estimation

Independence of Errors

To determine if there is serial independence of the residuals in the regression model, I conducted the Breusch-Godfrey test (a popular choice) as well as the Autocorrelation Function (ACF).

A test for first order serial correlation from the Breusch-Godfrey yielded no statistically significant evidence of residual autocorrelation (p-value = 0.3476; LM = 0.882).

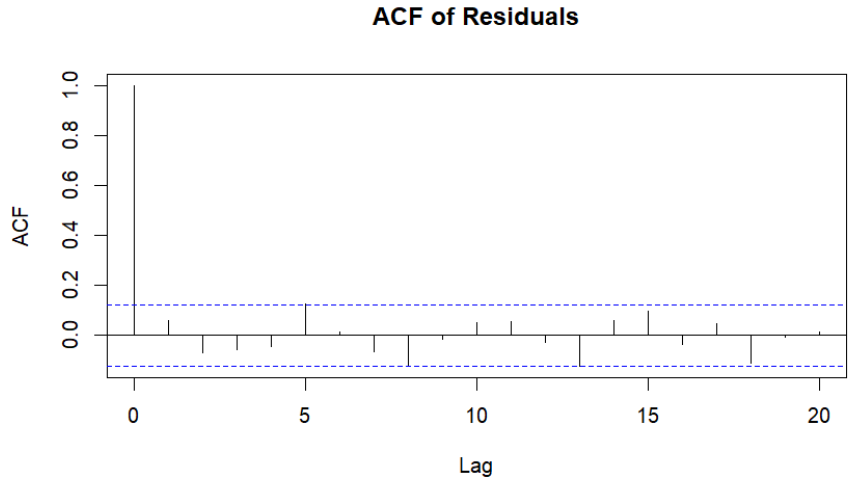

Figure 14: histogram with a kernel density estimation

An examination of the ACF plot (Figure 14) indicates that with one exception, each lagged sample autocorrelation lies outside the 95% confidence limits. There were five instances of lags where an autocorrelation approached but did not exceed the upper limit of the band (lags of approximately 5, 8, 13, 15 and 18). The size of these deviations was small relative to those typically seen in financial return data which frequently display weak forms of short range dependence, though most often such dependence is economically insignificant.

Overall, both formal and graphical evidence strongly support the assumption of independence.

Homoscedasticity

To assess whether the variance of the regression residuals remains constant across the range of fitted values, both formal and graphical diagnostics were examined. The Breusch–Pagan test provides statistically significant evidence of heteroskedasticity:

$$BP = 8.33, \quad df = 2, \quad p\text{-value} = 0.0156,$$

which leads to a rejection of the null hypothesis of constant error variance at the 5% level.

Although the data is presented graphically in a way that suggests an alternative view of the residuals than one of a simple increase or decrease in residual variability for increasing/decreasing levels of fitted value,

The Residuals vs Fitted plot (Figure 13), doesn't show what could be termed a "cone" shape which indicates a variation in the spread of the residual clouds as fitted values increase/decrease. The residuals are generally uniform, and there is no apparent widening or narrowing occurring as the fitted-values increase or decrease, except for notable departures from this uniformity at the lower limit of the fitted-value range; these fitted-values correspond to the known outlier

associated with the policy-induced shock in the market. Therefore, these observations cannot be considered representative of the behaviour of the model under normal conditions.

The *Scale–Location* plot (Figure 15) provides somewhat clearer evidence of mild heteroskedasticity. The smoothed trend line exhibits a slight upward slope for higher fitted values, indicating that the standardized residuals become only marginally more variable when the model predicts higher SOXX returns. Importantly, this increase in dispersion is minimal, and the variance remains relatively stable throughout the central, well-populated region of the fitted-value distribution.

