The neglected impact of measurement error on disaggregate transportation demand models.

David Brownstone, Department of Economics and Institute of Transportation Studies, U.C. Irvine

Dedicated to Charles Lave 1938 - 2008

- Econometricians have known for almost a century that using variables subject to measurement errors in regression models always biases inference and frequently leads to inconsistent estimation.
- Route choice, mode choice, and vehicle choice models all require information about non-chosen alternatives, and these data are frequently imputed (e.g. from network skims) with substantial error.

Gross Measurement Errors - Outliers

- Maximum likelihood estimators of discrete choice models very sensitive to outliers:

$$\max_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \log\left(P\left(y_{ij} = 1 \mid x_i, \theta\right)\right)$$

(contribution of *i* is unbounded)

- Alternative Nonlinear Least Squares:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(y_{ij} - P\left(y_{ij} = 1 \mid x_i, \theta\right)\right)^2$$

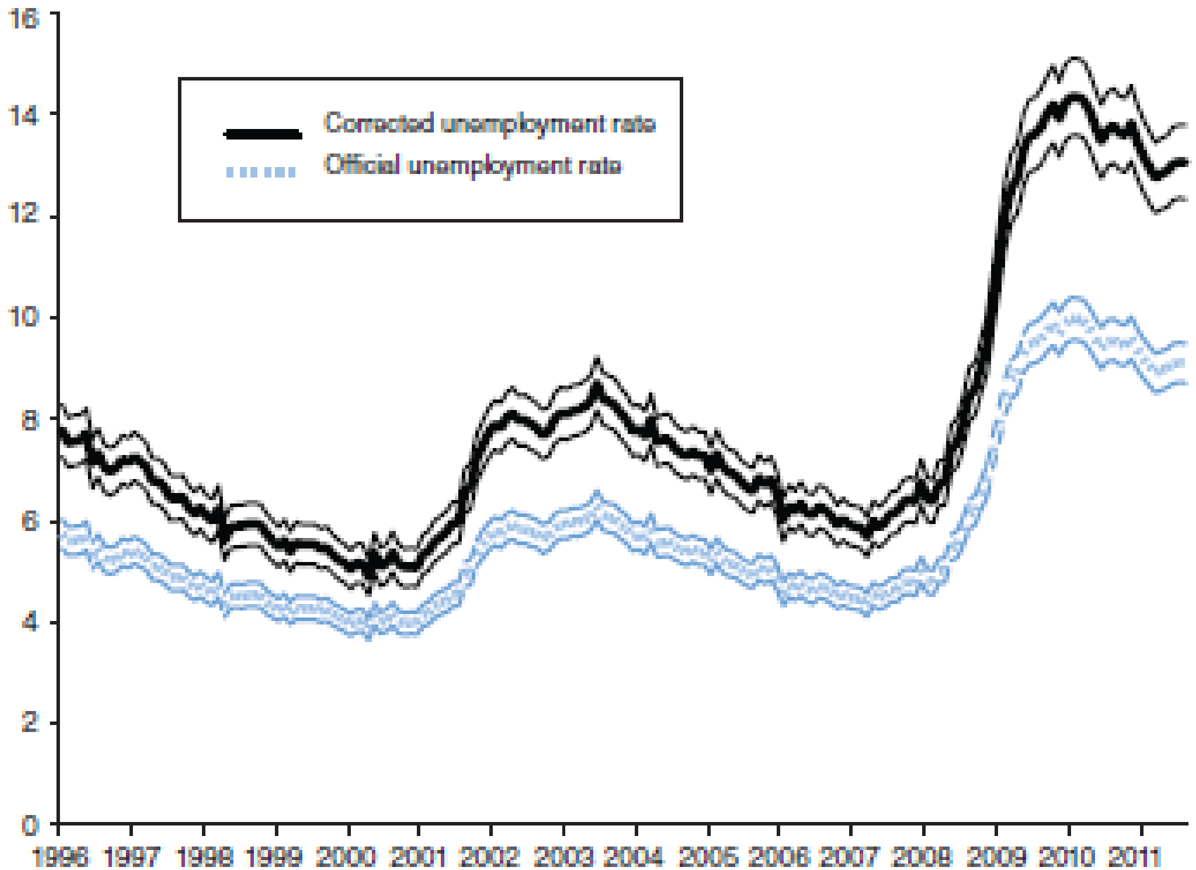


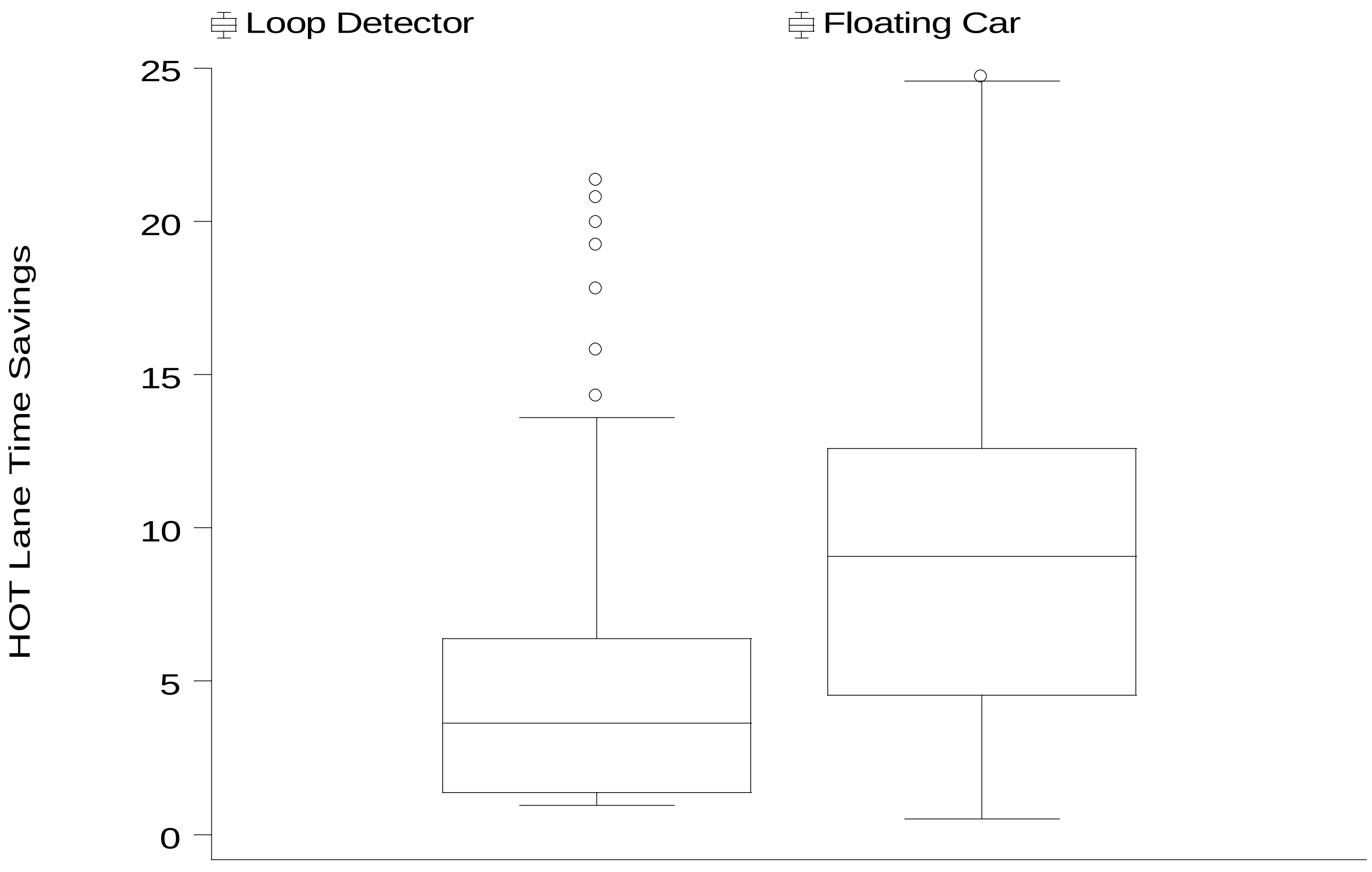
Figure 1. Corrected and Official (*reported*) Unemployment Rates

Feng and Hu, *American Economic Review* 103:2, 1054-1070, 2013. Based on repeated CPS panel observations and various Markov assumptions on reporting process.

Measurement Errors in Income

- Brownstone and Valletta (*Review of Economics and Statistics*, 78:4, 705-717, 1996) show that measurement errors in annual earnings are negatively correlated with potential experience (age – yrs of schooling – 6) and blue collar status.
- Corrected wage equations show higher returns to experience and no sensitivity to union or blue-collar status

