Aggregation Bias in Discrete Choice Models

Timothy Wong[*], David Brownstone[†] and David Bunch[‡]

[†]Department of Economics, National University of Singapore

[†]Department of Economics, University of California, Irvine

[‡]Graduate School of Management, University of California, Davis

Overview

- Multinomial choice models are popular in demand estimation because
  - unlike systems of demand equations, the number of parameters to be estimated is not a function of the number of products, removing the obstacle of estimating markets with many differentiated products.
- One challenge of choice modeling in application is determining the level of detail at which the choice set is defined.
  - modeling choices at their finest level of detail can cause the resulting choice set to grow so large that it exceeds the practical capabilities of estimation
  - Household choices are often not observed at their finest level, hence researchers aggregate choices to the level at which they are observed

Application

- Partially observed choices are particularly common in vehicle choice applications:

Table 3: Vehicle Specifications for 2009 Civic Hybrids – Ward's Automotive Data

Make & Series	Body Style	Drive Type	Length (ins.)	Width (ins.)	Weight (lbs.)	Horsepower		Trans Std.	MPG City/Hwy	Retail Price
						Hp	@RPM			
Hybrid	4-dr. sedan	FWD	177.3	69.0	2,875	110	6000	CVT	40/45	$24,320
Civic DX	4-dr. sedan	FWD	177.3	69.0	2,630	140	6300	M5	26/34	$16,175
Civic LX	4-dr. sedan	FWD	177.3	69.0	2,687	140	6300	M5	26/34	$18,125
Civic EX	4-dr. sedan	FWD	177.3	69.0	2,747	140	6300	M5	26/34	$19,975

Exact choices

Broad group I

Broad group II

Adapted from Brownstone and Lloro, 2015

- These applications are used to estimate consumer valuations of fuel efficiency, a quantity heavily debated in the energy literature.

Our contributions

- We generate a Monte Carlo setting to study the impact of aggregating alternatives on model estimates.
- We consider nine different methods of aggregation but focus our discussion here on three main approaches:
  - Average model: choice model that aggregates to broad groups of choices
  - McFadden Aggregation model: choice model that aggregates to broad groups of choices, then places distributional assumptions on the attributes in each aggregated group
  - Broad Choice Model: a choice model that take in to account the presence of broad choice data without aggregation.
- Findings:
  - Aggregation mis-specifies the choice model affecting point estimates which may lead to erroneous policy conclusions.

The Model

- Let $n = 1, \dots, N$ index households and j index products, $j = 1, \dots, J$ in the market.
- The indirect utility of household n from the choice of product j, U_{nj} follows the following specification:

$$U_{nj} = w'_{nj}\beta_w + \epsilon_{nj}$$

- Households select the product that yields them the highest utility:

$$y_{nj} = \begin{cases} 1 & if\ U_{nj} \geq U_{ni} \quad \forall\ i \neq j \\ 0 & otherwise. \end{cases}$$

The Model

- ϵ_{nj} follows a type I extreme value distribution. Therefore the probability that consumer n, chooses product j is:

$$P_{nj} = \frac{exp\left(w'_{nj}\beta_w\right)}{\sum_k exp(w'_{nk}\beta_w)} .$$

- The log-likelihood function of this conditional logit is as follows:

$$L(y;\delta,\beta) = \sum_n \sum_j y_{nj} \log(P_{nj})$$

Aggregation in choice models

- Define C as the exact choice set that contains all products, $j = 1, 2, \dots, J$.
- C is decomposed into B groups, denoted $C_b, b = 1, 2, \dots, B$.
- $C = \bigcup_{b=1}^{B} C_b$ and $\bigcap_{b=1}^{B} C_j = \emptyset$.

$$Y_{nb}^{*} = \begin{cases} 1 \; if \; y_{nj} \in C_b \\ 0 \; otherwise. \end{cases}$$

- Common solution: aggregate choices and choice attributes to the broad group level.

