A Dynamic Binary Response Models: Estimation for the S&P500 from 1993 to 2024

Bruno Sciascia

March, 2025

Contents

1 Introduction

We are looking at estimating a binary response (probit) model where the dependent variable is an indicator of positive versus negative excess stock market returns. The basic idea is:

1.1 Static Probit:

$$Y_t = \begin{cases} 1, & \text{if excess return} \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

with

$$\pi_t = \omega + X_{t-1}'\beta,$$

and

$$P(Y_t = 1 \mid I_{t-1}) = \Phi(\pi_t).$$

1.2 Dynamic Extensions:

We can extend this by including lagged dependent variables (or even lagged values of the linear predictor π_t) to account for dynamics:

$$\pi_t = \omega + \sum_{k=1}^{K} \delta_k Y_{t-k} + X_{t-1}'\beta,$$

or an autoregressive structure with lagged π_t

1.3 Our Objectives:

1. Data Selection: To decide on a stock market (e.g. S&P 500 for a U.S. market) and gather both stock return data and macro/industry predictors (like CPI, default spread, dividend yield, volatility, etc.).

2. Data Preparation: To convert continuous returns to a binary indicator (1 if return is non-negative, 0 otherwise) and create lagged variables for the predictors and the dependent variable as required by your model.

3. Model Implementation: First, to implement a basic static probit model using Python libraries (e.g. statsmodels). Then, to show how you would extend it to a dynamic (lagged) version.

4. Machine Learning Angle: Finally, to discuss how ML approaches (e.g. logistic regression, tree-based methods) could be used to capture non-linearities and interactions—and why this might be beneficial.

2 Data

We'll use the S&P 500 as an example market. Since the S&P 500 data is daily and the CPI data is monthly, we convert our stock data to a monthly frequency. Then merge the datasets by their date to macro variables from FRED API.
We'll include all the predictors suggested by the literature, so we'll cover:

- Lagged Excess Market Return (RE)
- Consumer Price Inflation (CPI)
- Default Spread (DSPR)
- Market Dividend Yield (MDY)
- Market Volatility (MVOL)

2.1 Lagged Excess Market Return (RE)

This is computed as the difference between the market return and a risk-free rate. For the market return, we use the S&P 500 returns (which you already have). For the risk-free rate, we can use the 3-Month Treasury Bill rate from FRED.

2.2 Consumer Price Inflation (CPI)

The CPI measures the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.

2.3 Default Spread (DSPR)

The default spread is typically the difference between yields on corporate bonds and government bonds, or between LIBOR and Treasury yields (TED spread). It is a measure of perceived credit risk in the market. We will use TED spreatd in this case.

2.4 Market Dividend Yield (MDY)

The dividend yield reflects the dividends paid relative to the market price and is often considered a measure of valuation. We can compute the dividend yield from an ETF that tracks the S&P 500 (such as SPY) by downloading its dividend history from Yahoo Finance.

2.5 Market Volatility (MVOL)

Market volatility is often measured using the VIX index, which reflects expected volatility in the S&P 500 over the next 30 days.

This comprehensive set of predictors aligns with the literature (e.g., Hong et al. (2007)) and should give us a strong basis for modeling stock market return directions

