

Estimation of Airline Itinerary Choice Models Using Disaggregate Ticket Data

Virginie Lurkin

Laurie Garrow, Matt Higgins, Jeff Newman, Michael Schyns

Transport and Mobility Laboratory (TRANSP-OR),
École Polytechnique Fédérale de Lausanne (EPFL)

AGIFORS SSP Group Meeting
May 16-18, 2017



Outline - Part 2

Introduction and motivation
Main contributions
Data
Methodology
Construction of choice sets
Price endogeneity
Generalized extreme value (GEV) models
Model results
Brief conclusion

Understanding demand

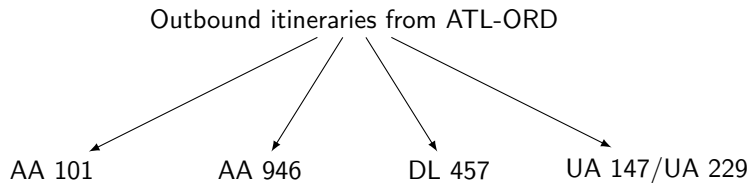
Challenging environment for airlines:

- ▶ Advent of the **Internet**
- ▶ Growth of **Low cost carriers**
- ▶ High fuel costs
- ▶ Terrorism threats, disease outbreaks
- ▶ Financial and economic crisis

→ Increasing interest in using discrete choice models to **model demand as the collection of individuals' decisions**

Itinerary choice model

$$y_{ni} = \begin{cases} 1 & \text{if individual } n \text{ chooses itinerary } i, \\ 0 & \text{otherwise} \end{cases}$$



$$U_i = V_i + \varepsilon_i$$

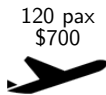
$$V_i = \alpha_i + \beta_1 \text{Cost}_i + \beta_2 \text{Time}_i + \dots$$

$$P_i = \frac{e^{V_i}}{\sum_j e^{V_j}}$$

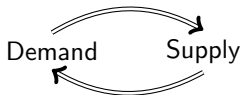
Factors influencing itinerary choice



The fundamental problem

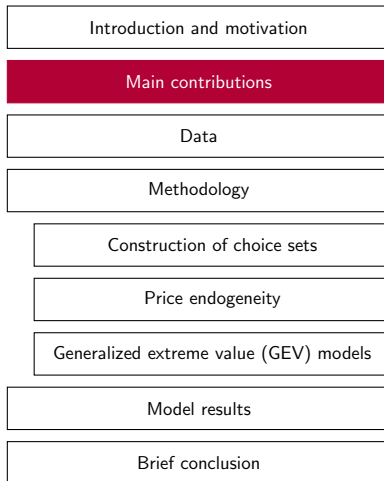


$$\text{demand} = \beta \times \text{price} + \dots + \varepsilon$$



$$\beta = +0.14$$

Outline - Part 2

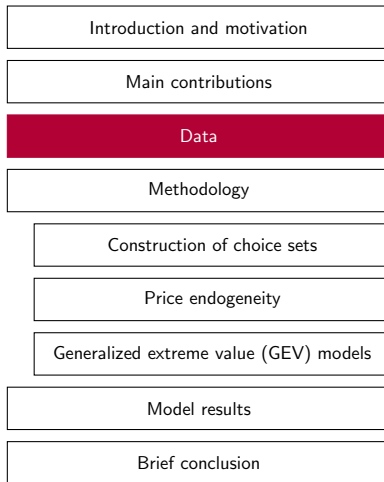


Contributions

Main **contributions**:

- ▶ Estimate a **baseline multinomial logit (MNL)** model that controls for **price endogeneity** for high-yield and low-yield fare products using the control-function method.
- ▶ Estimate **highly redefined time-of-day preferences**
- ▶ Test the **sensitivity of results from our baseline MNL** to different modeling assumptions:
 - Sensitivity to different time-of-day formulations
 - Sensitivity to different price formulations
 - Estimate advanced GEV models (2-level NL, 3-level NL, and OGEV)

Outline - Part 2



Data

- ▶ ARC ticketing data for May 2013 departures
- ▶ Restrict analysis to Continental U.S. markets
- ▶ Include simple one-way and round-trip tickets with at most 2 connections
- ▶ Eliminate tickets with fares $< \$50$ (employee and frequent flyers) or in top 0.1% (charter flights)
- ▶ 3,265,545 directional itineraries, representing 10,034,935 passenger trips