$$L(y; \delta, \beta) = \sum_n \sum_b Y_{nb}^{*} \log(P_{nb})$$

where $w_{nb} = \frac{1}{J} \sum_{j \in b} w_{nj}$

McFadden, 1978 method for aggregation

- When the number of dwellings within a community is large, and

$$w_{nj} \sim N(w_{nb}, \Omega_{nb}), \quad i.i.d. \quad j \in b$$

$$\tilde{P}_{nb} \xrightarrow[a.s.]{} \frac{\exp(w_{nb}'\beta + \frac{1}{2}\beta'\Omega_{nb}\beta + \log(D_b))}{\sum_k \exp(w_{nk}'\beta + \frac{1}{2}\beta'\Omega_{nk}\beta + \log(D_k))}$$

where D_k is the number of dwellings in community k.

- Consistent but inefficient estimates can be obtained by ignoring the non-linear constraint on β

McFadden, 1978 method for aggregation

$$\tilde{P}_{nb} = \frac{\exp(w_{nb}'\beta + \frac{1}{2}\beta'\Omega_{nb}\beta + \log(D_b))}{\sum_k \exp(w_{nk}'\beta + \frac{1}{2}\beta'\Omega_{nk}\beta + \log(D_k))}$$

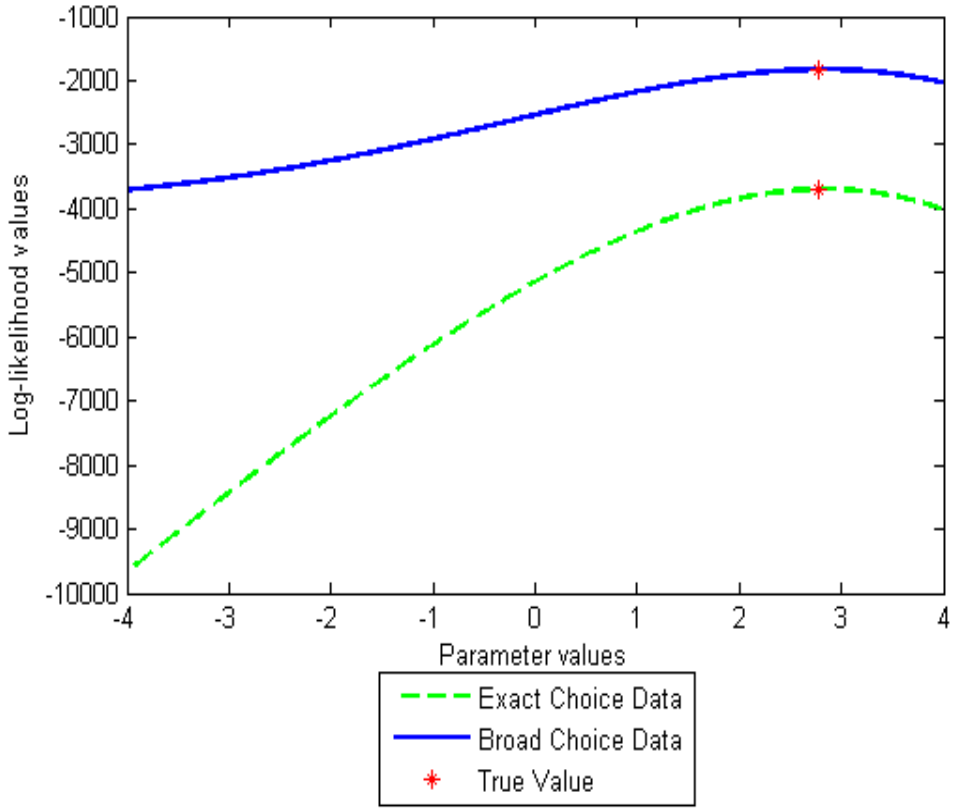
- The intuition for including Ω_{nb} is that community attributes with larger variances should have a greater impact on the probability that the community is selected.
- The $\log(D_b)$ term is a measure of community size. Other conditions being equal, a community with a large number of housing units should have a higher probability of being selected than a very small one.