Measurement Errors in Travel time savings



Measurement Errors in Value of Travel Time Savings

Value of Time ($/hour)	Corrected	Loop Data
95th Percentile	108.70	105.60
90th Percentile	72.12	73.63
75th Percentile	31.30	35.27
50th Percentile	18.71	23.37
25th Percentile	10.30	16.55
10th Percentile	-20.72	14.43
5th Percentile	-83.02	14.08
Mean	25.63	32.64

Steimetz and Brownstone, *Transportation Research B*, 39, 865-889, 2005

Urban Bus Fleet Efficiency

- UMTA – EPA approach: urban busses use about 30 Gal/100 Miles and cars about 4.4. Therefore breakeven is approximately 7 passengers per bus.
- This assumes only one person/car and that bus passengers stay on for entire run.
- John Naviaux (UCI Economics Honors Thesis 2011) rode OCTA busses for a week to collect data.

Route	Total Car CO_2(lbs)	Total Bus CO_2(lbs)	Total Sampled Distance (miles)	Total Inefficient Miles	Inefficient Miles as % of total
33	78.36	40.39	8.7	0.00	0.00
47	136.78	82.18	17.7	4.70	26.55
50	311.65	63.61	13.7	0.00	0.00
53	163.98	96.11	20.7	8.80	42.51
55	107.70	58.04	12.5	3.70	29.60
57	132.86	60.82	13.1	1.20	9.16
59	644.99	415.07	89.4	42.20	47.20
64	364.88	82.64	17.8	0.00	0.00
66	223.32	79.39	17.1	2.10	12.28
143	58.16	48.75	10.5	6.70	63.81
TOTALS:	2222.67	1027.00	221.20	69.40	

-

	Car w/ 1.1 Riders	Car w/ 1.6 Riders	Car w/ 2 Riders
Superior to CNG Bus	46.18 mpg	31.75 mpg	25.40 mpg
Superior to Diesel Bus	38.92 mpg	26.76 mpg	21.41 mpg

Errors in NHTS VMT measures

- Charles Lave (1994, http://escholarship.org/uc/item/5527j8dj) showed that big jump in VMT from 1983 – 1990 caused by switch from personal to telephone interviews. This led to bias towards newer vehicles.
- Lave also showed that NHTS self-reported VMT was very unreliable by comparing to California smog check data.

