Assessing the S&P 500 as a relevant predictor of various stocks and macroeconomic factors

IEOR E4150 Project

Joseph High
jph2185@columbia.edu

Introduction

Index funds are designed to track and represent the performance of a pool of underlying stocks. They attempt to duplicate the portfolio of a major stock market index. The *Standard & Poor's 500* (S&P 500) is such an index which tracks the performance of the U.S. stock market. In particular, the S&P 500 is an average of 500 largest companies on the U.S. stock exchange.[1] It is often used as an indicator of the state of the U.S. economy. The list of companies included in the S&P 500 is frequently updated to ensure that its representativeness of the U.S. stock market is as accurate as possible.

This project seeks to assess the relationship between the S&P 500 index fund and stocks of large companies that have yet to be included in the list of companies under the S&P 500 index. We will further this analysis by comparing the S&P 500 index fund to four stocks whose underlying companies are among the largest components that make up the S&P 500 index. Finally, we will assess the whether the S&P 500 is a good indicator (or predictor) of the U.S. macroeconomy. In particular, we will assess the index against the national Housing Price Index (HPI). For the sake of brevity, this will be the only macroeconomic factor assessed as as part of this Report.

Overview of Analysis

In evaluating the relationships between the S&P 500 and the various companies discussed in the introduction, we will asses the similarity between the distributions of the log-returns of the underlying stocks and that of the S&P 500 index fund, in addition to the pairwise correlation between stocks and the S&P 500. Correspondingly, normality of the underlying distributions will also be assessed. For each company, we will fit a simple linear regression model of the log-returns of a underlying stock against the log-returns of the S&P 500 index fund. In doing so, we will assess the predictive capabilities of the S&P 500 on the stock. We will perform a similar set of procedures in our assessment of the S&P 500 and the national HPI, with one minor addition: an assessment of the *price* of the S&P 500 index fund against HPI.

[1] As of the date of this Report, the S&P 500 includes 505 companies (Source: http://www.investopedia.com).

Data

The data used in this analysis includes dailly historical stock prices for the underlying companies assessed and historical S&P 500 index prices for every trading day between May 10, 2019 through April 26, 2022. The data also include monthly average prices for the S&P 500 and historical U.S. National HPI.

The stocks and underlying companies included in our analysis of the S&P 500 and companies not included in the S&P 500 include:

Table 1: Non-S&P 500 Companies

Ticker	Company Name
UBER	Uber
ZG	Zillow
SQ	Block, Inc.
SNAP	Snap, Inc.

The stocks and underlying companies included in the S&P 500 that are assessed include the following:

Table 2: S&P 500 Companies

Ticker	Company Name
AAPL	Apple, Inc
GOOG	Alphabet, Inc. (Google)
MSFT	Microsoft
AMZN	Amazon

All S&P 500 price data and HPI data were sourced from the St. Louis Federal Reserve Economic Database (FRED). Stock prices for each underlying company assessed were attained using the `pandas_data_reader` and `yfinance` libraries in Python.

Data Analysis

For each stock data set, we first compute the log-returns across all time points. Next, we assessed whether the data where sufficiently randomness by conducting a one-way Wald-Wolfowitz runs test in Python using the `runstest_1sample` function from the `statsmodels` library. The Wald-Wolfowitz runs test evaluates a null hypothesis that a given data set was sufficiently randomly sampled against the alternative hypothesis that the data were not randomly sampled. The p-values are as follows:

Data (log-returns)	W-W Runs Test p-value
S&P 500	0.1199
UBER	0.4992
ZG	0.1223
SQ	0.5175
SNAP	0.3382
AAPL	0.5050
GOOG	0.1424
MSFT	0.0662
AMZN	0.6190

At the $\alpha = 0.05$ level of significance, the p-value for each data set is above indicates that we may not reject null hypothesis. That is, the evidence suggests that each data set is sufficiently random.

We can assess the assumption of normality of the log-returns of each stock via visual analysis. In particular, the histogram of each stock is plotted against a superimposed normal density and subsequently compared.

Figures 1 and 2 suggest that the data are approximately normal, given that their histograms resemble a normal density. However, an evaluation of (standard) normal quantile-quantile plots for each stock tells a different story (see accompanied code). In particular, for each log-return data set, the corresponding Q-Q plots suggest that the data are not normally distributed. In particular, the data appear to have slightly heavier right and left tails compared to the (standard) normal distribution.

