Estimating Transaction Cost Models

November 11, 2024

1 Project Overview

- Transaction costs represent the difference between the expected and actual prices of trades. For example, when purchasing a stock, the price paid is typically higher than the prevailing market price at the time the order was submitted.

- This data science project focuses on analyzing transaction costs (or ”t-costs”) incurred during actual trade execution using a sample of 80,000 trades. Accurate estimation of t-cost models is crucial in active investment management, which often requires ongoing rebalancing and trading as views change. Many investors, not just those who are quantitative, consider t-costs to be a vital aspect of their investment process, on par with that of alpha forecasting. Only rarely can such costs be avoided, and they often have a first-order impact on performance. Hence a well-estimated t-cost model is essential for balancing the anticipated returns of an asset against the expected trading expenses necessary to capitalize on those returns.

- For this exercise, we use a dataset consisting of three variables: the trade size (measured in terms of estimated percent of ADV), the annualized volatility of the traded asset (from a risk model), and also the realized cost (measured as the difference between the immediate pre-trade and post-trade asset prices).

- Using these data, we fit four standard t-cost models to the data, and report parameter estimates and fit metrics. The models include non-linearities to reflect the increasing impact of larger trades.

- A preliminary survey of the data revealed gross outliers. From prior work we are aware that standard techniques (e.g., non-linear regression based on MSE) are vulnerable to and can be frustrated by outliers. Therefore, in addition to non-linear regression, we optimized all four models using an MAE metric.

- The model fit metrics include, for example, standard errors for the parameters and a pseudo R2.

2 The Four Models

If C_i, S_i, and V_i are the (realized) trading cost, trade size, and asset volatility as described above, the four models we estimate are:

$$C_i = \beta_0 + \beta_1 S_i^{0.5} + \epsilon_i \tag{a}$$
$$C_i = \beta_0 + \beta_1 S_i + \epsilon_i \tag{b}$$
$$C_i = \beta_0 + \beta_1 S_i^{0.5} + \beta_2 V_i + \beta_3 V_i S_i^{0.5} + \epsilon_i \tag{c}$$
$$C_i = \beta_0 + \beta_1 S_i + \beta_2 V_i + \beta_3 V_i S_i + \epsilon_i \tag{d}$$

These models are standard for estimating transaction costs from academic literature. The four parameters are typically interpreted as follows:

- β_0: Fixed transaction cost coefficient (bid-ask spread).
- β_1: Market impact coefficient, capturing cost increases with trade size.
- β_2, β_3: Coefficients representing the effect of volatility on bid-ask spread and market impact, respectively.

Of course, in a non-linear model, the interpretation of the precise role of an individual coefficient can be ambiguous, due to potentially complex interactions among variables.

3 Remarks on tcost.ipynb

The interactive python notebook, tcost.ipynb, estimates the four models using MSE and MAE loss metrics, and presents the results in a self-contained fashion. The pdf version of the final executed notebook is provided using the actual tcost.csv file with 80,000 records as input. However, because the t-cost data are proprietary to a trading firm, we cannot provide them for general use. Instead, to illustrate the operation of the notebook, we provide code that simulates random data that are consistent with model C and its optimized parameters under an MAE metric (ultimately our preferred option).

4 Results and Conclusions

To recap, our goal is estimate four non-linear models using two metrics to accurately predict cost based on trade size and volatility. While analyzing the data, the estimations, and the illustrative results, several points stand out:

1. Data.

   (a) The dominant feature of the cost data is that it is measured with significant noise in the form of market moves.

   (b) Certain observations of the other two variables (especially trade size) also appear anomalous. For example, one record shows a trade size with a whopping 22x ADV.

   (c) And the extremes are not limited to individual observations. Looking at the representative records and the sorts of the data according to each of the three variables, all columns in this data set are characterized by extreme outliers. For instance, the (excess) kurtosis values for C, S and V were circa 40, 75600, and 12, respectively.

   (d) Many of the input "costs" are negative, which some might see as counter-intuitive, as costs ought to always be positive. Of course, these are realized costs, not expected costs.

2. Estimation

   (a) We estimated optimal parameters for four models using two error metrics, MAE and MSE. Although virtually all parameters were highly significant, this was driven primarily by the extremely large observation count.

   (b) While models estimated using MAE were on average superior to those estimated using MSE, virtually all model fits were quite unimpressive, with no R2 exceeding 1%. Some even yielded negative R2 (which can occur in a non-linear model).

   (c) The eight plots illustrating t-costs for the estimated models at representative values of the two input variables, trade size and volatility, suggest some of the model (fits) are more intuitive than others. In our view, model C is the best of the bunch, as its estimated cost increases sensibly with increases of each of the two variables.

   (d) As noted our preferred model and metric are model C under MAE. Although Model C is also the most complicated of the models - it is nonlinear with an exponent of 0.5 - this formulation is also closest to the models recommended in the academic literature.

3. Issues and next steps

(a) My primary conclusion is that much more work is required before any of these models could be used with confidence. While model C estimated under MAE yielded intuitive results, none of the models nor metrics were acceptable.

(b) Additional data to reduce the evident noise in realized costs would be quite helpful in refining our estimates. Removing contemporaneous market moves for the duration of each trade would be a logical first step. If needed later, closer correlates (e.g., industry ETFs) could be used to refine the adjustments.

(c) The data show a wide range of volatility values. Does the data set blend trades from disparate asset classes? If so, parameters (and expected t-costs) might be estimated more precisely by separating records into buckets of shared origin.

(d) Presumably bid / offer could be measured directly because it is observable at time of trade. This could help refine cost estimates, though it might not necessarily be helpful when building portfolios since in many processes desired weights and exposures are created in advance of trading.