Fig. 1: NPTS vs ORNL Auto Registrations

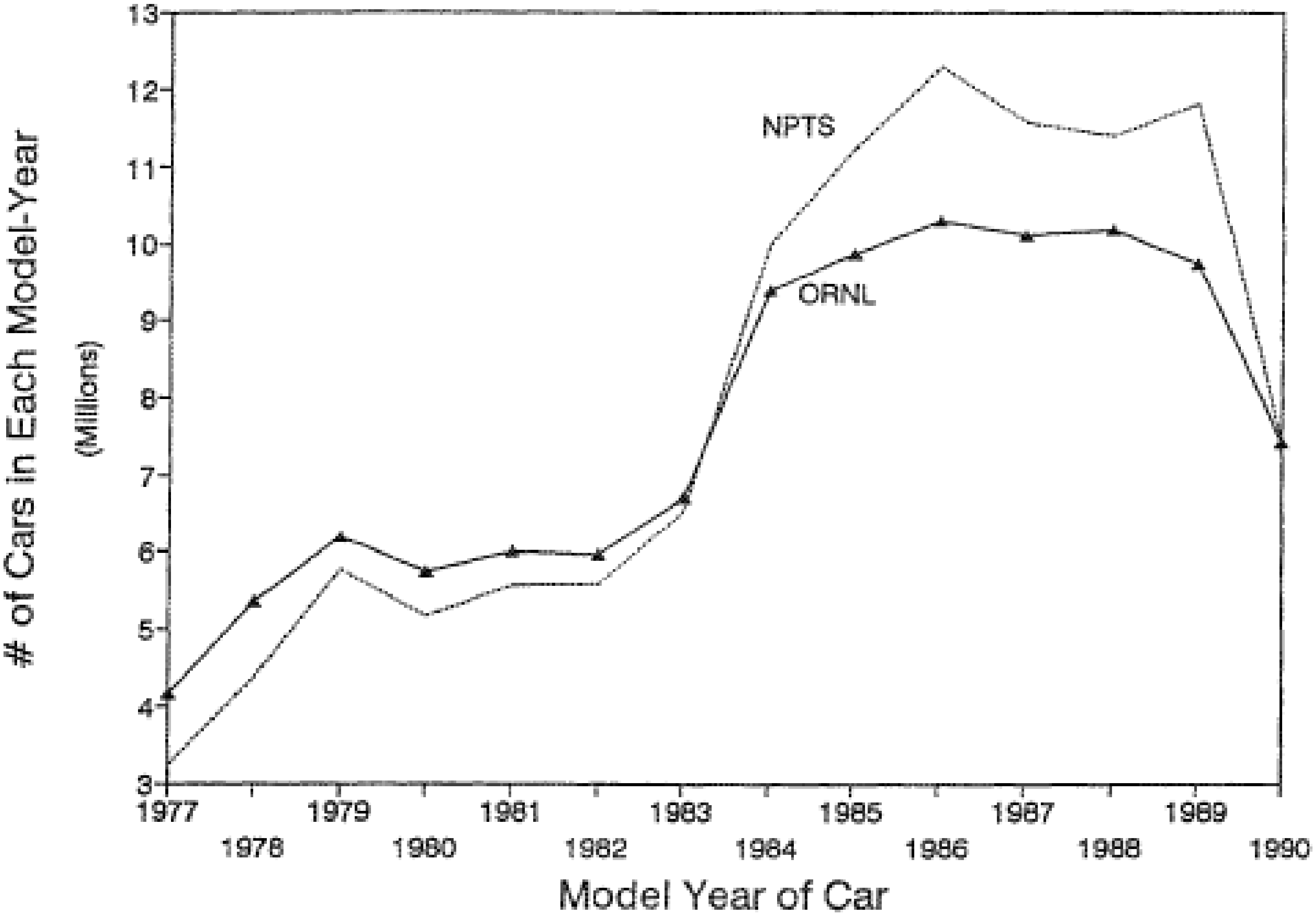
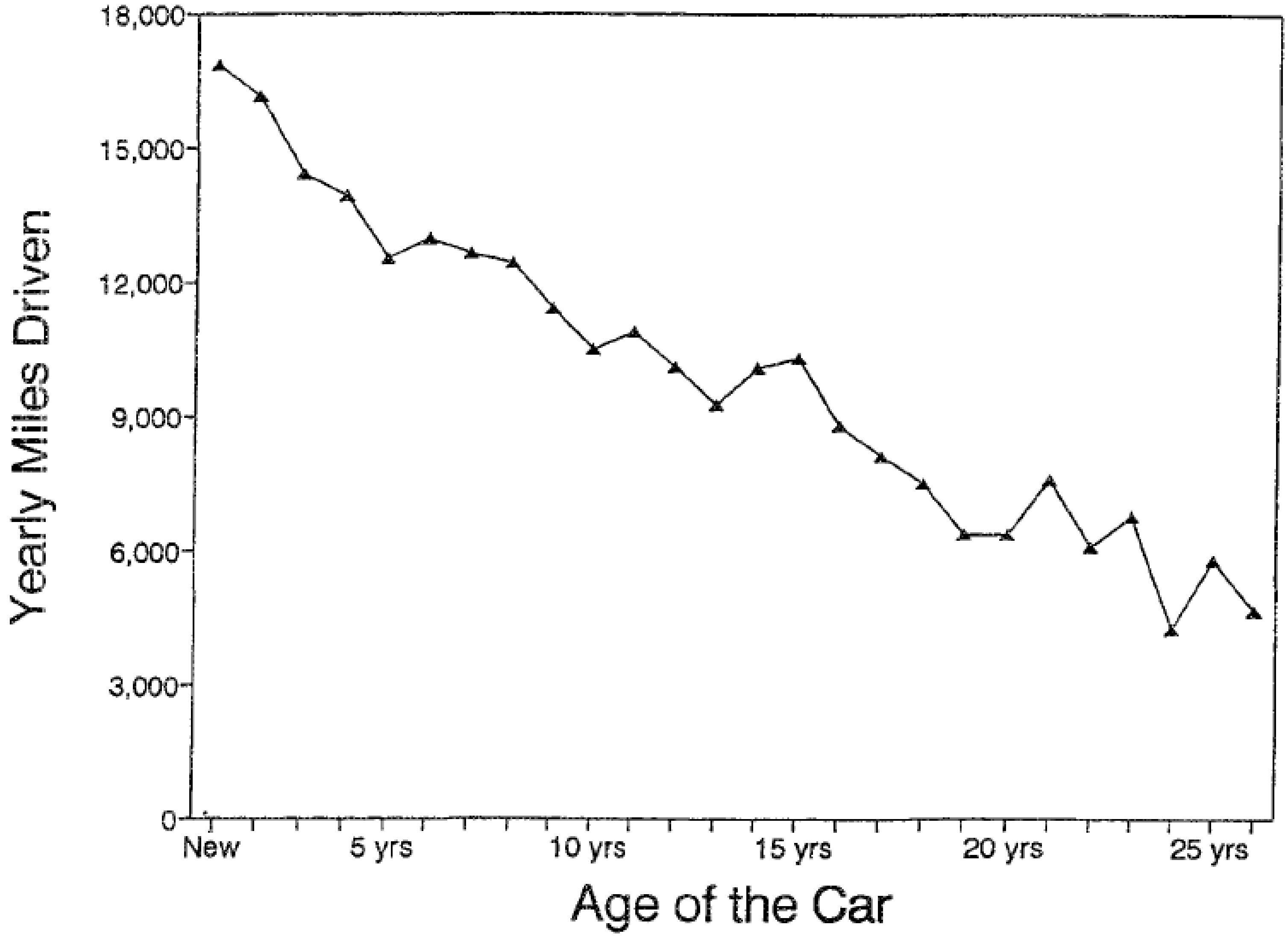


