The Market Impact Model™

Nicolo G. Torre, Ph.D.
Mark J. Ferrari, Ph.D.

REPRINTED FROM *HORIZON*, THE BARRA NEWSLETTER

BARRA

The Market Impact Model™

Nicolo G. Torre, Ph.D.
Mark J. Ferrari, Ph.D.

Nicolo Torre is the Managing Director, Research at BARRA, where he oversees BARRA's worldwide research. Under Nicolo's leadership, BARRA's research program continues to emphasize economic fundamentals and institutions over black-box statistical approaches.

A California native, Nicolo attended Harvard College, where he studied mathematics and history. He returned to California to earn a Ph.D. in Pure Mathematics from U.C. Berkeley. On completing the doctoral program, he joined BARRA in 1990. Beyond the classroom and workplace, Nicolo delights in exploring the natural world, and has made a number of extended journeys.

Mark J. Ferrari is a Senior Manager in Investment Strategies Research at BARRA. He was involved in the design and estimation of BARRA's Market Impact Model. Mark has also contributed to BARRA's research efforts in conditional volatility, trading systems, industry allocation, and market simulation.

Mark received his B.S. in Physics from Yale University and his Ph.D. in Physics from the University of California at Berkeley. Prior to joining BARRA, he was a principal investigator at Bell Laboratories, where he studied magnetic structures in superconductors.

The Market Impact Problem

Nicolo G. Torre, Ph.D.

REPRINTED FROM *HORIZON*, THE BARRA NEWSLETTER, WINTER 1998, PUBLICATION NUMBER 165

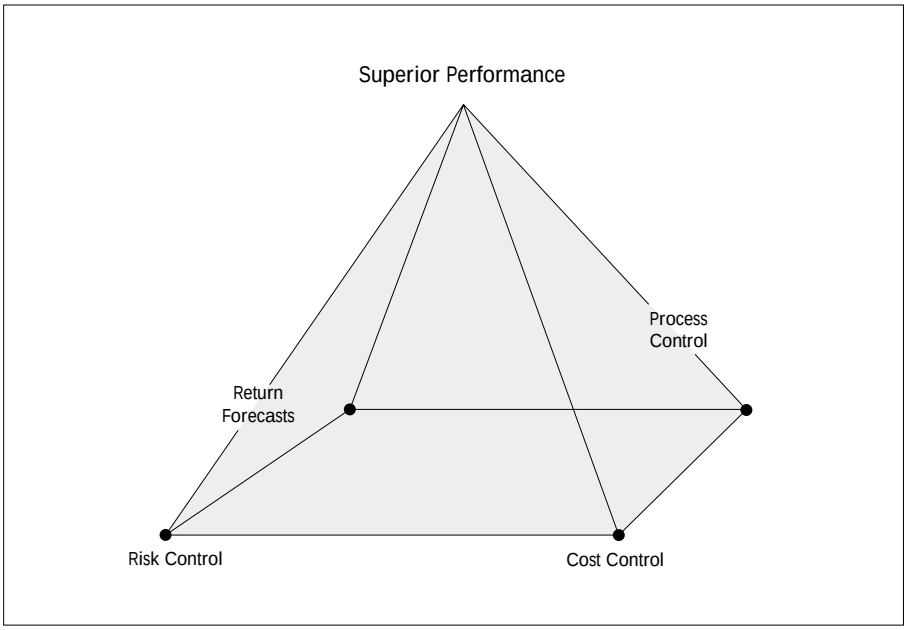
Superior investment performance is the product of careful attention to four elements:

- Forming reasonable return expectations
- Controlling risk so that the pursuit of opportunities remains tempered by prudence
- Controlling costs so that investment profits are not dissipated in excessive or inefficient trading
- Controlling and monitoring the total investment process to maintain a consistent investment program

These four elements are present in any investment management problem, be it a strategic asset allocation decision, an actively managed portfolio, or an index fund. The Market Impact Model focuses on cost control.

Figure 1.1

Elements of superior investment management



Components of Transaction Costs

Portfolio transaction costs are incurred each time assets are bought or sold in a portfolio. The total transaction cost is the sum of two components:

- Order processing costs
- Market impact cost

The *order processing* costs are all costs explicitly incurred to accomplish the transaction. The most important such cost is brokerage commissions, but

processing costs may also include taxes and other transaction-related costs. The *market impact cost* occurs because the transaction itself may change the market price of the asset. The difference between the transaction price and what the market price would have been in the absence of the transaction is termed the *market impact* of the transaction. The market impact is a price-per-share amount. Multiplying the market impact by the number of shares traded gives the market impact cost of the transaction.

Often an order is worked in multiple transactions. In this case, the market impact should be measured from the price that would have pertained in the absence of all the transactions, and the market impact cost of the order is the sum of the cost of the constituent components.

The result of both the processing costs and impact costs is that some portion of portfolio funds are expended in effecting the transaction, rather than being productively invested in the portfolio. Thus, there is also the loss of investment income which these funds would have produced had they been invested. This lost income is an indirect cost of transacting. However, we take the simpler course of restricting the meaning of transaction cost to direct costs.

Market Impact Is a Hidden Cost

Market impact is the difference between the transaction price and what the market price would have been in the absence of the transaction. Since we cannot observe both the occurrence and non-occurrence of the transaction, the market impact cost cannot be directly measured. The best we can do is estimate what the price would have been in the absence of the transaction and compare this estimated price with the actual transaction price.

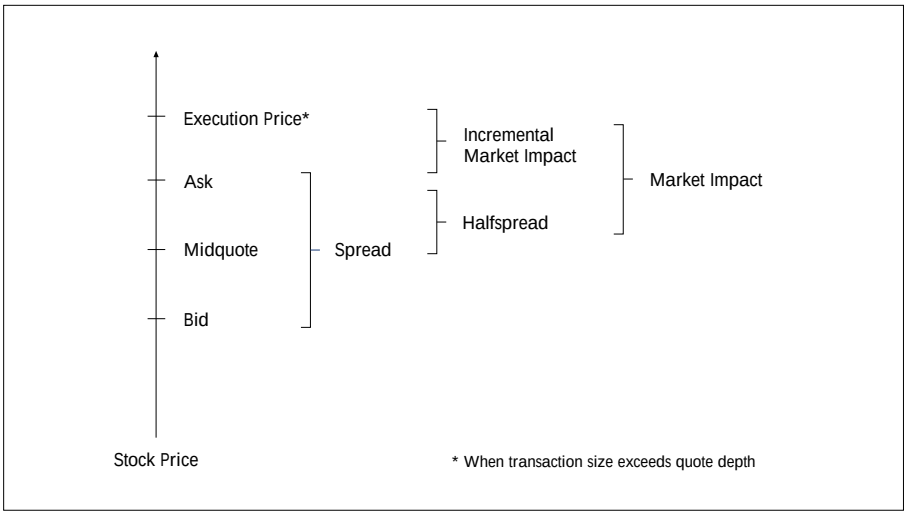
The simplest way to estimate market impact is to look at quoted prices just prior to the transactions. Quoted prices on a stock are the prices at which other market participants have offered to buy (*bid quotes*) or to sell (*ask quotes*). These offers are usually made with respect to fairly limited quantities of the stock, termed the *quote depth*. The average of the bid and ask prices is the *midquote*. The difference between the ask and bid prices is the spread. Hence the difference between either quote and the midquote is the *halfspread*.

If the transaction size is less than the quote depth, then the transaction will usually occur at the bid or ask price respectively. Taking the midquote as an estimate of the market price in the absence of the transaction, the halfspread

then becomes an estimate of the market impact. If the transaction size exceeds the quote depth, then the transaction price will usually be on less favorable terms than the quoted prices. The amount by which the execution price exceeds the halfspread is termed the *incremental market impact*.[1]

Figure 1.2

Price points and their relationship to market impact



Although comparing transaction prices to the midquote can give a useful estimate of market impact, this calculation is only an estimate of market impact and not the definition of market impact. Market impact occurs even if quotes are unavailable and the estimate cannot be made.

Typical Transaction Costs

As a first step toward understanding market impact, it is useful to form a rough estimate of these different transaction costs. The size of the order processing costs is the easiest to determine. In principle, it would be enough to look at brokerage house statements. In actuality, brokerage commissions often represent compensation for services which are not directly transaction-related, such as payment for security research. Thus, making an appropriate allocation of a portion of commissions to transaction cost may not be entirely straightforward. For U.S. institutional investors, however, a ball park figure for the order processing costs would be about 5¢ per share.

[1] This account simplifies somewhat, since there can be more volume available at the quote than is revealed in the quoted depth and since there can be volume available inside the quoted spread. The realized spread (i.e., what investors experience) is thus different from the quoted spread.

The magnitude of market impact depends very much on the type of asset traded. Let us consider the assets which form the S&P 500. For these assets the average halfspread is about 7¢. The median quote depth is typically 3,500 shares and the median share price is approximately $40. If we assume the typical institutional order is for 10,000 shares (i.e., about $400,000), we see that the order size will usually exceed the quote depth, and so some incremental market impact will be incurred. In fact, from the Market Impact Model we will learn that the average market impact of such a trade is about 20¢, which cost indicates an incremental impact of 13¢.

The transaction costs are summarized as:

Table 1.1
Transaction cost ledger

Processing cost	5¢/share
Halfspread	7¢/share
Incremental market impact	13¢/share
Total market impact	20¢/share
Total transaction costs	25¢/share
Memo: Lost income after one year, assuming a 12% annual return to stocks	3¢/share

We see at once that market impact costs represent the largest portion of total transaction cost. Furthermore, the incremental market impact by itself accounts for roughly half of the cost. Since it is difficult to estimate the amount of incremental market impact prior to trading, there is also uncertainty about what the exact cost will be. Clearly, if the cost is unknown ahead of time, it may be difficult to control these costs.

The Economic Significance of Transaction Costs

Next we ask how significant are costs of this magnitude. To answer this question, we need some standard of comparison. One possible comparison is the return to be earned from the assets. The expected annual return from the S&P 500 is about 12% per annum. The median share price is about $40, so the expected annual return is about $4.80 per share. If each transaction generates costs of 25¢ per share and the annual turnover is 100% (implying two transactions per year), then annual costs will be 50¢ per share, or about 10% of annual return.

Transactions are often undertaken to implement an active management strategy. Therefore, another valid standard of comparison is the value added by active management. The value added by a manager is measured by the information ratio, which is defined as:

$$information\ \ ratio = \frac{active\ return}{active\ risk}$$

Consider the case of an active manager who runs a large-cap U.S. equity portfolio. A common benchmark for such a manager would be the S&P 500 Index, and so we may suppose active risks and returns are measured with respect to this benchmark. Suppose that the manager is in the top quartile of active managers; then the manager's information ratio is about 0.5. Suppose also that the manager takes on an active risk of about 4% per annum. Given the assumed information ratio, we conclude that the manager's skill will contribute a 2% active return increment to the portfolio. For a median-priced stock, a 2% annual return amounts to 80¢ per share.