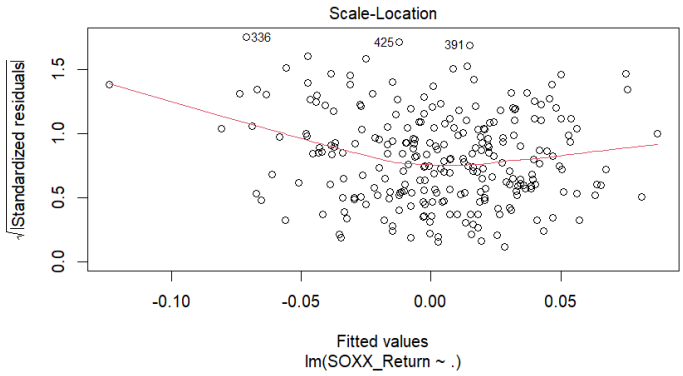

Figure 15: Scale–Location plot

Residual plots against individual predictors (Figure 16) further support this interpretation: the spread of residuals does not show systematic expansion or contraction across the ranges of `REMX_Return` or `HACK_Return`, apart from observations influenced by the same extreme event described above.

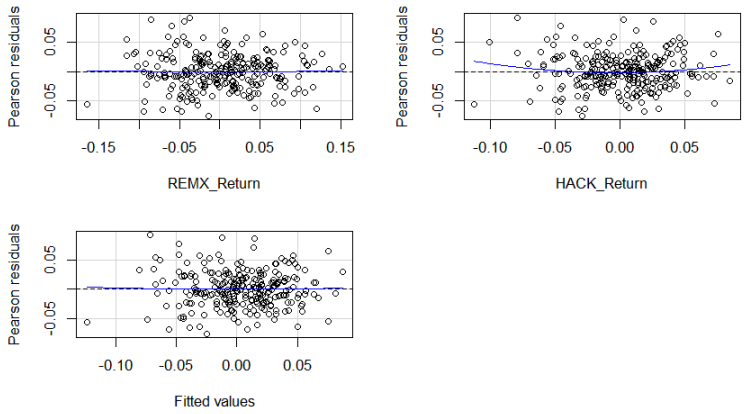

Figure 16: Residuals versus predictors

The diagnostic plots do not indicate any strong evidence of heteroscedasticity; however, the

Breush-Pagan test does confirm its statistical presence. The mild heteroscedasticity in variance is typical of financial returns where small amounts of variation are sufficient to bias OLS standard error estimates. Because there is no indication of biased OLS coefficient estimates due to heteroscedasticity.

Therefore, the graphs illustrate little if any distortion in heteroscedasticity, except for some minor increases in variability at higher fitted value levels; whereas the formal test confirmed the statistical existence of such subtle effects.

Multicollinearity

To assess whether the explanatory variables included in the final regression model exhibit linear dependence, we computed the Variance Inflation Factors (VIF) for the two predictors retained in the BIC-optimal specification: `REMX_Return` and `HACK_Return`. The resulting VIF values are:

$$\text{VIF}_{\text{REMX}} = 1.25, \qquad \text{VIF}_{\text{HACK}} = 1.25.$$

All values lie extremely close to 1, which is the theoretical minimum indicating complete absence of multicollinearity. These results fall well below commonly used rule-of-thumb thresholds. Therefore, no evidence of problematic linear dependence is present among the predictors in the final model.

The evidence from this finding fits the economics of the specification. `REMX_Return`, measures changes to rare earth metal prices and industrial input cost. `HACK_Return` measures the returns to technology firms (cybersecurity) which measure different components of overall tech sector movement. Although the two measure some aspect of tech sectors, they represent entirely different under lying factors of the economy. The VIF results of the empirical analysis also show that `REMX_Return` and `HACK_Return` each provide unique information and therefore do not create excessive inflationary variance in the model due to their lack of redundancy in a linear relationship.