Fig 2: Annual VMT, by Age of Car



NHTS data

- Large representative national sample including inventory of household vehicles and miles driven by each vehicle.
- Previously used for vehicle choice and utilization modeling (e.g. Bento et. al., 2009 used 2001 NHTS data)
- 2009 data include month of purchase and include about 8000 hybrids (most common are Prius, Civic and Camry)

Current NHTS VMT measures

- Lave showed that RTECS survey which used dual odometer readings was accurate, so in 2001 NHTS switched to dual odometer readings.
- Due to budget cuts, 2008 NHTS reverted back to one odometer reading.
- 2008 NHTS “BestMiles” variable is imputed from single odometer reading using model fit on 2001 NHTS.

Utilization Estimation for Model Year 2008 Vehicles in the 2009 NHTS

Dependent Variable: ln(Vehicle Miles Traveled)

Number of Observations: 6730

	Measurement Method					
	Odometer		Self-Reported		"BestMiles"	
Variable	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
ln(Cost per Mile)	-0.027	0.063	0.028	0.058	-0.020	0.059
hybrid	0.105	0.052	0.150	0.069	0.074	0.062
car	-0.234	0.103	-0.221	0.083	-0.232	0.066
truck	-0.322	0.111	-0.227	0.098	-0.110	0.090
van	-0.138	0.127	-0.121	0.107	-0.110	0.088
suv	-0.261	0.105	-0.236	0.091	-0.156	0.079
import	-0.116	0.039	-0.025	0.035	-0.009	0.040
household income (in $10,000)	0.014	0.005	0.010	0.005	0.004	0.006
distance to work	0.007	0.001	0.004	0.001	0.003	0.001
college	0.106	0.036	0.072	0.033	0.102	0.037
worker	0.133	0.048	0.144	0.048	0.064	0.054

Aggregation Bias in Discrete Choice Models with an Application to Household Vehicle Choice

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

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

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

With help from Alicia Lloro, Jinwon Kim, and Phillip Li

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.

Model Notation

$$
\begin{aligned}
U^*_{ij} &= \delta_j + x'_{ij}\beta + \epsilon_{ij}, \quad \epsilon_{ij} \overset{i.i.d.}{\sim} \text{Type 1 Extreme Value}, \\
Y^*_i &= j \quad \text{if} \quad U^*_{ij} > U^*_{ik}\, \forall\, k \in C = \{1, 2, \ldots, J\}, \\
Y_i &= m \quad \text{if} \quad Y^*_i \in C_m,
\end{aligned}
$$

decision makers $i = 1, \ldots, N,$

alternatives $j = 1, \ldots, J,$

groups $m = 1, 2, \ldots, M$

Likelihood Function

$$
\begin{aligned}
\tilde{P}_{im} &= \Pr(Y_i = m), \\
&= \Pr(Y_i^* \in C_m), \\
&= \sum_{c \in C_m} P_{ic},
\end{aligned}
$$

$$P_{ic} = \Pr(Y_i^* = c) = \frac{\exp\left(w_{ic}\theta\right)}{\sum_{j=1}^{J} \exp\left(w'_{ij}\theta\right)}.$$

$$L_B(\theta) = \sum_{i=1}^{N} \sum_{m=1}^{M} Y_{im} \log(\tilde{P}_{im}),$$

Score Function

$$S_B(\theta) = \frac{\partial L_B(\theta)}{\partial \theta} = \sum_{i=1}^{N} \left(\sum_{m=1}^{M} Y_{im} \sum_{c \in C_m} w_{ic} P_{ic|C_m} - \sum_{j=1}^{J} w_{ij} P_{ij} \right),$$

$$P_{ic|C_m} = \frac{\exp\left(w'_{ic}\theta\right)}{\sum_{s \in C_m} \exp\left(w'_{is}\theta\right)},$$

