Marketing Mix Modelling and Big Data

P. M Cain

1) Introduction

Big data is generally defined in terms of the volume and variety of structured and unstructured information. Whereas structured data is stored in conventional relational formats, unstructured data comprises multimedia content such as e-mail, text, social media feeds and web pages. The value of this type of information is typically exploited using data mining and machine learning techniques, uncovering patterns and predicting customer behaviour with results delivered in real or near-real time. Well known examples are Google Translate, Autocomplete and the targeting models of Netflix and Amazon, where products are recommended based on historical usage patterns.

In the marketing sector the growing wealth of data on individual purchase behaviour, online activity, social media and socio-demographic profiles is changing the face of media buying and analytics. Firstly, by consolidating the information into single data management platforms, media agencies can achieve more granular segmentation of viewing audiences, leading to increasing efficiency in digital media buying and targeting.[1] Secondly, the sheer volume and complexity of available data has prompted an increasing use of machine learning methods to generate marketing insights. This typically covers correlation based model building, network analysis, consumer segmentation, classification and forecasting.

Against this backdrop, more conventional 'small data' analytics such as econometrics and marketing mix modelling have taken something of a back seat. This is unfortunate since big data should not necessarily be viewed as more accurate with no need for conventional interpretation. Data mining techniques for example are ideal for association rule learning, where customer purchasing patterns on frequent and jointly purchased products can help guide marketing strategy. However, such methods shed little light on underlying data generation processes, marketing ROI and the *ceteris paribus* causal impacts central to simulation and scenario planning.[2]

Consequently, despite the significant role that unstructured big data can play in marketing analytics it should not overshadow the importance of traditional analysis of large structured data sets. The key is to build analytical frameworks that can harness the value of increasing data size, yet retain the benefits of sound economic theory and valid causal inference. This is the domain of Big Data econometrics.

[1] Improved targeting in this way is analogous to the Amazon and Netflix models. Note, however, that data management platforms are essentially treated as if they were single-source consumer panels across all off and online touchpoints. This is a strong assumption as we are attempting to combine a mass of highly granular disparate information based on varying consumer samples coupled with multiple online platform usage.

[2] Note, however, that machine learning techniques often play a key role in Bayesian network analysis to uncover off and online consumer purchase journeys. Results are often given a causal interpretation and used for simulation and marketing budget optimisation.

2) Structured data sets in marketing

The size of any data set is essentially driven by the level of disaggregation over products (depth) and time (frequency) together with the number of variables involved (variety). Figure 1 illustrates a typical case, representing various levels of granularity.

