February 18, 2018

Aggregation Biases in Discrete Choice Models[1]

Timothy Wong
National University of Singapore

David Brownstone[2]
U.C. Irvine

David S. Bunch
U.C. Davis

Abstract

This paper examines the common practice of aggregating choice alternatives within discrete choice models. We carry out a Monte Carlo study based on realistic vehicle choice data for sample sizes ranging from 500 – 10,000 individuals. We consider methods for aggregation proposed by McFadden (1978) and Brownstone and Li (2017) as well as the more commonly used methods of choosing a representative disaggregate alternative or averaging the attributes across disaggregate alternatives. The results show that only the "broad choice" aggregation method proposed by Brownstone and Li provides unbiased parameter estimates and confidence bands. Finally, we apply these aggregation methods to study households' choices of new 2008 model vehicles from the National Household Travel Survey (NHTS) where 1120 unique vehicles are aggregated into 235 make/model classes. Consistent with our Monte Carlo results we find large differences between the resulting estimates across different aggregation methods.

JEL Codes: C25, C35, L62

Keywords: discrete choice, aggregation, household vehicle demand

Introduction

This paper studies the practice of aggregating choice alternatives within discrete choice models. Assume that a researcher can establish the "ideal" level of detail for defining choice alternatives in a discrete choice problem, and denote these "exact choices." Researchers have often faced situations where exact choices yield choice set sizes that are too large for practical model estimation. In the literature for the main example considered here (vehicle choice), choice sets have frequently been defined at a lower level of detail (e.g., vehicle type), by aggregating over the relevant exact choices. In some cases, choices might be observed at a lower level of detail, so researchers estimate models by aggregating exact choices to the observed level. However, this practice of aggregation miss-specifies the true choice set of interest. Previous work (Brownstone and Li, 2017 and Wong, 2015) investigated this concern within the context of the Berry, Levinsohn, and Pakes (BLP) choice model for micro- and macro-level data. This paper compares commonly used methods for estimating choice models

[1] We acknowledge funding provided by the University of California, Irvine through the Department of Economics, Institute of Transportation Studies, School of Social Sciences and Center for Economics and Public Policy, the University of California Center for Energy and Environmental Economics, the University of California Multi-Campus Research Program in Sustainable Transport: Technology, Mobility and Infrastructure, and the University of California Center on Economic Competitiveness in Transportation. The authors are solely responsible for any errors or omissions.

[2] Corresponding author. Department of Economics, University of California, Irvine, 3151 SSPA, Irvine, CA 92697-5100 USA. Email: dbrownst@uci.edu

where choices need to be aggregated. In addition to the aggregation methods given by McFadden (1978) and Brownstone and Li (2017), we also investigate the common practices of averaging attributes across the aggregated alternatives (Lave and Train, 1979; Bento et al., 2009) and choosing one alternative to represent the alternatives being aggregated. Common candidates for representative alternatives include the modal alternative (Berry et al., 2004) and the "base" alternative, which often is the alternative with the most basic features. (Berry et al., 1995; Petrin, 2002.)

The key problem we are trying to solve is that the exact alternative chosen by a household is not observed. If it were and the true data generating process is multinomial logit, then fitting a standard logit model with a choice set containing the chosen alternatives and a sample of other non-chosen alternatives and would lead to consistent but inefficient estimation (due to the Independence of Irrelevant Alternatives property of the logit model). The method recently proposed by Fernandez-Antolin et al. (2017) averages across many such estimators and is therefore an improvement over just choosing one set of representative alternatives, but is inconsistent if the chosen alternative is not fully observed (as is the case with common U.S. data sources).

We investigate the issue of potential aggregation bias by generating results for a specific example (new vehicle choice) based on previous empirical work using the 2009 National Household Transportation Survey (NHTS). For specific time intervals, new vehicle purchases correspond to household choices of 2008 model year vehicles. In any recent model year there are well over 1000 vehicle configurations available, which can vary in important ways (e.g., fuel economy and performance) due to alternative engine, drivetrain, and transmission offerings. Modelling choice at the make/model level reduces the number to about 230 options, implying considerable aggregation. For example, there are 7 different trim lines under the Honda Civic label, and over 100 trim lines under the Ford F-150 make/model label, with notable variation in vehicle equipment that affects fundamental attributes such as purchase price and fuel consumption rate.

Using these data, we generate a much simpler Monte Carlo setting to isolate the impact of aggregating alternatives from the possible effects of model misspecification. The data generation process is specified as conditional logit with an outside good, 3 make/models grouped into "cars" and 3 make/models grouped into "trucks." There are 98 distinct alternative vehicles that are grouped into make/models. In one case there was only one vehicle assigned to a particular make/model which corresponds to some vehicles in the real marketplace (e.g. Toyota Prius in the early 2000s), and the remaining make/models correspond to between 2 and 55 vehicles. We consider alternative aggregation methods from the literature for estimating the 7-alternative choice model corresponding to only observing the make/model. As is true in the real U.S. vehicle market we assume that there are data on the attributes of each vehicle, but unlike BLP we do not assume that we have any macro-level market share information for the vehicles. McFadden's (1978) aggregation procedure only requires information on the mean and covariances of the attributes being aggregated, but it is only valid for aggregating alternatives at the lowest level of a Nested Logit choice model. We

examine the performance of the estimators with sample sizes ranging from 500 to 10,000. This range encompasses most of the applied literature.

The Monte Carlo results show that the only method that performs well for all sample sizes is the "broad choice" estimator described in Brownstone and Li (2017). This is not surprising since this estimator is the maximum likelihood estimator for this problem. Both the coefficient estimators and their covariance estimators are biased for the other methods. Simply averaging attributes clearly leads to measurement error, and this is not helped by including the logarithm of the number of vehicles being aggregated as is done in some studies. McFadden's (1978) method would be consistent if the joint distribution of the attributes that are aggregated are multivariate normal, but this approximation is not satisfied in our Monte Carlo design even if we relax the implied constraints on the parameters.