A model for broad choice data

- Brownstone and Li, 2014, propose the following model for broad choice data:

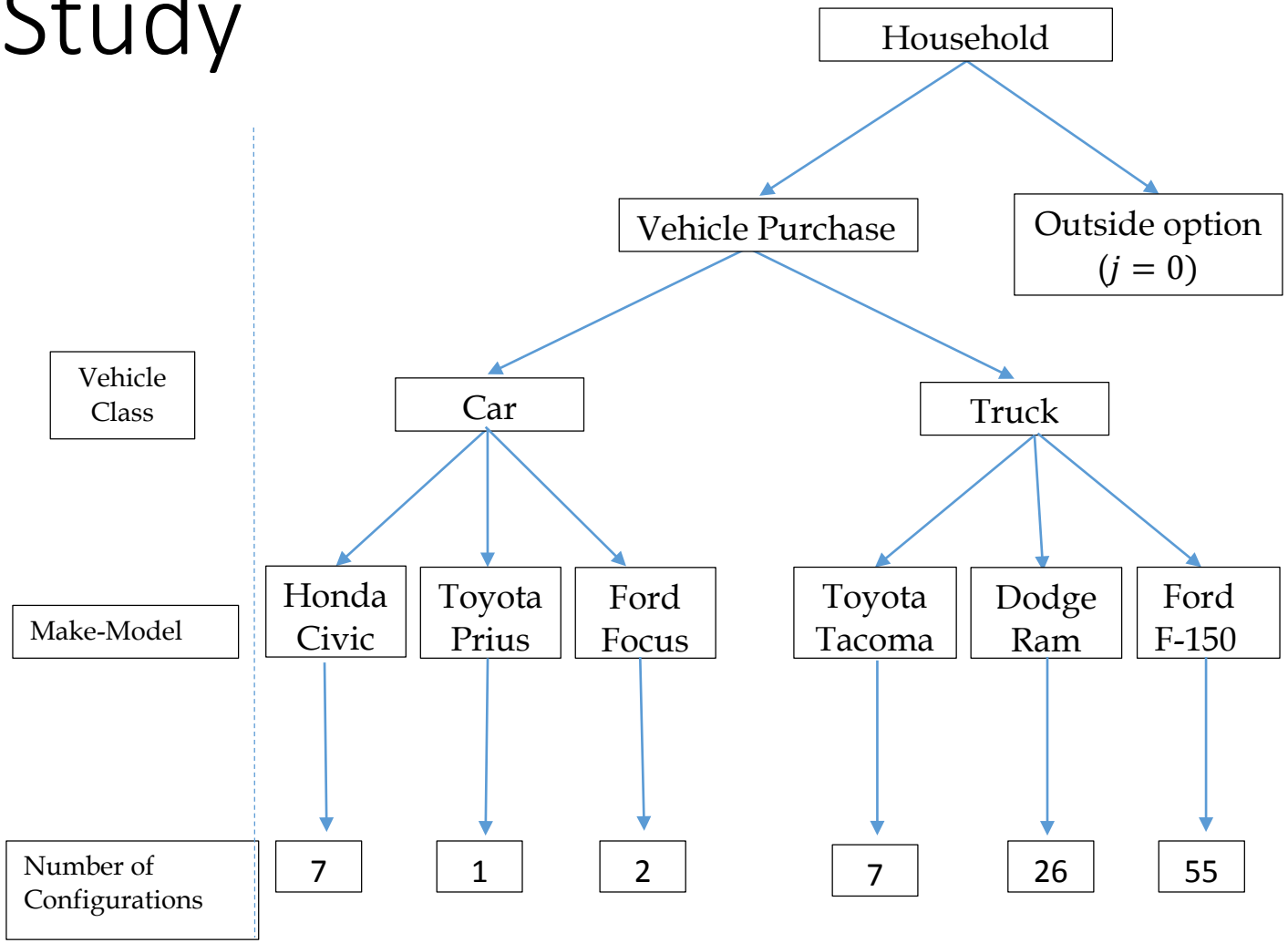
$$L(y;\delta,\beta) = \sum_n \sum_b Y_{nb}^* \log(P_{nb}^*)$$

where $P_{nb}^* = \sum_{j \in C_b} P_{nj}$ and P_{nj} is the standard logit choice probability formula.