Explanatory variables

Carrier characteristics

- ▶ Carrier preferences
- ▶ **Marketing relationships**

Itinerary characteristics

- ▶ **Price**
- ▶ **Departure time of day**
- ▶ Elapsed time
- ▶ Number of connections
- ▶ Direct flight indicator
- ▶ Equipment type

Marketing relationships



- ▶ **Online** = same marketing and operating carrier on all legs
- ▶ **Codeshare** = same marketing carrier, different operating carriers
- ▶ **Interline** = different marketing carriers, different operating carriers

Price

- ▶ ARC database = **ticket-level price information** linked to **specific itineraries** and the **time of purchase**

“Business” Prices	“Leisure” Prices
Average price for First, Business, and Unrestricted Coach fares	Average price for Restricted Coach and Other fares

- ▶ Average is taken by **origin, destination, carrier and level of service (NS, 1 CNX, 2 CNX)**
- ▶ Assume outbound (or inbound) price = total price/2
- ▶ Include taxes

Departure time of day

1. Departure time preferences vary by:
 - ▶ Length of haul
 - ▶ Direction of travel
 - ▶ Number of time zones
 - ▶ Day of week
 - ▶ Itinerary type (OW, OB, IB)
2. **Continuous time of day preferences formulation** is preferred over discrete formulation to avoid counter-intuitive forecasts

10 time of day classifications

Same time zone, ≤ 600 miles



Same time zone, > 600 miles



1 time zone WB, ≤ 600 miles

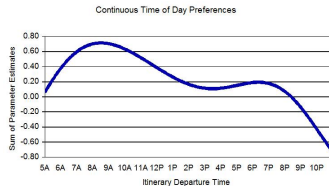
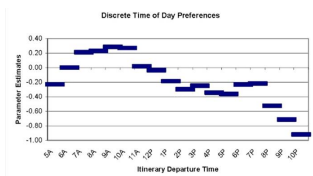


1 time zone WB, > 600 miles



For each classification, we estimate separate time of day preferences for **outbound**, **inbound** and **one-way** itineraries and **day of week**

Detailed departure time of day preferences



$$\beta_{1cmd} \sin\left(\frac{2\pi t}{1440}\right) + \beta_{2cmd} \cos\left(\frac{2\pi t}{1440}\right) + \beta_{3cmd} \sin\left(\frac{4\pi t}{1440}\right) \\ + \beta_{4cmd} \cos\left(\frac{4\pi t}{1440}\right) + \beta_{5cmd} \sin\left(\frac{6\pi t}{1440}\right) + \beta_{6cmd} \cos\left(\frac{6\pi t}{1440}\right)$$

where

c = time of day classification (1,...,10)

m = itinerary type (OW, OB, IB)

d = day of week (1,...,7)

t = departure time in minutes past midnight

1440 = number of minutes in a day

Data representativeness

Carrier	ARC Mkt Share	DB1B Mkt Share
Delta Air Lines (DL)	29.5%	23.4%
United Airlines (UA)	22.9%	17.1%
US Airways (US)	18.4%	10.0%
American Airlines (AA)	17.5%	19.0%
Alaska Airlines (AS)	3.3%	4.2%
JetBlue Airways (B6)	3.2%	3.0%
Frontier Airlines (F9)	2.2%	1.7%
AirTran Airways (FL)	1.4%	2.8%
Virgin America (VX)	1.3%	0.9%
Sun Country Airlines (SY)	0.3%	0.2%
Southwest Airlines (WN)	0.0%	17.7%
Total	100%	100%

- ▶ However, the sampling protocol is exogenous, so there is no need to correct the sample at the estimation stage

Outline - Part 2

Introduction and motivation
Main contributions
Data
Methodology
Construction of choice sets
Price endogeneity
Generalized extreme value (GEV) models
Model results
Brief conclusion