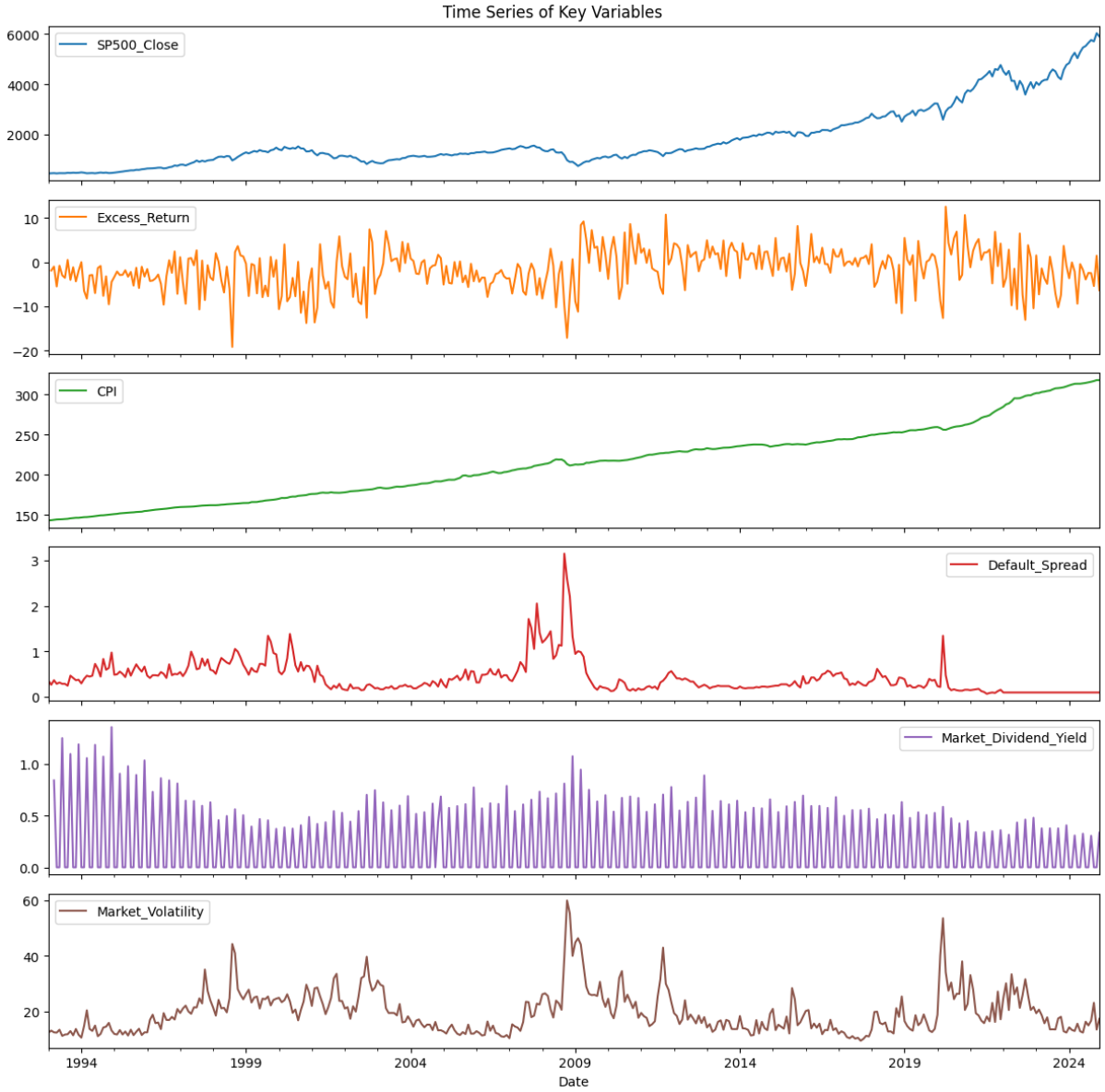

Figure 1: Time Series Plots

3 Model Construction

Create the Binary Response Variable:
Our dependent variable is defined as follows:

$$Y_t = \begin{cases} 1, & \text{if the excess return} \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Generate Lagged Predictor Variables: The model uses the information available at time $t-1$ (denoted as X_{t-1}). We need to shift our predictors by one period. Here, we create lagged versions for each predictor:

- CPI

- Default Spread
- Market Dividend Yield
- Market Volatility

3.1 Build the Static Probit Model

We will now specify the static probit model.Recall that in the static specification, the model is defined as:

$$\pi_t = \omega + X'_{t-1}\beta,$$

and

$$P(Y_t = 1 \mid I_{t-1}) = \Phi(\pi_t).$$

Table 1: Result of static probit model

Variable	**Coefficient**	**Erreur Std.**	**z**	**p-value**	**[0.025]**	**[0.975]**
const	-1.7577	0.431	-4.077	0.000	-2.603	-0.913
CPI_lag	0.0037	0.002	2.259	0.024	0.000	0.007
Default_Spread_lag	-0.6862	0.239	-2.874	0.004	-1.154	-0.218
Market_Dividend_Yield_lag	0.2977	0.222	1.339	0.181	-0.138	0.733
Market_Volatility_lag	0.0429	0.010	4.442	0.000	0.024	0.062

3.1.1 Detailed Analysis

We now provide a detailed analysis of the model output.