Figure 1: Longitudinal data structure

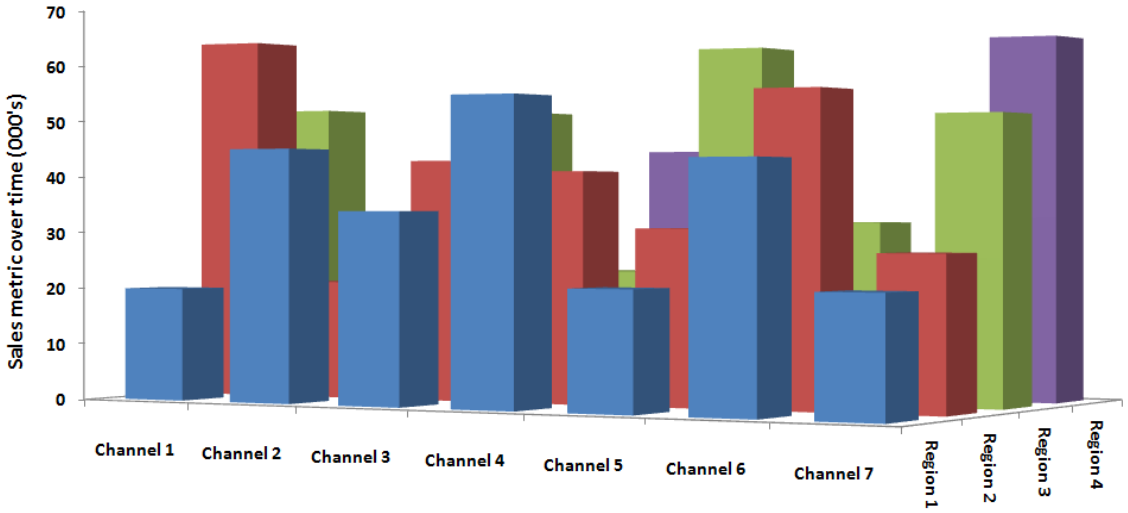

Figure 1 splits market level sales over time into cross-sections where each block depicts a time series of consumer demand for each product-channel-regional combination. Sales channels may be further divided into stores or individual customers. Coupled with information on pricing, promotion, offline media plus customer attributes and profiles, the size of the data set starts to expand significantly.

The digital revolution has only exacerbated this problem, leading to a proliferation in the volume and variety of online information such as web traffic by source, natural and paid search behaviour by platform, display exposure and social media feeds – much of which is available at individual consumer level by day. Merged with offline data to create a holistic view of consumer demand creation, we truly are in the realms of Big Data.

Econometric analysis of such data brings two broad challenges. Firstly, the consumer purchase journey is much more complicated, with an increasing number of endogenous outcome variables to deal with. As a result, accurate off and online marketing attribution is more demanding as set out in my previous article *Advanced Methods in Marketing Econometrics*. Secondly, as the number of dimensions illustrated in Figure 1 grows, so too does the number of detailed marketing response parameters that businesses require. Obtaining stable results using conventional methods is increasingly difficult. In this article, I look at some common approaches that help resolve this *dimensionality* problem.

3) Conventional disaggregated data analysis

Analysis of disaggregated data sets is very common in marketing analytics. Rather than estimate one product, one cross section at a time, it is preferable to take advantage of the longitudinal structure, stack the data and estimate all dimensions simultaneously. Typically, we are seeking to quantify cross sectional deviations (interaction effects) from the market mean or base level of response (main effect). This can be seen as analogous to data mining techniques such as CHAID and helps improve targeting of marketing investments. To illustrate the issues involved, consider the data structure for Figure 1 expressed in stacked form as follows:

$$y_{it} = \boldsymbol{x_{it}}\boldsymbol{\beta_i} + \varepsilon_{it} \quad (1)$$

Where *i = 1-N* denotes the defined cross-sectional unit, *t = 1-T* denotes the time period, y_{it}, denotes a vector of dependent variables for product or brand sales, $\boldsymbol{x_{it}}$ denotes a row vector of *K* current and lagged explanatory variables for cross-section *i*, $\boldsymbol{\beta_i}$ is a K-vector of response coefficients and ε_{it} represents a vector of error terms. Specific intercepts or fixed effects are typically added to each row to account for mean cross-sectional differences.

Appropriate estimation of (1) depends on the properties of the error structure, both within and across cross-sections. Classical Ordinary Least Squares (OLS) requires that the error covariance matrix of the i^{th} cross section satisfies the standard assumptions of constant variance and zero serial correlation:

$$E(\varepsilon_{it}\varepsilon_{it}^{'}) = \delta_i^2 I_T = \Pi_i \quad (2)$$

With zero contemporaneous error correlation across cross-sections:

$$\Omega = \begin{bmatrix} \Pi_1 & 0 & 0 \dots & 0 \\ 0 & \Pi_2 & 0 \dots & 0 \\ : & : & : & : \\ : & : & : & : \\ 0 & 0 & 0 \dots & \Pi_N \end{bmatrix} \quad (3)$$

OLS estimation of (1) provides the coefficient vector β_i, with marketing response estimates specific to each cross-section *i*. In circumstances where (2) and (3) do not hold, Generalised Least Squares (GLS) approaches are typically applied. This uses OLS to estimate the relevant error structures and transform the data such that (2) and (3) are then applicable.[3]

4) Big Data approaches

As modern data sets burgeon in size, obtaining reliable detailed marketing response information can be challenging. Full interaction time series-cross sectional approaches such as model (1) are often unstable as the number of dimensions and parameters increase, delivering many zero and/or incorrectly signed effects. A natural solution to this dimensionality problem is to reduce the number of estimated parameters, thereby increasing the available degrees of freedom.

