On Aggregation Bias in Structural Demand Models

Pei-Chun Lai and David A. Bessler

Department of Agricultural Economics, Texas A&M University, College Station, TX 77843

Poster prepared for presentation at the Agricultural & Applied Economics Association 2010 AAEA,CAES, & WAEA Joint Annual Meeting, Denver, Colorado, July 25-27, 2010

On aggregation bias in structural demand models

Pei-Chun Lai and David A Bessler
Department of Agricultural Economics, Texas A&M University, College Station, Texas 77843

Introduction

Consumer demand analysis attracts considerable attention. It remains an open question, however, whether estimating demand with aggregate data is reliable when disaggregate store-level data is given. Demand models may produce biased results when applied to data aggregated across stores with different pricing strategies. In this study, the graphical model is used to investigate the following question: *Do we find the same structure when we fit causal models on sub-groupings of stores, as we find when we fit models on aggregate data from all stores?*

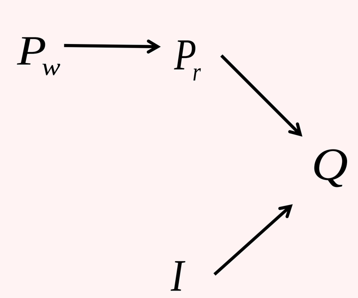
Graphical methods for the discovery of causal connection in structural equation models (SEM) provide interesting tools to justify causal claims between variables. Nevertheless, an observed relation among variables might reflect the influence of a hidden common cause, thus making the correlation spurious. Fast Causal Inference (FCI) algorithm is developed to explore the causal structural when latent confounders exist.

We apply constraint based FCI algorithm on the Dominick's scanner data and zip code information for the chain stores. The data set contains weekly sales information (03/ 02/ 95-03/ 06/ 96) of Coke 6 package with 12 fl oz about 74 supermarket chain stores in Chicago area. The sales information includes supermarket's retail price (P_r), manufacturer's wholesale price (P_w), weekly sold quantity (Q), and store-specific median family income (I).

Materials and methods

We do not impose an a priori causal flow among the four demand related variables studied here. The usual structure of demand has the following causal graph:

$$Q = \alpha_1 P_r + \alpha_2 I + \varepsilon_1$$
$$P_r = \beta_1 P_w + \varepsilon_2$$
$$P_w = \varepsilon_3$$
$$I = \varepsilon_4$$



If there is a association between the corresponding error terms (i.e. $\varepsilon_3 \leftrightarrow \varepsilon_4$), for SEM with correlated errors, the possible influence of latent (unobserved) confounders can be taken into account by implementing the FCI algorithm.

Since we attempt to detect the existence of aggregation bias, we classify the whole data into aggregate and disaggregate groups Figure 1 illustrates the processing flow of our analysis.