- **Model Fit:**
  The pseudo R-squared is 0.06909. While pseudo R-squared values are not directly comparable to the R^2 in OLS, this value provides a measure of the model's explanatory power relative to a null model.
- **Log-Likelihood:**
  The log-likelihood of the fitted model is -234.70, compared to the null model log-likelihood of -252.12. A higher log-likelihood (less negative) for the fitted model indicates an improvement over the null model.
- **Coefficient Estimates:**

- **Constant:** The constant is estimated at -1.7577 (standard error = 0.431) and is statistically significant with a p-value of 0.000. This indicates that when all predictors are at zero, the baseline log-odds of a positive excess return is significantly negative.
- **CPI_lag:** The coefficient for the lagged Consumer Price Index is 0.0037 (standard error = 0.002) with a p-value of 0.024. This positive coefficient suggests that an increase in the lagged CPI is associated with a higher probability of a positive excess return.
- **Default_Spread_lag:** The lagged default spread has a coefficient of -0.6862 (standard error = 0.239) and is statistically significant ($p = 0.004$). The negative sign implies that a higher default spread reduces the likelihood of a positive excess return.
- **Market_Dividend_Yield_lag:** The coefficient for the lagged market dividend yield is 0.2977 (standard error = 0.222), but it is not statistically significant at conventional levels ($p = 0.181$). Although the positive sign indicates a positive relationship, the lack of significance suggests that its effect may not be reliably different from zero.
- **Market_Volatility_lag:** The coefficient for the lagged market volatility is 0.0429 (standard error = 0.010), with a highly significant p-value (0.000). This indicates that higher market volatility is associated with an increased probability of a positive excess return.

- **Overall Model Significance:**
  The likelihood ratio test (LLR p-value) is 5.012e-07, which confirms that the model is statistically significant overall.

In summary, our static probit model shows that lagged CPI, default spread, and market volatility significantly influence the probability of a positive excess return, while the effect of market dividend yield is not statistically significant. The model provides a statistically significant improvement over the null model, as indicated by the likelihood ratio test.

3.2 Extending to a Dynamic Probit Model:

To account for dynamics—such as persistence in the stock market behavior—we can include lagged values of the dependent variable. For example, a dynamic specification might include the first lag of Y (denoted Y_{t-1}):
Table 2 summarizes the estimated coefficients from the dynamic probit model:

Table 2: Dynamic Probit Model Regression Results

Variable	Coefficient	Std. Err.	z	p-value	[0.025]	[0.975]
const	-1.8644	0.440	-4.236	0.000	-2.727	-1.002
CPI_lag	0.0031	0.002	1.876	0.061	-0.000	0.006
Default_Spread_lag	-0.6618	0.240	-2.753	0.006	-1.133	-0.191
Market_Dividend_Yield_lag	0.3433	0.226	1.522	0.128	-0.099	0.785
Market_Volatility_lag	0.0458	0.010	4.666	0.000	0.027	0.065
Y_lag1	0.3917	0.142	2.760	0.006	0.114	0.670

3.2.1 Detailed Analysis

We now discuss the key findings from the dynamic probit model:

- **Lagged Dependent Variable (Y_{t-1}):**
  The inclusion of Y_{t-1} allows us to capture the persistence in stock market behavior. The coefficient of Y_{t-1} is estimated at 0.3917 and is statistically significant ($p = 0.006$). This suggests that past market performance significantly increases the probability of a positive current excess return.

- **Lagged CPI (CPI_lag):**
  With a coefficient of 0.0031 (marginally significant at $p = 0.061$), the lagged CPI appears to have a positive relationship with the probability of a positive excess return, though the effect is relatively weak.

- **Lagged Default Spread ($Default\_Spread\_lag$):**
  The coefficient of -0.6618 ($p = 0.006$) indicates that a higher default spread reduces the likelihood of a positive excess return. This result is statistically significant and aligns with our expectations about market stress conditions.

- **Lagged Market Dividend Yield ($Market\_Dividend\_Yield\_lag$):**
  The estimated coefficient of 0.3433, although positive, is not statistically significant ($p = 0.128$), implying that its effect on the market direction may be less pronounced or subject to greater variability.

- **Lagged Market Volatility ($Market\_Volatility\_lag$):**
  With a coefficient of 0.0458 and a high level of significance ($p = 0.000$), market volatility plays a crucial role in increasing the probability of a positive excess return.

- **Overall Model Fit:**
  The log-likelihood of the fitted model is -230.59, and the pseudo R-squared is 0.08368. These statistics, along with the significant likelihood ratio test (LLR p-value = 5.584e-08), indicate that incorporating dynamic elements improves the model performance over a static specification.