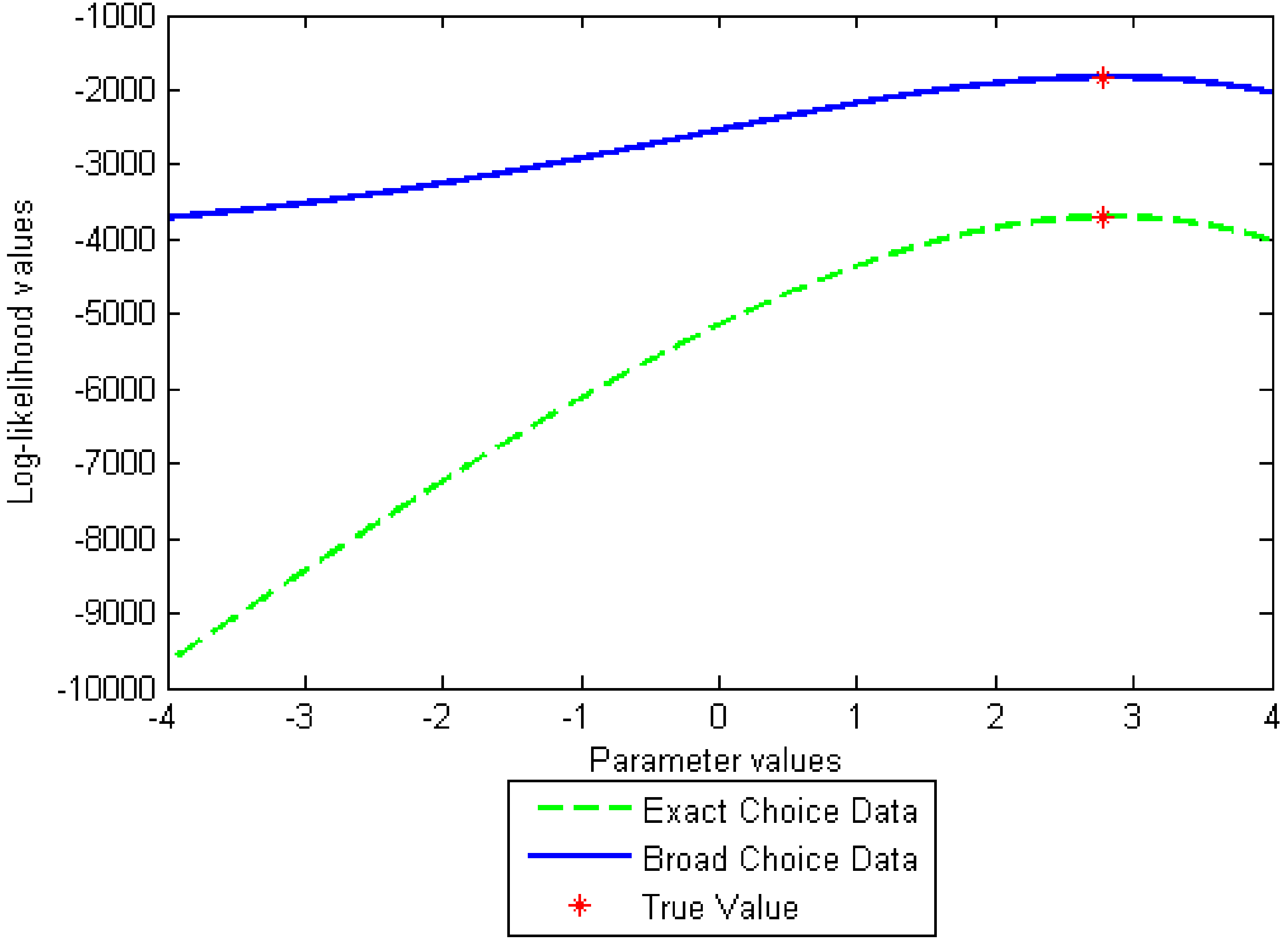
Hessian

$$H_B(\theta) = \frac{\partial L_B(\theta)}{\partial\theta\partial\theta'} = L - F,$$

$$L = \sum_{i=1}^{N}\left(\sum_{m=1}^{M} Y_{im} \sum_{c\in C_m}(w_{ic} - \sum_{s\in C_m} P_{is|C_m} w_{is}) P_{ic|C_m} (w_{ic} - \sum_{s\in C_m} P_{is|C_m} w_{is})'\right)$$

$$F = \sum_{i=1}^{N}\left(\sum_{j=1}^{J}(w_{ij} - \sum_{r=1}^{J} P_{ir} w_{ir}) P_{ij} (w_{ij} - \sum_{r=1}^{J} P_{ir} w_{ir})'\right)$$

With exact choice data, Hessian = -*F*


-1000
-2000
-3000
-4000
-5000
-6000
-7000
-8000
-9000
-10000
Log-likelihood values
-4
-3
-2
-1
0
1
2
3
4
Parameter values
Exact Choice Data
Broad Choice Data
True Value