Let us further suppose that the manager has an average holding period of one year. This means that each year the manager will make two decisions about an asset (to buy and to sell). Furthermore, the total value of these decisions is 80¢ per share, so each individual management decision contributes on average 40¢ per share to total portfolio performance. The estimate of 40¢ per share is the value added net of transaction costs. Based on the estimate of 25¢ per share of transaction cost, one concludes that the value before transaction costs of the manager's information is about 65¢ per share. Thus, transaction costs consume about 40% of the gross value of the manager's information. This comparison suggests that careful cost control may be an effective means for active managers to improve their performance.

We can also consider transaction costs in the context of passive management. The basic thesis of passive management is that it is difficult to forecast returns, while it is easy to pay transaction costs. Thus, passive management seeks to maintain a broad exposure to economic opportunity while incurring minimal costs. Typically, an ideal portfolio is defined. The actual portfolio will tend to deviate from this ideal as cash flows occur and the assets held experience different returns from the assets in the ideal portfolio. The basic management decision is, therefore, to decide when deviations from the ideal are sufficiently large as to justify the costs of bringing the portfolio composition back toward the ideal. Estimating and controlling transaction costs is thus central to the passive management strategy.

Cost Experience Reflects the Total Investment Process

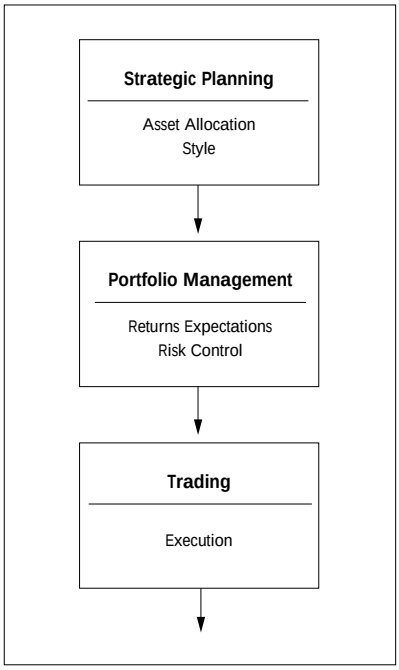
Having assessed the significance of transaction costs, our attention shifts to controlling them. Here a key insight is that the cost experience is the outcome of decision-making at many levels of the portfolio management process:

- At the *strategic level*, choices are made regarding asset classes and management style. Clearly, the choice between a large-cap passive strategy and a micro-cap high-turnover active strategy will have important consequences for the cost experience of the portfolio.
- At the *portfolio level*, specific asset holding decisions are taken and, in consequence, trade lists are generated. Generally, there are many different ways to implement investment ideas in portfolios, and some ways naturally lead to lower transaction costs.
- At the *trade level*, individual positions are transacted. Here there is some scope for affecting transaction costs, although it can be limited by time constraints and the need to efficiently process a large number of orders.

We see that each decision contributes to the total result and that each decision limits the scope of what can be achieved at the following level.

Figure 1.3

Levels of decision making



On further examination of the investment management process, we see that there is a significant opportunity to improve transaction cost control at the portfolio management level. At this level, a delicate balance must be struck between return, risk, and cost. Achieving this balance requires asset-level estimates of each quantity. The basic difficulty, therefore, is to develop and integrate the necessary information. In many investment management firms, cost expertise is concentrated in the trading desk, whereas return expertise is the domain of the portfolio manager. This division of knowledge can lead to inefficiencies in the integration process.

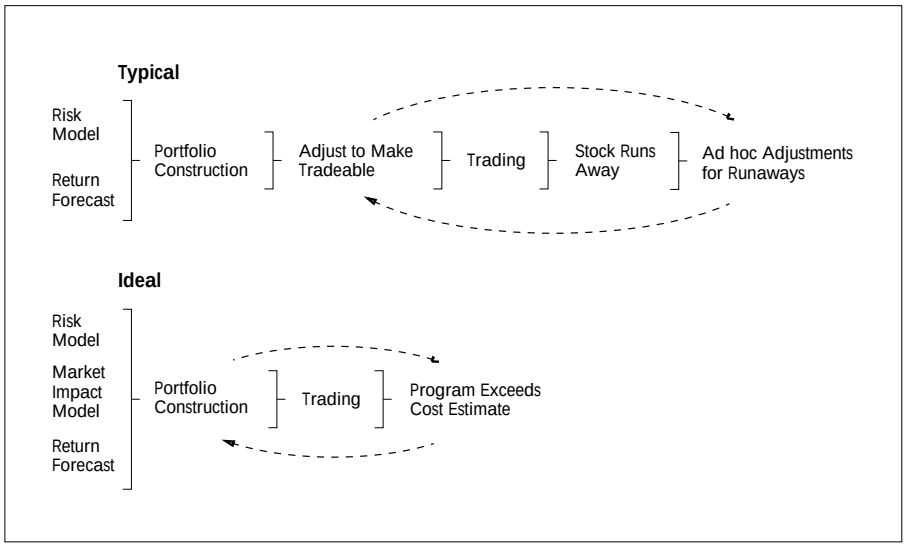
A hypothetical example illustrates the practical difficulties. Consider a manager who develops return expectations and constructs a portfolio with an optimal balance of risk and return characteristics. During the portfolio construction,

the manager may control transaction costs by placing a limit on total turnover, but no true cost-benefit analysis occurs. Transforming the current portfolio into the optimal portfolio requires a certain number of trades. The manager reviews the projected trade list, perhaps in consultation with a senior trader. Typically, the list is found to contain untradable positions, and so the manager adjusts the list by hand. The resulting target portfolio is now no longer optimal. Furthermore, it may be unclear at the time adjustments are made whether an asset is being purchased primarily for return reasons or for risk control. Hence, the best way to adjust the portfolio is unclear.

Once the manager decides on the target portfolio, the trading desk begins implementing the trades. As the trading progresses, market impact is observed to drive up the price of some assets—i.e., the assets run away from the traders. The manager then decides either to pay up or to abandon the attempted trades. Generally these decisions must be made on an asset-by-asset basis, so a portfolio viewpoint is lost. The portfolio that finally emerges from the trading process can differ from the original optimal portfolio in a fairly ad hoc way.

Figure 1.4

Portfolio management: typical vs. ideal



The weakness in this process is that transaction costs were not incorporated in the portfolio construction decision at the start and thus no true cost-benefit analysis occurred. Had the cost been incorporated into the portfolio construction from the beginning, the trade list would have been inherently more tradable. Hence the need for hand-tuning would have been reduced or eliminated. During the trading process, the handling of runaway assets could also be guided—e.g., by a simple rule: If the cost of implementing the program

reaches the forecast costs before completion of the program, then the target portfolio can be adjusted based on the new current holdings. Thus, incorporating transaction cost estimates into the portfolio construction decision should permit a more streamlined process that ensures the portfolio owned is actually the optimal portfolio.

Requirements for the Market Impact Model

The goal of the Market Impact Model is to provide transaction cost forecasts suitable for portfolio construction purposes. To realize this goal the model must meet the following criteria:

- The model must provide *a priori* forecasts rather than *ex-post* measurements.
- Based on information available at close of market today, the model must provide forecasts valid for the trading conditions likely to apply over the next few days.
- The model must cover a wide universe of assets with uniform accuracy.
- The model must cover a wide range of transaction sizes.

These requirements make construction of the Market Impact Model a difficult task. On the one hand, requiring forecasts to remain valid over a few days limits the data that can be applied to the problem. Thus, information which is hard to forecast, such as price trend, is not directly useful. On the other hand, the breadth requirements demand high performance from the model.

Econometric Challenges Faces by the Market Impact Model

At first, it might seem that forecasts of market impact costs could be developed in a fairly straightforward manner. Seemingly, it would be enough to monitor the costs of past trades. One might categorize the trades by asset traded and trade size. Then the mean of past costs at a given trade size would provide a forecast of future costs. Three econometric difficulties guarantee that such a methodology will fail, however:

- Market impact is not directly observable; rather it must be estimated. To reduce estimation error, large statistical samples are required, which implies gathering data over extended time periods. Thus, only time-averaged behavior can be measured with any precision. But the underlying phenomenon is a product of the highly dynamic market environment. Thus, long-term averages end up having little forecasting ability.

- Trade data suffers from a censoring problem. Since investors avoid costly trades, we never see the most costly trades in the record. Unless we somehow compensate for this circumstance, estimates derived from trade data will systematically underestimate market impact costs.
- There is an uneven level of information available for assets. We see market impact for an asset by watching how it trades. Therefore, we know more about heavily traded assets and less about thinly traded assets. Yet it is the thinly traded assets that tend to have the higher market impact costs, and it is the more costly-to-trade assets that are most important for the portfolio construction problem. Again, we need to adjust for the data problems to avoid introducing a methodological bias into the results.

Choice of Modeling Strategy

Of the three econometric challenges, the data censoring problem exerts the greatest influence over the choice of modeling strategy. The key assumption in the naïve modeling approach is that the data contain the information we need, and so all we have to do is to extract it. However, the data does not contain the information we want, at least not in any directly extractable form. As a result, the naïve approach cannot succeed.

A successful modeling strategy must begin with a more sophisticated approach to the data. Although the data may not directly answer our question, they can still answer many questions about market functioning. If we knew how the market mechanism determines impact, then we could hope to assemble the measurements of market function drawn from the data into an answer to the question of true interest—namely, what will be the impact of a prospective trade? For this approach to succeed we need an understanding of the economic fundamentals determining impact.

Our model-building exercise begins, therefore, with an economic analysis of the markets and trading. The result of this analysis is the identification of the factors determining market impact and the structure relating the factors. Once the model components have been identified, we turn to empirical analysis to complete their estimation.

The ultimate modeling strategy is thus a combination of the analytical and empirical approaches. We feel that a marriage of these two approaches is especially suitable to the market impact problem—since the economic analysis can overcome the limitations of the data, while the empirical approach ensures that the results are closely tied to the realities of the market.

Summary

We have seen that the portfolio management process benefits from timely a priori estimates of trading costs covering a wide range of assets and position sizes. Providing these estimates is the goal of the Market Impact Model. Because of econometric difficulties, a purely data-driven approach to this modeling problem cannot succeed; a combination of analytical and empirical methods is required. Future articles in this series will address these issues and describe the construction and implementation of the Market Impact Model.

The Economics of Equity Trading

Nicolo G. Torre, Ph.D.