Finally, we apply some of the estimation procedures used in the Monte Carlo design to real vehicle choice data. We do not apply the "representative alternative" method since it has no theoretical justification and performed poorly in the Monte Carlo experiments reported in Appendix B. We use data from the 2009 NHTS survey supplemented by detailed data on attributes for each of the 1120 2008 model year vehicles. Like Train and Winston (2007) we do not include an outside good since it would include both purchasing used vehicles and not buying any vehicles. We recognize that not including the outside good is only justified if the data generating process is Nested Logit with "Buy New", "Buy Used", and "No Buy" at the top level, but we do not have the data necessary to deal with used car attributes. The purpose of this empirical exercise is to show that the problems with aggregation methods found in the Monte Carlo results also apply with real data. Since the NHTS only collects data on make, model, and year for each vehicle, these 1120 vehicles need to be aggregated into 235 make/model classes. As expected from the Monte Carlo results, we find large differences between the estimates produced from the various methods. The confidence bands for willingness to pay estimates do not overlap, and the Broad Choice and McFadden's method yield larger willingness to pay estimates than averaging attributes or choosing representative vehicles.

This paper shows that it is critical to properly account for the biases introduced when aggregating alternatives. We have demonstrated the importance of these biases in both a Monte Carlo study and an empirical example using the conditional logit model in the simplest case where there is no external market share data available. Our earlier work shows that incorporating external market share data does improve the quality of the estimates, but does not alleviate the problems caused by aggregating alternatives. The Broad Choice maximum likelihood method is the only one that performs well in our Monte Carlo study, and we expect that it will continue to perform well in other applications with more flexible discrete choice models. The simple expedients of averaging attributes or picking a "representative" alternative perform very poorly in our studies, and we expect them to perform at least as poorly in other situations.

The Model

Let $n = 1, \dots, N$ index households which can either purchase any of J products, $j = 1, \dots, J$ in the market or not purchase any product, characterized by selecting the "outside good", $j = 0$. The indirect utility of household n from the choice of product j, U_{nj}, is assumed to follow the following linear specification:

$$U_{nj} = x_j'\beta_x + w_{nj}'\beta_w + \epsilon_{nj},$$

$$n = 1, \dots, N, \quad j = 0, 1, \dots, J,$$

where x_j is a vector of product attributes while w_{nj} is a vector of household attributes interacted with product attributes. β_x and β_w are the parameters associated with x_j and w_{nj} respectively, that are to be estimated and ϵ_{nj} is an error term with mean zero that captures all remaining elements of utility provided by product j to household n. For the purpose of identification, average utility of the "outside good," δ_0, is normalized to zero. Households select the product that yields them the highest utility:

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

Assuming that ϵ_{nj} follows a type I extreme value distribution, the probability that household n, chooses product j, P_{nj} is:

$$P_{nj} = \frac{exp(x_j'\beta_x + w_{nj}'\beta_w)}{\sum_k exp(x_k'\beta_x + w_{nk}'\beta_w)}.$$

One can obtain estimates of the model parameters, β_x and β_w by maximizing the following log-likelihood function:

$$L(y;\beta) = \sum_n \sum_j y_{nj}\, log\ (P_{nj}). \tag{1}$$

It is often the case that researchers do not observe household decisions at the finest level of detail. Instead, they only observe household choices among broad groups of products. We formally model such situations as follows:

Define C as the exact choice set that contains all products, $j = 1, 2, \dots, J$. C is decomposed into B groups, denoted $C_b, b = 1, 2, \dots, B$ so that each product, j, belongs to only one choice group. Individuals' exact choices, y_{nj}, are not observed. Instead, what is observed are individuals' choices among the broad choice groups:

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

In this section, we introduce two methods that address the concerns related to aggregation of products to broad levels. The first model, introduced in McFadden, 1978, proposes that the covariance matrix of the attributes within each broad group and the logarithm of the number of products within each broad group be included in the utility function of the choice model. The second model is a model for broad choice data, introduced in Brownstone and Li, 2017. In their model, equation (1) is defined in terms of the broad choice sets from which household

choices are observed and the broad choice probabilities are defined as the sum of the probabilities of the exact choices contained within each broad choice group.

McFadden's Method for Aggregation

McFadden, 1978, models households' choice of residential location. Here, the broad choice groups are communities where households are known to reside while the exact choice set contains the dwellings within these communities.

Let x_j denote the attributes of dwelling j and let w_{nj} denote the observed attributes of household n interacted with the attributes of dwelling j. Denote $X_{nj} = [x_j\ w_{nj}]$ and $\beta = [\beta_x\ \beta_w]$. When the number of dwellings within a community is large, and X_{nj} behaves as if it is normally distributed with known mean, X^*_{nb}, and known variance Ω_{nb}, then, McFadden, 1978, shows that for the conditional logit with linear utility specification, the probability that household n chooses community b converges to:

$$P_{nb} = \frac{\exp({X^*_{nb}}'\beta + \frac{1}{2}\beta'\Omega_{nb}\beta + \log(D_b)\,\beta_D\,)}{\sum_k \exp({X^*_{nk}}'\beta + \frac{1}{2}\beta'\Omega_{nk}\beta + \log(D_k)\,\beta_D)} \tag{2}$$

where D_b is the number of dwellings in community b.

To obtain estimates of β and β_D, maximize the following log-likelihood function:

$$L(y;\beta) = \sum\nolimits_n \sum\nolimits_b y_{nb}\, log\ (P_{nb})$$

The presence of the term, $\frac{1}{2}\beta'\Omega_{nb}\beta$, in (2) comes from the fact that the sample mean and sample sum of squared errors are sufficient statistics for the normal distribution with unknown mean and variance. To account for the distribution of characteristics of products within group b, it is necessary to condition on both quantities. The intuition here is that community attributes with larger variances should increase the probability that the community is selected since then it is more likely that the community contains a desirable dwelling for the household. A simpler approach to incorporate Ω_{nb}, is to relax the constraint that its associated parameter, β, is equal to the parameter on X^*_{nb}. This approach yields consistent estimates without the complexity of non-linear constraints.