[3] Incorporating digital media into equation (1) requires a model of the off and online consumer purchase journey. Consequently, endogenous outcome variables such as web traffic typically appear as explanatory variables. Under these circumstances, alternative instrumental variable estimation techniques are required.

The simplest approach is to aggregate and reduce the size of the data set. This is a fairly common practice: big data challenges often revolve around storage issues in the first instance but, once processed, data are aggregated to simplify analysis. In marketing analytics, behaviour is often summed over time to a weekly frequency and attention is focused on average relationships across brands and sales channels.

Valid aggregation, however, requires quite specific assumptions about consumer behaviour (Deaton and Muellbauer, 1980). Furthermore, if the demand relationships are non-linear at granular levels then application of these same forms to linearly aggregated data results in bias. Consequently, alternative methods are required to handle increasing data size and granularity that obviate the need for aggregation over product and consumer dimensions.[4]

i) Classical pooling

The most basic approach is to pool the data. This provides a single average response coefficient β for each relevant explanatory variable in model (1). The downside is that this ignores response heterogeneity over cross-sections and is of little use to media planners and budget holders seeking guidance on media targeting at regional level. This can be remedied by regional pooling, but at the cost of increasing the number of parameters whilst still imposing homogeneity across products and consumers.

ii) Hierarchical Bayes

A more flexible technique is to introduce random coefficients. Equation (1) is the typical structure of the classical or frequentist approach to statistics, where the model parameters are regarded as unknown fixed quantities to be estimated from the data. In the Bayesian approach, parameters are viewed as unknown outcomes of a random process determined by another higher level joint distribution. In the context of Figure 1, this assumes that each cross sectional coefficient is drawn from a population distribution shared by all the cross-sections. This is a strong assumption, but leads to a dramatic reduction in the number of estimated parameters. For example, consider model (1) re-written as follows.

$$y_{it} = \boldsymbol{x_{it}\beta_i} + \varepsilon_{it} \quad (1a)$$

$$\boldsymbol{\beta_i} = \boldsymbol{\beta} + \boldsymbol{v_i} \quad (4)$$

$$y_{it} = \boldsymbol{x_{it}\beta} + u_{it} \quad (5)$$

$$u_{it} = \boldsymbol{x_{it}v_i} + \varepsilon_{it} \quad (6)$$

Where equation 1(a) represents each cross-sectional (micro) model and equation (4) represents the higher level (macro) distribution for coefficients $\boldsymbol{\beta_i}$ with mean $\boldsymbol{\beta}$ and error $\boldsymbol{v_i}$ - denoting the random spread around the mean. It is this micro and macro view that gives the model its hierarchical structure. Combining 1(a) and (4) leads to model (5), with a composite error term u_{it}. Covariance matrix (2) then becomes:

[4] Given the increased storage needs and computational complexity involved in big data sets, these methods often deploy *sparse matrix* techniques to facilitate large scale modelling and optimisation.

$$E(u_{it}u_{it}^{'}) = E[(\boldsymbol{x_{it}v_i} + \varepsilon_{it})(\boldsymbol{x_{it}v_i} + \varepsilon_{it})^{'}] = \sigma_i^2 I_T + \boldsymbol{x_i \Gamma x_i^{'}} = \widetilde{\Pi}_i \quad (2a)$$

The error covariance matrix of each cross-section is now a function of both the variance (σ_i^2) and the parameter spread over cross-sections ($\boldsymbol{\Gamma}$). The balance between the two determines the estimated β_i coefficients and reveals the distinctly Bayesian nature of the approach. A high value of σ_i^2 relative to $\boldsymbol{\Gamma}$ implies relatively imprecise cross-sectional estimates. Consequently, the data have little to offer and the individual β_i values are *shrunk* towards the pooled (prior) mean value. Conversely, where $\sigma_i^2$is low relative to $\boldsymbol{\Gamma}$ so the sample data are more informative and the cross-section specific estimates dominate with minimal shrinkage. In this way, the HB estimator is essentially a weighted average of the pooled and cross-sectional estimates.[5]