REPRINTED FROM *HORIZON*, THE BARRA NEWSLETTER, SPRING 1998, PUBLICATION NUMBER 166

In our first installment in this series we discussed the motivation behind constructing the Market Impact Model as well as the challenges presented by the task. In this second installment we will describe the underlying economic structure of equity trading as it relates to the creation of market impact transaction costs.[1]

The Economic Structure of Equity Trading

Mismatched Order Flow Is Common

Let us consider a generic equity market to which investors send orders for execution. For executions to occur, offsetting buy and sell orders must be matched together. Some matching occurs automatically from buy and sell orders arriving in the market simultaneously. This natural match rate is surprisingly low, however. We can learn something about this match rate from a sample of orders made available by the New York Stock Exchange. In this data, incoming market orders cross with one another about 20% of the time. This average is somewhat driven by the more heavily traded companies, for which the cross rate is higher. For the typical stock, the match rate is less than 20%.

POSIT is a very different order-handling system from the New York Stock Exchange, but the match rate for POSIT is of comparable magnitude. One concludes, therefore, that a low match rate derives from the behavior of investors rather than from the specific structure of the markets.

Liquidity Providers and Demanders

Given the low natural match rate, some market participants must modify their behavior so as to accommodate mismatched order flow. We say that an investor who carries on with his trading plans independent of the presence or absence of matching order flow is a *liquidity demander*. An example of such an investor would be one who submits market orders. By contrast, we term an investor who adjusts his trading to accommodate mismatched order flow a *liquidity provider*. Examples of liquidity providers include specialists and limit order submitters.

[1] This is an excerpt from a longer section of the *Market Impact Model Handbook*. For additional details, contact BARRA.

The distinction between liquidity providers and demanders allows us to introduce a distinction between trades. For every trade there is, of course, a buyer and a seller. When the trade takes place between a liquidity demander and a liquidity provider, however, we shall give the trade the character of the role played by the liquidity demander. For instance, when the liquidity demander is the buyer, we say the trade is a buyer-initiated trade, or just a buy trade for short. Similarly, a sell trade is one in which the liquidity demander sells and the liquidity provider buys. Trades that happen to cross between two liquidity demanders are said to have indeterminate initiation, or to be indeterminate.

Table 2.1

Who takes the opposite side to market orders arriving at the NYSE

Trading crowd	38.6%
Limit orders	28.7%
Specialists	17.3%
Market orders	12.0%
Other (e.g., ITS)	3.4%

Proprietary trading by non-specialist member firms is comparable in volume to specialist trading but cannot be separately identified in the above classification. Presumably most proprietary trading falls in the the Trading Crowd category.

Source: BARRA analysis of TORQ data and NYSE Fact Book 1994.

The accommodation of mismatched order flow is a service the liquidity providers render the liquidity demanders. The compensation the liquidity providers receive for rendering this service is simply the market impact cost paid by the liquidity demanders.

Competition Among Liquidity Providers

When we examine the provision of liquidity, we see that there is significant competition among liquidity providers to render this service. The most visible liquidity provider is the specialist on the floor of the New York Stock Exchange. He competes for order flow with specialists on the regional exchanges, the limit order book, proprietary traders, and the various specialized trading services. The New York Stock Exchange as an institution is in competition with the Nasdaq market, and potentially in competition with foreign stock exchanges or innovative trading mechanisms such as Optimark.[2]

[2] Optimark is an electronic crossing mechanism operated by the Pacific Exchange that is expected to come online in August 1998. It allows users to express their willingness to trade away from the market price in return for a faster execution or larger fills. Crosses are arranged based on maximizing the satisfaction of participants with the resulting fills.

Economic analysis classifies businesses with respect to the degree of competition they face on a scale ranging from pure competition through monopolistic competition and oligopoly to pure monopoly. The basic factors determining a business's position in this classification are the opportunity for market segmentation and the barriers to entry of competitors. When we examine the business of providing liquidity in stocks, we see that there is some opportunity for segmentation. One form of segmentation is for a liquidity provider to concentrate in a few stocks. In fact, this is the route taken by the specialist. Another form of segmentation is to concentrate on meeting specific types of liquidity demands. This is the route taken by principal bid desks. Barriers to the entry of competitors are fairly low, however. We conclude, therefore, that the liquidity provider's business may be classified as intermediate between pure competition and monopolistic competition.

For a business facing either pure competition or monopolistic competition, microeconomics provides two basic insights:

- Market forces establish the price of goods. Profit-seeking suppliers of goods adjust their activity level so that marginal revenue equals marginal cost. During normal operating conditions, most firms will usually find that marginal costs are determined by variable costs.
- Again under normal operating circumstances, the firm's economic profits will be zero—i.e., the operating profits will equal the risk-adjusted cost of capital.

Let us apply these insights to the analysis of the liquidity provider's business. For the liquidity provider, the marginal revenue is just the market impact earned from the marginal trade accommodated. A liquidity provider's cost structure will generally include staff, equipment, and capital costs. Staff and equipment costs are fixed in the short term, however, so the main variable cost will be the risk-adjusted cost of capital. This suggests that one should measure the quantity of liquidity provided in terms of the amount of risk assumed by the liquidity provider. Hence we conclude that the market impact of a trade should be proportional to the risk borne by the liquidity provider. This relationship is the fundamental hypothesis upon which the Market Impact Model is built.

Modeling the Liquidity Provider's Risk

Fundamental Hypothesis of Market Impact

We may introduce some notation to express the fundamental hypothesis. Let κ represent the market impact cost of the trade and let ρ represent the risk borne by the liquidity provider. We shall represent the factor of proportionality between κ and ρ by τ. Then the basic relationship is: $\kappa = \tau\rho$.

For the moment, these symbols lack sufficient definition for this relationship to be useful. The goal of the model-building effort is to define τ and ρ independently from this relationship so that the relationship may be applied to calculate κ.

As a first step in this direction, we note that ρ is a measure of the quantity of a good (namely, liquidity), while κ is the cost. Hence, τ is the cost per unit of the good. In other words, τ is the price of liquidity. It follows that τ will be established by competitive conditions in the market for the supply of liquidity, rather than by the particular liquidity provider who accommodates a trade. In other words, τ should be a market-wide factor. A high value of τ corresponds to a market in which liquidity is tight, while a low value corresponds to a liquid market. Accordingly, we term τ the *market tone*.

The market tone can be estimated from trade results. Suppose that for a sample of roughly contemporaneous trades we have an estimate of their market impact κ_i and risks ρ_i. Then from the basic relationship, τ can be estimated by the mean of the ratios:

$$\frac{\kappa_i}{\rho_i}$$

Since τ is a market-wide number, it follows that we can draw on trades in a number of assets to perform the estimation. Thus, there is an opportunity to exploit the statistical power of cross-sectional estimation.[3]

The Liquidity Provider's Risk

Having considered market tone in some depth, let us analyze the liquidity provider's risk, which we have previously designated as ρ. Risk derives from uncertainty about the future. When the liquidity provider takes a position into

[3] In the first installment of this series we noted that one of the econometric difficulties of modeling market impact was the variability through time of the phenomenon. By estimating τ cross-sectionally at a point in time, we avoid using time-series data whose interpretation may be confounded by the variability of market conditions. Thus, the estimation of τ provides a first illustration of how the identification of economic structure can overcome the econometric challenges of the data.

inventory, he faces two basic uncertainties. First, he does not know the price at which he will be able to remove the position from his inventory. Second, he does not know how long the position will remain in his inventory. We refer to the length of time it takes the liquidity provider to work out of the position as the *time-to-clear*. A long time-to-clear is undesirable since:

- A longer holding period for a fixed dollar profit implies a lower rate of return.
- A longer holding period implies a greater exposure to the risk of adverse price movement.
- Longer holding periods imply a lower rate of capital utilization.

The combination of price uncertainty and time-to-clear risk leads to market impact cost, which the liquidity provider seeks to transfer to the liquidity demander.

When a liquidity demander approaches the market with a large order, a common procedure would be to break up the order into trades that are worked gradually in order to mitigate market impact. There are two consequences to working the order in this manner:

- Each trade presents a liquidity provider with less risk than the total order, so the impact of each trade is less. Later in this series we shall see that the cost is a nonlinear function of the trade size, so that the sum of the costs of all the trades is less than the cost of the full order.
- This cost reduction is simply a reflection of the second aspect of working an order—namely, that the liquidity demander is retaining some of the risk that otherwise would be transferred to the liquidity provider.

The trades that have been set aside to be worked later remain in the investor's inventory and create risk. In fact, the investor bears the same risks as the liquidity provider: Prices may move against him and positions may move out of inventory slowly.

Time-to-Clear in a Toy World

To isolate a key issue in the liquidity provider's risk environment, consider a toy version of the real world. In this simplified setting we consider a market for a single asset. We shall suppose that there is a single market maker in this asset and that all investors trade only with the market maker. We wish to model the state of this asset in equilibrium conditions. Equilibrium is usually taken to mean that there is no imbalance in buyer and seller interest. Hence we shall

suppose that incoming orders are to buy and sell with equal probability and that the size and arrival rate of the orders are independent of order direction.

For the sake of simplicity, let us assume that each order is for 100 shares and that orders arrive at the rate of one per hour. Initially price does not enter into our analysis, but for the sake of concreteness let us assume that the price is constant. Let us further suppose that the dealer begins with zero shares in inventory and that the first order to arrive is a sell order, so the dealer begins by acquiring an inventory of 100 shares.

Let us denote by T the amount of time that must pass before the dealer's inventory first returns to zero. The value of T is a random variable that depends on the sequence in which buy and sell orders arrive in the market. Thus, the dealer cannot in general know ahead of time what the value of T will be. However, the dealer can seek to form some expectation of what T will be; that is, the dealer can seek to know the average value of T.

The situation we have hypothesized is sufficiently simple that the entire distribution of T and its average value may be explicitly calculated. This calculation is of some interest, as it demonstrates in a simple case the methodology that applies in more complex cases. Accordingly, the mathematically inclined may wish to read through the details of the calculation given in Exhibit 2.1. For those not so inclined, it is enough to know the somewhat surprising result that the average value of T is infinite!

This is bad news for the dealer; it means that if the dealer commits capital to just 100 shares, he can have no expectation of eventually freeing up that capital. Additionally, if we ask how high the dealer's inventory is expected to go before returning to zero, we again get an infinite answer. Thus, the world we have described is one in which a market maker requires infinite patience and resources. Clearly, in such a world, no one will wish to be a security dealer.

Search for an Explanation

Yet in the real world, some people voluntarily choose to be security dealers. Furthermore, empirical studies have found that, at least for NYSE specialists, holding periods are typically a week. We conclude that our initial model is too simple; it must omit some essential feature of security markets. This is a valuable result, for it indicates that if we can determine what is missing from our model, then we will have learned something about the real world.