Lerman (1977) explains that the $\log(D_b)$ term is a measure of community size. "Other conditions being equal, a large tract (i.e., one with a large number of housing units) would have a higher probability of being selected than a very small one, since the number of disaggregate opportunities is greater in the former than the latter." Here, β_D is assumed to be one, because it is assumed that the logit model applies to each product in the exact choice set. Should this assumption not hold, then the coefficient on $\log(D_b)$ will differ from one.

A Model for Broad Choice Data

When the researcher observes individuals' choices among broad choice groups but alternative attributes are observed at the exact choice level, Brownstone and Li (2017) propose estimating household choice at the broad group level, defining the probability of choosing a broad group

as the sum of the probabilities of the exact choices contained within the group. This involves replacing the likelihood function in equation (1) with the following:

$$L(y;\delta,\beta) = \sum_n \sum_b Y_{nb}\, log\ (\tilde{P}_{nb})$$

where $\tilde{P}_{nb} = \sum_{j \in C_b} P_{nj}$ and P_{nj} is defined as in equation (1).

Given full knowledge of the universe of exact choice alternatives, and assuming the exact choice model is correctly specified, this is the correctly defined likelihood function given the information available to the analyst. In contrast to the other methods, it makes no additional assumptions and does not rely on an ad hoc representation of aggregated alternatives. Given that there is less-than-full information, all of the alternative methods considered here will generally contain less information for estimating parameters, so that estimates may be poorly identified and have larger variances than in the full-information case.

Brownstone and Li (2017) derive the Hessian and Information matrices for equation (3), which include terms that correspond to the full-observability model. These demonstrate the potential loss in concavity due to partial versus full observability. Parameter estimates using this approach can be poorly identified, particularly when the broad choice groups are very large.

There are numerous advantages to estimating the broad choice model over McFadden's method. The broad choice model avoids aggregation altogether. Also, the broad choice model does not require asymptotic distributional assumptions be placed on the variables of the exact choice set within each broad group as is the case with McFadden's method. There may not always be an intuitive way to partition the exact choice set into groups that are all large to best approximate the asymptotic normality assumption required for consistency when using McFadden's method.

However, the model for broad choice can be poorly identified, particularly when the broad groups are very large. One solution to overcome this problem is to supplement the data with macro-level market share data at the exact choice level and take a BLP approach to estimating the model. Wong (2015) and Brownstone and Li (2017) explore choice set aggregation within that context.

