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


Copyright 2010 by Pei-Chun Lai and David A. Bessler. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.
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-groups of stores, as we find when we fit models on aggregate family income data? 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 structure when latent confounders exist. We apply constraint based FCI algorithm on the Domino’s scanner data and zip code information for the whole supermarket chain stores in Chicago area. The sales data includes supermarket’s retail price (P), manufacturer’s wholesale price (Pw), 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:

\[ Q = \alpha_1 P + \alpha_2 J + \varepsilon_1 \]
\[ P = \beta_1 P + \varepsilon_2 \]
\[ I = \varepsilon_3 \]

The output of the FCI algorithm is a partial ancestral graph (PAG) and the edges in a PAG can be interpreted as follows:
- \( a \rightarrow b \): it is a cause of \( b \)
- \( a \leftarrow b \): there is a latent common cause of \( a \) and \( b \) so that \( a \) does not cause \( b \) and \( b \) does not cause \( a \).
- \( a \leftarrow b \): it is a cause of \( b \). There is a latent common cause of \( a \) and \( b \) or both.
- \( a \rightarrow b \rightarrow c \): either a latent cause of \( b \) or a cause of \( a \), or there is a latent common cause of \( a \) and \( b \) or both.

If there is a association between the corresponding error terms (i.e., \( \varepsilon_1 \), \( \varepsilon_2 \), \( \varepsilon_3 \)), for SEM with correlated errors, the possible influence of latent (unobserved) confounders cannot be taken into account by implementing the FCI algorithm.

In the aggregate and low median household 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 and disaggregate-level data. Our findings show that:
- For the variables \( P_{\text{AG}} \), \( P_{\text{DG}} \), 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 groups, 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 \( P_{\text{AG}} \) and \( P_{\text{DG}} \) or there is a combination of these.
- For stores that face median family income greater than $48,705, there is no relation between \( P_{\text{AG}} \) and \( P_{\text{DG}} \).
- We find agreement in 3 edges and directions but we miss one edge.

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 and disaggregate-level data. Our findings show that:
- For the variables \( P_{\text{AG}} \), \( P_{\text{DG}} \), 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 groups, 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 \( P_{\text{AG}} \) and \( P_{\text{DG}} \) or there is a combination of these.
- For stores that face median family income greater than $48,705, there is no relation between \( P_{\text{AG}} \) and \( P_{\text{DG}} \).
- We find agreement in 3 edges and directions but we miss one edge.

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 a 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 consumer’s family income may have effects on purchase behavior. We found this result without imposing the causal structure a priori.

Literature cited
Ko, D. 2007. Consumer and Aggregation Economics: the Use of High Dimensional Panel Data in Microeconomics and Macro-
     Economics. Ph. D. Dissertation, Department of Agricultural Economics, Texas A&M University.


Acknowledgments
The authors gratefully acknowledge the invaluable assistance by the faculty of the Marketing Department at the University of Chicago, and particularly, J. M. Kilts, whom we thank for his insight and encouragement throughout the research.

For further information
Please contact PCLai@ag.tamu.edu.