Exhibit 2.1

Evaluating time-to-clear in the toy world

We want to calculate the mean holding period for a dealer who has an initial inventory of 100 shares and has equal 50% probabilities of either buying or selling 100 shares. Let *p(t)* be the probability that the dealer holds no shares at time *t* but had a positive inventory at every previous time. (We want to calculate the expected amount of time before he first returns to holding zero shares.) For example, $p(0) = 0$ as the dealer has 100 shares at time $t = 0$. We see that

$$p(1) = \frac{1}{2} \quad \text{and} \quad p(2) = 0$$

To have arrived at time 2, the dealer must have bought more shares at time 1, giving him 200 shares at time 1. He can sell only 100 shares at time 2, so he can never have zero holdings at time 2.

Similarly, there is only one way to first have zero holdings at time 3; he must buy at time 1, and then sell at times 2 and 3. This happens with probability:

$$p(3) = \frac{1}{2}.\frac{1}{2}.\frac{1}{2} = \frac{1}{8}$$

Much like time 2, he can never return to zero at time 4, or indeed at any even time. At time 5 the calculation is slightly more complicated, as there are two possible ways of first returning to zero then. Either the dealer can buy twice (leaving him at 300 shares) and then sell three times, with probability 1⁄32; or he can buy once, sell once, buy once, and then sell twice, also with probability 1⁄32. The sum of these two, 1⁄16, is $p(5)$. Figure 2.1 shows the possible paths the dealer's inventory could have taken for times $t \leq 5$.

We can continue to calculate probabilities in this way. The mean holding period *m* is the average of all possible times *t* weighted by *p(t)*:

$$m = \sum_{t=0}^{\infty} \cdot p(t)$$

It is possible (although unpleasant) to evaluate this sum directly. As it happens, we can show that it is infinite without directly evaluating the series. After one period, the dealer either holds no shares or 200 shares, with equal probability. If he holds 200 shares, his mean holding time is $2m$ (m periods to return to 100 shares, and another m to return to 0). Thus, his mean holding time is:

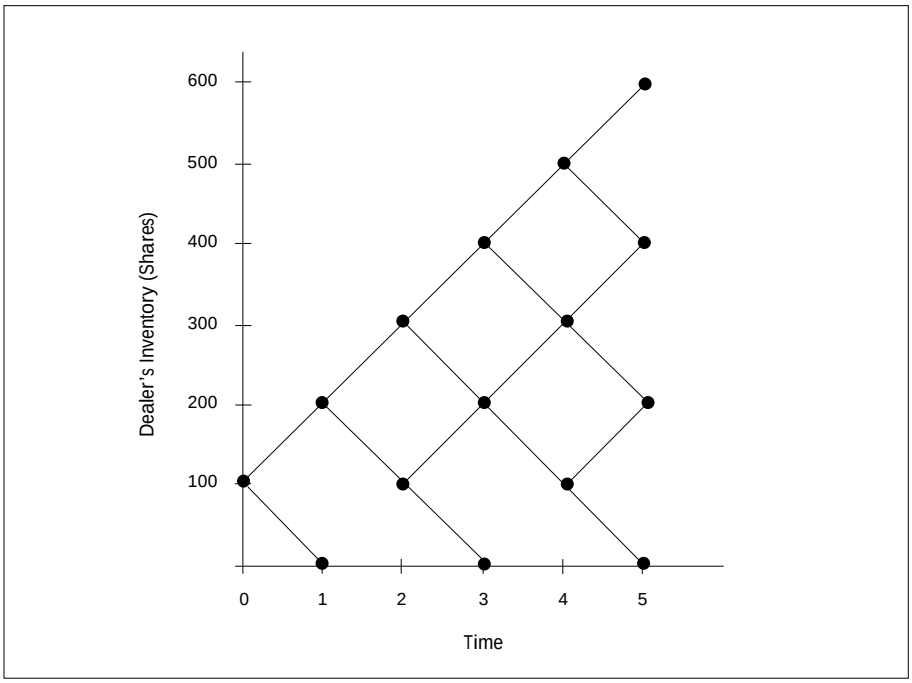
$$m = \left(\frac{1}{2} \cdot 1\right) + \left(\frac{1}{2} \cdot 2m\right), \text{ so} \qquad m = \left(\frac{1}{2} + m\right)$$

This is clearly not possible for any finite m, so the mean holding time must be infinite.

The problem we have identified is that *T* has infinite average value. Working through the math, we see that this result arises from a single circumstance: namely, that buy and sell transactions are equally likely at all times. Would it solve the problem if there were some permanent imbalance in the probabilities—if, say, the probability of buy orders were always greater than sell orders? This adjustment to the model would ensure that the dealer's inventory, when long, returns to zero on average in a finite time. Unfortunately, no short inventory would return to zero on average in a finite time. Thus, introduction of a permanent imbalance in the buy and sell probabilities cannot fix the model.

Figure 2.1

Possible paths followed by the dealer's inventory



Instead, we must make the imbalance in the probabilities contingent on the dealer's inventory level. As the dealer's inventory rises above zero, buy orders must become more probable, and as the inventory falls below zero, sell orders must become more probable. Yet how can investors be induced to adjust their order submissions based on the dealer's inventory? Clearly the necessary inducement must be a change in the stock price.

The dealer must adjust his trading price based on his inventory level and investors must respond to this price signal. As the dealer's inventory rises, he should lower the market price and investors should respond by buying more often and selling less often. Similarly, as the dealer's inventory goes below zero, he should raise the market price and investors should respond by selling more often and buying less often. Thus, in the real world, the market price must

fluctuate in response to order imbalance if the dealer's holding period is to remain finite.

But the movement of the market price in response to order imbalances is what we mean by market impact. Hence we see that market impact is a necessary component of real markets to moderate investor actions so that chance fluctuations in order flow do not cause a market breakdown. Or to put it another way, a real market can only eliminate market impact by securing the services of a dealer with infinite resources and patience.[4]

Red Herrings

We have discovered why market impact occurs. Equally important, we should note what did not enter into our model. Our model makes no mention of adverse selection, bid-ask spreads, price discretization, price volatility, dealer profit objectives, the cost of capital, or any other details of market microstructure. All of these factors are important for determining the magnitude of market impact. However, they are not the cause of market impact.

By implication, tinkering with market microstructure cannot eliminate market impact. In fact, we can see easily what the likely influence of these various market microstructure factors will be on the magnitude of market impact. Market impact is moderated by dealers committing capital to market-making. Dealers are led to commit capital by the profitability of the market-making business. Changes in market microstructure that improve the risk/ reward characteristics of the dealer's business thus tend to reduce market impact. For instance, police action against insider trading reduces the dealer's risk of catastrophic losses and thus attracts capital into the market-making business, lowering market impact.

The Law of Supply and Demand

We can recast our conclusions thus far in economic language that makes the results seem rather natural. The dealer performs a service for investors by carrying temporary inventories. Transaction prices thus represent a payment for both the asset and for the liquidity service. In general, therefore, the transaction price will differ from the asset's equilibrium price. This price differential is the price of liquidity. As with any economic good, the price of liquidity must

[4] Governments occasionally attempt to provide markets with such dealers in the form of buffer stock managers who are mandated to stabilize some market price. A common experience is that such schemes are prone to sudden collapse. The assumptions of our model are sufficiently simplifying that the model may not apply directly to any particular price stabilization scheme. However, the model does serve to indicate that the mathematics of the situation may render a price stabilization scheme unsound in principle.

fluctuate in response to the conditions of supply and demand. Hence we observe this fluctuation in the form of market impact.

Conclusion

How can we identify the underlying factors that contribute to the liquidity provider's risk? Essentially these are the features we excluded from our toy world. In the real world, we must allow for varying transaction sizes that arrive at a variable rate. Both sources of variation will contribute to uncertainty in the time-to-clear. Additionally, we must allow for the variability in asset prices, which creates uncertainty in the liquidity provider's return. Further, we should note that these sources of variation do not simply add up to produce the variability of return. Rather there is a complex functional relationship between the various quantities. Modeling these factors and their relationship will be the goal of the next installment.

The Market Impact Model™

Nicolo G. Torre, Ph.D.

REPRINTED FROM *HORIZON*, THE BARRA NEWSLETTER, SUMMER 1998, PUBLICATION NUMBER 167

This is the third part of our continuing series on the Market Impact Model.[1] In the first two parts we defined the problem that the model addresses and applied economic reasoning to establish the framework of the solution. Now we shall turn to empirical data analysis to provide the details of the solution. To organize the discussion, we begin by stating the conclusion. We then begin building to this conclusion. First we treat data upon which our work rests. Next we treat each model factor in turn, surveying the principal findings of data analysis. Finally, we show how the information is assembled into the final forecast of market impact.

The Market Impact Model

To anticipate our conclusion, we state the final model as:

$$\kappa = F\left(V, \varepsilon, \sigma, \phi, \zeta, \tau, \chi\right) \qquad \text{Equation 3.1}$$

where:

- κ is the *forecast* cost of the trade measured by value
- V is the *trade volume* measured by value
- There are four parameters characterizing the asset being traded:
  ε is the *elasticity*; it describes the response of order flow to price signals
  σ is the *volatility* of the asset; it describes the variability of the asset's price
  ϕ is the *intensity*; it describes how often the asset trades
  ζ is the *shape*; it describes the distribution of trade sizes
- There is a single market parameter:
  τ is the *market tone*; it is the price of liquidity
- There is an investor-specific parameter:
  χ is the *skill*; it captures the effect of the investor's trading process on the cost experience
- $F()$ is a complicated function that integrates these different sources of information into the final forecast

It should be noted that several of these parameters are actually multidimensional quantities. We consider next the data upon which the model is estimated.

[1] This is an excerpt from the *Market Impact Model Handbook*. For additional details, contact BARRA.

The Data

There are a number of different data sources that we have used to investigate market impact, among them:

- The market data feed
- Trading results achieved by clients
- Specialized data sets made available by third parties
  For instance, the New York Stock Exchange has released a sample of data on trades, orders, reports, and quotes, known as the "TORQ dataset."

For actually updating the model, however, only the market data feed is available on a timely basis. Thus, our comments will focus on this data source.

The market data feed is a complex and voluminous data source. The lineal descendant of the ticker tape, the feed reports trade and quote revisions as well as special trading conditions (e.g., trading halts due to pending news releases.) The volume of data has been growing at a compound rate for years, and as of 1997 it averaged about a gigabyte of raw data per day.

Much of this data originates with the specialist's clerk, who types it in by hand, often under considerable time pressure. Thus, data errors are not uncommon. Some of these errors are corrected, possibly hours after the first report. Other errors remain uncorrected. Adding to the complexity are the cryptic encoding of the data, the reporting of data from multiple markets, and the opportunities for the time order of events to become garbled. A significant, highly technical effort is required to put the market data feed in a form suitable for analysis.