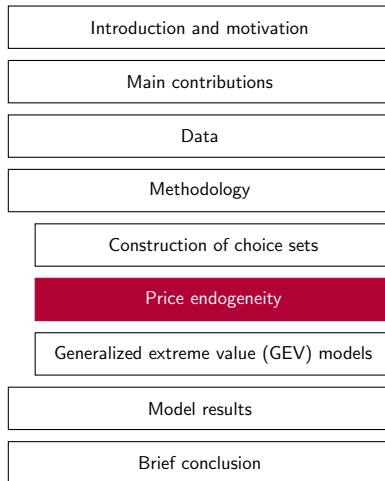
Define choice sets

- ▶ Construct **choice sets** for **each OD city pair that departs on day of week d**
- ▶ Create a **representative weekly schedule** as the Monday after the 9^{th} of the month [May 13 - May 19]
- ▶ Define a unique itinerary by $org_1, dest_1, opcarr_1, opfltnum_1, deptdow_1$ for legs $l = 1, 2, 3$
- ▶ Map all demand to representative schedule/unique itinerary
- ▶ Eliminate choice sets with demand < 30 pax/month



Mapping process is **98%** accurate for all variables and screening rule changes MNL parameter estimates by **4.4%**

Outline - Part 2



Two-stage control-function (2SCF) method

- ▶ Stage 1: Estimate price by ordinary-least-square (OLS)

$$p_{ni} = \alpha_1 IV_{ni}^1 + \dots + \alpha_k IV_{ni}^k + \gamma_i' \mathbf{x}_{ni} + \mu_{ni}$$

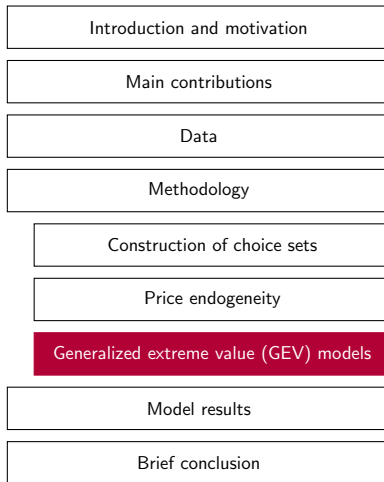
- ▶ Stage 2: Estimate the choice model using the residuals δ from first stage

$$U_{ni} = \beta_{\hat{\delta}} \hat{\delta}_{ni} + \beta_p p_{ni} + \beta_i' \mathbf{x}_{ni} + \varepsilon_{ni}$$

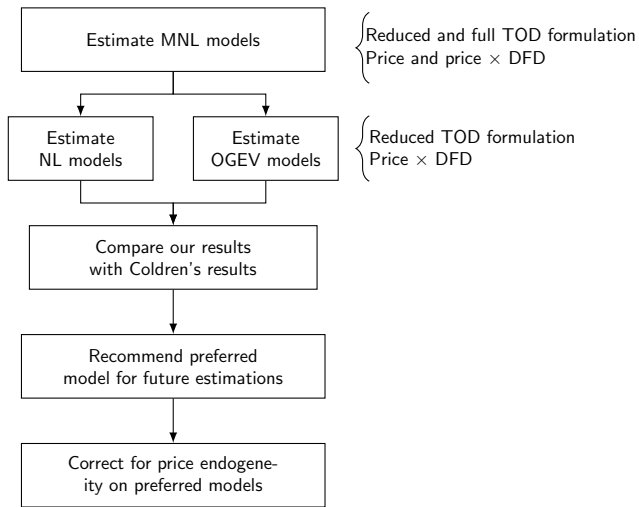
- Test: Estimate the choice model using the residuals δ from first stage and one instrument

$$U_{ni} = \beta_{\hat{\delta}} \hat{\delta}_{ni} + \beta_p p_{ni} + \beta_i' \mathbf{x}_{ni} + \varepsilon_{ni} + \alpha_1 IV_{ni}^1$$