The Monte Carlo Study

- We construct the dataset for the Monte Carlo study based on a vehicle choice application.
- The structure of the choice set is illustrated in tree form in Figure 1.



The Monte Carlo Study

- We use the following vehicle attributes for these vehicles, obtained from the Volpe Center:
  - Vehicle price
  - transmission (manual or automatic)
  - fuel consumption rate (gallons of fuel per mile)
- We choose to use real vehicles and their corresponding attributes to ensure that when we consider aggregation across realistic distributions of attributes.
- We also use real household data from the U.S. National Household Transportation Survey.
  - We make use of their income data and average gasoline price in their state of residence in 2008.

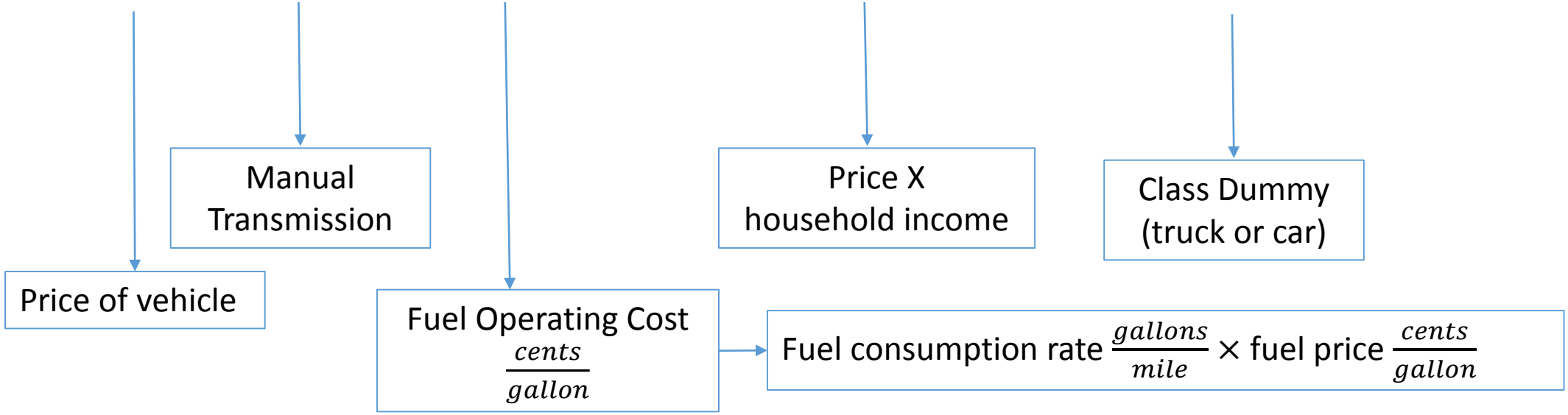
Summary Statistics of Vehicle Configuration Attributes by Make/Model

Vehicle Make/Model	Number of Configurations	Price (US$ '000)				Fuel Consumption Rate (gallons/ 100miles)				Manual Transmission	
		Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.
Honda Civic	7	26.00	3.16	23.10	30.05	3.30	0.11	3.01	3.93	0.43	0.53
Toyota Prius	1	22.04	0.00	22.04	22.04	1.52	0.00	1.52	1.52	0.00	0.00
Ford Focus	2	15.85	0.00	15.85	15.85	2.73	0.00	2.69	2.77	1.00	0.00
Toyota Tacoma	7	21.82	2.15	18.79	23.80	4.16	0.25	3.40	4.79	0.57	0.53
Dodge Ram	26	27.19	3.02	22.06	31.87	5.29	0.09	4.51	5.61	0.23	0.43
Ford F-150	55	25.25	2.31	22.89	29.35	5.25	0.16	4.70	6.23	0.09	0.29