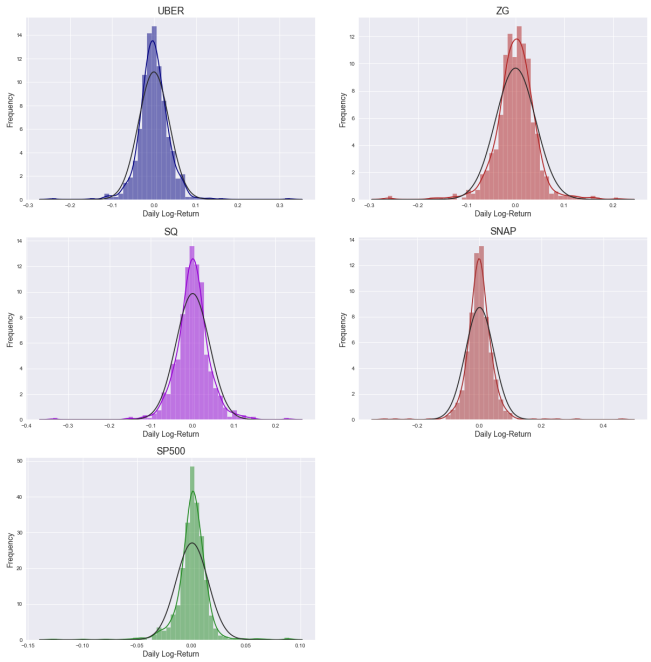

Figure 1: Non S&P 500 Stock Log-Returns Histogram w/ superimposed normal densities

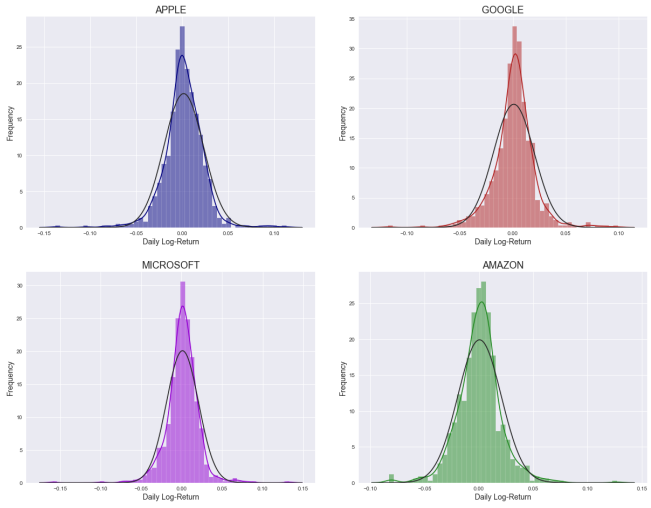

Figure 2: S&P 500 Stock Log-Returns Histogram w/ superimposed normal densities

Evaluating the correlation between the log-returns of each stock and the log-returns of the S&P 500 should give some indication of the existence of a linear relationship between the two, or lack thereof. The Pearson correlation coefficient for each is listed in the table below. We note that the correlation between the companies included in the S&P 500 are substantially greater than that of the non S&P 500 companies. This suggests that linear relationship between the S&P 500 and companies included in its index is much stronger than those that are not. However, this is not a surprising result as these companies contribute to the S&P 500 index fund price.

Table 3: Pearson Correlation Coeff. with SP500 log-returns

Stock (log-returns)	ρ
UBER	0.51
ZG	0.49
SQ	0.62
SNAP	0.41
AAPL	0.80
GOOG	0.79
MSFT	0.85
AMZN	0.61

Testing Equivalence of Means

To test the equivalence of the population means of each stock against that of the S&P 500, we conducted both a one-way ANOVA test and a two-sample t-test. The results are displayed in the table below.

Data sets	t-stat	$Pr(>\|t\|)$	F-stat	p-value
$UBER, SP500$	-0.604	0.546	0.365	0.546
$ZG, SP500$	-0.240	0.811	0.057	0.811
$SQ, SP500$	-0.042	0.966	0.002	0.966
$SNAP, SP500$	0.483	0.629	0.234	0.629
$AAPL, SP500$	1.134	0.257	1.285	0.257
$GOOG, SP500$	0.524	0.600	0.275	0.600
$MSFT, SP500$	0.612	0.541	0.375	0.541
$AMZN, SP500$	0.026	0.979	0.001	0.979

The large p-values for each stock-SP500 pair indicate that we cannot reject the null hypothesis that the means from the two distributions are equivalent, at all reasonable levels of significance.

Linear Regression Analysis

For each stock, we fit a simple linear regression model of the log-returns of the stock against time, i.e.,

$$\log\left(\frac{Stock_t}{Stock_{t-1}}\right)_i = \beta_0 + \beta_1 t + \epsilon_i$$

and a regression model of the log-returns of the stock against the log-returns of the S&P 500. That is, the S&P 500 was the independent variable and each stock the dependent variable, i.e.,

$$\log\left(\frac{Stock_t}{Stock_{t-1}}\right)_i = \beta_0 + \beta_1 \log\left(\frac{SP500_t}{SP500_{t-1}}\right)_i + \epsilon_i$$

The results from the OLS computations of the log-returns of each stock on time are below. The p-values for each well above any reasonable level of significance, indicating that we cannot reject the null hypothesis that the slope on time is zero. That is, there is not enough evidence to conclude a linear relationship between the log-returns of each stock and time, i.e., no linear trend.