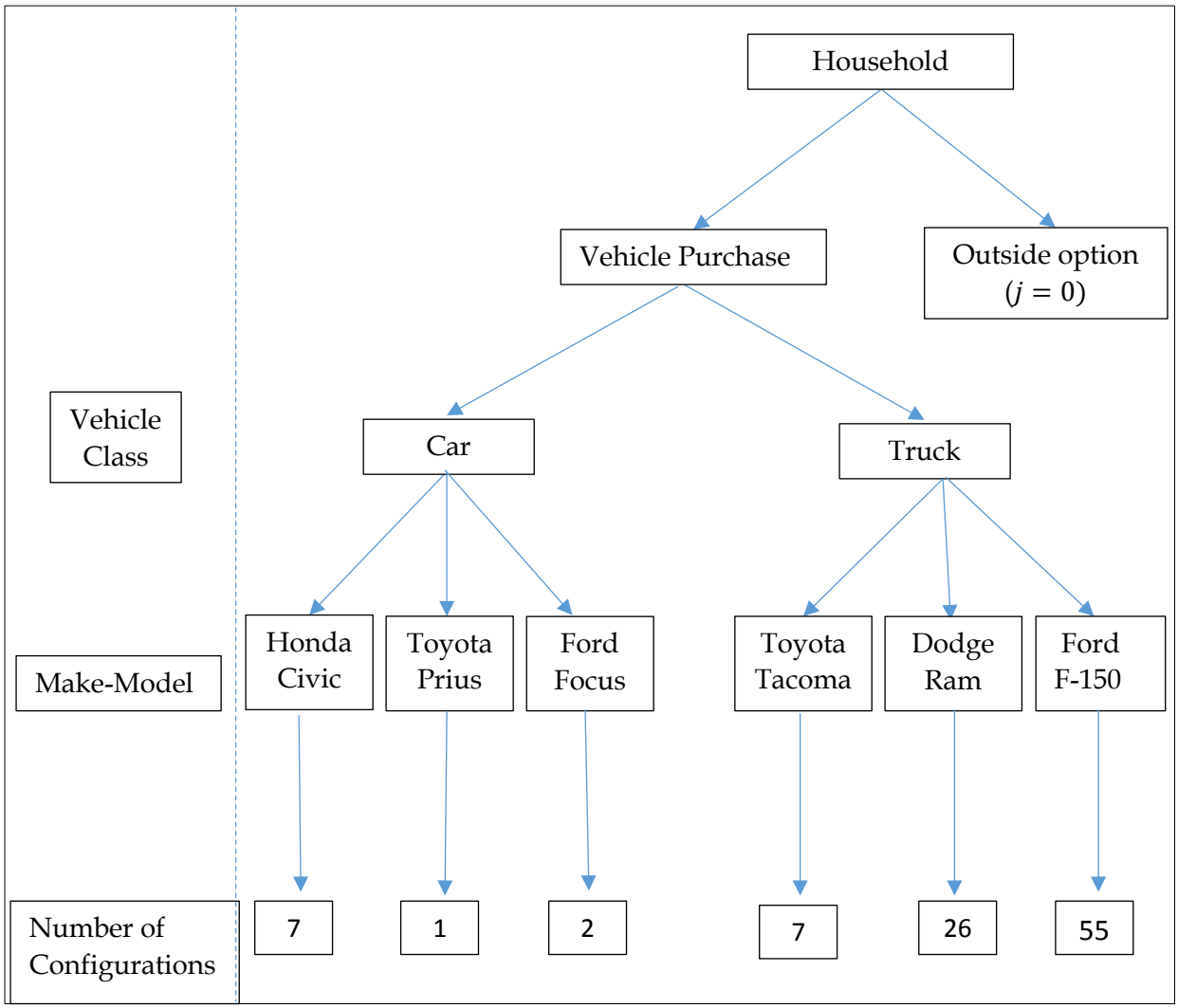

Figure 1: Vehicle Choices for Monte Carlo Experiment structured in tree-form

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. Households have the option of purchasing a vehicle or selecting the outside option. We split all vehicles into two vehicle classes: cars and trucks. Then, we consider three make/model vehicles within each class. Here we use data on real vehicles. We obtain Model Year 2008 vehicle attribute data from the Volpe Center, selecting three cars and three light trucks for the study. The corresponding number of configurations for each make/model vehicle is shown at the bottom of the tree in Figure 1. We use the price, transmission (manual or automatic) and gallons of fuel per one hundred miles variables for this study. We choose to use real vehicles and their corresponding attributes even in the Monte Carlo stage, rather than generate independent variables on our own to ensure that when we consider aggregation of vehicles to broader levels, we aggregate across realistic distributions of attributes. The summary statistics of configuration attributes by make/model are given in Table 1.

We also use real household data from the National Household Transportation Survey for our Monte Carlo design. There are a total of 145,732 households in our sample and we make use of their income data and average gasoline price in their state of residence in 2008. Households are divided into two income bins, those with annual incomes below $75,000 and those with incomes above $75,000.

Vehicle Make/Model	Number of Configurations	Price (US$ '000)				Fuel Consumption Rate (gallons/100 miles)				Manual Transmission			
		Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max	Mean	Std. Dev.	Min	Max
Honda Civic	7	26.00	3.16	23.10	30.05	3.30	0.11	3.01	3.93	0.43	0.53	0.00	1.00
Toyota Prius	1	22.04	0.00	22.04	22.04	1.52	0.00	1.52	1.52	0.00	0.00	0.00	0.00
Ford Focus	2	15.85	0.00	15.85	15.85	2.73	0.00	2.69	2.77	0.50	0.71	0.00	1.00
Toyota Tacoma	7	21.82	2.15	18.79	23.80	4.16	0.25	3.40	4.79	0.57	0.53	0.00	1.00
Dodge Ram	26	27.19	3.02	22.06	31.87	5.29	0.09	4.51	5.61	0.23	0.43	0.00	1.00
Ford F-150	55	25.25	2.31	22.89	29.35	5.25	0.16	4.70	6.23	0.09	0.29	0.00	1.00

Table 1: Summary Statistics of Vehicle Configuration Attributes by Make/Model

From these data, we generate the following choice model. Let $n = 1, \ldots, N$ index households who can purchase a particular vehicle configuration indexed by $j = 1, \ldots, J$, or choose the outside option, represented by $j = 0$. The indirect utility of household n from the choice of product, j, U_{nj} is assumed to follow the following linear specification:

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

where p_j is the price of the vehicle j, T_j is a binary variable indicating whether a vehicle has manual transmission, foc_{nj} is the fuel operating cost of the vehicle, inc_n is a dummy variable indicating that household n is a high-income household, and D_j is a class dummy (car or truck). $\beta = \{\beta_p, \beta_T, \beta_{foc}, \beta_{p-inc}\}$ are the parameters associated with these household and vehicle attributes. ϵ_{nj} is an error term that follows the type I extreme value distribution. We normalize the utility of the outside option to zero.

The fuel operating cost variable is in cents per mile. It is created by multiplying the fuel consumption rate of the vehicle (in gallons per mile) with the average price of fuel (in cents per gallon) in the state of residence of the household. We set the true choice model parameters so that about 47.5% of households choose cars, 47.5% of households choose trucks and 4% of households choose the outside option.

We choose the following true values for β: $\beta_p = -0.4, \beta_T = -0.1, \beta_{foc} = -0.2, \beta_{p-inc} = 0.1$. The values for β_p and β_{foc} yield a willingness to pay of \$500 for a 1 cent/mile improvement in fuel operating cost for low-income households and \$670 for high-income households. Assuming vehicles are held for 14 years with annual mileage of 10,000 miles, this works out to an implied discount rate of about 18% for low-income households and 11% for high-income households.

We consider four models. The first is a full observability model. Here we estimate the conditional logit assuming that the researcher observes the exact choices made by households. The remaining three models consider the scenario where the researcher only partially observes vehicle choice at the Make-Model level but has vehicle attribute data at the configuration level.

In the first, which we call the "average configuration method," we average the attributes of configurations within each Make-Model and assign these average attributes as the properties of each Make-Model vehicle. The second is the "McFadden aggregation method," where choices are also aggregated and then estimated at the Make-Model level, but also included in the model are the variances of configuration attributes and the log of the number of configurations within each Make-Model. The third model is the "broad choice model" where choices are modeled at the configuration level even though choices are only observed at the Make Model level.

We estimate each model 1000 times with varying sample sizes. The models are estimated by maximizing the respective likelihood functions using the "fminunc" function in Matlab. Given that the true model is multinomial logit, we provide both the analytic gradients and Hessians for all methods, except the McFadden method for which the Hessian calculations slow down estimation tremendously. In all but two cases (the McFadden and Broad Choice

approaches) the model is a linear-in-parameters multinomial logit, which is known to be globally concave and optimization is straightforward. The McFadden method (equation 2) has a nonlinear utility function, and the broad choice model is not necessarily globally concave. In any of these cases there is the possibility that the partially observable choices could lead to poor enough identification that the likelihood would be flat. However, we observed no difficulties with convergence due to singular Hessians, and testing with multiple starting values suggested no problems with alternative local optima.

Presented here are the results for 500 and 10,000 households. To obtain 10,000 households, we sample without replacement from the NHTS sample.

Table 2 presents the results obtained from estimating the choice model 1000 times for each of the aggregation methods. Presented are the mean estimate of the parameters across these 1000 iterations as well as the mean of the standard errors of the parameters and the proportion of time the 90% confidence intervals captured the true value.