Identification

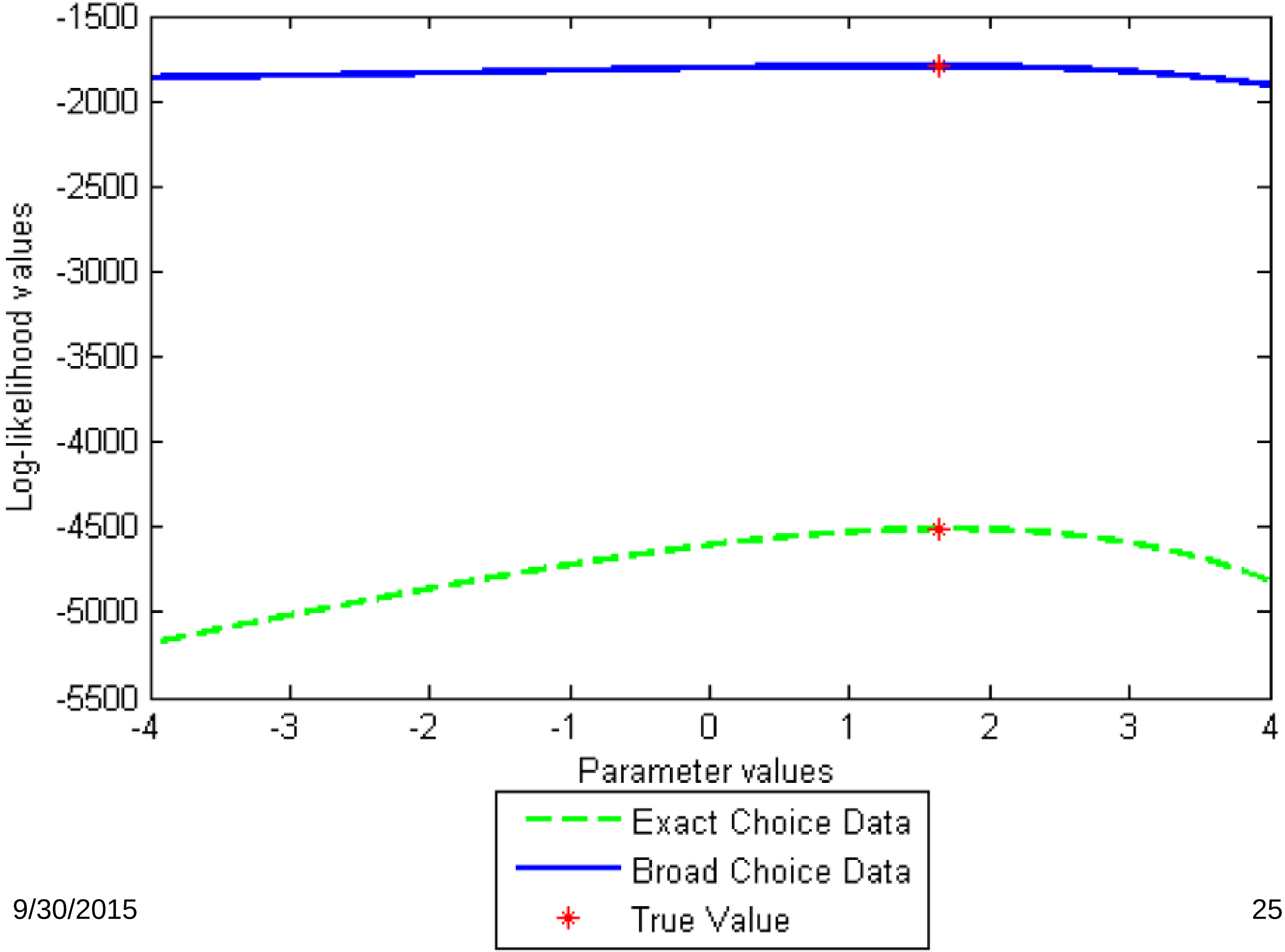
$$\begin{aligned}\mathbb{I}_B(\theta) &= -\mathbb{E}(H_B(\theta)),\\ &= F - IL,\\ &= \mathbb{I}_E(\theta) - IL,\end{aligned}$$

$$IL = \sum_{i=1}^{N}\left(\sum_{m=1}^{M}\tilde{P}_{im}\sum_{c\in C_m}(w_{ic} - \sum_{s\in C_m}P_{is|C_m}w_{is})P_{ic|C_m}(w_{ic} - \sum_{s\in C_m}P_{is|C_m}w_{is})'\right).$$

Note that IL=0 for exact choice data.

Model is locally identified by functional form unless M=1, but weak identification is likely as group size gets large.

Alternative-specific constants cannot be identified except at group level!


-1500
-2000
-2500
-3000
-3500
-4000
-4500
-5000
-5500
Log-likelihood values
-4
-3
-2
-1
0
1
2
3
4
Parameter values
Exact Choice Data
Broad Choice Data
True Value

Multiple Imputations

- Previous work typically assigns average values over the possible vehicles. This introduces measurement error and biases inference
- Multiple Imputations randomly chooses a vehicle and assigns it to household, and then repeats this multiple times. Provides consistent inference only if estimation on each imputed data set is consistent.

$$\hat{\theta} = \sum_{j=1}^{m} \tilde{\theta}_j \Big/ m \qquad \hat{\Sigma} = U + \left(1 + m^{-1}\right) B,$$

where $$B = \sum_{j=1}^{m} \left(\tilde{\theta}_j - \hat{\theta}\right)\left(\tilde{\theta}_j - \hat{\theta}\right)' \Big/ (m-1)$$

$$U = \sum_{j=1}^{m} \tilde{\Omega}_j \Big/ m.$$

$$\left(\theta - \theta^0\right)' \hat{\Sigma}^{-1} \left(\theta - \theta^0\right) \Big/ K$$ is asymptotically distributed $F_{K,\nu}$

