MATH-GA 2708.001 Algorithmic Trading and Quantitative Strategies

Homework 1 – Market Impact & Optimal Trading

Haiyi Yue (hy2777)

1. Overview

We developed and analyzed a market impact model using publicly available Trade and Quote (TAQ) data. Our goal is to estimate the permanent and temporary components of market impact by conducting cross-sectional non-linear regression across a broad set of liquid stocks.

This model can help inform execution strategies by quantifying how trading size and timing affect cost. It provides a data-driven foundation for improving algorithmic trade execution and minimizing market impact risk.

2. Assumptions

i. Trade Direction Inference

We use the tick test to infer trade direction, assuming classification errors are uncorrelated and cancel out over time, especially during noise-dominated trading days.

ii. Value over Volume

To avoid distortions from stock splits, we use dollar-based metrics (VWAP × target) for traded value and imbalance, ensuring consistency across stocks and time.

iii. Model Uniformity

We assume a common impact model applies across all stocks, enabling cross-sectional regression rather than stock-specific models.

iv. Stable Market Conditions

The model is assumed to be most valid on low-volatility days. We test sensitivity by excluding volatile days where assumptions may not hold.

v. Intraday Stationarity

Metrics like arrival price, VWAP, and volatility are assumed to be stable over the intraday window (9:30–4:00), allowing for consistent input computation.

vi. No Order Anticipation

We assume trades are not anticipated or front-run by the market, so the observed impact reflects only the trade itself.

3. Data Preprocessing Pipeline

The data preparation process consists of two main stages: feature extraction from TAQ quote/trade data (dataLoader.py) and final dataset construction (dataPreprocessor.py). Together, they generate the input for cross-sectional regression estimation of the market impact model.

i. Data Loading

We apply a liquidity filter to select a universe of stocks that ensures meaningful and reliable impact modeling. This step is implemented in *stockList.py*. We calculate the average daily volume for each stock across all dates, and ranked them by average volume. The top 1500 most liquid stocks are selected as the candidate universe for model building.

We then read binary TAQ files using given reader classes (TAQTradesReader, TAQQuotesReader). The pipeline iterates through each stock and trading date, skipping over those without valid quote or trade files.

ii. Feature Extraction

For each valid stock-day pair, we extract the following metrics:

- **Mid-quote Return**: Computed from 2-minute buckets between 9:30 AM and 4:00 PM using ReturnBuckets. This serves as a volatility proxy.

- **Total Daily Volume**: Aggregated by summing trade sizes from the entire day.

- **Arrival Price**: Calculated as the average of the first five mid-quote prices after 9:30 AM.

- **Terminal Price**: Average of the last five mid-quote prices before 4:00 PM.

- **VWAP (Value Weighted Average Price)**:

    - **VWAP_1**: From 9:30 AM to 3:30 PM, used for model regression.

    - **VWAP_2**: From 9:30 AM to 4:00 PM, used to scale dollar imbalance and daily traded value.

- **Imbalance**: Computed using the TickTest classifier. Each trade is assigned a direction (buy/sell), and the net signed volume is computed. This raw imbalance is later scaled by VWAP to obtain dollar imbalance.

Feature data is stored in dictionaries and then exported as CSV files, one per feature (e.g., arrival_price.csv, VWAP_1.csv). These CSV files are used as input for the dataPreprocessor.py. Stocks with missing values were dropped (30 out of 1491).

In the second stage, the preprocessed CSV feature files are loaded and transformed into a unified tabular format suitable for regression analysis. This step is handled in dataPreprocessor_new.py.

iii. Feature Computation

To align with the modeling framework and avoid distortions from corporate actions like stock splits, we compute:

- **Daily Traded Value** = VWAP_2 × Daily Volume

- **Dollar Imbalance** = VWAP_2 × Imbalance

- **Volatility** = Standard Deviation of 2 minutes mid-quote returns computed using last 10 days of data and scaled to a daily value

iv. Data Integration

Each cleaned feature matrix is reshaped from wide (stocks × dates) to long format (stock, date, value). The fully assembled dataset is exported as a single CSV file (input.csv), with one row per stock-day. This dataset serves as the input for nonlinear cross-sectional regression used to estimate the model parameters η and β.

4. Model Estimation

In impactModel.py, we estimate the parameters of the market impact model using cross-sectional nonlinear regression:

$$h = \eta\sigma\left(\frac{X}{(6/6.5)\mathrm{V}}\right)^{\beta}$$

where:

- h : Temporary impact, calculated from VWAP_1 (9:30–3:30) and the arrival price.

- σ: Intraday volatility

- X : Dollar imbalance

- V : Dollar volume traded

- T : Fraction of trading day

- η, β: Parameters to estimate

We use *scipy.optimize.curve_fit* to perform nonlinear least squares estimation of η, β.The regression is performed on the full dataset across all stocks and dates. Initial parameter guesses were set to η = 0.142 , β = 0.6.

The estimated parameter values are:

Estimated eta = 2.3489, beta = 0.3007

If we separate the volatile days (volatility>=0.038, 8116 out of 79128) with the normal days, we got the estimated parameter values as:

Normal days: eta = 3.0983, beta = 0.3546

Volatile days: eta = 1.0850, beta = 0.1738

We observe that η and β are significantly higher on normal days compared to volatile days. This suggests that on normal days, the market is more sensitive to trade imbalances. The higher β on normal days also implies a more nonlinear relationship, where large trades incur disproportionately higher costs. In contrast, during volatile periods, the market appears to absorb trades more easily, likely because noise and informational volatility dominate price movements, reducing the relative effect of individual trades.