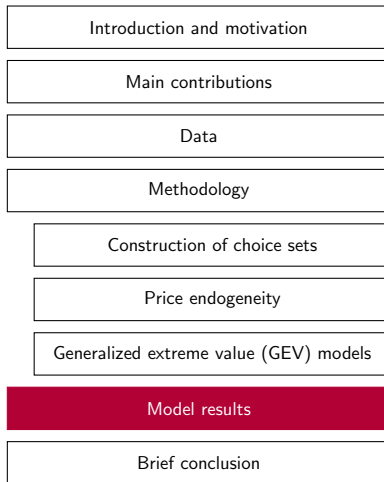
Outline - Part 2



Overview of modeling approach



Outline - Part 2



Departure time preferences

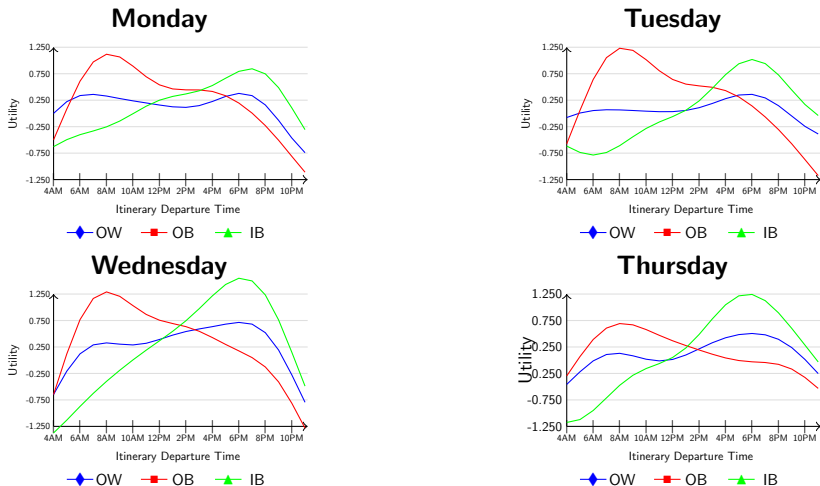


Figure: One TZ WB, distances ≤ 600 mi.

Departure time preferences

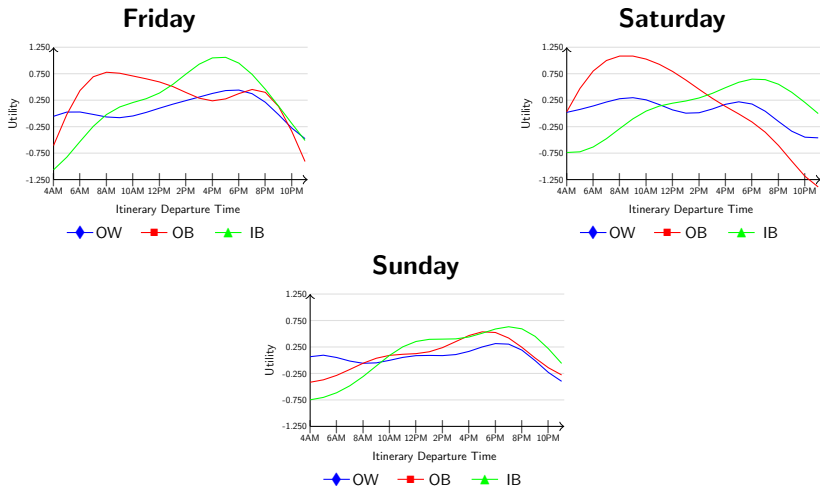


Figure: One TZ WB, distances ≤ 600 mi.

Model results - MNL model

Variable	Uncorrected model		Corrected model	
	Parameter	t-stat	Parameter	t-stat
Average high yield fare (\$)	-0.0025	-258	-0.0036	-265
Average low yield fare (\$)	-0.0049	-383	-0.0069	-327
Elapsed time (min)	-0.0053	-503	-0.0050	-455
Number of connections	-2.4892	-1,194	-2.5582	-1,179
Direct flight	-2.2624	-375	-2.3311	-384
Regional jet or propeller (ref.)	0	-	0	-
Wide- or narrow-body	0.4150	384	0.3889	353
Online (ref.)	0	-	0	-
Codeshare	0.2742	208	0.2825	214
Interline	-0.2342	-35.4	-0.1297	-19.4
δ (residuals)	-	-	0.0020	118.60
LL(0)	-32,652,846.05		-32,652,846.05	
Final LL	-26,239,664.32		-26,232,323.64	
Adj. ρ^2	0.1964		0.1966	

Note: TOD and carrier constants are not reported in this table.