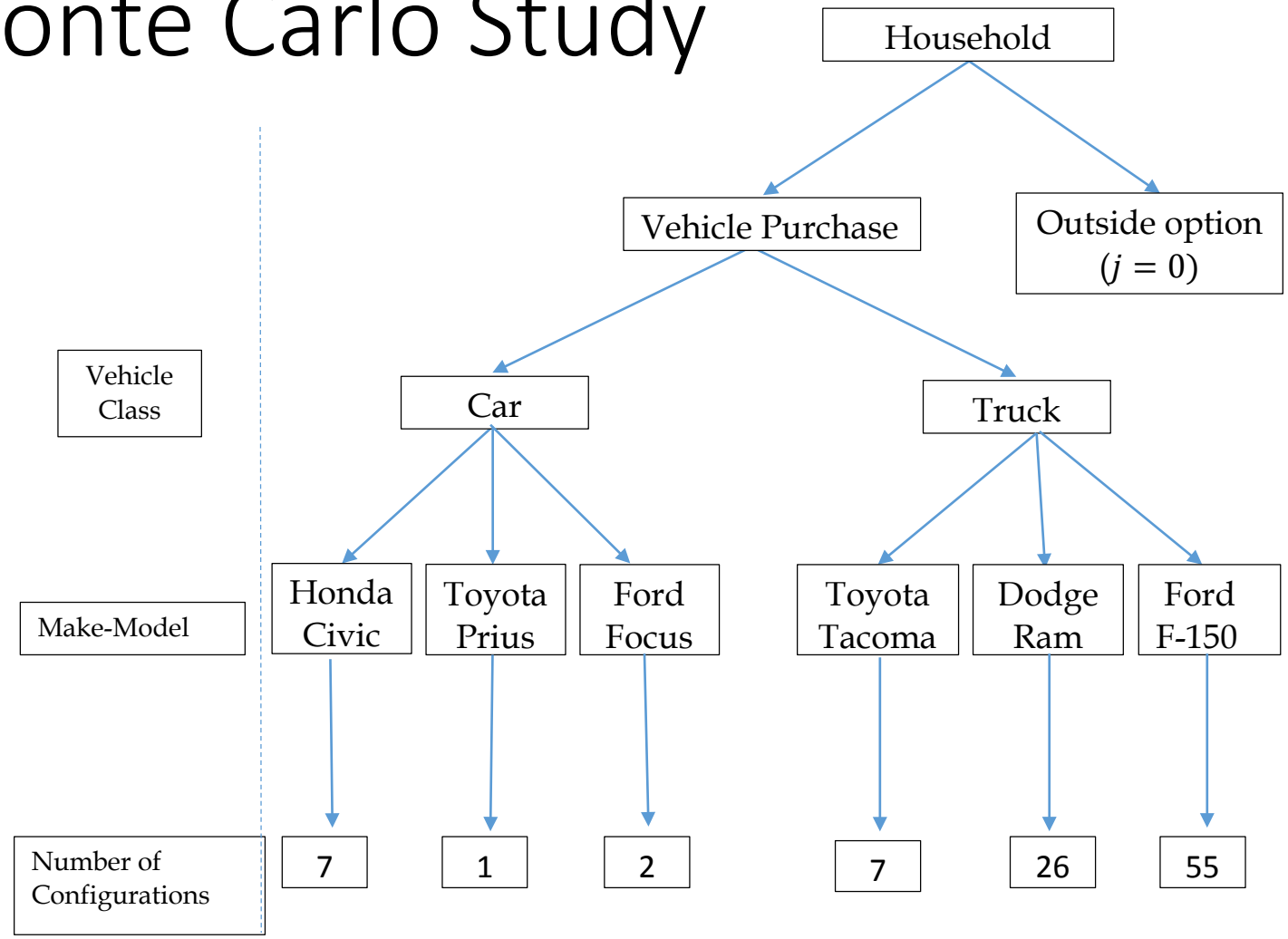
The Monte Carlo Study

- From the data above, we generate the following choice model.
- The indirect utility of household n form the choice of product, j, U_{nj} is assumed to follow the following linear specification:

$$U_{nj} = p_j\beta_p + T_j\beta_T + foc_{nj}\beta_{foc} + (p_j \times inc_n)\beta_{p-inc} + D_j + \epsilon_{nj}$$