The implications of the findings of this study also have implications on the interpretation of the estimated coefficients. Since there is no multicolinearity, the estimated coefficients will be relatively precise and stable. Therefore, since the standard error of the estimates will not be artificially inflated and since the statistical significance of the predictors can be reliably interpreted, the multicolinearity is consistent with the parsimonious nature of the final model. Additionally, the low multicolinearity is also consistent with the fact that the best subset selection method would naturally eliminate variables that were either redundant or highly correlated to other variables.

Overall, the multicolinearity diagnostic results clearly support the conclusion that the final regression model is well conditioned; the two retained predictors are economically significant,

empirically independent and can be used as a reasonable base for making inferences.

Summary

Taking all diagnostics into consideration, the final regression model has met most of the classical requirements to support valid inference with only small exceptions that are not critical enough to undermine its interpretation.

The residual plots exhibit approximate normal distributions, with some small deviations at the tails (as is common in financial return data). The residual plots clearly show linearity and no systematic structure or curvature, indicating that the regression model has properly captured the functional form of the relationship between SOXX returns and the two predictor variables selected.

Independence is also upheld. There is no evidence of serial correlation in either the Breusch-Godfrey test or the residual autocorrelation function (ACF), thus the residuals do not contain any predictable temporal structure; this will help to ensure the reliability of the coefficient estimates as well as the stability of the error process over time.

The primary exception to these assumptions is heteroskedasticity, although it is relatively mild in nature. Heteroskedasticity indicates that the variance of the error term is not constant across the fitted value range; while this may not significantly impact the estimated coefficients, it could potentially lead to inaccuracies in measuring their standard errors. These inaccuracies could result in inaccurate t-statistics and p-values being calculated for each individual predictor variable; thereby, affecting the strength of the inference made regarding each predictor variable.

Lastly, there is no indication of multicolinearity in the model, since both predictor variables have VIFs close to 1, therefore, they appear to add distinct explanations to the model, an their respective coefficient estimates are not inflated nor destabilized due to linear dependence.

Overall, the final regression model is generally consistent with the OLS assumptions. Although, there is one material deviation from the OLS assumptions (a mild form of heteroskedasticity) that may introduce variability in the accuracy of standard errors, however, this is not critical enough to diminish the overall ability of the model to assess the economic factors driving weekly SOXX returns during the post-Covid period.

5.4 MCAR & MNAR

Missing data mechanisms can change the statistical analysis of the data. To evaluate how robust the Brown–Forsythe test is depending on the type of missing data that was added to the data set, we used two types of missing data mechanisms: missing completely at random (MCAR) and missing not at random (MNAR). These two types of missing data mechanisms represent different ways that the probability an observation will be removed from the data set may depend on the characteristics of the observations that were removed or retained in the data set.

MCAR: Random Removal of Observations

To assess the effect of MCAR missingness on the linear regression model, we removed observations at random at varying levels (10%, 20%, 30%, 50% and 60%) and estimated the model each time. Table 4 summarizes the resulting coefficient estimates. We found that the coefficients for REMX return and HACK return remained statiscally significant across all missingness levels. Although the estimates show minor natural fluctuations, there is no directional drift as missingness increases. This behavior is consistent with theoretical expectations, as MCAR reduces sample size but does not introduce bias into regression coefficients. Furthermore, both R^2 and adjusted R^2 remain close to their values in the complete dataset, indicating that the underlying relationships are preserved and that random deletion does not systematically distort explanatory power.

Table 5: Coefficient Estimates Under MCAR Missingness

Missingness	**Intercept**	**REMX_Return**	**HACK_Return**
10%	0.001423	0.206971	0.797944
20%	0.002186	0.208200	0.708527
30%	-0.001732	0.166099	0.962164
50%	-0.003334	0.206770	0.794603
60%	-0.004884	0.238207	0.832329

MNAR: Selective Removal Based on Return Magnitude

In financial markets, trading halts, circuit breakers, and regulatory suspensions can affect ETFs such as SOXX, REMX, and HACK, particularly during periods of heightened volatility or market-moving events. For example, semiconductor stocks can be halted when news related to US-China chip export ban is suddenly disseminated. Similarly, large scale cyber attack on US banks may result in temporarily ban of stocks in HACK ETF. When trading is halted, daily returns for those assets may be absent from the dataset. Since these events are typically associated with unusually large price movements, liquidity shocks, or sector-specific disruptions, the probability that a return is missing depends directly on the magnitude of the (unobserved) return itself. This mechanism corresponds to Missing Not At Random (MNAR) and can substantially affect our regression model.

