Regression Analysis of S&P 500 Index Return

Nikhila Kulukuru, Zhi Yun Yap
Submitted for MSCF Investment Empirical Project

1 Intuition behind Choice of Predictors

Under the law of one price, index prices are expected to exhibit mean-reverting nature. In other words, shocks to the equity market should be temporary and the mean of the asset's price is to remain stationary over time. The historical or **lagged return** therefore can be interpreted as the support and resistance of the price fundamental.

As the Fed continues to combat inflation this year, we observed a notable shift in the equity investment landscape. By June 2022, the S&P 500 index slipped into a bear market as higher interest rates incentivized investors to sell assets and take profits. In an efficient market, fair asset prices can be interpreted as the future discounted valuations and **interest rates** should have a negative relationship with securities return.

Based on the Gordon Growth Model, Cutler, Poterba, and Summers (1991) [1] study employed dividend yield, the inverse of the **price-dividend ratio**, as a statistically significant predictive power for international equity return, particularly in the United States where the volatility of real dividend growth is low. JP Morgan Research (2014) [2] further revealed that, for every one percent increase in dividend yield, market prices rise by approximately five percentage points.

2 Data Preparation and Exploratory Analysis

In this project, we extracted monthly S&P index price data from YahooFinance, spot interest rate from FRED, and price-dividend ratio from Shiller's website. From Table 1.1, we observed that the spot interest rate and price-dividend ratio are positively skewed, with their mean larger than the median.

No clear pattern or linear relationship is observed between the monthly index return and the individual predictors (Figure 1.1, rescaled with Min-Max normalization), but this does not conclude their collective predictability on the response. Over time, fluctuations in the spot interest rate and price dividend ratio are associated with the large movement of index prices (Figure 1.2). In particular, the declining spot interest rate in early 2020 is followed by a sharp increase in index return. In mid-2022, a hike in interest rate and a decline in the price dividend ratio led to a downswing in the S&P return.

	Return	Lagged Return	Spot Interest Rate	Price Div Ratio
count	93.000000	93.000000	93.000000	93.000000
mean	0.008514	0.009481	0.875806	4.784236
std	0.045277	0.044041	0.864662	0.724727
min	-0.124871	-0.124871	0.040000	3.770706
25%	-0.016744	-0.015706	0.100000	4.336545
50%	0.014113	0.014113	0.410000	4.572031
75%	0.036198	0.036198	1.600000	4.904108
max	0.126984	0.126984	3.080000	6.670873

Table 1.1: Statistics of the descriptive features and response

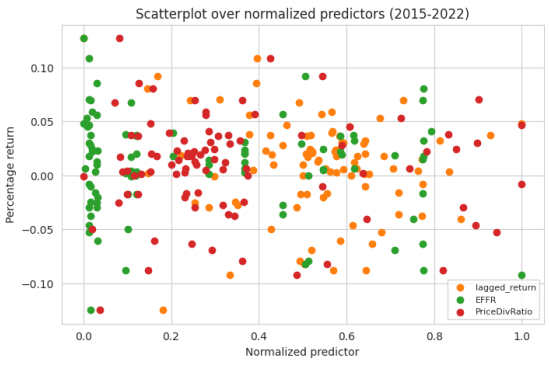


Figure 1.1: Scatterplot of response vs normalized predictor terms

Figure 1.2: Time series plots of predictors and response over time

3 Regression Analysis

From Ordinary Least-square (OLS) regression (Table 1.2), lagged return, with the largest coefficient appears to possess the strongest relationship with the S&P price movement. Concerning the lagged returns, we see that the t-statistic and p-value do not reject the null hypothesis of zero. However, the lagged returns (historical returns) do influence the response variable as per our initial hypothesis. Although both the spot interest rates and the price-dividend ratio variables do not have significant t-statistic and p-value to reject the null hypothesis of zero, their negative coefficients, which denote an inverse relationship with the monthly index return, match our initial hypothesis.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.037
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	1.138
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	0.338
Time:	05:25:10	Log-Likelihood:	158.12
No. Observations:	93	AIC:	-308.2
Df Residuals:	89	BIC:	-298.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0275	0.033	0.826	0.411	-0.039	0.094
lagged_return	-0.1415	0.107	-1.317	0.191	-0.355	0.072
EFFR	-0.0070	0.006	-1.268	0.208	-0.018	0.004
PriceDivRatio	-0.0024	0.007	-0.361	0.719	-0.016	0.011

