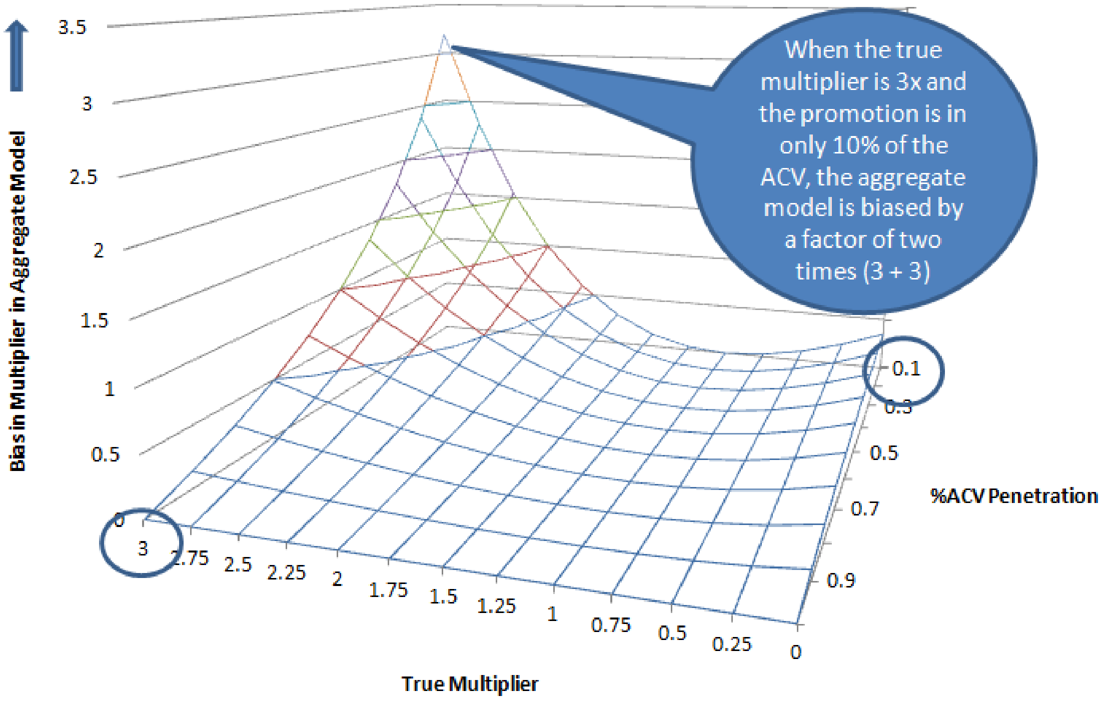
in4mation insights

Engineering Your Competitive Advantage

Methods to Eliminate Bias in Models using Aggregate Data

When the econometric model is non-linear (e.g. as when we take the natural log of the dependent variable) and the data is aggregated to the chain or market level then the price and trade parameters are biased compared to a model fit to store-level data. The paper by Markus et al. (1997) is a comprehensive discussion this kind of ***aggregation bias***. This problem emerged as an issue in the 1990's because Nielsen and IRI maintained the business practice of releasing chain and market level data to their customers (and their customers' third party consultants) but used store-level data for their own models. Part of this practice was a strategy to deny access of the "best" data to competitive analysts and part of it was a legal safeguard to not release data that might be used to identify sample stores. The store-level model became the gold standard for non-linear specifications both because it contained no aggregation bias at the weekly level and maximized the proximity of cause and effect. Analysts could avoid the aggregation bias by using linear (additive) models with interaction terms on aggregate data but these models were shown to be inferior in other ways.