To investigate the affects of MNAR, we force large positive and negative returns more likely to be removed from our dataset and run the regression model with this new dataset. The results obtained from the regression are summarized in Table 5. Relative to the complete-case model, MNAR introduces systematic distortions in the coefficient estimates: the REMX coefficient increases by approximately 9% while the HACK coefficient decreases by roughly 7–8%. Despite only a modest reduction in sample size (31 removed observations), the direction of these shifts demonstrates the biasing effect of MNAR missingness.

We also notice that although the R^2 value remains close to the full-data result (0.5565 vs. 0.547), the residual standard error decreases, falsely suggesting improved fit. This reduction occurs because the MNAR mechanism selectively removes extreme observations, not because the model has become more accurate. Thus, even limited MNAR missingness leads to meaningful distortions in inference and highlights the need for robust missing-data strategies.

Some of the strategies we can use to handle MNAR are selection models, pattern mixture models and shared parameter models which require explicit modeling of missingness mechanism. For example, selection models jointly model the outcome and missingness equations while pattern mixture models stratify the data by missingness patterns and recombine the resulting distributions. Shared parameter models on the other hand introduce a latent variable that influences both the outcome and the missingness. It is important to note however that all of these approaches rely on strong parametric assumptions that cannot be empirically validated.

Table 6: Regression Estimates Under MNAR Missingness

	Estimate	Std. Error	p-value
Intercept	0.0002307	0.0019644	0.907
REMX_Return	0.280034	0.0405884	6.44×10^{-7}
HACK_Return	0.8307145	0.0703917	$< 2 \times 10^{-16}$
Residual Std. Error	0.02946 (df = 222)		
R^2	0.5565		
Adjusted R^2	0.5525		
Observations Used	223		
Observations Removed (MNAR)	31		

Implications

Thus under MCAR, where deletions are random, our simulations confirm that the linear regression estimates remain unbiased. In contrast, MNAR resulted in significantly biased estimates, and therefore requires careful handling.

5.5 Conclusion

We built a linear regression model to explain the drivers of SOXX ETF returns. We incorporated a range of explanatory variables including industry-specific ETFs, volatility measures, fluctuations in currency and commodity prices, and several polynomial features derived from original variables. Through hypothesis tests and exhaustive subset selection, we identified the returns of REMX and HACK as the most influential predictors of SOXX returns. We then evaluated key model assumptions such as normality, homoscedasticity, independence, linearity, and multicollinearity and found the model to be largely consistent with these conditions, aside from mild heteroskedasticity, which is typically expected in financial return series.

We also investigated the effects of two missing-data mechanisms: MCAR and MNAR. Under MCAR, the regression remained unbiased, with coefficient estimates and model fit showing minimal deviation from the complete-case model. Under MNAR, however, the parameter estimates became significantly distorted, thus demonstrating that missingness driven by extreme market events or trading restrictions can bias inference even when sample loss is small. We suggested some common strategies for handling such scenarios and emphasized why careful treatment of missing data is essential for reliable model interpretation in explaining financial returns.

BIBLIOGRAPHY

[1] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: With Applications in R.
Springer Texts in Statistics. Springer.
Material consulted: Chapter 3 (pp. 74–117) and Chapter 6 (pp. 218–228).

[2] Tamhane, A. C. and Dunlop, D. D. (2000).
Statistics and Data Analysis: From Elementary to Intermediate.
Prentice Hall, Upper Saddle River, NJ.
General use of the book; no specific chapters referenced.