The first thing to note from Table 2 is that the estimates of the parameters on the interaction terms (Price*High Income and Fuel Operating Cost) are relatively close to the true values across all three methods of aggregation, including the average configuration method. The reason the average configuration estimates perform well is that the household characteristics are providing sufficient information to allow the estimation of these parameters, in spite of the obfuscation from the aggregation of configuration attributes.

However, the remaining estimates from the average configuration method perform poorly. The willingness to pay value is almost four times larger than the true value in the sample of 500 households and three times larger in the sample of 10,000 households. The standard errors do not perform any better. The 90% coverage probability column indicates the fraction of times the 90% confidence intervals generated using the means and standard errors capture the true value. We see that for the average configuration method the confidence intervals capture the true value far fewer than 90% of the time, with the exception of the Price*High Income variable that still shows large variability across the different sample sizes.

Besides the manual transmission variable, all the estimates display downward bias, similar to what happens in miss-specified models. We see no improvement in the mean of the estimates when the household sample size is increased from 500 to 10,000, while smaller standard errors only give a false sense of confidence in the precision of estimates.

The McFadden aggregation method also does poorly, despite the fact that the model incorporates the variance of variables as well as the number of configurations within each broad group to properly account for the averaging of configurations. The mean estimates for the price variable and vehicle class dummies are closer to the true values than the average configuration model but farther away for the remaining value. The 90% coverage probabilities do not perform well in either case. They do better in the small household sample than in the larger sample. The smaller samples cause standard errors to be larger, thus generating wider confidence intervals that overlap the true values more often. As sample size increases, the standard errors shrink, generating tighter confidence intervals that almost never capture the true value. The estimate of the coefficient on the log (counts) variable comes close to the true

value of one, as it should, since the logit model does apply to the configuration level, however, the standard errors obtained for this estimate are imprecise. Surprisingly, the point estimates of the willingness to pay measure come very close, because both the price and fuel operating cost estimates are biased in the same direction, thus the bias diminishes when the ratio of both estimates are taken. However, the standard errors of the willingness to pay measures are too large.

500 households		Full Observability Model			Average Configuration Method			McFadden Aggregation Method			Broad Choice Model		
Variable	True Value	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability
Manual Transmission	-0.10	-0.10	0.14	0.90	0.48	0.27	0.25	0.33	0.32	0.59	-0.14	0.45	0.86
Price ('000)	-0.40	-0.40	0.02	0.91	-0.18	0.03	0.00	-0.32	0.03	0.18	-0.41	0.04	0.86
Price*High Income	0.10	0.10	0.02	0.90	0.10	0.02	0.85	0.08	0.02	0.71	0.10	0.02	0.89
Fuel Operating Cost (cents/mile)	-0.20	-0.20	0.04	0.90	-0.16	0.04	0.71	-0.16	0.03	0.66	-0.20	0.04	0.90
Car Dummy	8.00	8.07	0.50	0.89	4.33	0.73	0.00	6.14	0.85	0.29	8.19	1.14	0.85
Truck Dummy	7.50	7.57	0.56	0.91	5.22	0.80	0.11	5.69	0.88	0.33	7.68	1.14	0.86
Log (counts)	1.00							0.97	0.09	0.85			
Willingness to Pay*	0.67	0.68	0.02	0.90	2.51	5179.05	0.89	0.67	0.03	0.97	0.69	0.03	0.89

10,000 households		Full Observability Model			Average Configuration Method			McFadden Aggregation Method			Broad Choice Model		
Variable	True Value	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	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	-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

Table 2: Mean estimates, mean standard errors and 90% coverage probabilities from the four choice models.

*Willingness to pay (in thousands of dollars) for a 1 cent/mile improvement in fuel operating cost for high income households.

The broad choice model performs well in both large and small sample sizes. Point estimates are close to the true values and the standard errors are consistent. The 90% coverage probabilities perform as expected. An important comparison to make is the magnitude of the standard errors of estimates in the broad choice model relative to the full observability model. This tells us the loss in precision that stems from the lack of observability. The standard errors for the estimates associated with the vehicle attribute variables (manual transmission, price, car dummy, truck dummy) are 2-3 times larger in the broad choice model than in the full observability model. In contrast, the standard errors of the parameters on the w_{nj} variables (price*high income and fuel operating cost) are the same in both the full observability model and the broad choice model. The added variation that household characteristics supply increases the precision of estimates, diminishing the cost of information loss.

Comparing across the three models, we see that the estimates of β_w are closer to the true values and less variable across the three models. Again, the variation that household characteristics supply provide stability to the estimates.

We also re-ran the Monte Carlo study using different "true" parameter values in the data generation process. None of our results are sensitive to changing the true parameter values.

Section B of the appendix gives Monte Carlo results from other methods of aggregation commonly used by choice model practitioners, such as using the most commonly purchased vehicle to represent each broad group and averaging the attributes within each broad group using macro-level market shares as weights. These other methods generally perform worse than McFadden's method described above.

Empirical Application

Next, we estimate the same four models on an actual vehicle choice data set. Here we model vehicle choice conditional on vehicle purchase, that is, we do not include a "no buy" option. Vehicle attributes are provided by the Volpe Center and supplemented with data from Polk, the American Fleet Magazine, and the National Automobile Dealers Association. Vehicle price data are adjusted adding the gas guzzler tax for some vehicles and subtracting estimated purchase subsidies for hybrid vehicles. Vehicle attribute data are available at the Make/Model/Fuel-type/configuration level, however, NHTS household choices are only observed at the Make/Model/Fuel-type level.

There are 10,500 NHTS households in the dataset who purchase at least one new model year 2008 vehicle during the sample period. All household characteristic variables are categorical in nature. Because of this, we aggregate to 4157 unique "household types," with between one and forty-one households within each type. There are 235 broad groups of vehicles that households choose from, and 1120 vehicles in the exact choice set.

Table 3 provides some descriptive statistics about the NHTS household sample. Table 4 summarizes the utility specification that is used in the models.