Omnibus:	8.171	Durbin-Watson:	1.997
Prob(Omnibus):	0.017	Jarque-Bera (JB):	7.765
Skew:	-0.627	Prob(JB):	0.0206
Kurtosis:	3.657	Cond. No.	115.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 1.2: Summary of OLS Regression with all 3 predictors

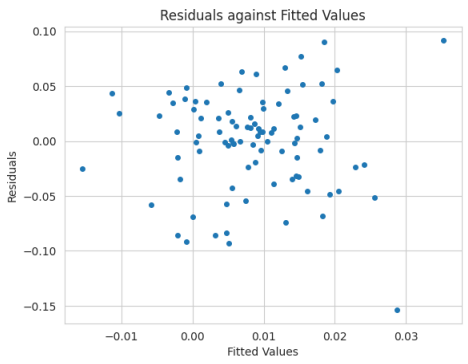


Figure 1.3: Residual plot shows homoscedasticity

The low R-squared value (an R-squared value larger than 0.6 is considered a good fit industry-wide) of the initial multivariate regression model shows that movements of S&P monthly return are not well-explained by the three predictors proposed. In the following sections, we conducted experiments to improve the overall regression model. Additionally, there is no clear pattern observed in the plot of residuals against fitted values (Figure 1.3) and in the plots against all three predictors (Appendix 1.b). The random scattering of error terms confirms that the homoscedasticity assumption of OLS regression holds and that the model is a valid fit. The scatterplots also reveal potential outliers which can be highly influential to the model fit.

We further employed the Cook's distance measure (Table 1.3) to identify particular observations which may be highly influential in OLS regression. We identified 3 exceptionally influential observations with approximately four times larger Cook's distance value than ordinary observations. The high influence of the 2020 data points can be attributed to the unexpected intensity of the Covid pandemic on the market as a whole, while the 2022 data point can be attributed to the aggressive interest Fed rate hike, followed by the job growth that fell short of market expectations this September.

In predicting an asset's return, high or perfect multicollinearity can skew model results, and often lead to invalid beta calculations. One simple way to measure the correlation between predictors in a model is through the Variance Inflation Factor (VIF). It is assumed that if the VIF is relatively high for 2 predictor variables, they are correlated with each other. From the VIF computed (Table 1.3), there is no single variable with an exceptionally large VIF which shows us the model is free from the multicollinearity problem.

Date	Return	Lagged Return	Spot Interest Rate	Price Div Ratio	Cook's Distance
2020-03-31	-0.124871	-0.079166	0.09	3.877358	0.278825
2020-04-30	0.126984	-0.124871	0.04	4.010343	0.183061
2022-09-30	-0.092446	-0.040802	3.08	5.182476	0.140879
2022-07-31	0.092087	-0.082460	1.58	5.352889	0.067783
2019-01-31	0.080065	-0.088048	2.40	4.229892	0.065609
2022-04-30	-0.087769	0.037590	0.33	6.149213	0.061697
2018-12-31	-0.088048	0.018549	2.40	4.197695	0.048061
2020-05-31	0.047645	0.126984	0.04	4.212962	0.042031

Table 1.3: Top 7 Cook's distances in descending order

	Feature	VIF
1	Lagged Return	1.075380
2	Spot Interest Rate	1.915603
3	Price Div Ratio	2.471363

Table 1.4: VIF value of each predictor

4 Model Variation

In this section, we attempted to improvise the initial OLS regression model by introducing additional interactive terms to capture the potential relationship between predictors (Appendix 1a). The new model has an improved R-squared - better explainability of the response variable, but also results in a greater AIC, suggesting that model might be overfitting.

