More on Micro Data
BLP (2004) ("Micro BLP")

Phil Haile
Yale University

August 2018

Data Types (discrete choice, review)

1. Market Level Data (e.g., BLP, Nevo): market and product characteristics, prices, shares
2. Micro Data (e.g., Goldberg 95): individual characteristics *matched* to individual consumer choices
3. Panel Data (e.g., Nielsen Homescan): choices of same csr on multiple choice occasions

Hybrids (e.g., Petrin).

“Rank Data”

- consumer ranks 2 or more choices
- examples
    - marketing surveys: “conjoint analysis” (hypothetical)
    - survey of recent buyers, (partially hypothetical) MicroBLP
    - school choice mechanisms (if truthful ranking–otherwise must “invert” strategies to learn true preferences)
- standard discrete choice: we see only first rank
- second, third etc. directly reveal which products are close substitutes, i.e., where consumers would go if their favorite product disappeared or its price rose enough for them to switch away.

Rank Data = Panel Data

Rankings on one choice occasion are equivalent to having panel data with specific variation in the choice set: each rank is discrete choice solution when higher ranks are removed from the choice set.

In fact, an ideal version of panel data because no ambiguity about whether $(\xi_t, \nu_{it}, \epsilon_{it})$ should be viewed as fixed across the "two" choice settings.

Value of Rank Data

- no need for IV to create exogenous changes in choice sets: answer to "what is your second choice" can directly reveal what would have been chosen under a different choice set
- some changes in choice sets more informative than others. *best possible change* is removal of the consumer's first choice
- note: substitution patterns in particular, "diversion ratios"

$$\frac{\partial s_k}{\partial p_j} \Big/ \frac{\partial s_j}{\partial p_j},$$

  are always about 2nd choices.

Rank Data: An Important Limitation

Do we really believe answers to questions about hypothetical choice?

- all else equal, revealed preference clearly more reliable than answers to hypothetical questions
- questions of recent buyers may be more reliable
- strong tradition of conjoint analysis in marketing; economists usually more skeptical; seemingly little/no work validating
- an interesting possible strategy (?): interpret hypothetical choices formally as noisy measures of what actual choices would be.

Micro BLP

Cars again

- demonstrate use of combined market-level data and micro/rank data (similar to Petrin. . .)
  - emphasize the identifying power of "second-choice" data wrt *substitution patterns*
- demonstrate importance of stochastic component of random coefficients (vs. only consumer characteristics)
- counterfactuals: outcomes when add/drop some products.

MicroBLP Data

Market Data Plus Some Rank Data

1. GM survey of 37,000 recent car buyers (1993)
   - what actually bought
   - second choice (hypothetical)
   - HH char
2. CPS: demog of US pop
3. aggregate market shares
4. vehicle char and prices

(2-4 as in BLP95, but for a single year!)

Modal Second Choices

TABLE 5
Examples of Second Choices

Model	n_j	Modal Second Choice	Number Choosing	Next Second Choice	(Modal+ Next)/n	Number of Different Choices
Metro	188	Escort	22	Geo Storm	.22	49
Cavalier	238	Escort	16	LeBaron	.12	59
Escort	166	Tempo	16	Taurus	.18	53
Corolla	250	Civic	42	Camry	.33	55
Sentra	203	Corolla	34	Civic	.31	60
Accord	223	Camry	58	Taurus	.35	61
Taurus	147	Camry	18	Sable	.22	45
Legend	119	Lexus ES300	19	Lexus SC300	.24	40
Seville	243	DeVille	38	Lincoln MK8	.26	49
Lexus LS400	148	DeVille	33	Infiniti Q45	.39	27
Caravan	166	Voyager	31	Aerostar	.32	36
Quest	232	Caravan	50	Villager	.43	31
Grand Cherokee	137	Explorer	75	Blazer	.59	34
Trooper	137	Explorer	43	Rodeo	.41	27
GMC Full-Size Pickup	469	Chevy Full-Size Pickup	222	Ford Full-Size Pickup	.55	29
Toyota Pickup	113	Ford Ranger	29	Nissan Pickup	.43	25
Econovan	90	Chevy Full-Size Van	20	Suburban	.44	23

Model

(no t)

$$
\begin{aligned}
u_{ij} &= \delta_j + \sum_{k,r} x_{jk} z_{ir} \beta^o_{kr} + \sum_k x_{jk} \nu_{ik} \beta^u_k + \epsilon_{ij} \\
\delta_j &= x_j \gamma + \xi_j \\
& \quad x_j \text{ includes price} \\
u_{i0} &= \sum_k z_{ir} \beta^n_{kr} + \nu_{i0} + \epsilon_{i0}
\end{aligned}
$$

Assume ν_i independent of z_i *in population*
For now, no assumption on (x, ξ) distn or on instruments.