Model results - MNL model

Variable	Uncorrected model		Corrected model	
	Parameter	t-stat	Parameter	t-stat
Average high yield fare (\$)	-0.0025	-258	-0.0036	-265
Average low yield fare (\$)	-0.0049	-383	-0.0069	-327
Elapsed time (min)	-0.0053	-503	-0.0050	-455
Number of connections	-2.4892	-1,194	-2.5582	-1,179
Direct flight	-2.2624	-375	-2.3311	-384
Regional jet or propeller (ref.)	0	-	0	-
Wide- or narrow-body	0.4150	384	0.3889	353
Online (ref.)	0	-	0	-
Codeshare	0.2742	208	0.2825	214
Interline	-0.2342	-35.4	-0.1297	-19.4
δ (residuals)	-	-	0.0020	118.60
LL(0)	-32,652,846.05		-32,652,846.05	
Final LL	-26,239,664.32		-26,232,323.64	
Adj. ρ^2	0.1964		0.1966	

Note: TOD and carrier constants are not reported in this table.

Value of time- MNL model

Value of Time	Base Model	Control Function
High-yield (\$/hr)	126	83
Low-yield (\$/hr)	65	43

Table: Value of time results

Elasticities - MNL model

Low Yield

Segment	Mean fare	Uncorrected	Corrected
Same TZ, distance \leq 600 mi.	221.42	-0.8251	-1.1551
Same TZ, distance $>$ 600 mi.	207.96	-0.7732	-1.0826
One TZ Westbound, distance \leq 600 mi.	221.32	-0.8134	-1.1387
One TZ Westbound, distance $>$ 600 mi.	248.76	-0.9413	-1.3180
One TZ Eastbound, distance \leq 600 mi.	219.15	-0.8260	-1.1564
One TZ Eastbound, distance $>$ 600 mi.	251.80	-0.9567	-1.3396
Two TZ Westbound	265.62	-0.9995	-1.3992
Two TZ Eastbound	263.82	-0.9975	-1.3963
Three TZ Westbound	289.58	-1.1161	-1.5618
Three TZ Eastbound	290.41	-1.1586	-1.6218
Average for All Segments	240.20	-0.8976	-1.2567

Elasticities - MNL model

Low Yield

Segment	Mean fare	Uncorrected	Corrected
Same TZ, distance \leq 600 mi.	221.42	-0.8251	-1.1551
Same TZ, distance $>$ 600 mi.	207.96	-0.7732	-1.0826
One TZ Westbound, distance \leq 600 mi.	221.32	-0.8134	-1.1387
One TZ Westbound, distance $>$ 600 mi.	248.76	-0.9413	-1.3180
One TZ Eastbound, distance \leq 600 mi.	219.15	-0.8260	-1.1564
One TZ Eastbound, distance $>$ 600 mi.	251.80	-0.9567	-1.3396
Two TZ Westbound	265.62	-0.9995	-1.3992
Two TZ Eastbound	263.82	-0.9975	-1.3963
Three TZ Westbound	289.58	-1.1161	-1.5618
Three TZ Eastbound	290.41	-1.1586	-1.6218
Average for All Segments	240.20	-0.8976	-1.2567

Elasticities - MNL model

High Yield

Segment	Mean fare	Uncorrected	Corrected
Same TZ, distance \leq 600 mi.	290.73	-0.5281	-0.7459
Same TZ, distance $>$ 600 mi.	293.41	-0.4949	-0.6936
One TZ Westbound, distance \leq 600 mi.	320.53	-0.5757	-0.8111
One TZ Westbound, distance $>$ 600 mi.	329.03	-0.5735	-0.8051
One TZ Eastbound, distance \leq 600 mi.	315.33	-0.5769	-0.8112
One TZ Eastbound, distance $>$ 600 mi.	344.59	-0.5929	-0.8330
Two TZ Westbound	364.98	-0.6243	-0.8828
Two TZ Eastbound	347.66	-0.5773	-0.8136
Three TZ Westbound	503.74	-0.8416	-1.2106
Three TZ Eastbound	498.00	-0.8646	-1.2423
Average for all Segments	343.60	-0.5877	-0.8307

Results for MNL model that includes price \times DFD

Variable	Uncorrected model		Corrected model	
	Parameter	t-stat	Parameter	t-stat
Average low-yield fare (\$)				
0-6 days from departure	-0.0044	-310	-0.0067	-287
7-20 days from departure	-0.0057	-277	-0.0080	-289
21+ days from departure	-0.0071	-326	-0.0094	-329
Average high-yield fare (\$)				
0-20 days from departure	-0.0020	-174	-0.0035	-208
21+ days from departure	-0.0045	-223	-0.0058	-255
δ (residuals)	-	-	0.0024	126
LL(0)	-32,652,846.05		-32,652,846.05	
Final LL	-26,176,476.72		-26,168,218.55	
Adj. ρ^2	0.1983		0.1986	

Note: not all coefficient estimates are reported in this table.