The first stage of analysis aims at detecting data errors and assigning trade initiation. This is essentially a rule-based endeavor, involving such logic as:

- If a trade is reported at a price that is ten times the prior trade price, the decimal point was probably misplaced, so this is a data entry error. Infer that a trade took place at probably one-tenth of the reported price.
- If the trade price is above the latest ask, and the ask was reported recently, then this trade was probably buyer-initiated.

The logic can become quite intricate. While such rules cannot achieve complete accuracy, we may control the accuracy by comparison with some of the other data sources we mentioned. This control is important, for it allows us to describe the error process in our data and to adjust for it by appropriate statistical corrections. The output of the first stage of analysis is thus a cleaned-up trade record, with trades identified as buyer- or seller-initiated. We use this data to estimate the model.

Model Factors

In the previous article of this series[1] we identified four characteristics of an asset that would control the risks of a liquidity provider taking a position in that asset. Our purpose now is to capture these characteristics in quantitative terms. We refer to such a quantification of an asset characteristic as a model factor. In general, each factor is itself the product of a submodel. It is convenient to assign a name to each factor which conveys the basic characteristic being quantified. As noted at the commencement of this chapter, the asset factors are elasticity, volatility, intensity, and shape.

Besides the asset-related factors, the model also has a market factor—the market tone—and an investor-related factor—the skill. To discuss these latter two factors, it is first necessary to partially construct the market impact cost function. Accordingly, discussion of these factors will be deferred to a later section.

Elasticity

As we saw in the previous article, it is necessary to model how the direction of order flow responds to price changes.

We define the *equilibrium* price to be the price at which an order arriving in the market is equally likely to be a buy order or a sell order. While the equilibrium price would be difficult to estimate in real time, in retrospective analysis it can be determined with fair accuracy. We define the *price deviation* δ to be the difference between the logarithms of the midquote Q and the equilibrium price P_0:

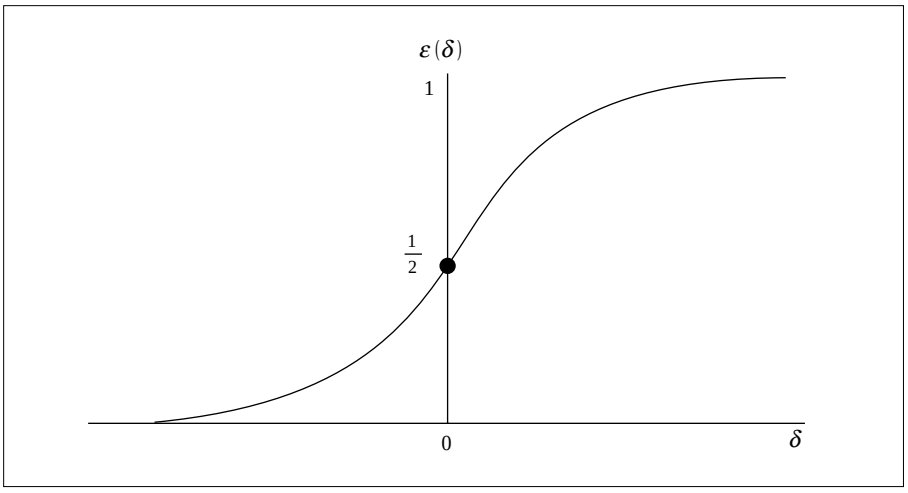
$$\delta = 1n\left(Q\right) - 1n\left(P_0\right) \qquad \text{Equation 3.2}$$

The *elasticity function* $\varepsilon(\delta)$ gives the probability that an incoming market order will be a sell order, conditional on the current price deviation. By definition, we have:

$$\varepsilon\left(0\right) = \frac{1}{2}$$

Figure 3.1

Expected form of the elasticity function



We might expect that, as δ becomes large, $\varepsilon(\delta)$ should move toward 1, and that, as δ becomes a large negative number, then $\varepsilon(\delta)$ should go to 0. Thus, we might expect the shape of the function to resemble Figure 3.1.

Actually, however, when we measure the market's response to price deviation, we find that the elastic function is compressed within narrower vertical limits. This compression is probably due to the presence of *price-takers* in the market —investors who are content to trade at whatever price is quoted in the market, and who, therefore, do not respond to price signals.

We find it convenient to parameterize the elasticity function as:

$$\varepsilon\left(\delta\right) = \frac{1}{2} + \frac{f}{2}\tanh\left(2c\delta\right) \qquad \text{Equation 3.3}$$

where:

f is the *elastic fraction*; it describes the maximum strength of the market's response to price signals

c is the *elastic coefficient*; it describes the sensitivity of the market to price deviations

tanh() is the hyperbolic tangent function

We also define the *elasticity S* as the slope of the elasticity function at the point $\delta = 0$. For small deviations δ:

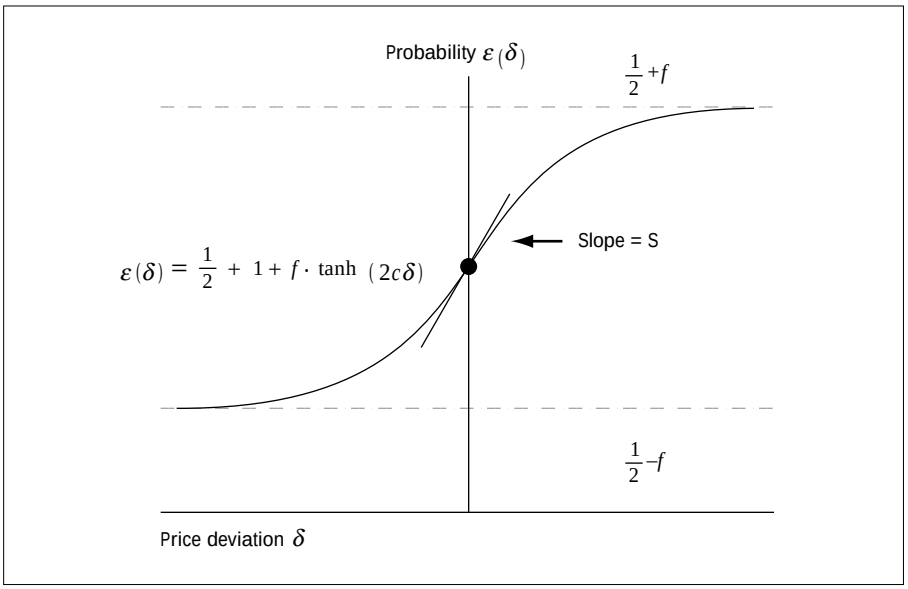
$$\varepsilon\left(\delta\right) = \frac{1}{2} + S\delta \qquad \text{Equation 3.4}$$

The slope is related to the elastic fraction and coefficient by $S = fc$. S is the elasticity reported in the Market Impact Model.

Referring to Figure 3.2, we see that the vertical range of the function is $2f$, and S is the steepness of the function at the origin. The elastic coefficient c determines how quickly the function flexes in response to price deviations.

Figure 3.2

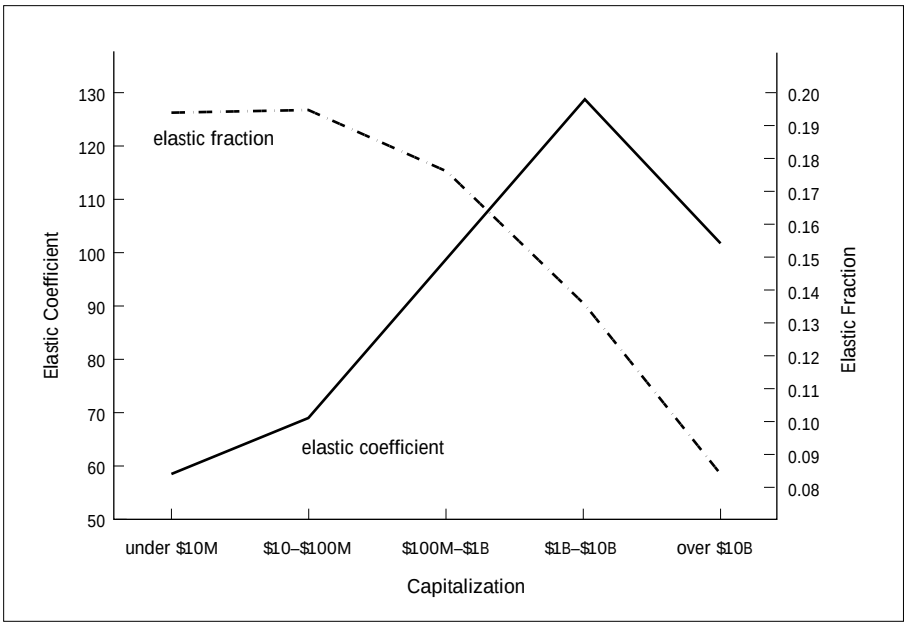
Actual form of the elasticity function



When we measure these various parameters for different assets (Figure 3.3), we find some interesting tendencies. For the larger companies the elastic fraction tends to be lower, while the elastic coefficient tends to be higher. In other words, more of the orders for these assets are generated by price-takers, but the non-price-takers seem to be fairly sensitive to price deviations.

Figure 3.3

Elasticity parameters as a function of capitalization



By contrast, for smaller companies it seems that fewer orders are generated by price-takers, but larger price deviations are required to generate a response. Possibly investors are less certain of what the "right" price is for these assets, and so they are slower to perceive a bargain.

Although it requires fairly sophisticated econometric technique to back the elasticity function out of the available data, we may note that the specialist has a considerable informational advantage in this regard. Simply by looking at the limit order book, he can gauge the strength of near market demand. Furthermore, by moving his quotes around he can test the market's response to price changes. Thus, he can assess the likely reaction of order flow to price change.

Volatility

Elasticity describes the reaction of order flow to deviation of the market price from the equilibrium price. However, the equilibrium price itself varies due to the changing level of demand for an asset. The volatility factor provides a description of this variability.

The standard model for asset prices is that they follow a lognormal diffusion process. This process is characterized by two parameters μ and σ, such that if P_t is the price of the asset at time t, then for

$$r = ln\left(\frac{P_{t+\Delta t}}{P_t}\right)$$

r is normally distributed with mean $\mu\Delta\tau$ and variance $\sigma^2\Delta\tau$. Several improvements can be made on this basic model:

- In general σ is not a fixed constant, but will vary through time.
- For short holding periods $\Delta\tau$, the size of r will typically be small, and so the distribution of r will deviate from normality due to the discreteness of price changes.
- Empirically, return distributions deviate from normality in that the probability of large returns, albeit rare, is somewhat greater than a normal distribution would allow.