The authors in Markus (1997) show that fitting a chain- or market-level model gives highly biased estimates of trade merchandising effects and somewhat biased effects for price. The bias for trade merchandising was found to be biggest when the tactic was very effective (i.e. had a high multiplier) and occurred in a small %ACV of the chain/market:



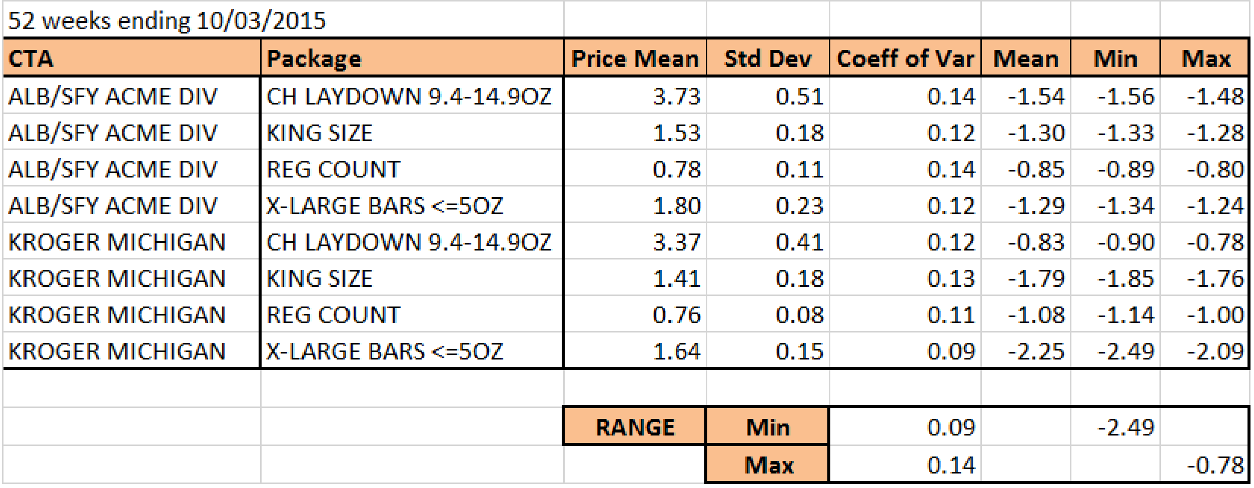
When the true multiplier is 3x and the promotion is in only 10% of the ACV, the aggregate model is biased by a factor of two times (3 + 3)

This turned out to be a serious problem since it was very common for highly effective promotions such as features with displays to be distributed in the 10%-50% range of ACV.

The paper also focused on ways to "de-bias" the estimates given that aggregated data in one form or another was bound to be used far more often than store-level data. ***Bias can be substantially corrected by slicing the data so that each slice is homogeneous with respect to trade merchandising conditions.*** Ross Link showed that coefficients ("lifts") for feature and display could be almost completely de-biased by slicing the chain data into separate observations ("rows" in the data) corresponding to mutually exclusive tactics such as "feature only," "display only," "feature and display only," "price discount only" and "no promo." He called this the Store-Group-Condition method since the data were structured so that a group of stores with the same condition (e.g. feature only) were separately aggregated. In this scheme, all price and distribution measures are restated at the condition level, while trade %ACV measures are replaced with dummy variables {0, 1} as they would be coded in a store-level model. This essentially solves the problem for trade tactics. So, instead of rising up sharply in the corner of low %ACV/high multiplier, the bias for trade merchandising can be made near-zero and flat.

