The world runs on data. Every decision, no matter the magnitude, is based on data. They can take many forms, but are broadly classified as numbers or words, or better yet, quantitative and qualitative, respectively.

Data, of course, are not very useful because they’re usually raw and chaotic. They have to be organized, cleaned, often aggregated or summarized, and used in a model to draw out important elements – information – useful for decision making. This certainly holds for every business function, not the least of which is pricing.

Pricing analysts typically build models, usually regression models, from internal databases such as a data warehouse or “data mart” organized into tables of customer purchases. However, oftentimes these data sets, because they are based on purchasing decisions, do not reflect the buying behaviors of the entire buyer market. The data sets actually represent self-selected samples, and because of self-selection, the samples are not random, but biased, as the author explains. Walter R. Paczkowski, Ph.D. is President of Data Analytics Corp. He can be reached at info@dataanalyticscorp.com.

For example, the marketing department of a national seller of exercise equipment maintains a data mart of all the equipment sold to private gyms, schools, institutions, and households. National marketing campaigns in a highly competitive market are developed and tracked using this data mart. Pricing models are also built by a team of pricing analysts using this data.

The equipment seller’s product sales are in tables called Orders, Products, and Customers. The Orders table contains data about each order with fields for an order number, order date, ship date, order quantity, product ID number (PID), and customer ID number (CID). The Products table has fields for a linking PID, product description, current price, and maybe price change history. The Customers table has a linking CID, customer name, address, and maybe some demographics.

By linking, filtering, and aggregating the tables, perhaps with SQL commands, the pricing analyst can build a table of quantity ordered of a product (say, treadmills by schools), the relevant prices, and pertinent aggregate customer information. The pricing analyst could sample the records using an appropriate sampling technique such as simple random sampling or stratified random sampling dividing the data into training, validation, and test tables. In the end, a Pricing Data Table (PDT) ready for modeling will be produced.

The pricing analyst would typically build a regression model (using OLS – Ordinary Least Squares) which is usually straightforward with quantity sold or sales on the left-hand side and prices (own and cross prices), maybe a time trend if the data are times series, and some demographic variables such as income on the right-hand side. Some corrections for heteroskedasticity or autocorrelation (depending on the data type) may also be done. Finally, elasticities, the key piece of information needed for pricing decisions, would be estimated and proposals for price changes or price setting for a new offering will be made. This process is illustrated in Fig. 1.

Since the data in the PDT are observed, it is natural to believe they reflect the “real world” – what people actually do – and are therefore the correct data, the only data, needed for pricing. What is better for estimating elasticities than “real data”?

Figure 1: A Typical Pricing Modeling Process

![Diagram of a Typical Pricing Modeling Process]

Orders Table → Products Table → Customers Table → SQL Engine → Aggregator and Sampler Functions → Pricing Data Table → Statistical Model
Real World Data
It’s true that the data in the PDT reflect what people really did in the market; their preferences were revealed by their actions and we observed their purchases. Hence, the data could be called revealed preference or observational data. They show the choices people made when presented with a market price and product attributes. But for developing elasticities and making critical pricing decisions, this observational data may not be the best to use and OLS may not be the best estimation method. There are some problems associated with observational data. For example, the data may suffer from:

- Multicollinearity
- Lack of variation
- Timeliness
- Existence

There is another, more subtle problem: the people in the data mart self-selected to be there.

The Self-Selection Problem
Self-selection means that people chose to be in the data mart because of their purchase. If they purchased elsewhere or chose not to purchase at all because the price is too high, they would not be in the company’s data mart as a buyer; there would be no record of them. Yet they are there because they liked the product’s features (including price) better than a competitor’s product.

For those in the data mart, all we know is what they bought and the price they paid. We don’t know what they compared the product to in order to make their final purchase decision (basically, we don’t know the competitive price and features they saw).

For all those in the data mart who made a purchase decision, there are many others not in the data mart either because they chose not to buy anything or because they chose to buy the competitive product at a better price; we just don’t know which one. In short, we don’t know the true, total population of buyers.

The pricing analyst could argue that at least the competitive prices are known because the competitive assessment group in marketing tracks this data using input from the sales force, advertisements, and surveys so this is not an issue. Unfortunately, because the pricing analyst knows the prices and features does not mean the customers (and non-customers) knew those same prices when they made their purchase decision. We also can’t say with certainty what prices people saw because we simply don’t know.

Impact of Self-Selection
Since people self-selected to buy and thus appear in the data mart, the OLS estimation will be impacted. Basically, because of self-selection, the sample is not a random sample; it’s biased. A random sample would provide a description of the entire purchasing population, not just the one in the company’s data mart.

The OLS estimates will be inconsistent because the models fail to account for this self-selection; they’re misspecified. To put it briefly, inconsistency means that if we could allow the sample size to become larger, the OLS estimates would not converge to the true parameter value, in our case the elasticities; they would be biased. Intuitively, we want the estimates to be consistent so that in very large samples we get the correct answer.

With a nonrandom sample, no matter how large the sample becomes it will still not reflect the true population of buyers. If the estimated elasticities based on the nonrandom, self-selected sample are biased, then how can we make the correct pricing decision?

The whole purpose of modeling is to take the raw and chaotic data and extract the information needed for pricing to “beat the competition” in a competitive market. But if the estimated elasticities are wrong, then how can we beat anyone? In fact, the competition will win if they use correctly specified models based on a random sample.

The Basics of Misspecification
The reason for the model misspecification is that in self-selecting, people are making two decisions. The first is a vendor choice decision – who they will buy from – and the second is a quantity decision – how much they will buy. The typical model outlined above focuses only on the second of the two decisions, the first being ignored. But the first cannot be ignored since they’re not independent. They’re functions, typically, of the same key drivers such as price and product attributes. There are two models!

The vendor decision model can be summarized in a key ratio that is incorporated in the quantity decision model. Getting this ratio is a complicated modeling issue, but with the right data that reflects the population and sophisticated software, it can be done. This is a relevant variable that captures the first decision. To ignore this variable is to omit a relevant variable and thus misspecify the model, and misspecified models lead to biased results.

Handling the Problem
This is a complicated problem, but there are two recommendations. First, recognize that a problem exists because of self-selection. This is not an insurmountable problem or a killer of data warehouses or data marts. But not to recognize it or, worse, ignore it is to help the competition win.

Second, don’t rely solely on data warehouses and data marts as the source of all data. Aside from self-selection, the four other problems mentioned above still hold and have to be addressed. These are characteristics of observational data. Consider using discrete choice experiments in which these problems, including self-selection, are avoided. The self-selection problem, for example, is a non-issue because the sample used in an experiment is a random sample of all potential buyers. If the data mart is used, consider supplementing the data with a random sample of people who did not buy the product.