$\nu = (m - 1)(1 + r_m^{-1})^2$ and $\qquad r_m = (1 + m^{-1})\,\mathrm{Trace}(BU^{-1})/K$

Hybrid Pairs Logit Choice Model from 2008 NHTS

	Partial Observability		Average		Random Assignment w/ Multiple Imputation (M=30)	
	coeff	std error	coef	std error	coef	std error
(price-fedTax)/income	-5.31	1.88	-4.13	2.32	-2.03	1.97
hp/weight	11.19	39.74	-71.43	48.29	-13.67	21.06
cost per mile	-0.139	0.053	0.107	0.054	0.100	0.054
hybrid	-0.747	0.593	-1.998	0.648	-1.639	0.494
hyb x college	0.546	0.182	0.583	0.181	0.620	0.180
hyb x urban	-0.124	0.224	-0.101	0.223	-0.104	0.223

Vehicle Choice Modeling

- We consider the Berry, Levinsohn and Pakes (BLP) choice model for micro- and macro-level data. This allows use of aggregate market share data to improve identification and estimation.
- Compare the results across three models:
  - a choice model that aggregates to broad groups of choices
  - a choice model that aggregates to broad groups of choices, then places distributional assumptions on the attributes in each aggregated group
  - a choice model that accounts for the presence of broad choice data without aggregation.
- Findings: Aggregation misspecifies the choice model affecting point estimates and seriously understates standard errors.

BLP Estimation issues

- The Berry, Levinsohn and Pakes (BLP) choice model for micro- and macro level data is commonly estimated sequentially
- Standard errors obtained from this approach are inconsistent
- Consistent standard errors for the BLP model for micro- and macro- level data, have not been formally derived.
- We use a Generalized Method of Moments (GMM) framework to derive consistent analytic standard errors
- We find that the inconsistent standard errors from sequential estimation are downward biased.

The BLP Model for Disaggregate Data

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

$$U_{nj} = \delta_j + w_{nj}{}'\beta + \epsilon_{nj},$$

δ_j is a product specific constant that captures the "average" utility of product j

- 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 BLP Model for Disaggregate Data

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

$$P_{nj} = \frac{exp\left(\delta_j + w_{nj}'\beta\right)}{\sum_k exp(\delta_k + w_{nk}'\beta)} \ .$$

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

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

The BLP Model for Disaggregate Data

- The estimates from maximum likelihood estimation of this model match the predicted shares from the model, $\frac{1}{N}\sum_n \hat{P}_{nj}$ to the sample shares, $\frac{1}{N}\sum_n y_{nj}$.
- An innovation of BLP is to match the predicted shares to aggregate market share data, A_j.
- Finally, the product specific constants are a linear combination of product attributes:

$$\delta_j = {x_j}'\alpha_1 + p_j'\alpha_2 + \xi_{1j},$$
$$p_j = {z_j}'\gamma + \xi_{2j}$$
$$where\ E\left(\xi_{1j} \middle| z_j\right) = 0.$$

Sequential Estimation Procedure

- First stage: Iterate between two steps until convergence:
  - Maximum likelihood estimation over β
  - Enforcing the aggregate market share constraint through δ
    - BLP contraction mapping algorithm:

$$\delta_{j,t+1} = \delta_{j,t} + \ln(A_j) - \ln(\widehat{S}_j), \quad \forall\, j = 1, \ldots, J$$

- Second stage: IV estimation:

$$p_j = z_j{}'\gamma + \xi_{2j}$$

$$\widehat{\delta}_j = x_j{}'\alpha_1 + p_j{}'\alpha_2 + \xi_{1j},$$

Estimates from the first stage

BLP Inference

- The IV standard errors for α from the second stage are downward biased because they ignore the uncertainty inherent in $\widehat{\delta_j}$.
- The standard errors for β derived from the Hessian of the log likelihood function are inconsistent because $\widehat{\beta}$ is not a maximum likelihood estimate unless the sample is representative.
- To correct these standard errors, recast the model within a GMM framework.

Estimation Procedure

- The following moments correspond to the sequential process detailed earlier:

$$G_1(\beta, \delta) = \frac{1}{N} \sum\nolimits_n \sum\nolimits_j y_{nj} (w_{nj} - \sum\nolimits_i P_{ni} w_{ni})$$

$$G_2(\beta, \delta) = A_j - \frac{1}{N} \sum\nolimits_n \sum\nolimits_j P_{nj} \, .$$

$$G_3(\delta, \alpha) = \frac{1}{J} \sum\nolimits_j z_j \left(\delta_j - x_j \alpha \right) .$$

- The standard GMM covariance matrix formula is applied

Monte Carlo Study on Standard Errors

Parameter	$N = 2500$		$N = 10000$		$N = 60000$	
	Sequential	GMM	Sequential	GMM	Sequential	GMM
$\widehat{\beta_1}$	0.390	0.907	0.371	0.839	0.382	0.807
$\widehat{\beta_2}$	0.606	0.883	0.672	0.806	0.700	0.805
$\widehat{\alpha_0}$	0.789	0.813	0.791	0.796	0.810	0.810
$\widehat{\alpha_{11}}$	0.747	0.797	0.794	0.806	0.806	0.806
$\widehat{\alpha_{12}}$	0.597	0.858	0.746	0.805	0.781	0.797
$\widehat{\alpha_2}$	0.807	0.809	0.829	0.827	0.802	0.802

Table 1: Coverage probabilities of 80% confidence intervals for β and α.