Estimation Options (1 of 2)

Estimate $(\beta^o, \beta^u, \delta)$ or $(\beta^o, \beta^u, \gamma)$?

They focus mainly on $(\beta^o, \beta^u, \delta)$.

- limited uses
  - describe heterogeneity
  - could simulate shares with new product j^* with given δ_{j^*}, holding all other δ fixed
  - but no derivatives of demand wrt prices, so no eqm counterfactuals possible
- but no IV needed
- emphasis here really is on learning about the "substitution patterns" (i.e., parameters β^o, β^u) from the first + second choices; estimation of γ would require instruments as usual (or some other "trick") and they don't really have any because they have only one market.

Estimation Options (2 of 2)

MLE (MSL) or GMM (MSM)?

Likelihood exploits all features ("all moments") of the data. But as discussed last week, MSL can be computationally difficult, especially with small choice probabilities.

GMM (MSM)

Attractive here regardless of computational burden of MSL, since easy to combine micro and market level data. Note that score of likelihood could be one set of moments. But because score of likelihood still introduces computational challenges, they use...

Three Sets of Moments

1. aggregate shares
2. correlation between consumer and car characteristics
3. correlation between characteristics of 1st and 2nd choices

(1 + 2 similar to Petrin moments; 3 replaces exogenous variation in choice sets).

Moments (1 of 3)

Aggregate shares:

$$\Pr(y_i^1 = j|\beta, \delta)$$

As in BLP, simulate, drawing ν_i from normal, z_i from CPS (not from micro data, which is nonrepresentative)

Note: they *solve* for $\delta = \delta(\beta)$ that yields *perfect fit* given β (contraction). Estimation is not trading off fit of the aggregate moments vs. fit of the micro moments. This is justified by asymptotics $N/n \longrightarrow \infty$ (N=$U.S.$ pop) even as survey size $n \longrightarrow \infty$. This means that this first "moment" is treated a *constraint*, just as in BLP.

Moments (2 of 3)

Covariance between z_i and $x_{y_i^1}$ (char of first choice)

$$\frac{1}{n}\sum_i x_{y_i^1} \underbrace{\left(z_i - \hat{E}\left[z_i | y_i^1, \beta, \delta\right]\right)}_{\substack{x \text{ includes intercept, so they also} \\ \text{match this part alone}}}$$

$$\begin{aligned}
E\left[z_i | y_i^1, \beta, \delta\right] &= \int_z z dF\left(z | y_i^1, \beta, \delta\right) \\
&= \int_z z \frac{\Pr\left(y_i^1 | z, \beta, \delta\right)}{\Pr\left(y_i^1 | \beta, \delta\right)} dF(z) \quad \text{(Bayes' rule)} \\
\hat{E}\left['' \right] &: \quad \text{use simulation to approx integral and} \\
& \quad\ \text{choice probs in integrand.}
\end{aligned}$$

loosely: pins down β^o

Moments (3 of 3)

Predicted second choice characteristics given first choice characteristics:

$$\frac{1}{n}\sum_i \left(x_{y_i^2} - \hat{E}\left[x_{y_i^2}|y_i^1, \beta, \delta\right]\right)$$

(similar calculations to those above)

loosely: given β^o, this pins down β^u.

Estimation and Inference

Estimate by MSM (similar to BLP), adjust standard errors for simulation error. Details in paper. Note: no instruments. Will not learn the effect of price on anything, but learning a lot about substitution patterns.

TABLE 11
Most Popular Second Choices: A Comparison among Models and to the Data

Vehicle	Full Model	Rank	Logit First	Rank	Logit First and Second	Rank	$\beta^o \equiv 0$	Rank
Metro	Chevy Geo Storm	2	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Tercel	12
Cavalier	Sun Bird	3	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Ford Escort	1
Escort	Tempo	1	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Tempo	1
Corolla	Escort	6	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Civic	1
Sentra	Civic	2	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Civic	2
Accord	Camry	1	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Camry	1
Taurus	Mercury Sable	2	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Accord	4
Legend	Civic	10	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Town Car	≥25
Seville	Deville	1	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	Deville	1
Lexus LS400	MB 300	3	Ford Full-Size Pickup	≥25	Ford Full-Size Pickup	≥25	DeVille 2	1
Caravan	Voyager	1	Ford Full-Size Pickup	≥25	Voyager	1	Voyager	1
Quest	Aerostar	7	Ford Full-Size Pickup	≥25	Caravan	1	Caravan	1
Grand Cherokee	Explorer	1	Chevy Full-Size Pickup	≥25	Chevy Full-Size Pickup	≥25	Explorer	1
Trooper	Explorer	1	Chevy Full-Size Pickup	22	Chevy Full-Size Pickup	22	Rodeo	2
GMC Full-Size Pickup	Chevy Full-Size Pickup	1	Chevy Full-Size Pickup	1	Ford Full-Size Pickup	2	Chevy Full-Size Pickup	1
Toyota Pickup	Ranger	1	Chevy Full-Size Pickup	4	Chevy Full-Size Pickup	4	Ranger	1
Econovan	Chevy Van	1	Ford Full-Size Pickup	6	Ford Full-Size Pickup	6	Chevy Van	1

Estimate gamma?