NHTS Socioeconomic Attribute Variables	Sample Value (%)
Percent retired with no children	34.17
Percent whose children is under the age of 15	26.89
Percent living in urban areas	68.07
Percent of household respondents with college degree	48.12
Average gasoline price at time of vehicle purchase ($/gallon)	3.46
Household Income Distribution†:	
Less than $25,000	5.98
$25,000 - $75,000	35.36
$75,000 - $100,000	16.62
Greater than $100,000	35.28
Income Missing	6.76
Household Size Distribution	
1	10.96
2	49.31
3	17.14
4+	22.58
Market share of MY2008 vehicle purchases by Manufacturer	Share (%)
General Motors	21.76
Toyota	18.82
Honda	15.53
Ford	13.87
Other Japanese	8.87
Chrysler	8.53
European	6.46
Korean	4.30

Table 3: Descriptive Statistics of the NHTS sample and market shares

†Although five household income categories are observed, we use only four in the empirical application. We combine the lowest two categories into one for purposes of identification as we find the results for the two categories are very similar.

x_j	w_{nj}
Price	(Price) × (75,000<Income<100,000)
Horsepower/Curb Weight	(Price) × (Income>100,000)
Hybrid	(Price) × (Income Missing)
Curb Weight	(Prestige) × (Urban)
Wagon	(Prestige) × (Income>100,000)
Mid-Large Car	(Performance Car) × (Income>100,000)
Performance Car	(Japan) × (Urban)
Small-Medium Pickup	(Van) × (Children under 15)
Large Pickup	(Large SUV) × (Children under 15)
Small-Medium SUV	(Small SUV) × (Children under 15)
Large SUV	(Korea) × (Rural)
	(Seats≥5) × (Household Size≥4)
	(Mid-Large Car) × (Retired)
	(Prestige) × (Retired)
	(Import) × (College)
	(Prestige) × (Japan) × (College)
	(Prestige) × (Europe) × (College)
	(Prestige) × (Japan) × (Urban)
	(Performance Car) × (College)
	Fuel Operating Cost (cents per mile)
	(Fuel Operating Cost) × (College)

Table 4: Vehicle attributes, x_j, and vehicle-household attribute interactions, w_{nj}, included in the estimated model

Note**:** Fuel operating cost is the product of gallons per mile and fuel cost (in cents per mile)
"Korea," "Japan," and "Europe" are dummy variables that equal 1 if the vehicle is made in that region and 0 otherwise.
"Prestige" is a dummy variable that equals 1 if the vehicle is classified as a "prestige brand" by the American Fleet Magazine.
The following vehicle classes were adopted from the American Fleet Magazine: Mid-Large Car, Performance Car, Small-Medium Pickup, Large Pickup, Small-Medium SUV and Large SUV.

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 ***

Table 5: Select estimates across the three models.

*Notes: * denotes significance at the 10% level. ** denotes significance at the 5% level. *** denotes significance at the 1% level.*

We estimate the same three partial observability models as in the Monte Carlo study on this data. Table 5 presents the estimates of key variables. Full results are in Section A of the appendix. Only the fuel operating cost variable exhibits robustness across models. This is consistent with the findings from the Monte Carlo Study where the w_{nj} estimates were closer

to the true value and closer to each other across models because of the variation from household characteristics. The price interacted with income coefficient estimates do not exhibit similar robustness because of the discrete nature of the income variable, which renders it less informative. The vast differences across the other estimates suggests that the manner in which one aggregates a choice set can have very significant impacts on model estimates. To see this more clearly, we turn to the estimates of a more policy relevant quantity: willingness to pay for improvements in vehicle fuel efficiency.

Table 6 presents the implied willingness to pay estimates for a 1 cent/mile improvement in fuel operating cost, in thousands of dollars for households with incomes less than $75,000. A 1 cent/mile improvement in fuel cost is a 7.4% improvement over the average fuel operating cost of households in the sample. The final column of Table 5 provides the implied discount rate assuming vehicles are held for 14 years (Greene, 2010), with an annual mileage of 11,000 (FHWA, 2014). The negative discount rates in this column indicate that across all three models, households overvalue future fuel savings compared to present day investments in vehicle fuel efficiency.

Willingness to pay for a 1 cent/mile improvement in fuel efficiency (thousands)†	Estimate	Standard Error	Implied Discount Rate
Average Configuration Model	3.998	0.090 ***	-10.73
McFadden Aggregation Model	6.503	0.287 ***	-15.13
Broad Choice Model	5.449	0.185 ***	-13.59

Table 6: Willingness to pay estimates across the three model specifications

*Note: * denotes significance at the 10% level. ** denotes significance at the 5% level. *** denotes significance at the 1% level. † Willingness to pay for a 1 cent/mile reduction in fuel operating costs for households' income below $75,000 (in thousands of dollars).*

We see that when the researcher simply averages configurations to the Make/Model/Fuel-type level, treating these choices as exact, the estimate of willingness to pay are much smaller than the other methods. The estimates from the McFadden aggregation method are 60% larger while the broad choice estimates are 35% larger.

As is the case in the Monte Carlo study, McFadden's Aggregation Model does not perform well because the normality assumption within each broad group is not satisfied. The Make/Model/Fuel-type groups contain between one and fifty-five configurations, with 47.7% of the groups containing three configurations or less. We should thus consider the Broad Choice Model willingness to pay estimates from the most plausible of the three.

Conclusion

We examine the implications of choice set aggregation on parameter estimates in multinomial choice models, a common practice when household choices are not fully observed and when modelling choices at the most detailed level renders the model too large for estimation in standard computing environments. In addition to just averaging attribute values across elemental alternatives we examine two models that account for choice set

aggregation. The first is a method for aggregation by McFadden, 1978, that places distributional assumptions on the elements within each aggregated alternative and uses the higher order moments of the distribution in the utility specification. The second is a model for broad choice by Brownstone and Li, 2017, that defines the choice probability of a broad group of products as the sum of the probabilities of products within that group, circumventing the need for aggregation. We carry out a Monte Carlo study based on real data, and we find that only the broad choice estimator is reliable. We also consider additional methods used in the literature (choosing a representative elemental alternative) and show that these other methods perform worse than McFadden's method.