4 Machine Learning Discussion

In addition to traditional econometric models like the static and dynamic probit, machine learning (ML) methods offer a powerful alternative for modeling binary outcomes in complex financial environments. This section explains both how to implement such an approach and presents arguments for its use.

Arguments for Using Machine Learning

- **Capturing Non-linear Relationships:** Traditional probit models assume a linear relationship in the predictor space. ML methods, however, can naturally uncover complex non-linearities and interactions, which are often present in financial market dynamics.
- **Handling High-dimensional Data:** Financial datasets may include a large number of predictors. ML algorithms such as regularized regressions (e.g., LASSO) and tree-based methods effectively perform feature selection, reducing overfitting and improving model robustness.
- **Enhanced Predictive Accuracy:** By using ensemble methods and non-linear models, ML approaches often achieve superior out-of-sample forecasting performance. This is particularly valuable in financial markets, where predictive accuracy is critical.
- **Flexibility and Adaptability:** ML techniques impose fewer assumptions about the underlying data distribution. This flexibility allows them to adapt to various data patterns and incorporate alternative data sources (such as news sentiment or social media signals) that might further enhance prediction.
- **Integration with Traditional Methods:** The combination of ML and classical econometric models provides a comprehensive toolkit. While ML improves predictive performance, econometric models facilitate a clear interpretation of parameter effects, ensuring that both accuracy and insight are achieved.

How to Implement a Machine Learning Model

4.0.1 1. Data Preparation and Feature Engineering:

- **Preprocessing:** Convert the continuous excess return data into a binary variable (1 if return ≥ 0; 0 otherwise). Ensure that data from different sources (e.g., daily S&P 500 returns and monthly CPI) are aligned by aggregating or resampling as needed.
- **Generating Lagged Variables:** Just as in the probit models, create lagged versions of key predictors (such as CPI, default spread, dividend yield, and market volatility) and,

if applicable, the dependent variable. Additionally, consider interaction terms or rolling statistics (e.g., moving averages, rolling volatility) to capture dynamic patterns.

4.0.2 2. Model Selection and Training:

- **Baseline Model – Logistic Regression:** Begin with logistic regression as a benchmark, since it closely mirrors the binary outcome framework.
- **Non-linear and Ensemble Methods:**
  - **Decision Trees and Random Forests:** These models automatically capture non-linearities and interactions, providing feature importance insights.
  - **Gradient Boosting Machines:** Tools like XGBoost or LightGBM offer robust performance, especially in capturing complex patterns in financial data.
  - **Support Vector Machines (SVM):** SVMs with non-linear kernels can be useful for high-dimensional feature spaces.
  - **Neural Networks:** Feed-forward neural networks or deep learning architectures are capable of modeling very complex relationships, although they require larger datasets and more computational resources.

4.0.3 3. Model Validation and Tuning:

- **Data Splitting:** Divide the dataset into training, validation, and test sets (or use cross-validation) to ensure robust performance evaluation.
- **Hyperparameter Optimization:** Employ grid search or random search methods to tune hyperparameters, thereby enhancing model performance.
- **Performance Metrics:** Use classification metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to assess the model's predictive power.

4.0.4 4. Model Interpretation and Integration:

- **Interpretability Tools:** Leverage techniques like feature importance plots, SHAP values, or LIME to interpret the influence of individual predictors.
- **Hybrid Models:** Combine ML methods with traditional econometric approaches to benefit from both high predictive accuracy and interpretability. For instance, ML models can be used for variable selection or to capture non-linear effects, while econometric models can provide insights into the magnitude and statistical significance of predictors.

5 Conclusion

This report has presented an empirical application of dynamic binary response models to the S&P500, focusing on modeling the direction of excess stock market returns. The analysis demonstrated that key predictors such as the Consumer Price Index, default spread, and market volatility significantly influence market behavior, with dynamic models capturing the persistence in returns more effectively than static specifications.

The incorporation of machine learning approaches was also discussed as a complementary strategy, offering the potential to capture non-linearities and interactions that traditional models might overlook. These techniques could enhance forecasting performance, especially in complex and high-dimensional financial environments.

In summary, while the traditional probit framework provides a strong basis for understanding the determinants of stock market movements, integrating machine learning techniques presents an exciting opportunity for further improvement in predictive accuracy and model robustness. Future work could focus on developing hybrid models that leverage the interpretability of econometric methods alongside the flexibility and power of machine learning, thereby offering a more comprehensive toolkit for financial market analysis.