The Monte Carlo Study



Results

10,000 households		Full Observability Model			Average Configuration Method			McFadden Aggregation Method			Broad Choice Model		
Variable	True Value	Mean Estimate	Mean Std. Err.	90% Coverage Probability	Mean Estimate	Mean Std. Err.	90% Coverage Probability	Mean Estimate	Mean Std. Err.	90% Coverage Probability	Mean Estimate	Mean Std. Err.	90% Coverage Probability
Manual Transmission	-0.10	-0.10	0.03	0.91	0.48	0.06	0.00	0.34	0.07	0.00	-0.11	0.11	0.90
Price ('000)	-0.40	-0.40	0.01	0.91	-0.17	0.01	0.00	-0.32	0.01	0.00	-0.40	0.01	0.90
Price*High Income	0.10	0.10	0.01	0.90	0.09	0.01	0.61	0.08	0.04	0.00	0.10	0.01	0.90
Fuel Operating Cost (cents/mile)	-0.20	-0.20	0.01	0.90	-0.16	0.01	0.00	-0.16	0.01	0.00	-0.20	0.01	0.89
Car Dummy	8.00	8.01	0.11	0.91	4.26	0.16	0.00	6.11	0.19	0.00	8.02	0.26	0.90
Truck Dummy	7.50	7.51	0.13	0.92	5.14	0.18	0.00	5.66	0.19	0.00	7.52	0.26	0.89
Log (counts)	1.00							0.96	0.02	0.35			
Willingness to Pay*	0.67	0.67	0.00	0.90	2.03	0.05	0.00	0.66	0.00	0.96	0.67	0.00	0.89

Empirical Application

- Next, we estimate the same three models on an actual vehicle choice data set.
- We model vehicle choice conditional on vehicle purchase.
  - we do not include a “no buy” option.
- Vehicle attributes are provided by the Volpe Center
  - supplemented with data from Polk, the American Fleet Magazine, and the National Automobile Dealers Association.
- Household data are obtained from the NHTS
  - There are 10,500 NHTS households in the dataset who purchase at least one new model year 2008 vehicle during the sample period.

Empirical Application

- The need for aggregation arises because vehicle data is observed at the Make/Model/Fuel-type/configuration level while household choices are only observed at the Make/Model/Fuel-type level.
- There are 1120 vehicles at the configuration level, that we have to aggregate to 235 broad groups of vehicles that households choose from.

Results

Variable	Average Configuration			McFadden's Aggregation			Broad Choice		
	Estimated Parameter	Standard Error		Estimated Parameter	Standard Error		Estimated Parameter	Standard Error	
(Price) × (75,000<Income<100,000)	0.038	0.003	***	0.014	0.003	***	0.019	0.003	***
(Price) × (Income>100,000)	0.075	0.003	***	0.050	0.003	***	0.044	0.003	***
(Price) × (Income Missing)	0.066	0.004	***	0.041	0.004	***	0.039	0.003	***
Fuel Operating Cost (cents per mile)	-0.240	0.012	***	-0.255	0.012	***	-0.250	0.012	***
(Fuel Operating Cost) × (College)	-0.081	0.007	***	-0.058	0.007	***	-0.016	0.007	**
Price	-0.060	0.003	***	-0.039	0.003	***	-0.046	0.003	***
Horsepower / Curb weight	0.187	1.773		0.088	0.821		17.729	1.647	***
Curb Weight	0.391	0.037	***	0.554	0.035	***	0.510	0.040	***

Results

Willingness to pay for a 1 cent/mile improvement in fuel efficiency (thousands)†	Estimated Parameter	Standard Error		Implied Discount Rate
Average Configuration Model	3.998	0.090	***	-5.83
McFadden Aggregation Model	6.503	0.287	***	-11.43
Broad Choice Model	5.449	0.185	***	-9.50

Conclusion

- The existing evidence on consumer valuation of fuel efficiency is varied and inconclusive. Part of this may be a result of modeling errors because
  - aggregation affects point estimates
  - Researchers should pay careful attention to how they define choice sets when implementing discrete choice models.
- We highlight two models that can be used to address aggregation.