5. Diagnostics and Sensitivity Analyses

i. Bootstrap

We considered two bootstrap methods for estimating standard errors and testing the significance of our model parameters:

- Paired Bootstrap: This method resamples full observation tuples (Z, σ, h), preserving the joint distribution and allowing for arbitrary forms of heteroskedasticity. Since this approach does not assume any specific error structure, it is robust to heteroskedasticity, making it the appropriate choice for our data.

- Residual Bootstrap: This method resamples residuals and re-adds them to fitted values. However, it assumes homoskedastic and identically distributed errors. Given the strong evidence of heteroskedasticity in our residuals, using residual bootstrap would violate key assumptions and produce unreliable inference. Therefore, we did not use it.

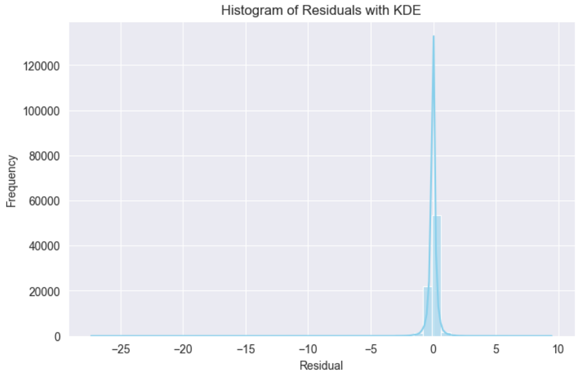
Based on 500 bootstrap samples, we estimated t-statistics of 4.22 for η and 3.66 for β. Both values are well above conventional thresholds for significance (|t| > 2), indicating that the estimated parameters are statistically different from zero. This supports the model's core assumption that trade imbalance and volatility are key drivers of temporary market impact.

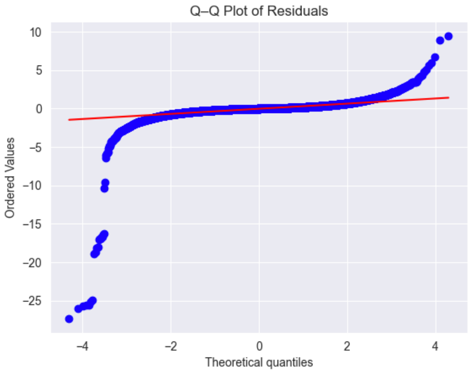
ii. Residual Analysis

To assess the validity of our market impact model, we analyzed whether the residuals satisfy the standard assumptions required for reliable inference in nonlinear least squares estimation.

a. Normality

We plotted histograms and Q-Q plots of the residuals.





The residuals deviate from a normal distribution, showing a sharp peak at zero and heavy tails, particularly on the left side. The Q-Q plot confirms this by highlighting significant departures from the diagonal in both tails, indicating the presence of extreme values.

While the distribution is not symmetric and exhibits leptokurtosis, such behavior is typical in high-frequency financial data. These deviations do not invalidate the model but reinforce the importance of using robust standard errors and bootstrap methods to ensure valid inference.

b. Zero Mean

Mean of residuals: -0.02595. The mean was close to zero, indicating that the model does not systematically overestimate or underestimate the temporary impact, and the residuals are unbiased on average.

c. Homoskedasticity

Visual inspection of residuals plotted against fitted values initially revealed increasing variance with higher predicted impact, suggesting potential heteroskedasticity. To formally assess this, we applied White's test, which yielded a test statistic of 17.84 with a p-value of 0.0032. This result leads us to reject the null hypothesis of homoskedasticity, indicating that the residual variance is not constant.

Overall, while the residuals exhibit some deviations from ideal assumptions in terms of variance, they are sufficiently well-behaved for model inference using robust techniques.

iii. Active Stocks vs. Inactive Stocks

We split the dataset into equally sized subsets based on daily traded value to compare market impact dynamics between active and inactive stocks.

Inactive stocks: eta = 2.6948, beta = 0.5515
Active stocks: eta = 4.9144, beta = 0.3760

The estimated η for active stocks (4.91) is substantially higher than that for inactive stocks (2.69). This suggests that, for the same normalized trade size, active stocks exhibit a stronger temporary price impact. One possible explanation is that active stocks are more tightly monitored and rapidly repriced by market participants, so even moderate order imbalances can move prices more. The estimated β is higher for inactive stocks (0.55) than for active stocks (0.38), indicating that market impact grows more nonlinearly with trade size in less active stocks.

iv. Heteroskedasticity Test

(Done and interpreted in ii)

6. Challenges and Issues

i. Winsorize Extreme Inputs:

Market impact data often contain extreme outliers, particularly in normalized imbalance (Z) and volatility (σ). Without winsorization, a small number of large trades or data errors can exert undue influence on parameter estimates. We addressed this by trimming the top and bottom 10% of Z and σ, improving both interpretability and statistical robustness.

ii. Taking Absolute Value in Fitting

The model form uses $\eta \times \sigma \times |Z|^\beta \times \mathrm{sign}(Z)$ to handle nonlinear scaling while preserving trade direction. However, applying the absolute value alters the shape of the response and may obscure potential asymmetries between buy and sell side impacts. Although this form is numerically stable and common in the literature, it implicitly assumes symmetric impact curvature, which may not fully reflect real market behavior, especially in less liquid names.

iii. Cross-Sectional Model Assumption

Our model assumes a single functional form (same η and β) applies across all stocks in the cross-section, which simplifies estimation but may ignore stock-specific microstructure effects. In reality, market impact dynamics can differ across sectors, tick sizes, and trading regimes. Although we partially address this by segmenting on volatility and activity, the global model may still mask heterogeneity, especially in tail behavior.