**[S1]** Hello, and welcome. Today, I’ll present two machine learning models used to predict future stock volatility.

The dataset is divided into 10-minute chunks, called time\_IDs. But because volatility can shift even within those intervals, I subdivided them into 30-second buckets to capture more detailed changes. The models use 80% of the data train, finding patterns, and 20% to test, evaluating performance. This is the industry standard.

The first model, is a linear regression model, tuned with the default settings in R. This model predicts a dependent variable – here, volatility – by fitting a linear equation to the independent variables. For example, an equation to evaluate a quant trader’s skill could be: 0.4×Math + 0.3×Programming + 0.2×Finance. Applied to this dataset, the model uses a price, bid ask spread, and number of orders, to predict volatility.

**[S2]** These variables were derived from bid and ask price and sizes in the dataset. [**S3]** After training, the resulting formula is: volatility = -7.6×10-3\*price + 1×10-7×order – 4.8×10-2×BidAskSpread + 8×10-3.

**[S4]** The model’s average correlation coefficient – a measure of how well it explains variance in the data – is 0.24, indicating weak positive correlation. This may be because the model assumes data points are independent, which isn’t true for time series data. The model also assumes linear relationships between independent and dependent variables.

**[S5]** The second modela Generalized Additive Model (GAM), relaxes this constraint. It fits smooth functions for each variable, instead of a linear equation, helping to capture more complex patterns. I used the mgcv package, with “REML” smoothing, to avoid overfitting. The GAM yields a slightly higher correlation coefficient of 0.3, which is still weak.

**[S6]** Furthermore, when visualising a select GAM model’s smooth function, they are mostly linear, suggesting the extra complexity may not be necessary.

**[S7] In terms of** Mean Squared Error (MSE), regression has a slightly lower error. So the regression model is simpler, performs better, and suits the data. Hence, I choose it as the final model.

That said, both models still had low correlation scores. To mitigate this, time-aware models—like ARMA-GARCH—might be a better fit.

Thank you.