The poor performance of commonly used methods in the Monte Carlo study is not an artefact of our Monte Carlo design. We estimated models using real data from the 2009 NHTS and found large differences in parameter estimates and willingness-to-pay estimates across the 3 aggregation methods included in the Monte Carlo study.

More generally, these findings send a cautionary message to choice model practitioners on the importance of giving due consideration to how choice sets are defined. Aggregating choices without accounting for the variation across aggregated alternatives may lead the researcher to flawed and invalid conclusions from model estimates. Given the popularity of multinomial choice models across a variety of fields, including transportation, industrial organization and marketing, such practices may have widespread consequences. The "broad choice" method performs much better than other commonly used alternatives.

References

Bento, A. M., Goulder, L. H., Jacobsen, M. R., von Haefen, R. H., 2009. Distributional and Efficiency Impacts of Increased US Gasoline Taxes. American Economic Review. 99(3):667 – 699.

Berry, S.T., Levinsohn, J., Pakes, A., 1995. Automobile Prices in Market Equilibrium. Econometrica. 63(4), 841-890.

Berry, S.T., Levinsohn, J., Pakes, A., 2004. Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market. Journal of Political Economy. 112(1), 68-105.

Brownstone, D., Li, P., 2017. A model for broad choice data. Journal of Choice Modelling, https://doi.org/10.1016/j.jocm.2017.09.001, forthcoming.

Federal Highway Administration, 2014. Highway Statistic Series 2010.

Fernández-Antolín, A., M. de Lapparent, and M. Bierlaire. Modeling purchases of new cars: an analysis of the 2014 French market. Theory and Decision, forthcoming 2017. http://dx.doi.org/10.1007/s11238-017-9631-y

Greene, D., 2010. How Households Value Fuel Economy: A Literature Review. Office of Transportation and Air Quality, U.S. Environmental Protection Agency, Report EPA-420-R-10-008.

Lave, C. and Train, K. 1979. A Disaggregate Model of Auto-type Choice. Transportation Research. 13A:1-9.

Lerman, S. R., 1977. Location, Housing, Automobile Ownership, and Mode to Work: A Joint Choice Model. Transportation Research Board Record. 610:6-11.

McFadden, D., 1978. Modeling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickars, and J.Weibull (Eds.), *Spatial Interaction Theory and Planning Models*. North-Holland, Amsterdam: 75-96.

Petrin, A., 2002. Quantifying the Benefits of New Products: The Case of the Minivan. Journal of Political Economy. 110, 705-729.

Train, K. and Winston, C. ,2007. Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers. International Economic Review, 48, 1469–1496.

Wong, T. J. *Econometric Models in Transportation***,** Ph.D. thesis, Department of Economics, University of California, Irvine, June, 2015.

Appendix: Section A

w_{nj} w_{nj}	Estimated Parameter	Standard Error	
(Price) × (75,000<Income<100,000)	0.038	0.003	***
(Price) × (Income>100,000)	0.075	0.003	***
(Price) × (Income Missing)	0.066	0.004	***
(Prestige) × (Urban)	-0.359	0.064	***
(Prestige) × (Income>100,000)	-0.167	0.074	**
(Performance Car) × (Income>100,000)	0.055	0.095	
(Japan) × (Urban)	0.265	0.027	***
(Van) × (Children under 15)	1.034	0.058	***
(Large SUV) × (Children under 15)	0.112	0.051	**
(Small SUV) × (Children under 15)	0.992	0.103	***
(Korea) × (Rural)	-0.795	0.092	***
(Seats≥5) × (Household Size≥4)	0.062	0.061	
(Mid-Large Car) × (Retired)	0.346	0.054	***
(Prestige) × (Retired)	-0.086	0.078	
(Import) × (College)	0.005	0.037	
(Prestige) × (Japan) × (College)	0.070	0.132	
(Prestige) × (Europe) × (College)	-0.244	0.101	**
(Prestige) × (Japan) × (Urban)	-0.556	0.099	***
(Performance Car) × (College)	0.723	0.090	***
Fuel Operating Cost (cents per mile)	-0.240	0.012	***
(Fuel Operating Cost) × (College)	-0.081	0.007	***
x_j	Estimated Parameter	Standard Error	
Price	-0.060	0.003	***
Horsepower/Curb weight	0.187	1.773	
Hybrid	-1.740	0.073	***
Curb weight	0.391	0.037	***
Wagon	-1.033	0.114	***
Mid-Large Car	0.487	0.036	***
Performance Car	-0.486	0.086	***
Small-Medium Pickup	0.466	0.073	***
Large Pickup	2.143	0.063	***
Small-Mid SUV	0.457	0.039	***
Large SUV	0.394	0.092	***

Table A1: Average configuration model: parameter estimates

*Notes: * denotes significance at the 10% level. ** denotes significance at the 5% level. *** denotes significance at the 1% level.*

w_{nj} w_{nj}	Estimated Parameter	Standard Error	
(Price) × (75,000<Income<100,000)	0.014	0.003	***
(Price) × (Income>100,000)	0.050	0.003	***
(Price) × (Income Missing)	0.041	0.004	***
(Prestige) × (Urban)	-0.230	0.062	***
(Prestige) × (Income>100,000)	-0.056	0.067	
(Performance Car) × (Income>100,000)	0.080	0.094	
(Japan) × (Urban)	0.470	0.029	***
(Van) × (Children under 15)	0.558	0.068	***
(Large SUV) × (Children under 15)	0.088	0.053	
(Small SUV) × (Children under 15)	0.349	0.124	***
(Korea) × (Rural)	-0.733	0.087	***
(Seats≥5) × (Household Size≥4)	0.103	0.014	***
(Mid-Large Car) × (Retired)	0.610	0.046	***
(Prestige) × (Retired)	-0.165	0.059	***
(Import) × (College)	0.109	0.038	***
(Prestige) × (Japan) × (College)	0.050	0.123	
(Prestige) × (Europe) × (College)	-0.302	0.100	***
(Prestige) × (Japan) × (Urban)	-0.246	0.089	***
(Performance Car) × (College)	0.111	0.085	
Fuel Operating Cost (cents per mile)	-0.255	0.012	***
(Fuel Operating Cost) × (College)	-0.058	0.007	***
x_j	Estimated Parameter	Standard Error	
Price	-0.039	0.003	***
Horsepower/Curb weight	0.088	0.821	
Hybrid	-1.453	0.069	***
Curb weight	0.554	0.035	***
Wagon	-1.080	0.101	***
Mid-Large Car	0.182	0.035	***
Performance Car	-0.172	0.062	***
Small-Medium Pickup	0.530	0.070	***
Large Pickup	1.453	0.067	***
Small-Mid SUV	0.172	0.038	***
Large SUV	-0.295	0.090	***