(recall $\delta_j = x_j \gamma + \xi_j$)
Why?

- need for price elasticities and, therefore, most counterfactual simulations
- why not estimate by assuming $\xi_j \perp x_j$ as usual?
  - too little data: one year, no geographic markets
  - foundation for estimating cross-elasticities from a single cross section is shaky!
- 2 possible alternatives
  - add supply side
  - calibration: choose α so that industry price elasticity is -1 (industry wisdom)

  (use the former in what follows; note: a lot of random coefficients in this specification!).

TABLE 6
ESTIMATES OF INTERACTION TERMS, β^o

Vehicle Characteristic and Household Attribute	Full Model (1)	Logit: First (2)	Logit: First and Second (3)
Price:			
Constant	−2.18	.092	.139
	(.142)	(.0001)	(.0003)
Income × (income <75th percentile)	.714	.299	.344
	(.044)	(.002)	(.001)
Income × (income >75th percentile)	1.17	.466	.603
	(.083)	(.091)	(.007)
Family size	−.565	−.144	−.143
	(.010)	(.001)	(.006)
Minivan: Kids (kids have age ≤16)	1.973	.765	.771
	(.242)	(.098)	(.323)
Pass:			
Adults (adults have age >16)	.203	.018	−.067
	(.095)	(.0004)	(.009)
Family size	.536	−.055	−.006
	(.052)	(.003)	(.0002)
Age (of household head)	.019	.002	.005
	(.003)	(.00001)	(.00001)
HP: Age	−.002	−.010	−.012
	(.001)	(.0004)	(.0001)
Acc:			
Age	.0004	.001	−.002
	(.001)	(.00001)	(.0001)
Age^2	.0001	.000	.000
	(.00001)	(.00001)	(.00001)
PUPayl:			
Age	.0174	−.003	.000
	(.002)	(.0001)	(.00001)
Rural dummy	1.075	.512	.376
	(.179)	(.005)	(.008)
Safe: Age	.013	.015	.016
	(.0006)	(.001)	(.0004)
SUV:			
Age	−.219	−.043	−.043
	(.010)	(.003)	(.004)
Rural dummy	.332	.403	−.016
	(.156)	(.007)	(.002)
Allw: Rural dummy	.278	.142	.734
	(.247)	(.005)	(.246)
Outside good:			
Total income	5.151	−.228	−.305
	(.228)	(.096)	(.063)
Family size	−.007	.532	−.346
	(.002)	(.057)	(.004)
Adults	−.428	.851	1.953
	(.766)	(.112)	(.148)

TABLE 7
ESTIMATES OF INTERACTION TERMS, $\boldsymbol{\beta}^u$

Parameter Name	Full Model (1)	$\boldsymbol{\beta}^o \equiv 0$ (2)
Price	.449	.055
	(.026)	(.004)
HP	.030	.183
	(.016)	(.020)
Pass	2.74	1.444
	(.147)	(.055)
Sport	.002	2.763
	(.0004)	(.068)
Acc	.554	.515
	(.078)	(.055)
Safe	.260	.376
	(.130)	(.093)
MPG	.488	.430
	(.018)	(.017)
Allw	.740	.431
	(.179)	(.049)
Minivan	4.787	6.641
	(.353)	(.113)
SUV	3.076	3.231
	(.292)	(.114)
Van	1.713	6.888
	(.289)	(.266)
PUPayl	2.160	4.301
	(.092)	(.210)
SUVPayl	.356	.015
	(.072)	(.013)
Chrysler	1.689	1.383
	(.058)	(.051)
Ford	.915	1.410
	(.072)	(.051)
GM	1.885	1.844
	(.057)	(.105)
Honda	.329	.086
	(.128)	(.043)
Nissan	.506	1.588
	(.142)	(.071)
Toyota	.169	.576
	(.134)	(.094)
Small Asian*	1.467	2.155
	(.068)	(.022)
European*	.454	1.883
	(.084)	(.034)
Outside good	27.858	10.256
	(1.004)	(.506)