The last point merits elaboration, for it sheds light on the fundamental economics of the liquidity provider's business. The basic presumption of any investment business is that a fluctuating level of profitability will be observed day to day, but over time the fluctuations will average out to produce a reasonable mean return. This presumption, however, is only valid under the assumption that the business survives long enough to permit some degree of averaging. Many liquidity providers operate on a highly leveraged basis, and they remain in the business only as long as minimum net equity is maintained. The significance of rare large returns is that they can put a leveraged liquidity provider out of business, with the result of course that the long-term mean profitability of the business is not realized. Thus, while the rarity of large returns renders them of modest importance for determining the mean, they are of greater importance in analyzing the risks faced by a liquidity provider.

Intensity

There is considerable variability in how quickly orders for a given asset flow to the market. Figure 3.4 illustrates this variation for GE on a daily time scale, and Figure 3.5 shows the variation on an intraday time scale.

Figure 3.4

Number of trades per day for General Electric

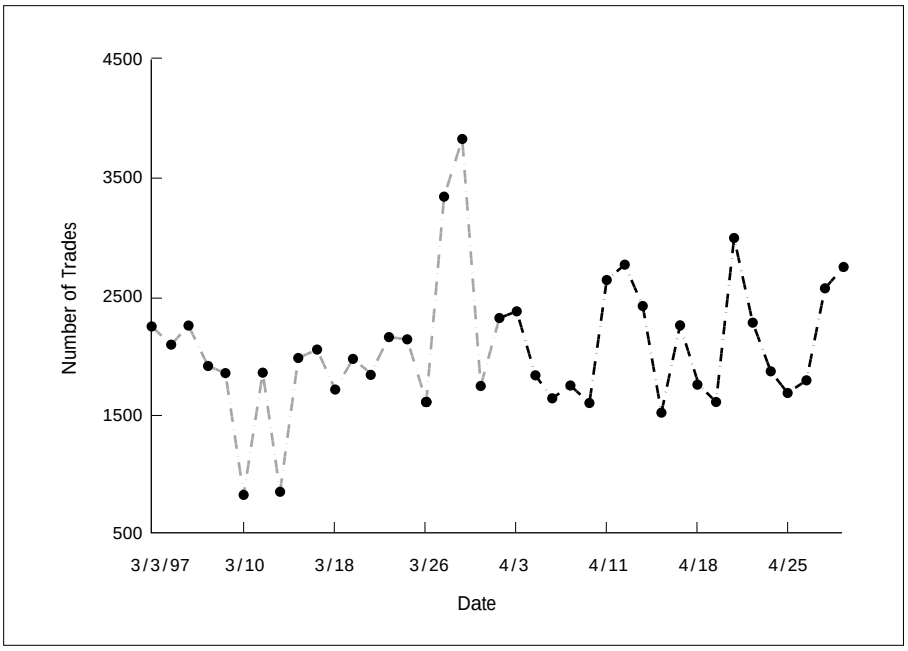
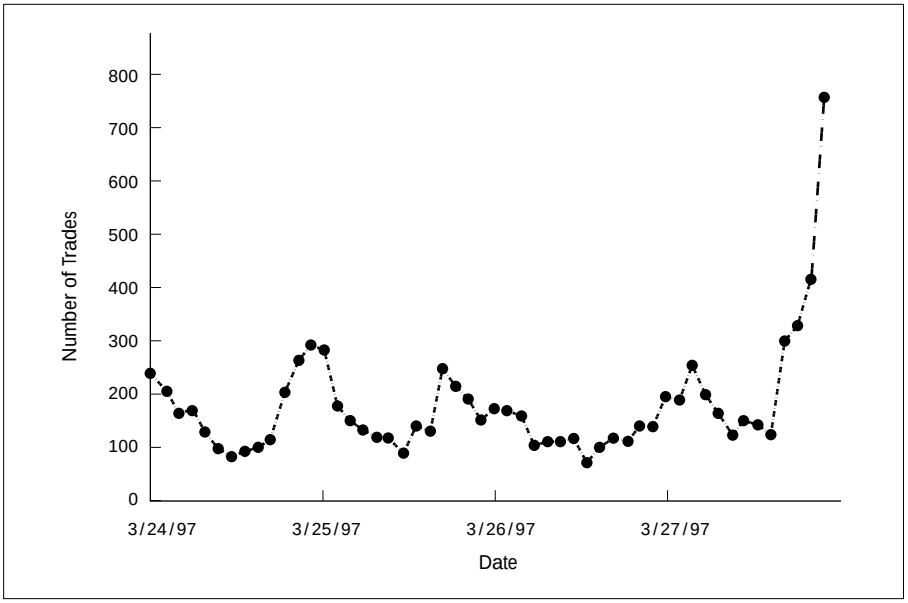


Figure 3.5

Number of trades per half-hour for General Electric



This variability may at first seem random, but careful data analysis reveals a rich degree of structure.

For instance:

- Trading activity exhibits calendar effects, tending to be light near certain holidays and heavy on days when exchange-traded options and futures contracts expire (i.e., “witching hours”).
- Trading activity has a characteristic daily and weekly tempo.
- Increasing price volatility correlates with increasing trading activity.
- On a daily time scale, trading activity exhibits long-term trends punctuated by spurts of activity which decay over a few days.
- On an hourly time scale, we observe a tendency for bursts of activity to cluster together, possibly as a result of a single order generating multiple trades.

When traders are asked what special skills they have developed, they often make reference to a sense of the market’s tempo. Having a sense for the rhythm of the market and being able to anticipate its response to emerging developments is judged an important trading skill. Our research fully confirms that there is a rich structure to the pace of trading.

The intensity factor is actually an entire submodel of trading activity. It provides a forecast of the level of trading activity and of the typical variation around that level.

Shape

For a given asset, there is considerable variation in trade sizes. Again we turn to GE to illustrate this effect (see Figure 3.6).

Obviously only the largest portfolios can place the largest orders and hence give rise to the largest transactions. We are led, therefore, to hypothesize a relationship between the distributions of portfolio sizes, order sizes, and transaction sizes. Figure 3.7 makes the comparison. In fact, we see that there is a rough similarity between these distributions. We note also, however, that there is a progressive shift leftward. Basically, as positions are moved through the market, they get broken up into multiple orders and an order may give rise to multiple transactions. Thus, a process of fragmentation shifts the distributions leftward. Here we are most probably observing the efforts of traders to minimize realization of market impact costs.

Figure 3.6

Trade size distribution for General Electric, showing percent of trades made

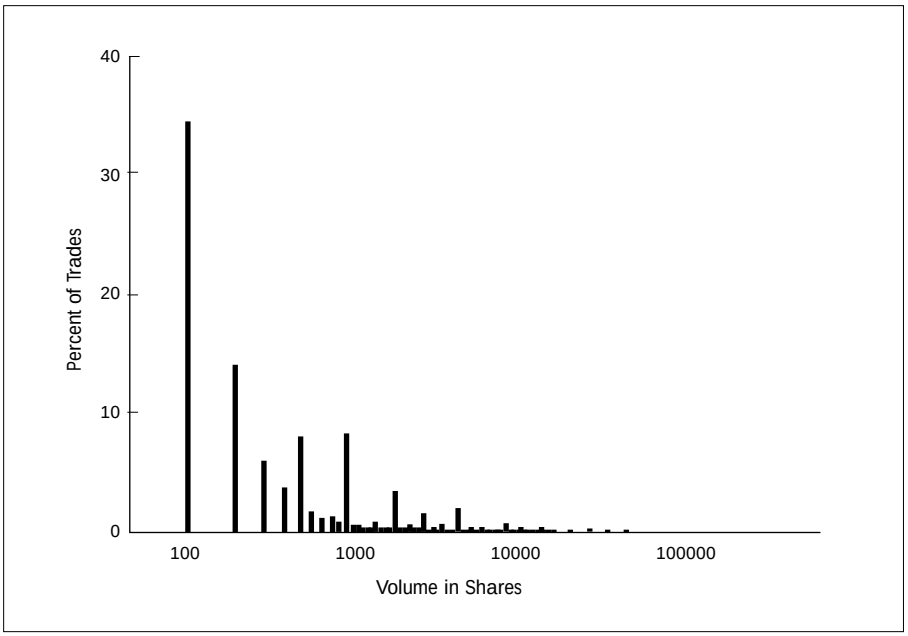
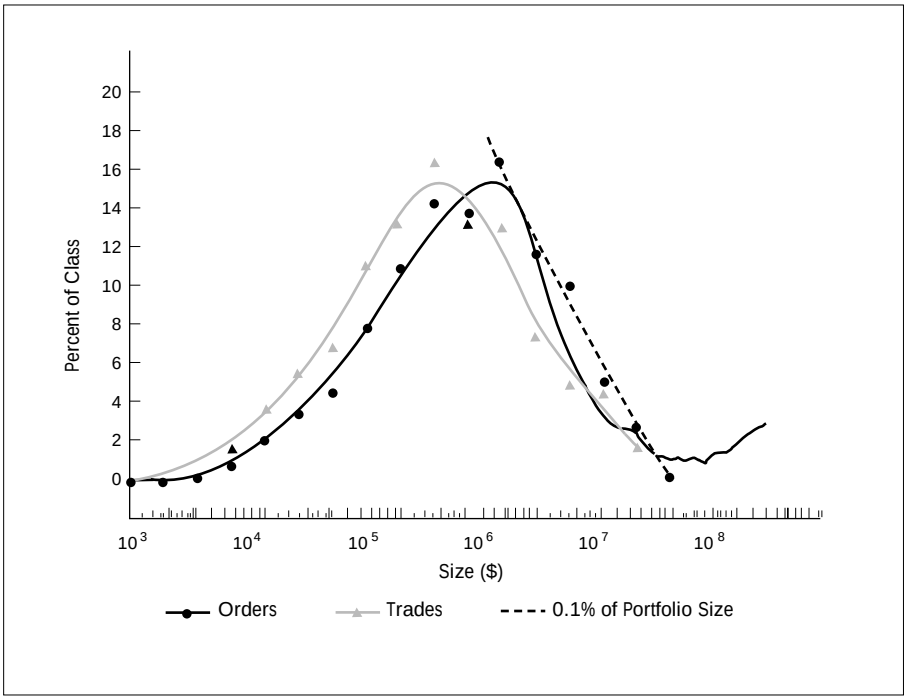


Figure 3.7

Distribution of order sizes for NYSE system orders, fill sizes for those orders and portfolio sizes for U.S. equity portfolios of $500 million or more



Another interesting feature of the data is the preference for round numbers, as indicated by the peaks at 100, 1000, and 5000 shares. We can also measure sizes in value rather than share terms, and then we find some preference for positions of $1 million. Probably these preferences arise from the experience that it is easier to find counterparties at these sizes.