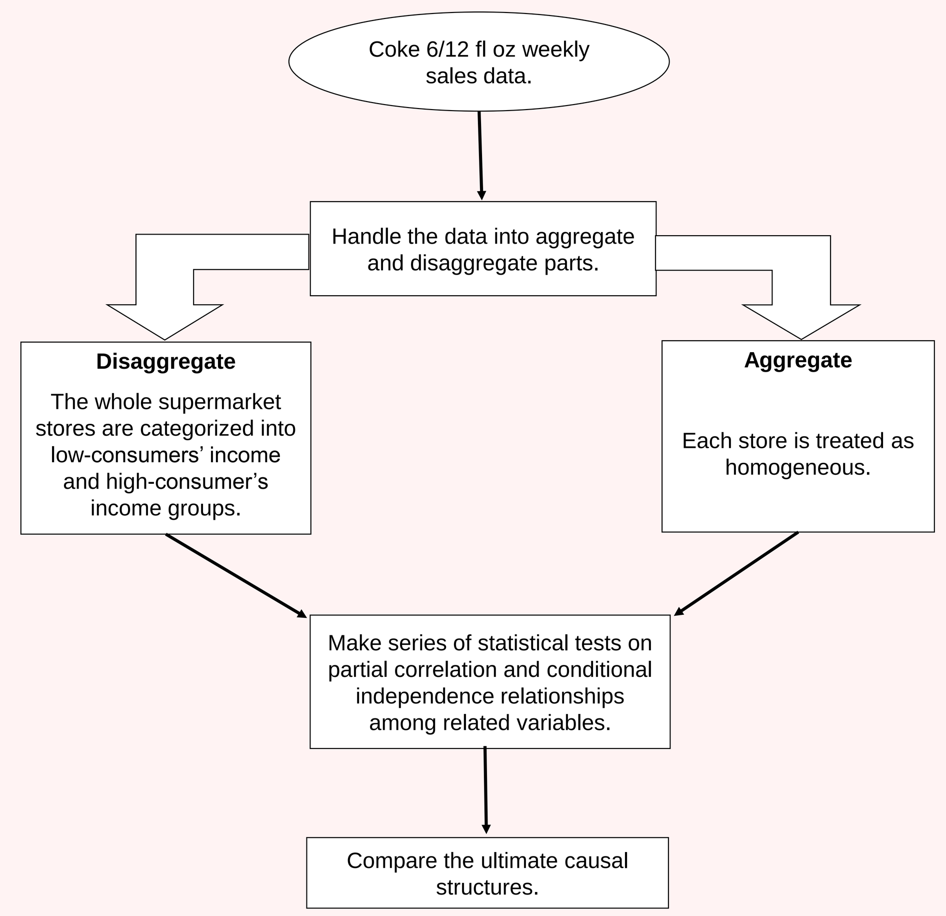


Figure 1. Flow of model processing

The output of the FCI algorithm is a partial ancestral graph (PAG) and the edges in a PAG can be interpreted as follows:

$a \rightarrow b$: a is a cause of b.
$a \leftrightarrow b$: there is a latent common cause of a and b so that a does not cause b and b does not cause a.
$a \circ\!\!\rightarrow b$: a is a cause of b, or there is a latent common cause of a and b, or both.
$a \circ\!\!-\!\!\circ b$: either a is a cause of b or b is a cause of a, or there is a latent common cause of a and b, or there is a combination of these.

Results

The disaggregate-level data is defined by using 1990 U.S. Census information. The stores that fall into group one are those that face a consumer base whose median family income is less than $35,597 (first quartile). Stores that reside in zip codes characterized by median household incomes greater than $48,705 define our second disaggregate group (third quartile). We ignore stores where median family incomes are between the first and third quartiles. Figures 2 and 3 display the PAGs of aggregate-level and disaggregate-level data. Our findings show that:

- For the variables P_w , P_r , and I, they have a direct effect on sold quantity, or their relation with Q is due to a common cause, or a combination of both.
- In the aggregate and low median household income graphs, either manufacturer may have more pricing power over supermarket retailer, or supermarket retailer has more pricing power over manufacturer, or there is a latent common cause of P_w and P_r, or there is a combination of these.
- For stores that face median family income greater than $48,705, there is no relation between P_w and P_r .
- We find agreement in 3 edges and directions but we miss one edge.

Mean	Median Value	First Quartile	Third Quartile
42486.7	42065	35597	48705

Table 1. Statistics of store-specifically median household income. The first quartile and third quartile are used to make the disaggregate groups.

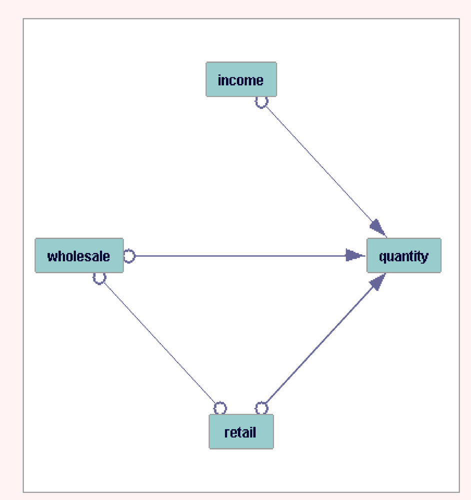


Figure 2. PAG of aggregate-level data (p=0.01).