TABLE 9
Price Substitutes for Selected Vehicles: Estimates from the Full Model

Vehicle	Price	Semi-elasticity	Best Substitute	Price	Movers* (%)	Second Best	Price	Movers* (%)	To Outside† (%)
Metro	7.84	−1.77	Tercel	9.70	14.96	Festiva	7.41	10.57	17.96
Cavalier	11.46	−4.08	Escort	11.49	8.62	Tempo	10.78	6.80	6.81
Escort	11.49	−4.02	Tempo	10.78	8.21	Cavalier	11.49	7.29	6.56
Corolla	14.51	−3.92	Civic	14.00	8.08	Escort	11.49	7.91	5.00
Sentra	11.78	−3.79	Civic	14.00	13.36	Escort	11.49	4.70	6.55
Accord	17.25	−3.92	Camry	18.20	8.60	Civic	13.00	4.47	5.06
Taurus	17.65	−3.73	Accord	17.25	6.25	Mercury Sable	18.66	6.09	3.97
Legend	32.42	−3.73	Accord	17.25	3.96	Camry	18.20	3.87	4.38
Seville	43.83	−3.16	DeVille	34.40	10.12	El Dorado	35.74	8.04	5.57
Lexus LS400	51.29	−3.43	Mercedes 300	47.71	7.97	Lincoln Town Car	35.68	6.29	5.87
Caravan	17.56	−3.32	Voyager	17.59	35.11	Aerostar	18.13	10.19	5.20
Quest	20.55	−3.98	Aerostar	18.13	12.50	Caravan	17.56	10.38	5.48
Grand Cherokee	25.84	−3.06	Explorer	24.27	17.60	Cherokee	20.10	9.51	6.38
Trooper	22.78	−3.96	Explorer	24.27	17.53	Grand Cherokee	25.85	8.50	5.42
GMC Full-Size Pickup	16.76	−3.78	Chevy Full-Size Pickup	16.78	43.74	Ford Full-Size Pickup	16.68	13.56	6.03
Toyota Pickup	13.77	−3.34	Ranger	11.74	20.53	Nissan Pickup	11.10	11.93	9.35
Econovan	24.54	−2.86	Chevy Van	25.96	12.90	Dodge Van	23.71	9.73	5.38

* Of those who substitute away from the given good in response to the price change, the fraction who substitute to this good

TABLE 10
PRICE SUBSTITUTES FOR SELECTED VEHICLES: A COMPARISON AMONG MODELS

		Logit		
Vehicle	Full Model	First	First and Second	Sigma Only
Metro	Tercel	Caravan	Ford Full-Size Pickup	Civic
Cavalier	Escort	Caravan	Ford Full-Size Pickup	Escort
Escort	Tempo	Caravan	Ford Full-Size Pickup	Ranger
Corolla	Escort	Caravan	Ford Full-Size Pickup	Civic
Sentra	Civic	Caravan	Ford Full-Size Pickup	Civic
Accord	Camry	Caravan	Ford Full-Size Pickup	Camry
Taurus	Accord	Caravan	Ford Full-Size Pickup	Accord
Legend	Town Car	Caravan	Ford Full-Size Pickup	Town Car
Seville	DeVille	Caravan	Ford Full-Size Pickup	DeVille
Lexus LS400	Mercedes 300	Econovan	Ford Full-Size Pickup	Seville
Caravan	Voyager	Voyager	Voyager	Voyager
Quest	Aerostar	Caravan	Caravan	Aerostar
Grand Cherokee	Explorer	Caravan	Chevy Full-Size Pickup	Explorer
Trooper	Explorer	Caravan	Chevy Full-Size Pickup	Rodeo
GMC Full-Size Pickup	Chevy Full-Size Pickup	Caravan	Chevy Full-Size Pickup	Chevy Full-Size Pickup
Toyota Pickup	Ranger	Caravan	Chevy Full-Size Pickup	Ranger
Econovan	Dodge Van	Caravan	Ford Full-Size Pickup	Dodge Van

Model Evaluation

(summary)

- full model
  - coefficients sensible
  - good at fitting first and second choices
  - β^u critical for matching second choices, precisely estimated
- with no random coefficients:
  - some counterintuitive coefficients
  - poor in predicting second choice characteristics *even when using 1st and 2nd choices*
- with no consumer observables
  - substition patterns (1st - 2nd choices) look pretty good
  - poor in predicting match between consumer char and car char.

Counterfactual Simulations in the Paper

1. Add 2 high end SUVs (toyota, mercedes)
   - observables of ford explorer
   - firm-specific average ξ
   - limitation: don't let prices re-equilibrate (in principle easy, but requires good estimates of price coefficients)
   - like Petrin, no change in other model entry/exit in the CF
2. Kill Oldsmobile (similar to Petrin, but eliminating many cars at once now)

(see the paper for details).

Conclusion

- high value of even limited panel data or rank data as supplement to market level data
- evidence on importance of unobs csr heterogeneity (i.e., random coefficients), even when have rich consumer observables; contradicts Nevo conclusion for cereal (but recall the problems with the evidence supporting his conclusion).