Results for MNL model that includes price \times DFD

Variable	Uncorrected model		Corrected model	
	Parameter	t-stat	Parameter	t-stat
Average low-yield fare (\$)				
0-6 days from departure	-0.0044	-310	-0.0067	-287
7-20 days from departure	-0.0057	-277	-0.0080	-289
21+ days from departure	-0.0071	-326	-0.0094	-329
Average high-yield fare (\$)				
0-20 days from departure	-0.0020	-174	-0.0035	-208
21+ days from departure	-0.0045	-223	-0.0058	-255
δ (residuals)	-	-	0.0024	126
LL(0)	-32,652,846.05		-32,652,846.05	
Final LL	-26,176,476.72		-26,168,218.55	
Adj. ρ^2	0.1983		0.1986	

Note: not all coefficient estimates are reported in this table.

Stability of MNL Model Results (1)

	6 TOD price	1260 TOD price	6 TOD price x DFD	1260 TOD price x DFD
Average low-yield fare (\$)	-0.006872	-0.006925		
0-6 days from departure			-0.006671	-0.006734
7-20 days from departure			-0.007894	-0.007997
21+ days from departure			-0.009263	-0.009438
Average high-yield fare (\$)	-0.003591	-0.003605		
0-20 days from departure			-0.003418	-0.003490
21+ days from departure			-0.005706	-0.005763
Elapsed time (min)	-0.004762	-0.005004	-0.004789	-0.005052
Nonstop (ref)	0.0000	0.0000	0.0000	0.0000
Direct flight	-2.3500	-2.3311	-2.5020	-2.4860
# connections	-2.5837	-2.5582	-2.7080	-2.6820
WB or NB (ref)	0.0000	0.0000	0.0000	0.0000
Regional jet or prop	0.3923	0.3889	0.3965	0.3925
Online (ref.)	0.0000	0.0000	0.0000	0.0000
Codeshare	0.2928	0.2825	0.3000	0.2885
Interline	-0.1257	-0.1297	-0.1317	-0.1341

Note: not all coefficient estimates are reported in this table.

MNL Results - Value of time (VOT)

	6 TOD price	1260 TOD price	6 TOD price x DFD	1260 TOD price x DFD
Value of time (\$/hr)				
0-6 DFD (high/low)			84 / 43	87 / 45
7-20 DFD (high/low)	80 / 42	83 / 43	84 / 36	87 / 38
21+ DFD (high/low)			50 / 31	53 / 32

MNL Results - Value of time (VOT)

	6 TOD price	1260 TOD price	6 TOD price x DFD	1260 TOD price x DFD
Value of time (\$/hr)				
0-6 DFD (high/low)			84 / 43	87 / 45
7-20 DFD (high/low)	80 / 42	83 / 43	84 / 36	87 / 38
21+ DFD (high/low)			50 / 31	53 / 32

NL models

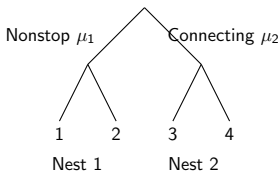


Figure: Two-level NL model

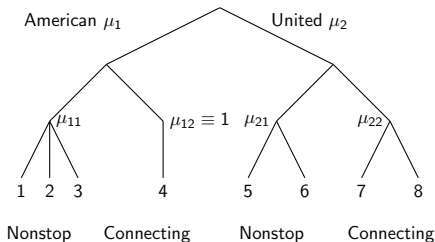


Figure: Three-level NL model

NL results and comparison with Coldren's models

► 2-levels NL Models

	MNL	Carrier	LOS	TOD
Our Models				
$\mu_{carrier}$		0.9612		
μ_{LOS}			0.9706	
μ_{TOD}				0.7869
LL Conv.	-26,699,555	-26,697,995	-26,699,163	-26,671,137
Adj. ρ^2	0.1823	0.1824	0.1823	0.1832
Coldren's				
$\mu_{carrier}$		0.8498		
μ_{LOS}				
μ_{TOD}				0.8167
Adj. ρ^2	0.2830	0.2837		0.2833

