Predicting Volatility Using Regression

Cindy Xu and Rohan Shukla

GitHub: `https://github.com/rohanshukla5/time-series-analysis`

1 Project Overview

Volatility is a measure of how much an asset's price changes over time. In particular, the higher the volatility is for an asset, the more we can expect its price to change more frequently and drastically. In contrast, the lower the volatility is, we expect to see smaller changes in the asset's price and more stable price movements. Understanding volatility is a important for investors and traders to better understand and assess the risk and uncertainty of an asset.

Our project intends to use statistical models to predict the volatility of the stock market using two different measures of volatility: the implied volatility and the realized volatility. The implied volatility is a measure of how much we expect the market to change in the future, while the realized volatility is a measure of how much the market has already changed. In other words, implied volatility is looking forwards in time while realized volatility is looking backwards in time and summarizing what we've seen changed in the market. *(More details about the specifics of these measures in the next section.)*

The project is split into two parts. The first part is about how we can predict implied volatility (VIX) by solely regressing on the realized volatility. Here, we fit regression models that don't take the time dependency into consideration to see whether time plays a role in the relationship between the VIX and the realized volatility. These models include ordinary linear regression models, generalized additive models (GAMs), and kernel regression models. The second part is about how we can predict realized volatility by regressing on both trailing realized volatilities and the VIX. Here, since we are including trailing realized volatilities as our additional predictor variables, we are adding back the time dependency into our regression models to see which model can best capture the dynamic behavior of the realized volatility. These models include seasonal autoregressive integrated moving average (SARIMA) models, ordinary linear regression models, and lasso regression models.

1.1 VIX Context and Overview

The Volatility Index (VIX) is a measure of the implied volatility of the stock market based on S&P500 index options. Specifically, there are four main VIX tickers tracked on CBOE's global markets database: VIX (30 days), VIX3M (90 days), VIX6M (180 days), and VIX1Y (365 days). Each is calculated using option prices with expiration dates approximately matching the horizon.

The VIX represents the market's expectation of annualized volatility over the next 30 days. For example, a VIX reading of 20 suggests an expected $\pm 20\%$ move in the S&P500 index on an annualized basis, corresponding to roughly $\pm 5.7\%$ over the next month.

In addition to implied volatility, we also consider *trailing realized volatilities*, which summarize past variability in returns, using measures like rolling window standard deviations and exponentially weighted moving averages (EWMA).

2 SARIMA Model

Autoregression is when we try to regress a time dependent variable X_t on its past values. While there are many different types of autoregressive models, we will be focusing primarily on the SARIMA models for this project to capture both trend and seasonal patterns in the VIX data.

A SARIMA model extends ARIMA by incorporating seasonal components. It models the data as:

$$\Phi(L^s)\,\phi(L)\,\Delta^d\Delta_s^D X_t = \Theta(L)\,\Theta(L^s)\,\varepsilon_t,$$

where L is the lag operator, Δ^d denotes differencing, s is the seasonal period, and ε_t is white noise. This allows us to capture both short-term autocorrelation and repeating seasonal patterns like weekly cycles.

3 Other Regression Models

Other than fitting autoregressive models to predict VIX, we also used linear regression as well as some non-parametric regression models. Non-parametric models are ones that don't assume a predetermined form of the relationship between the predictor variables and the response variable. This allows for more flexibility and complex relationships without imposing as many or any assumptions about the data. While these regression models don't take the time dependency of the data into consideration, they are still powerful tools we can use to analyze the relationship between VIX and the volatility of the S&P500 prices.

3.1 Generalized Additive Model

One of the non-parametric regression models used in this project is a generalized additive model (GAM). Recall that in a linear regression model, if we have predictor variable $X = \begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}^T$ and $\mu(x) = \mathbb{E}[Y \mid X = x]$ is the true regression function where $Y = \mu(X) + \epsilon$, then we are imposing the assumption that μ is of the form $\mu(X) = \beta_0 + \sum_{t=1}^n \beta_t X_t$. In other words, we are assuming that we have a linear and additive relationship between our predictors and the response variable. In a GAM, we loosen these assumptions by only assuming additivity. In other words, we are now assuming that there exists some smooth functions $f_1, \ldots, f_n$ such that $\mu(X) = \sum_{t=1}^n f_t(X_t)$. So while in linear regression, we are trying to estimate the coefficients β_t, in a GAM, we are instead trying to estimate the functions f_t which may or may not be linear functions.

When used in practice, we aren't able to output an estimated formula or definition of $\widehat{f_t}$ the same way we are able to output an estimate of $\widehat{\beta_t}$ in linear regression. However, we are able to plot the estimated function $\widehat{f_t}$ for each variable X_t to visualize the relationship of each X_t with Y.

3.2 Kernel Regression

Another non-parametric regression model we used is kernel regression, otherwise known as kernel smoothing or Nadaraya-Watson smoothing. This is a type of linear smoother.

Linear smoothers are a popular way of estimating the true regression function. By definition, $\widehat{\mu}$ is a linear smoother that estimates μ if there exists a function $w : \mathbb{R}^2 \to \mathbb{R}$ such that $\widehat{\mu}(x) = \sum_{i=1}^n y_i w(x_i, x)$. In other words, $\widehat{\mu}$ is estimating μ by taking a weighted average of the observed data and w is known as the weight function. Ordinary linear regression is a famous example of a linear smoother where

$$w(x_i, x) = \frac{1}{n} + \frac{(x_i - \bar{x})(x - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

Kernel regression is another example of a linear smoother where we define its weight function to be

$$w(x_i, x) = \frac{K(\frac{x - x_i}{h})}{\sum_{j=1}^n K(\frac{x - x_j}{h})}$$

Here, $h > 0$ is the bandwidth which helps scales the distance $x_i - x$ that we are inputting and K is some kernel function. In practice, K is typically chosen to be a probability density function (PDF). The most commonly used kernel function is the standard Gaussian PDF, which is

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

Note that for kernel smoothing, we don't make any assumptions about the data. This approach is fully non-parametric and allows the data to determine the shape of the fit.

3.3 Lasso Regression

In addition to standard regression, we implement Lasso regression, which introduces an L_1 penalty on the regression coefficients:

$$\min_{\beta} \sum_{t} (Y_t - \beta_0 - \sum_{j} \beta_j X_{t,j})^2 + \lambda \sum_{j} |\beta_j|.$$

The penalty term shrinks some coefficients toward zero, effectively performing feature selection and reducing multicollinearity. This improves model interpretability and can prevent overfitting when many predictors are involved. We select the optimal penalty parameter λ via time-series cross-validation.

References

[1] V. Lyubchich and Y. R. Gel. *Time Series Analysis: Lecture Notes with Examples in R.* 2023-09 edition, 2023.

[2] C. R. Shalizi. *Advanced Data Analysis from an Elementary Point of View.* 2025.