Therefore, we conducted further experiments on simpler models by fitting the data to a combination of bivariate models (Figure 1.5). Amongst them, the combination of spot interest rate and lagged return variables yields the best R-squared, but still "underperforms" the initial model that leverages all three predictors.

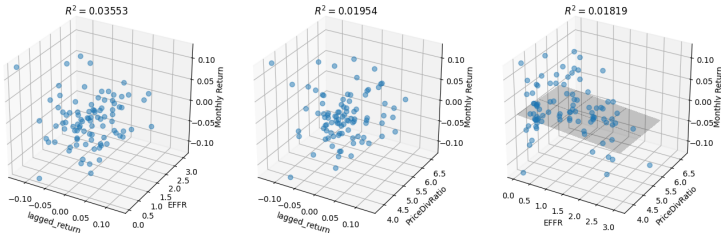


Figure 1.5: Visualizing bivariate regression using combinations of predictors

5 Prediction Analysis

The true test of whether a model is a good fit is to check its predictive power over other observations. To test whether our predictive model is a good fit, we split the initial dataset into training and testing data with a typical 80:20 train-test ratio. Upon re-training the regression model on the training set, the model performance is evaluated using the testing set, against the actual response values (Figure 1.6). In most cases, the model tends to underestimate the variation in the index return, which may be due to the sparsity of data as the sampling period is set at 1 month. This may be resolved by taking more data points or a higher sampling frequency.

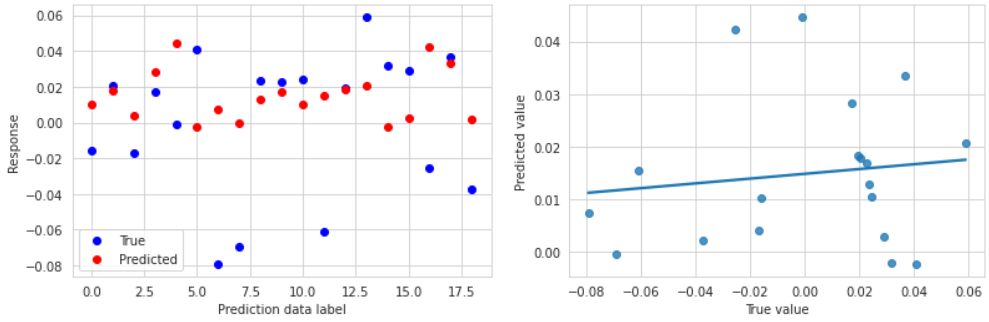


Figure 1.6: Visualizing predicted versus actual response values

We further try to improve our model by introducing second and third degree polynomial predictor variables in the OLS regression model (Appendix 1c). Polynomial variables model the curvature in the data which usually improves line fit. From the OLS regression model, we observe that the new model has an improved R-squared and lower AIC. Some of the new polynomial predictors are also highly significant with p-values less than 0.05 (95% confidence interval). We can thus conclude that including polynomial variables in this model improves its fit.

6 Conclusion

This project is an attempt to empirically investigate the effect of (1) Lagged Return, (2) Interest Rates, and (3) Price Dividend Ratio on the S&P 500 index return using the Linear Regression model. While the model created in this study might not be the best fit for the data, we need to remember that the returns on the S&P 500 are influenced by various macro and micro economic conditions such as natural calamities, political upheaval, exchange rate fluctuations, etc, which cannot be fully captured by only three explanatory variables. In traditional asset pricing, the Fama French three/five-factor model improved CAPM by constructing additional factors to describe stock returns. For future work, we propose adding sector-specific metrics and polynomial attributes to better predict variation in the sector-wide return in the S&P index.

7 References

[1] Cornell, B. (2014). Dividend-price ratios and stock returns: *international evidence*. *The Journal of Portfolio Management*, *40*(2), 122–127. https://doi.org/10.3905/jpm.2014.40.2.122

[2] Cutler, David M., et al. “Speculative Dynamics.” *The Review of Economic Studies*, vol. 58, no. 3, 1991, pp. 529–46. *JSTOR*, https://doi.org/10.2307/2298010. Accessed 11 Oct. 2022.