Note: LL Zero = -32,652,846 ; LL Conv = Log likelihood at convergence; LOS=level of service; TOD= time of day (nesting by three departure time periods: 12 midnight - 9:59 AM; 10 AM - 3:59 PM; 4 PM - 11:59 PM); No results = not theoretically valid. All logsum parameters are significant at the 0.01 level.

NL results and comparison with Coldren's models

► 3-levels NL Models

	TOD LOS	TOD Carrier	LOS Carrier	Carrier LOS	Carrier TOD	LOS TOD
Our Models						
$\mu_{carrier}$		0.7535	0.9730	0.9262		
μ_{LOS}	0.7666		0.9691	0.9765		
μ_{TOD}	0.8416	0.8499				
LL Conv.	-26,667,493	-26,658,469	-26,699,152	-26,697,109		
Adj. ρ^2	0.1833	0.1836	0.1823	0.1824		
Coldren's						
$\mu_{carrier}$		0.7370				
μ_{LOS}	0.8080					
μ_{TOD}	0.8248	0.9375				
Adj. ρ^2	0.2833	0.2848				

Note: LL Zero = -32,652,846 ; LL Conv = Log likelihood at convergence; LOS=level of service; TOD= time of day (nesting by three departure time periods: 12 midnight - 9:59 AM; 10 AM - 3:59 PM; 4 PM - 11:59 PM); No results = not theoretically valid. All logsum parameters are significant at the 0.01 level.

NL results and comparison with Coldren's models

► 3-levels NL Models

	TOD LOS	TOD Carrier	LOS Carrier	Carrier LOS	Carrier TOD	LOS TOD
Our Models						
$\mu_{carrier}$		0.7535	0.9730	0.9262		
μ_{LOS}	0.7666		0.9691	0.9765		
μ_{TOD}	0.8416	0.8499				
LL Conv.	-26,667,493	-26,658,469	-26,699,152	-26,697,109		
Adj. ρ^2	0.1833	0.1836	0.1823	0.1824		
Coldren's						
$\mu_{carrier}$		0.7370				
μ_{LOS}	0.8080					
μ_{TOD}	0.8248	0.9375				
Adj. ρ^2	0.2833	0.2848				

Note: LL Zero = -32,652,846 ; LL Conv = Log likelihood at convergence; LOS=level of service; TOD= time of day (nesting by three departure time periods: 12 midnight - 9:59 AM; 10 AM - 3:59 PM; 4 PM - 11:59 PM); No results = not theoretically valid. All logsum parameters are significant at the 0.01 level.

OGEV models

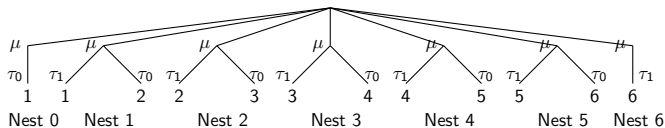


Figure: Ordered GEV model with one adjacent time period

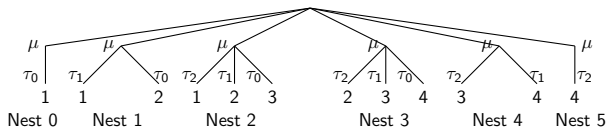


Figure: Ordered GEV model with two adjacent time periods

OGEV results

	Two Alloc Hourly	Two Alloc Two Hour	Two Alloc Three Hour	Three Alloc Hourly	Three Alloc Two Hour
Our Models					
μ_{OGEV}	0.7725	0.7386		0.7249	0.6916
τ_0	0.5	0.5		0.3333	0.3333
τ_1	0.5	0.5		0.3333	0.3333
τ_2				0.3333	0.3333
LL Conv	-26,647,373	-26,652,600		-26,641,896	-26,654,913
Adj. ρ^2	0.1839	0.1838		0.1841	0.1837
Coldren's					
μ_{OGEV}			0.7932		0.7586
τ_0			0.7785		0.6752
μ_1			0.2215		0.0728
μ_2					0.2520
Adj. ρ^2			0.2834		0.2836