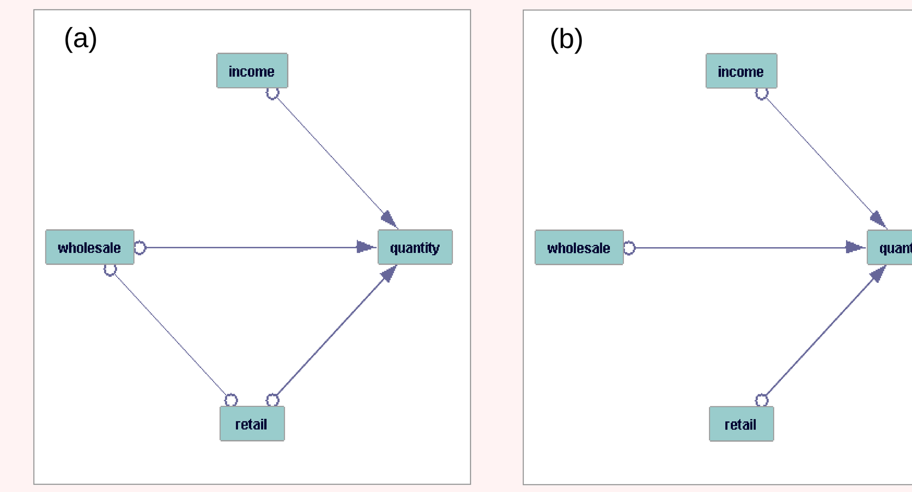


Figure 3. PAGs of disaggregate-level data (p=0.01). The two disaggregate-level groups are defined along the lines of median family income: (a) PAG of group one (b) PAG of group two.

Conclusions

Demand estimates based on aggregate data is possibly biased when stores are heterogeneous. In this study, we use FCI algorithm to test if an aggregation bias exists when aggregating data across stores with different geographical population distribution.

The question we ask is: does aggregation across stores give us the same result as disaggregate analysis? The answer is no! The aggregate result is not precisely consistent with disaggregate result, but they are similar to each. Our result suggests that when aggregating data, some association between variables may spuriously exist. However, how to obtain a properly modified aggregate demand framework to avoid this problem is not answered in this poster.

Unlike traditionally statistical method, we detect the causal patterns between variables to examine the existence of aggregation bias. Causal discovery techniques usually assume that all causes are observed and known a *priori*. This is the so-called causal sufficiency assumption. However, this presumption is not always true. FCI algorithm is helpful to check the possible unobserved latent confounders between variables when there is causal insufficiency.

Finally, as several previous studies in marketing indicate, our results show retail price and consumers' family income may have effects on purchase behavior. We found this result without imposing the causal structure *a priori*.

Literature cited

Kwon, D.H. 2007. *Causality and Aggregation Economics: the Use of High Dimensional Panel Data in Micro-Econometrics and Macro-Econometrics*, Ph. D. Dissertation, Department of Agricultural Economics, Texas A&M University.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2nd edition.

Temme, D. 2006. Constraint-Based Inference Algorithms for Structural Models with Latent Confounders-Empirical Application and Simulations. *Computational Statistics* 21:151-182.

Tenn, S. 2006. Avoiding Aggregation Bias in Demand Estimation: a Multivariate Promotional Disaggregation Approach. *Journal of Machine Learning Research* 9:1437-1474.

Acknowledgments

We gratefully acknowledge the provision of panel scanner data by the University of Chicago, James M. Kilts Center, Graduate School of Business.

For further information

Please contact *PCLai@ag.tamu.edu*.