A third feature of the data we may remark on is its tendency to carry through time. As we have noted, ultimately there is a connection between the distribution of portfolio sizes and of trade sizes. The mean of the distribution of portfolio sizes changes, of course, with the rise and fall of the overall market. Aside from this effect, however, the shape of the portfolio size distribution is quite stable. We expect the order and transaction size distributions also to be stable through time. Indeed, this is basically the case. However, two factors lead to greater variability in these measures:

- We have remarked on the fragmentation process, which involves a leftward shift in the distributions. The degree of fragmentation can vary with time, most likely in response to the overall level of liquidity in the market, and thus some variability in the distributions can arise.
- There can be some variation in the relative participation of institutions and individuals in the marketplace. Since the largest trades are placed by institutions, the right-hand tail of the distribution responds to the level of institutional activity in the market. If, for instance, institutions have a tendency to trade significantly more at the end of quarter than at the beginning, then the greater weight will fall into the tail of the trade size distribution.

When we look at a distribution such as that shown in Figure 3.6, the frequency of large orders seems quite slight. Such a presentation does not give a fair impression of the importance of such orders, however. In Figure 3.8 we show the percentage of traded value that is accounted for by orders of different sizes. We see that large orders, while rare, account for a non-negligible fraction of volume. Attention to modeling the tail of the distribution is, therefore, warranted.

The shape factor models the distribution of possible trade sizes. We have chosen to model this distribution as a mixture of continuous and discrete distributions. The continuous part captures the overall shape of the distribution, and in particular the tail. The discrete part captures the extra weight that occurs at round numbers.

The Liquidity Provider's Risk

Collectively elasticity ε, volatility σ, intensity ϕ, and shape ζ constitute a careful model of the uncertain trading environment faced by the liquidity provider. Suppose the liquidity provider assumes a position of size V at a price that represents a deviation δ from the equilibrium. As trading occurs, there are many possible paths by which the liquidity provider can clear this position out of inventory. By applying our model we can calculate the probability of each individual path, much as was done in Exhibit 2.1. In this way, we can find the distribution of possible returns that might be earned by the liquidity provider. This distribution is characterized by a certain mean $\alpha(V,\delta,\varepsilon,\sigma,\phi,\zeta)$ and standard deviation $\rho(V,\delta,\varepsilon,\sigma,\phi,\zeta)$. The term $\rho(V,\delta,\varepsilon,\sigma,\phi,\zeta)$ is what we have meant by the liquidity provider's risk.

Recall that the liquidity provider seeks to maximize the ratio of return (α) to risk (ρ) subject to competitive pressures. For $\delta = 0$ we have $\alpha(V,\delta,\varepsilon,\sigma,\phi,\zeta) = 0$, and as δ increases, $\alpha(V,\delta,\varepsilon,\sigma,\phi,\zeta)$ increases. Furthermore, as δ increases, $\rho(V,\delta,\varepsilon,\sigma,\phi,\zeta)$ decreases. This combination causes the return/risk ratio to increase with δ, which at a unique δ^* for each stock will yield a target ratio (τ) characteristic of the market as a whole:

$$\frac{\alpha\left(V,\delta^*,\varepsilon,\sigma,\phi,\zeta\right)}{\rho\left(V,\delta^*,\varepsilon,\sigma,\phi,\zeta\right)} = \tau \qquad \text{Equation 3.5}$$

We denote this unique δ^* as $\delta(V,\varepsilon,\sigma,\phi,\zeta,X)$. Since ε, σ, ϕ, and ζ are all characteristics of the particular asset, we abbreviate $\delta(V,\varepsilon,\sigma,\phi,\zeta,X)$ as $\delta_i(V,X)$ for asset i. As we shall now see, the ratio τ is the market tone.

Market Tone

The market tone (τ) is a market-wide number. To estimate it, we form a sample Ω of trades in a cross-section of assets on a given day. Since we require separate market tones for buyer- and seller-initiated transactions and for different market mechanisms, we restrict the sample accordingly. For V, a trade in the sample, we suppose that $K(V)$ is the estimated impact of the trade and $i(V)$ is the asset traded.

Then we set τ by the requirement that it minimize:

$$\sum\left|K(V) - \delta_{i(V)}(V,\tau)\right| \qquad \text{Equation 3.6}$$

where the sum is taken over all V in the sample Ω. The estimate of τ is based on data from a single day. From day to day, however, τ tends to change by only a few percent (at least during reasonably steady market conditions). The implication is that today's estimate of the market tone provides an adequate forecast of tomorrow's market tone.

Preliminary Calculation of Impact

Given the market tone τ and the asset parameters ε, σ, ϕ, ζ, we may calculate the market impact cost of a trade of size V as:

$$\delta(V,\varepsilon,\sigma,\phi,\zeta,\tau)V \qquad \text{Equation 3.7}$$

These estimates are made under the assumption that the market is at equilibrium at the time of the trade and that the liquidity provider holds a neutral inventory position. If the liquidity provider was looking to unwind a certain position and this trade provided such an opportunity, then the impact would probably be overestimated. Similarly, in non-equilibrium conditions, this estimate could be an underestimate.

We make the assumption of trading at equilibrium for practical reasons. Our goal is to provide forecasts valid for one to a few days ahead. Over this time frame there is little ability to predict in which direction (buy or sell) market disequilibrium will lie. Furthermore, to the extent there is any ability to predict the market's momentum, our research shows that it is already being priced by the market. This finding is entirely consistent with market efficiency. Thus, an equilibrium assumption is appropriate for the application we have in mind. For a randomly constructed sample of trades, we would expect the discrepancy between the hypothesis and reality to balance out over time. Thus, the impact estimates should retain their value, provided they are interpreted as expectations rather than on a trade-by-trade basis.

Skill

Investors' actual processes do not generate random samples of orders, however. For instance:

- "Contrarian" investors may tend to systematically trade in the opposite direction from the market's overall momentum.
- A trade list that is being worked creates liquidity that can be provided to the market opportunistically. The investor's trading operation may have the ability to efficiently manage this opportunity so that trading costs are reduced.

  The investor may hold non-consensus preferences, with the result that the opportunity costs and risks of working orders or using search market mechanisms are judged lower than the direct cost of market orders.

All of these factors could reduce the investor's cost experience below the preliminary estimate. Similarly, an investor who follows the herd, trades sloppily, and is very risk-averse could incur higher costs. Some adjustment is, therefore, required to reflect these process-level determinants of trading cost.

Accordingly, we introduce a skill factor χ that is used to adjust the forecast impact as:

$$(1-\chi)\ \delta(\mathrm{V},\varepsilon,\sigma,\phi,\zeta,\tau) \qquad \text{Equation 3.8}$$

and the impact cost is adjusted similarly. Thus, a skill of 0 corresponds to the preliminary model forecast, while a skill of 1 implies the ability to eliminate market impact costs through skillful exploitation of one's trading opportunities.

Naturally, the skill factor needs to be estimated for each investor separately by comparing average trade costs to the forecast for the generic investor.

The Market Impact Function

We have completed the description of the model. In summary, we see that the market impact cost function *F()* is:

$$F(V,\varepsilon,\sigma,\phi,\zeta,\tau,\chi) = (1-\chi)\ \delta(V,\varepsilon,\sigma,\phi,\zeta,\tau)V \qquad \text{Equation 3.9}$$

Although the functional form of *F()* is quite complex, we can easily determine the effect of each variable on the forecast cost.

- As the volume V increases, the liquidity provider has to carry inventory longer, so impact rises. Cost is the product of impact and volume, so cost also rises and at a faster than linear rate.
- As the elastic coefficient increases, order flow responds more quickly to price signals, so time-to-clear decreases and thus cost decreases. As the elastic fraction increases, the maximum response of order flow increases, also leading to lower cost. Thus, in general, as elasticity increases, cost decreases.
- As volatility σ increases, a liquidity provider's risk increases and cost increases.
- As the intensity of trading ϕ increases, time-to-clear shortens, and so cost decreases.
- As the mean of the shape distribution ζ increases, time-to-clear shortens and so cost decreases. As the dispersion of the shape distribution increases, the uncertainty in the time-to-clear increases and so cost rises.
- As the market tone τ increases, liquidity becomes more costly and cost increases.
- As the skill χ increases, the investor either purchases less liquidity or purchases it more economically, so costs decrease.

The Piecewise-linear Approximation

The full market impact function *F()* is very time-consuming to compute. Accordingly, we often replace it by an approximation that can be computed quickly. There are many different ways of constructing an approximating function. For some applications, however, it is convenient for the approximating function to be piecewise-linear, and so we use a function of this class. Constructing an approximation always means accepting some discrepancy between the base function and its approximation. We have constructed the approximation by requiring that:

- The approximation is always equal to or greater than the base value
- The approximation is within a fixed tolerance of the base value

We require the approximation to be biased high because in an application setting this will usually be the preferred deviation. We set the tolerance to a few percent. As a result, the discrepancy introduced by the approximation is less than the statistical estimation error involved in creating *F()*. Hereafter, when we refer to the cost function *F()* we will usually mean the piecewise-linear approximation to *F()*.

The Realized Spread

For the smallest trades, we find that the predicted impact is generally less than the minimum tick size. Accordingly, we need to increase the cost function to reflect the reality of minimum impacts. We rely on a measure of the spread for this adjustment. Many trades actually occur inside the quoted spread. Accordingly, we deflate the quoted spread by a factor that measures the volume of trading inside the quoted spread. We term this deflated spread the realized spread, since it reflects the spread investors actually pay. For the smallest trades, the forecast impact is rounded up to the realized spread, and the piecewise-linear cost function is adjusted accordingly.

Summary

There is considerable fine detail to market conditions for an asset. We model this detail with attention to the features that are of particular significance to a liquidity provider. Thus, we produce four submodels of elasticity, volatility, intensity, and shape that characterize market conditions for an asset. From this information, we can evaluate the liquidity provider's risk. Combining the quantity of risk with the market tone leads to a forecast cost for the generic investor. A process-level adjustment, the skill, then adapts the forecast to a specific investor. Finally, a piecewise-linear approximation is introduced to achieve a computationally efficient implementation. The final result is a forecast of market impact costs.

Testing the Market Impact Model™

Nicolo G. Torre, Ph.D.
Mark J. Ferrari, Ph.D.

REPRINTED FROM *HORIZON*, THE BARRA NEWSLETTER, FALL 1998, PUBLICATION NUMBER 168

Testing the Market Impact Model

In previous articles in this series, we have described the purpose of the Market Impact Model, the economic intuition behind it and the model itself. Now it is time to explore how well the model actually works. There is a wide variety of tests that may be made of the model. We shall begin with the simplest test and work up to the more complicated ones, sharpening our understanding of cost modeling as we go.

The simplest test one might think to make of the Market Impact Model is to forecast the cost of a particular trade, then work the trade and compare the resulting cost to the model's forecast. Unfortunately, this test is not particularly informative because the model forecasts the expected or average cost of trading. Any one trade, however, may happen to arrive in the market simultaneously with a large competing or complementary order. Consequently, the cost experience for a single trade will depart, often significantly, from the mean experience. Testing the Market Impact Model against an individual trade lacks power, for the same reason that a valuation model cannot be evaluated by looking at the performance of one name on its buy list; little is learned by comparing the mean of a distribution with just one draw from that distribution.