8 Appendix

(A) Summary of OLS Regression with additional interactive terms

OLS Regression Results

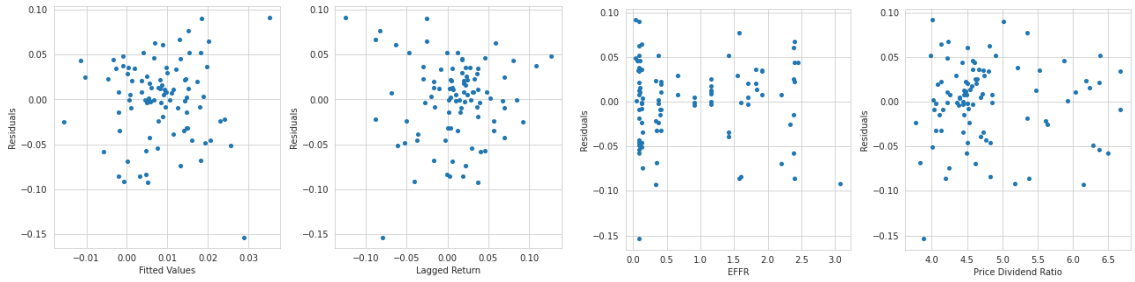
Dep. Variable:	y	R-squared:	0.075
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	1.163
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	0.334
Time:	05:22:36	Log-Likelihood:	160.00
No. Observations:	93	AIC:	-306.0
Df Residuals:	86	BIC:	-288.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0032	0.039	-0.081	0.935	-0.080	0.074
lagged_return	0.6184	0.688	0.899	0.371	-0.749	1.986
EFFR	0.0766	0.061	1.251	0.214	-0.045	0.198
PriceDivRatio	0.0039	0.008	0.497	0.621	-0.012	0.019
lagged_return:EFFR	-0.0567	0.110	-0.515	0.608	-0.275	0.162
lagged_return:PriceDivRatio	-0.1498	0.142	-1.057	0.294	-0.432	0.132
EFFR:PriceDivRatio	-0.0177	0.013	-1.368	0.175	-0.043	0.008

Omnibus:	4.282	Durbin-Watson:	2.036
Prob(Omnibus):	0.118	Jarque-Bera (JB):	3.568
Skew:	-0.406	Prob(JB):	0.168
Kurtosis:	3.511	Cond. No.	1.06e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.06e+03. This might indicate that there are strong multicollinearity or other numerical problems.

(B) Homoscedasticity assumption - plots of residuals versus predictors



(C) Summary of OLS Regression with additional polynomial terms (2nd and 3rd degree terms)

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.309
Model:	OLS	Adj. R-squared:	0.225
Method:	Least Squares	F-statistic:	3.672
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	0.000438
Time:	11:23:42	Log-Likelihood:	173.58
No. Observations:	93	AIC:	-325.2
Df Residuals:	82	BIC:	-297.3
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.4328	1.550	-1.569	0.120	-5.517	0.651
X[0]	-0.2585	0.167	-1.548	0.125	-0.591	0.074
X[1]	-0.0341	0.048	-0.707	0.481	-0.130	0.062
X[2]	1.4215	0.914	1.555	0.124	-0.397	3.240
X[3]	-0.6219	0.155	-4.015	0.000	-0.930	-0.314
X[4]	6.7445	1.701	3.966	0.000	3.362	10.127
X[5]	-27.8806	19.638	-1.420	0.159	-66.947	11.186
X[6]	0.0217	0.043	0.510	0.611	-0.063	0.107
X[7]	-0.0051	0.010	-0.505	0.615	-0.025	0.015
X[8]	-0.2697	0.178	-1.519	0.133	-0.623	0.084
X[9]	0.0168	0.011	1.476	0.144	-0.006	0.039

Omnibus:	1.148	Durbin-Watson:	1.909
Prob(Omnibus):	0.563	Jarque-Bera (JB):	1.135
Skew:	-0.143	Prob(JB):	0.567
Kurtosis:	2.541	Cond. No.	6.37e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.37e+05. This might indicate that there are strong multicollinearity or other numerical problems.