The hierarchical structure of model 1(a) – (6) indicates that coefficients by cross-section can be obtained simply through knowledge of the mean and variance of the macro distribution (4) plus the error variance σ_i^2 of the micro model 1(a).[6] This is far more parsimonious than the classical approach and often seen as a distinct advantage in the face of modern large data sets. Parameter estimation sets the mean of the macro distribution to the market level (pooled) estimate and the variance is derived from the global spread of the individual cross-section parameter estimates.[7] Alternatively, where *systematic* regional differences are known to exist, it is preferable to set regional mean priors via pooling across products and chains, with variances estimated using intra-regional coefficient spreads.[8]

iii) Attribute based models

Our third example is based on the economics of how consumers shop for products. Marketing mix models are based on conventional microeconomic demand theory, where consumer preferences are defined over the individual products themselves. However, an alternative approach defines preferences across higher level product attributes and characteristics (Lancaster, 1971). For example, the television category can be divided into brand, screen size and technology and further divided into brand name, dimensions and LCD/LED/Plasma/3D. Provided that there is a sufficient level of commonality in attributes and characteristics across the category, a complete product (SKU) level data set can be fully described over a significantly reduced number of dimensions.

---

[5] Note that as the number of time series observations increases, so the Hierarchical Bayesian result converges to the classical cross-sectional specific estimates.

[6] The Hierarchical Bayesian model essentially assumes that cross sectional differences are driven by chance. However, practitioners typically interpret and use them in exactly the same way as standard *systematic* fixed coefficients. Strictly speaking this is invalid, but can be alleviated by introducing time invariant fixed factors for cross-section *i* into equation (4) as discussed in Western (1998).

[7] The Hierarchical Bayesian model is identical to the fixed effects model in calculating cross sectional specific coefficients. However, these estimates are purely an intermediate step and only used to calculate the covariance matrix $\boldsymbol{\Gamma}$.

[8] This estimation approach is known as Empirical Bayes. Pure Bayesian approaches set priors independently of the data and represent an increasingly popular method of introducing user-control into marketing mix modelling. A prior for $\boldsymbol{\beta}$ allows us to set the mean value of the macro distribution to externally given values. This is particularly useful if we wish to constrain parameters to be positive or negative and/or set values consistent with previous studies. Priors for the coefficient dispersion (covariance) matrix $\boldsymbol{\Gamma}$ then allow control over the degree of shrinkage around the mean.

An application of the characteristics approach to demand in the marketing literature can be found in Fader and Hardie (1996) and represents a highly efficient method of parameter reduction. In practice, it is common to combine this approach with the Hierarchical Bayesian technique for even greater parsimony as data sets expand in size.

5) Concluding remarks

In the wake of the Big Data revolution, analytical methods such as data mining and machine learning have taken centre stage in unstructured data analysis. Although such methods are highly useful in marketing analytics, it is important not to lose sight of the distinct advantages of 'small' structured data methods such as econometrics which play a key role in marketing ROI, simulation and causal based inference.

In order to handle increased data size and complexity, econometric analysis often aggregates the data to more manageable proportions. However, this loses the value that modern data granularity has to offer. This article has looked at three common techniques that help avoid excessive aggregation. Each approach essentially constitutes an alternative method of data pooling, enabling a significant reduction in the number of specified parameters. Coupled with sparse matrix forms for efficient storage and estimation, such model structures are well placed to handle the Big Data challenge.

References

Deaton, A. and Muellbauer, J. (1980), Economics and consumer behavior, *Cambridge University Press, Cambridge.*

Fader, P. and Hardie, B. (1996), Modeling Consumer Choice among SKUs, *Journal of Marketing Research* (Vol XXXIII) pp 442-452.

Lancaster, K. (1966), A New Approach to Consumer Theory. *Journal of Political Economy*. Vol 74 (2) pp 132-157.

Western, B. (1998), Causal heterogeneity in comparative research: A Bayesian hierarchical modelling approach. *American Journal of Political Science*, Vol 42 (4) pp 1233-1259.