Empirical Application: Sequential vs. GMM Standard Errors

Variable	BLP with Aggregated Choices			
	Estimated Parameter	Uncorrected Standard Error	Corrected Standard Error	Ratio of Corrected to Uncorrected Standard Errors
(Price) × (75,000<Income<100,000)	0.065	0.004 ***	0.014 ***	3.067
(Price) × (Income>100,000)	0.102	0.004 ***	0.015 ***	3.556
(Price) × (Income Missing)	0.094	0.005 ***	0.015 ***	3.140
Fuel Operating Cost (cents per mile)	-2.877	0.053 ***	0.953 ***	18.064
(Fuel Operating Cost) × (College)	-0.061	0.009 ***	0.020 ***	2.225
Price	-0.116	0.019 ***	0.026 ***	1.368

The effect of price and gallons per mile variables on utility
*Notes: * denotes significance at the 10% level. ** denotes significance at the 5% level. *** denotes significance at the 1% level.*

Aggregation in BLP 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 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

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

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

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

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

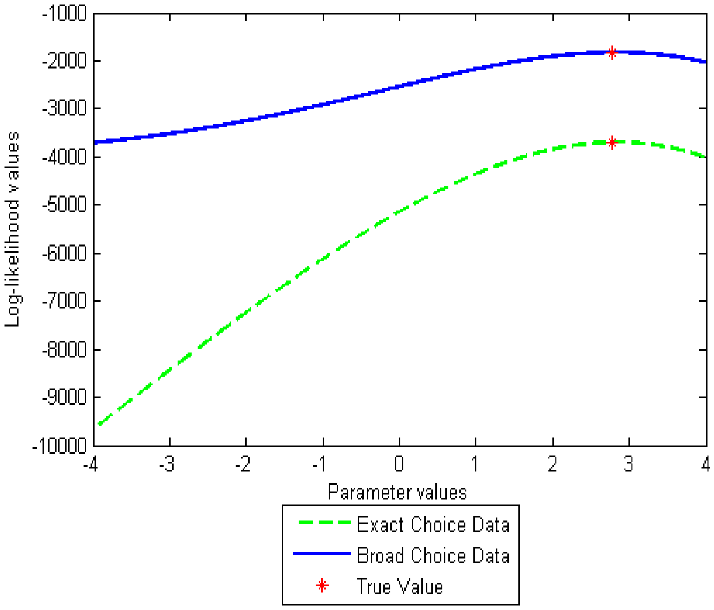
- The intuition for including Ω_{nb} is that community attributes with larger variances should have a greater impact on the probability that the community is selected.
- The $\log(D_b)$ term is a measure of community size. Other conditions being equal, a community with a large number of housing units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



Empirical Application: Choice Set Aggregation

Variable	BLP with Aggregated Choices		BLP with McFadden's Method		BLP for Broad Choice Data	
	Estimated Parameter	Corrected Standard Error	Estimated Parameter	Corrected Standard Error	Estimated Parameter	Corrected Standard Error
(Price) × (75,000<Income<100,000)	0.065	0.014 ***	0.001	0.067	0.038	0.052
(Price) × (Income>100,000)	0.102	0.015 ***	0.004	0.056	0.123	0.100
(Price) × (Income Missing)	0.094	0.015 ***	0.011	0.080	0.079	0.056
Fuel Operating Cost (cents/mile)	-2.877	0.953 ***	-2.946	0.263 ***	-0.599	2.044
(Fuel Operating Cost) × (College)	-0.061	0.020 ***	-0.027	0.466	-0.057	0.076
Price	-0.116	0.026 ***	-0.064	0.120	-0.098	0.097

Modelling vs Ignoring Broad Choice: The effect of price and gallons per mile variables on utility

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

Willingness to pay for fuel efficiency

Willingness to pay for a 1 cent/mile improvement in fuel efficiency (thousands)†	Estimated Parameter	Uncorrected Standard Error	Corrected Standard Error‡	Ratio of Corrected to Uncorrected Std. Errors	Implied Discount Rate
BLP Model with Aggregated Choices	24.695	4.090 ***	10.128 **	2.477	-23.675
BLP Model with McFadden's Method	46.083	14.663 ***	83.105	5.667	-28.132
BLP Model for Broad Choice Data	6.123	0.683 ***	22.706	33.234	-10.785

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 with no college education and income below $75,000 (in thousands of dollars).

‡ calculated using the delta method:

$$Var(williness\ to\ pay) = Var\left(\frac{\beta_{fuelop}}{\alpha_{price}}\right) = \frac{\beta_{fuelop}^2}{\alpha_{price}^4}\sigma_{price}^2 + \frac{1}{\alpha_{price}^2}\sigma_{fuelop}^2 - \frac{2\beta_{fuelop}}{\alpha_{price}^3}\rho_{fuelop,price}\sigma_{price}\sigma_{fuelop},$$

$$\sigma_{price}^2 = var(\alpha_{price}), \sigma_{fuelop}^2 = var\left(\beta_{fuelop}\right), \rho_{fuelop,price} = corr(\beta_{fuelop}, \alpha_{price})$$

Conclusion 1

- The existing evidence on consumer valuation of fuel efficiency is varied and inconclusive. Part of this may be a result of modelling errors because:
  - The use of sequential standard errors understate the uncertainty in estimates
  - Ignoring aggregation understates the uncertainty in parameter estimates

Overall Conclusions

- Measurement errors are first order problems for many applications.
- Modeling the error process leads to nice econometrics and publishable papers, although this usually leads to big confidence regions.
- But no amount of fancy modeling can replace good data – and we need to put more energy into getting better data.