Table 4: Summary of OLS Results

Model	Intercept	Slope	$Pr(>\|t\|)$	R^2
$SP500 \sim t$	0.0008	-7.798×10^{-7}	0.756	0.000
$UBER \sim t$	0.0002	-1.619×10^{-6}	0.796	0.000
$ZG \sim t$	0.0049	-1.295×10^{-5}	0.065	0.005
$SQ \sim t$	0.0042	-9.669×10^{-6}	0.160	0.003
$SNAP \sim t$	0.0058	-1.185×10^{-5}	0.128	0.003
$AAPL \sim t$	0.0028	-3.145×10^{-6}	0.391	0.001
$GOOG \sim t$	0.0015	-1.54×10^{-6}	0.640	0.000
$MSFT \sim t$	0.0021	-2.874×10^{-6}	0.397	0.001
$AMZN \sim t$	0.0017	-3.281×10^{-6}	0.337	0.001

The results from the OLS computations of the log-returns of each stock on the log-return of the S&P 500 are below:

Table 5: Summary of OLS Results

Model	Intercept	Slope	$Pr(>\|t\|)$	R^2
$UBER \sim SP500$	-0.0010	1.2789	0.00	0.264
$ZG \sim SP500$	-0.0006	1.3830	0.00	0.244
$SQ \sim SP500$	-0.0003	1.7040	0.00	0.386
$SNAP \sim SP500$	0.0007	1.2842	0.00	0.171
$AAPL \sim SP500$	0.0010	1.1749	0.00	0.646
$GOOG \sim SP500$	0.0004	1.0396	0.00	0.629
$MSFT \sim SP500$	0.0005	1.1517	0.00	0.727
$AMZN \sim SP500$	0.0001	0.8321	0.00	0.373

We first note that the p-value for each was found to be very small (less than 2×10^{-16}), which indicates that we reject the null hypothesis that the corresponding predictor variable (log-returns of S&P 500 in this case) is zero, at all reasonable levels of significance. That is, we can conclude that there is a linear relationship between the log-returns of each stock and the log-returns of the S&P 500 index fund prices.

To test the normality of the residuals, we plot the residuals against the fitted values for each model fit. In all cases, a majority of the residuals are clustered around a single location (approximately), raising suspicion about the validity of each regression model.

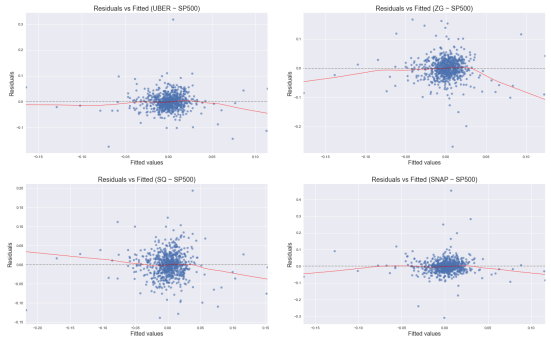

Figure 3: Residual Plots for non S&P 500 companies

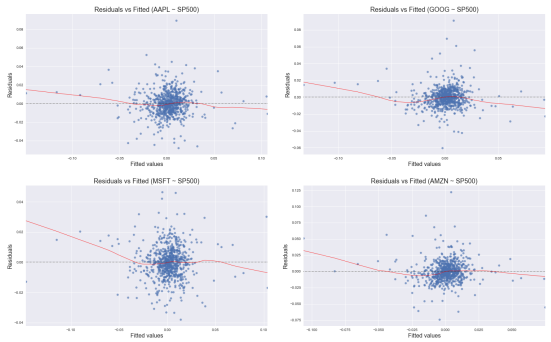

Figure 4: Residual Plots for S&P 500 companies

Before making any conclusions, standard normal Q-Q plots were computed for each set of residuals (see code). In agreement with our findings above, the resulting Q-Q plots indicate that the residuals are not normally distributed. Indeed, each exhibits heavier right and left tails as compared to the normal distribution. Therefore, the assumption that the residuals are normally distributed is not satisfied, and may be indicative of an invalid model fit. However, we will not investigate this any further as part of this project, as it is out-of-scope.