Table A2: McFadden aggregation model: parameter estimates

*Notes: * denotes significance at the 10% level. ** denotes significance at the 5% level. *** denotes significance at the 1% level.*

w_{nj} w_{nj}	Estimated Parameter	Standard Error	
(Price) × (75,000<Income<100,000)	0.019	0.003	***
(Price) × (Income>100,000)	0.044	0.003	***
(Price) × (Income Missing)	0.039	0.004	***
(Prestige) × (Urban)	-0.482	0.062	***
(Prestige) × (Income>100,000)	0.023	0.067	
(Performance Car) × (Income>100,000)	-0.046	0.094	
(Japan) × (Urban)	0.721	0.029	***
(Van) × (Children under 15)	0.741	0.068	***
(Large SUV) × (Children under 15)	0.688	0.053	***
(Small SUV) × (Children under 15)	0.036	0.124	
(Korea) × (Rural)	-0.530	0.087	***
(Seats≥5) × (Household Size≥4)	0.765	0.014	***
(Mid-Large Car) × (Retired)	0.588	0.046	***
(Prestige) × (Retired)	-0.120	0.059	*
(Import) × (College)	0.306	0.038	***
(Prestige) × (Japan) × (College)	0.032	0.123	
(Prestige) × (Europe) × (College)	-0.601	0.100	***
(Prestige) × (Japan) × (Urban)	-0.278	0.089	***
(Performance Car) × (College)	0.571	0.085	***
Fuel Operating Cost (cents per mile)	-0.250	0.012	***
(Fuel Operating Cost) × (College)	-0.016	0.007	**
x_j	Estimated Parameter	Standard Error	
Price	-0.046	0.003	***
Horsepower/Curb weight	17.729	0.821	***
Hybrid	-0.242	0.069	***
Curb weight	0.510	0.035	***
Wagon	-1.832	0.101	***
Mid-Large Car	0.109	0.035	***
Performance Car	-0.410	0.062	***
Small-Medium Pickup	0.254	0.070	***
Large Pickup	0.149	0.067	**
Small-Mid SUV	0.339	0.038	***
Large SUV	-0.033	0.090	

Table A3: Broad choice model: parameter estimates

*Notes: * denotes significance at the 10% level. ** denotes significance at the 5% level. *** denotes significance at the 1% level.*

Appendix: Section B

In this section, we present the results from a Monte Carlo study of three other commonly practiced methods of aggregation. One key difference between the methods considered in this section and those in the main body of the paper in that the methods here require the availability of macro-level data on market shares in constructing the choice set.

The first uses the most commonly purchased configuration within each broad group as the representative configuration for that broad group. We call this the "representative configuration method." This method yields consistent estimates only if the vehicles within each group have the exact same vehicle attributes. Table 1 shows summary statistics for the price variable across the six make-models

As one can see from Table 1, there is significant variance in the prices of vehicles within each make-model group. In particular, the Honda Civics and and Dodge Rams have large variations in prices. The other variables (gallons of fuel per mile and manual transmission) exhibit large variations within make-model groups too. This means that it is unlikely that the representative configuration model would perform well on this dataset.

The second aggregation method that we consider averages the configurations within each Make-Model group, using the macro-level market shares as weights. The third adds to the second method the log of the number of configurations within each Make-Model group to account.

The results from this Monte Carlo study are shown in Table B1. All three methods perform generally worse than the McFadden aggregation method. For example, the estimates of the parameter associated with Manual Transmission are positive and larger than 1 in magnitude across all three models. In comparison, the McFadden aggregation methods estimates the parameter to be 0.34.

The representative configuration method estimates perform particularly poorly. Its estimate of the parameter on manual transmission is the most inflated. This model's estimates of the parameters associated with vehicle price and the car and truck dummies are of the wrong sign.

Of these three methods, the final one performs the best. With the exception of the manual transmission variable, it produces the most reasonable parameter estimates. In particular, its estimate of willingness to pay is closest to the true value, though still almost three times larger. The coverage probabilities of this model never captures the true value, which means the standard errors obtained from this model are unreliable.

10,000 households		Representative Configuration Model			Weighted Average Configuration Model			Weighted Average Configuration Plus Log (counts) Model		
Variable	True Value	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability	Mean Estimate	Mean Std. Error	90% Coverage Probability
Manual Transmission	-0.10	3.51	0.07	0.00	1.21	0.08	0.00	1.67	0.11	0.00
Price	-0.40	0.14	0.01	0.00	-0.12	0.01	0.00	-0.20	0.01	0.00
Price*High Income	0.10	0.11	0.01	0.30	0.11	0.01	0.56	0.10	0.01	0.00
Fuel Operating Cost (cents/mile)	-0.20	-0.20	0.01	0.88	-0.18	0.01	0.30	-0.20	0.01	0.00
Car Dummy	8.00	-3.42	0.22	0.00	2.77	0.20	0.00	3.77	0.26	0.00
Truck Dummy	7.50	-1.72	0.21	0.00	3.46	0.20	0.00	3.91	0.26	0.00
Log (counts)	1.00							0.04	0.00	0.00
Willingness to Pay*	0.67	-0.82	0.00	0.00	14.92	14511394.24	0.47	1.90	0.05	0.00

Table B1: Monte Carlo Results

*Note: * Willingness to pay for a 1 cent/mile reduction in fuel operating costs for households' income below $75,000 (in thousands of dollars).*