Note: LL Zero = -32,652,846; Coldren's Two Allocation model includes six nests (5-6:59 AM; 7-9:59 AM; 10 AM -12:59 PM; 1-3:59 PM; 4-6:59 PM; 7-10:59 PM). Coldren's Three Allocation model includes six nests (5-6:59 AM; 7-8:59 AM; 9-10:59 AM; 11 AM-12:59 PM; 1-2:59PM, 3-4:59PM, 5-6:59 PM and 7-10:59PM).

OGEV results

	Two Alloc Hourly	Two Alloc Two Hour	Two Alloc Three Hour	Three Alloc Hourly	Three Alloc Two Hour
Our Models					
μ_{OGEV}	0.7725	0.7386		0.7249	0.6916
τ_0	0.5	0.5		0.3333	0.3333
τ_1	0.5	0.5		0.3333	0.3333
τ_2				0.3333	0.3333
LL Conv	-26,647,373	-26,652,600		-26,641,896	-26,654,913
Adj. ρ^2	0.1839	0.1838		0.1841	0.1837
Coldren's					
μ_{OGEV}			0.7932		0.7586
τ_0			0.7785		0.6752
μ_1			0.2215		0.0728
μ_2					0.2520
Adj. ρ^2			0.2834		0.2836

Note: LL Zero = -32,652,846; Coldren's Two Allocation model includes six nests (5-6:59 AM; 7-9:59 AM; 10 AM -12:59 PM; 1-3:59 PM; 4-6:59 PM; 7-10:59 PM). Coldren's Three Allocation model includes six nests (5-6:59 AM; 7-8:59 AM; 9-10:59 AM; 11 AM-12:59 PM; 1-2:59PM, 3-4:59PM, 5-6:59 PM and 7-10:59PM).

Recommended models

- ▶ Three-level nested logit by time and carrier
- ▶ OGEV with three allocations - hourly (constrained version)

Best models results

	MNL 6 TOD, price × DFD		NL, TOD and Carrier 6 TOD, price × DFD		OGEV, Three Alloc. hourly 6 TOD, price × DFD	
	<i>uncorrected</i>	<i>corrected</i>	<i>uncorrected</i>	<i>corrected</i>	<i>uncorrected</i>	<i>corrected</i>
VOT LY						
0-6 DFD	\$73	\$43	\$73	\$42	\$71	\$42
7-20 DFD	\$57	\$36	\$56	\$35	\$55	\$35
21+ DFD	\$46	\$31	\$45	\$30	\$44	\$30
VOT HY						
0-20 DFD	\$161	\$84	\$157	\$81	\$157	\$84
21+ DFD	\$72	\$50	\$71	\$49	\$73	\$51

Conclusions

- ▶ Importance to correct for price endogeneity
 - ▶ Over-estimation of customer's value of time
 - ▶ Biased price elasticities
 - ▶ Sub-optimal business decisions
- ▶ Highly refined departure time of day preferences
- ▶ Strong correlation across itineraries that share similar departure times

Directions for future research

- ▶ Generate choice sets as a function of consumer characteristics and search behaviors
- ▶ Expand itinerary choice models to include consumer characteristics (BLP approach)
- ▶ Use these models to assess the effects of industry consolidation on firms and consumers

Directions for future research

- ▶ Generate choice sets as a function of consumer characteristics and search behaviors
- ▶ Expand itinerary choice models to include consumer characteristics (BLP approach)
- ▶ Use these models to assess the effects of industry consolidation on firms and consumers

References

- ▶ Lurkin, V., Garrow, L. A., Higgins, M., Newman, J. P, and Schyns, M. (2017). Accounting for Price Endogeneity in Airline Itinerary Choice Models: An Application to Continental U.S. Markets, *Transportation Research Part A: Policy and Practice* 100:228-246.
- ▶ Lurkin, V., Garrow, L. A., Higgins, M., Newman, J. P, and Schyns, M. (2017). Modeling Competition among Airline Itineraries , *Under review in Transportation Research Part A: Policy and Practice*.

THANK YOU !