Impact on Price Coefficients

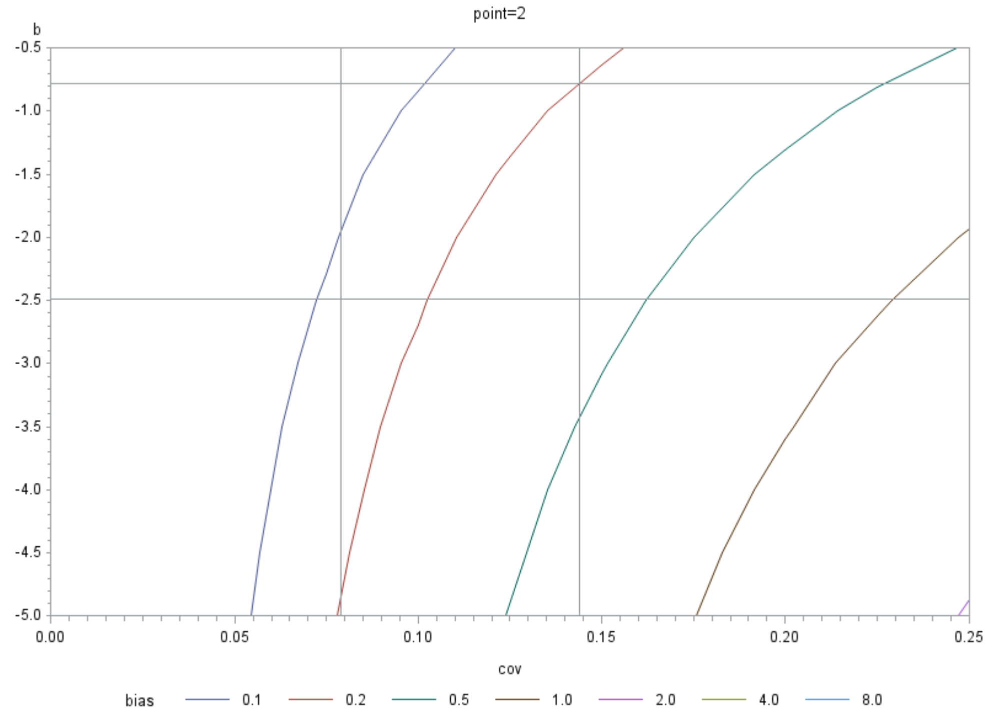
Price elasticity is also biased in aggregate non-linear models but only to a very small extent. The direction of bias in price elasticity is slightly downward (i.e. less extreme) and is proportional to the how much the price varies relative to the mean price. To show this we looked at the price of four major packs in two different chains over the course of a year as follows:



52 weeks ending 10/03/2015							
CTA	Package	Price Mean	Std Dev	Coeff of Var	Mean	Min	Max
ALB/SFY ACME DIV	CH LAYDOWN 9.4-14.9OZ	3.73	0.51	0.14	-1.54	-1.56	-1.48
ALB/SFY ACME DIV	KING SIZE	1.53	0.18	0.12	-1.30	-1.33	-1.28
ALB/SFY ACME DIV	REG COUNT	0.78	0.11	0.14	-0.85	-0.89	-0.80
ALB/SFY ACME DIV	X-LARGE BARS <=5OZ	1.80	0.23	0.12	-1.29	-1.34	-1.24
KROGER MICHIGAN	CH LAYDOWN 9.4-14.9OZ	3.37	0.41	0.12	-0.83	-0.90	-0.78
KROGER MICHIGAN	KING SIZE	1.41	0.18	0.13	-1.79	-1.85	-1.76
KROGER MICHIGAN	REG COUNT	0.76	0.08	0.11	-1.08	-1.14	-1.00
KROGER MICHIGAN	X-LARGE BARS <=5OZ	1.64	0.15	0.09	-2.25	-2.49	-2.09
		RANGE	Min	0.09		-2.49	
			Max	0.14			-0.78

The coefficient of variation of price (=standard deviation/mean) varies from 0.09 to 0.14. If price does not vary within each homogeneous slice then the coefficient of variation would be zero and there would be no price bias. The more price varies across stores, the higher the coefficient of variation and the greater the bias. The bias in price elasticity is also slightly greater when the true elasticity is higher. The contour plot below shows exactly how biased price elasticity is as a function of both the variation in price and the underlying price elasticity. The square highlighted on the plot shows the typical range of these variables for the representative set of products in the table above:

Standard deviations range from 0.08 to 0.51 across this set. Roughly, a standard deviation of 0.51 on a price of $3.73 for the CH LAYDOWN 9.4-14.9OZ package in Safeway ACME means that 95% of the price points are within +/- $1.02 of $3.73, so from $2.73 to $4.73. The chart below shows how far downwardly biased we can expect the price elasticity to be based on the standard deviation:



So in most cases, the bias is less than 2%. If Nielsen ran a store-level model and got an own price elasticity of -2.5, the i4i chain-level model would show -2.5 x 98% = -2.45. This is close enough for any practical purpose.

Christen, Markus, Sachin Gupta, John C. Porter, Richard Staelin and Dick R. Wittink (1997), "Using Market-Level Data to Understand Promotion Effects in a Nonlinear Model", Journal of Marketing Research, Aug 1997, 322-334