The evident solution to this problem is to look at a larger number of trades. Comparing the mean cost of a sample of similar trades with the mean forecast by the model increases the power of the test. The comparison can be made for a number of different samples of different types of trades. Estimating the mean cost of the executed trades in each sample is itself an interesting problem. We will return to this momentarily, but for now, let us suppose that we have only a rough notion of our trade costs, such as the trading desk's impression of what it costs to do the trades. From such rough estimates we will not be able to develop any very precise measurement of model performance. However, we can still conduct some basic validation of the model. We might begin by sorting a sample of trades into two groups. These groups can be distinguished in various ways—for instance, the date on which the trade occurred, the type of assets traded, buys or sells, how the trades were worked, trade sizes, and so forth. We could choose the groups to be identical with respect to all characteristics save one. For instance, we might take one group to be trades in large capitalization stocks and the other to be trades in small capitalization stocks. For the costs of each group we will have only our rough estimate, but as long as the error in our estimate is independent of the dimension in which

the groups differ, we should have a reasonable measurement of their relative cost to trade (e.g., the small capitalization stocks are three times more expensive to trade, all other things being equal). Next, using the Market Impact Model, we can forecast the cost to trade both groups and by taking the ratio arrive at a forecast of the relative cost. Comparing the forecast of relative cost to the estimated realized relative cost then provides a test of model performance along the dimension that distinguishes the two groups. Performing tests of this sort along multiple dimensions allows one to build a basic confidence that the model is behaving reasonably.

Next, one would like to move towards a more quantitative measurement of model performance. For this purpose, one requires a good measure of realized cost. A conventional choice is to measure cost as the difference between the price paid and a pre-trade reference price. It must be noted, however, that the number that results is not a pure measure of cost because it includes the asset's return due to factors other than one's own trading activity. For large orders worked over a long time horizon, such factors may dominate the measured value. Thus, it must be realized that costs are not directly observed, but rather are estimated with a varying level of measurement error. When applied to data containing measured error, many standard techniques (such as regression analysis) can lead to erroneous conclusions. Fortunately, the application of appropriate statistical techniques can compensate for these measurement errors.

The first step in analyzing data containing measurement error is of course to reduce the error as far as possible. A major source of error in estimating cost as the difference between a pre-trade reference price and the trade price is the confounding effect of other's trading. However, for the purpose of testing the Market Impact Model, there is no necessity to limit our analysis to trades originated by a single investor. Instead, one could aggregate trades placed by several investors and look at the cumulative impact. The natural extension of this idea is to aggregate all trading which occurs in a time interval. By doing so, we may eliminate the confounding effect of trading by others. We have found a half-hour period to be a suitable aggregation interval as it balances the need for multiple measurements during a trading day with the need for each measurement to be based on a reasonable level of activity.

The second step in analyzing the data is to summarize the measurements in a form suitable for comparison with the forecast. The forecast relationship between impact and volume is very nearly a square-root relationship.

Accordingly, we summarize the data by fitting the curve:

$$\text{Cost} = \gamma_{emp}\sqrt{volume}$$

to the measured data. Similarly, we fit the curve

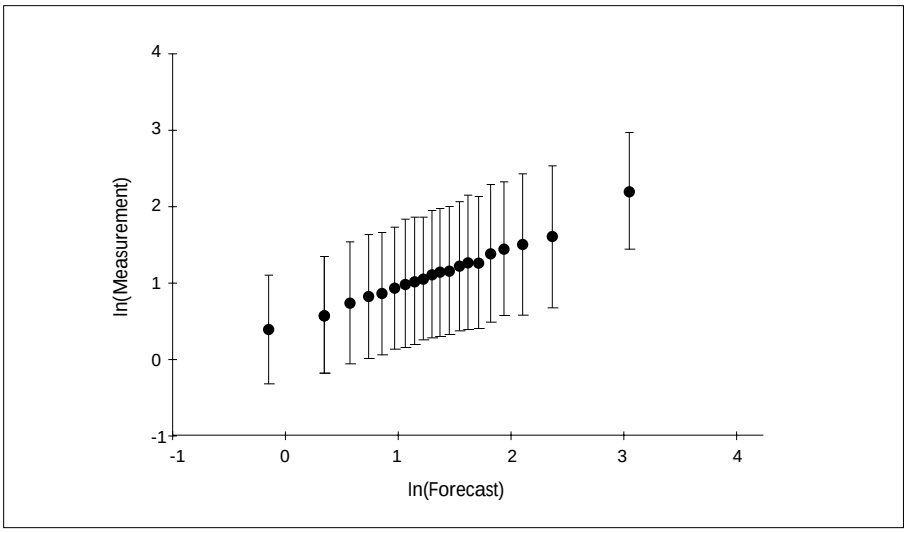
$$\text{Forecast} = \gamma_{mdl}\sqrt{volume}$$

to the model forecast. Then comparing γemp to γmdl gives an evaluation of model performance.

The third step in our analysis is to take a quick look at the data to see the basic relationship. For each stock i on day t we have a model forecast γmdl(i, t) and a measured result γemp(i, t). To summarize the data, we bin these observations into one of 20 bins on the basis of γmdl(i, t). Let $\overline{\gamma}$mdl(n) and $\overline{\gamma}$emp(n) denote the averages of γmdl(i, t) and γemp(i, t) respectively over the nth bin. Figure 4.1 then shows a graph of $\overline{\gamma}$mdl(n) versus $\overline{\gamma}$emp(n). As can be seen there is a clear linear relationship between the mean of the forecast and the mean of the observation. It can be also observed, however, that the measured cost is generally less than the forecast cost. To understand this discrepancy between forecast and realization, we must probe the data more extensively.

Figure 4.1

Comparison of Aggregated Forecasts and Measurements on Log-Log Axes



There are essentially two hypotheses that might explain the discrepancy. One is the possibility of systematic error in the model. Were this to be the case, one could solve the problem simply by scaling the model forecast down. The other possibility is a systematic error in the measured cost, which results in the measurement underreporting the true cost. Clearly it is important from an

investment viewpoint to distinguish between these hypotheses. We would not want to adjust the model to an erroneously low estimate of cost and thus stimulate investors to overtrade and so incur the actually higher true cost.

To distinguish between model error and measurement error, we must introduce a third estimate of transaction costs to arbitrate between the other two. Technically this third estimate is known as an *instrumental variable*. Consider the situation where we have three estimates of the transaction cost: the model estimate γ_{mdl}, the empirical measurement γ_{emp}, and the instrumental estimate γinst. Each of these estimates is related to the true cost γ_{true} by

$$\gamma_{mdl} = c_{mdl}\gamma_{true} + \varepsilon_{mdl}$$

$$\gamma_{emp} = c_{emp}\gamma_{true} + \varepsilon_{emp}$$

$$\gamma_{inst} = c_{inst}\gamma_{true} + \varepsilon_{inst} \qquad \text{Equation 4.1}$$

The departure of the coefficients c_{mdl}, c_{emp} , and c_{inst} from 1.0 captures the systematic bias in the estimates and the variables ε_{mdl}, ε_{emp}, and ε_{inst} represent random (unsystematic) errors. Although the true value γ_{true} is unobserved, one can still solve the system of equations to find c_{mdl}, c_{emp} , and c_{inst} provided the random errors are uncorrelated. To see how this is possible, consider the computation of the covariances implied by Equation 4.1. For instance

$$\mathrm{cov}(\gamma_{mdl,}\, \gamma_{emp}) = c_{mdl}\, c_{emp}\ \mathrm{var}(\gamma_{true})$$

This equation relates an observable quantity on the left hand side to three unobservable quantities on the right hand side. Calculating all the other observable covariances in this manner produces a set of equations which could in principle be solved to find the unknown coefficients c_{mdl}, c_{emp} , and c_{inst}. In practice, one deals with finite data samples. Thus one does not solve the equations exactly, but rather finds the statistical estimates c_{mdl}, c_{emp} , and c_{inst} which are most consistent with the Equation 4.1 and the actual data.

To apply this methodology, we require an instrumental variable. The method will work even if the variable is not a particularly good estimate of the true cost, thus we need not assume that c_{inst} =1 or that var(ε_{inst}) is small. However, we do require that the errors einst in the instrument be uncorrelated with the errors emdl and eemp in the other two estimates. For an instrumental variable we use the square root of the ratio of trade size to capitalization. Thus, the assumption is that a trade of fixed size will cost more in a small cap stock than in a large cap stock. This assumption is intuitively appealing and is borne out

by actual trading experience. In the construction of the Market Impact Model we do not use capitalization as a variable, so it is reasonable to expect that the model errors will be uncorrelated with the errors in the capitalization instrument. Similarly, we can expect the observational errors in the empirical estimate to be uncorrelated with the capitalization estimate. Hence, capitalization can serve as a suitable instrumental variable for our analysis.

Applying this methodology we can decide between the hypothesis of systematic error in the model or in the measurement. The full details are presented in the *Market Impact Model Handbook*, so here we shall just summarize the results. There is no evidence that the Market Impact Model systematically overpredicts transaction costs. If anything, the data are consistent with a slight underprediction (approximately 6%). On the other hand, there is compelling evidence that the empirical estimate is systematically biased low, by about 35%. This is a very interesting result! Intuitively we understand how this bias could arise. Survivorship bias deletes the expensive trades from our sample, because investors will not trade in the face of excessive costs. Until now we had no way of knowing whether this bias was large enough to matter. Now we have concrete evidence that the effect is real and significant. In particular, investors who extrapolate from past trading experience to forecast their future costs are setting themselves up for a significant unpleasant surprise. Their actual costs will prove higher because past executed trades represent a biased sample of market conditions, and biased samples produce inaccurate forecasts. In contrast, forecasts based on the Market Impact Model have proven to be largely free of systematic errors.

Having verified that the Market Impact Model is free of systematic error, we next turn our attention to estimating the size of the random error. The answer is simply the variance of emdl. On its own, however, this quantity is not very interpretable. We seek, therefore, a figure of merit that speaks to the practical usefulness of the model. We take as our standard of accuracy the cost estimate for a list of trades in fifty randomly selected stocks. The attraction of this standard is that it corresponds to the application for which the model was designed, namely evaluating the cost-benefit tradeoff of a portfolio rebalancing. Based on our data analysis, we find that the median error in the model cost estimate for such a list is generally less than 10%. To put this level of performance in perspective, this is the same accuracy we estimate for BARRA risk models applied to similar sized portfolios. In retrospect, perhaps it is not surprising that the performance of the Market Impact Model should turn out comparable to the risk model, because the Market Impact Model itself has a risk model at its heart.