S&P 500 as a Predictor of HPI

In this section we evaluate the monthly average S&P 500 price *and* the log-returns against the U.S. national HPI. Because the data for this analysis are less frequent (monthly), 10 years of data were used to assure a sufficient sample size. More, because we are evaluating whether S&P 500 is an adequate measure of the economy, the time period covered by the data

should include at least one full economic cycle. The national HPI is a measure of housing price movement across the U.S., and is computed on a monthly cadence. Because the S&P 500 is treated as an indicator of the state of the present-state of the economy, it is expected that the S&P 500 and HPI will have a strong relationship. Correspondingly, it is expected that the S&P 500 will be a good predictor of HPI. Figure 5 (below) plots the HPI and the monthly average prices and log-returns of the S&P 500. We note that the monthly S&P 500 prices appear to be substantially more correlated with the U.S. HPI than the log-returns.

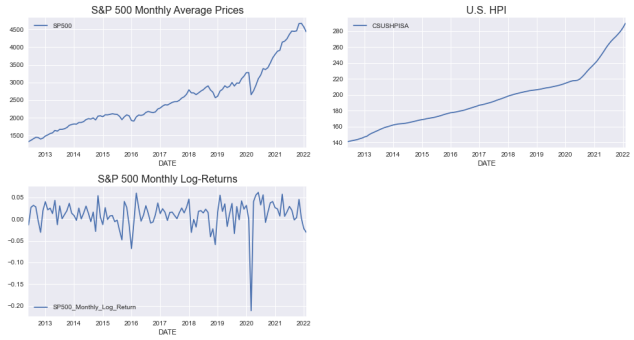

Figure 5: S&P 500 Monthly Avg. Prices, Log-Returns, and HPI (2012 - 2022)

Next, we fit two simple linear regression models, one of HPI on the monthly average S&P 500 prices and another of HPI on the monthly average log-returns of the S&P 500 index fund.

Model	**Intercept**	**Slope**	$Pr(>\|t\|)$	R^2
$HPI \sim SP500$	89.32	0.041	0.00	0.978
$HPI \sim \log\left(\frac{SP500_t}{SP500_{t-1}}\right)$	193.72	4.50	0.965	0.00

In alignment with points made above, the OLS results indicate that the monthly average price of S&P 500 is a much stronger predictor of HPI than is the log-returns of the S&P 500 index fund. Indeed, a R^2 of 0.978 is an indication of a very strong linear relationship compared to the small R^2 exhibited by the regression model on the log-returns. We also note that the p-value corresponding to the hypothesis test that checks whether the dependent variable and the independent (predictor) variable exhibit a linear relationship indicates that SP500 and HPI have a strong linear relationship. That is, we reject the null hypothesis that $\beta_1 = 0$ for the SP500 variable. On the other hand, the p-value for the slope corresponding to the $HPI \sim \log\left(\frac{SP500_t}{SP500_{t-1}}\right)$ model fit is very high (= 0.965). Therefore, we cannot reject the null hypothesis, indicating that there is no linear relationship between HPI and the log-returns of the S&P 500.

The plot of the residuals against the fitted values for the $HPI \sim SP500$ model show some randomness about the horizontal axis at $y = 0$, but not completely. That is, there appears to be a systematic pattern about the $y = 0$ axis, indicating that the assumption of normality of the residuals may not be satisfied.

Figure 6: $HPI \sim SP500$ Residual Plot

The standard normal Q-Q plot of the residuals clears this up, showing that the residuals are not normally distributed. Indeed, the heavy tails at both ends is indicative of deviation from normality. Therefore, while our linear regression model of HPI on SP500 is indicative of strong goodness-of-fit, the model may not be valid. While it is beyond the scope of this project, the strong performance may simply be due to spurious correlations.

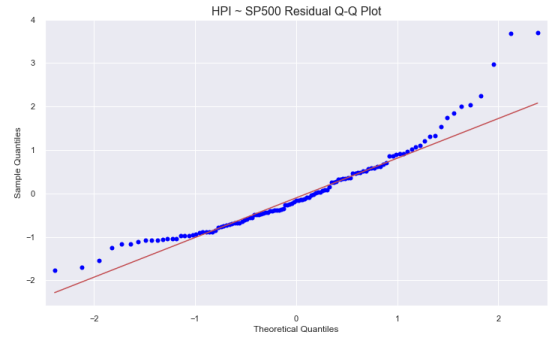

Figure 7: Residual Q-Q Plot for $HPI \sim SP500$

Conclusion

Our analysis revealed that there is a strong (linear) relationship between companies included in the S&P 500 and the S&P 500 itself compared to companies that are *not* included in the S&P 500. While this is not a surprising result, our analysis also indicates that the S&P 500 has some predictive capabilities for companies not included in its index. An even more surprising result was that the S&P 500 index price (rather than the log-returns) has a very strong linear predictive relationship with the national HPI. While we only evaluated the one macro variable, it does give some indication that the S&P 500 is an indicator for the U.S. economic as a whole. Nonetheless, there may be spurious correlation present between the S&P 500 and HPI, giving rise to its strong linear relationship; however, this is beyond the scope of this project.