Forecasting Repeat Buying for New Products and Services

Peter S. Fader
University of Pennsylvania
www.peterfader.com

Bruce G. S. Hardie
London Business School
www.brucehardie.com

15th Annual Advanced Research Techniques Forum
June 13–16, 2004

Motivation

Given a customer-level transaction database, what level of sales should we expect in future periods, both collectively and individually?

- Forecasting the sales of a new product given test market/rollout panel data
- Identifying "active" customers from a list of past customers
- Serving as a key input to a CLV exercise

Philosophy of Model-Building

- Keep it as simple as possible
- Minimize cost of implementation
  - Use of readily available software (e.g., Excel)
  - Use of data summaries

Outline

Part 1: Forecasting Aggregate Repeat-Buying

Part 2: Forecasting Individual-Level Repeat-Buying

Part 3: Links to the Broader Literature

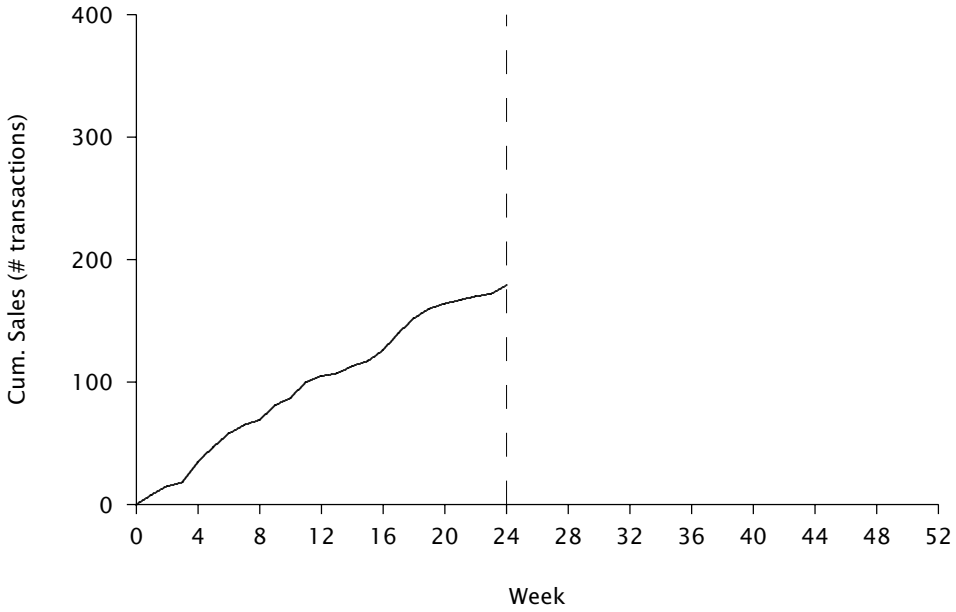
Part 1
Forecasting Aggregate Repeat-Buying

Setting

Ace Drinks has developed a new shelf-stable juice product, aimed primarily at children, called Kiwi Bubbles. Before deciding whether or not to "go national" with the new product, the marketing manager for Kiwi Bubbles has decided to commission a year-long test using IRI's BehaviorScan® service, with a view to getting a clearer picture of the product's potential.

The product has now been under test for 24 weeks. The marketing manager would like a forecast of the product's year-end performance in the test market.

Initial Test-Market Sales

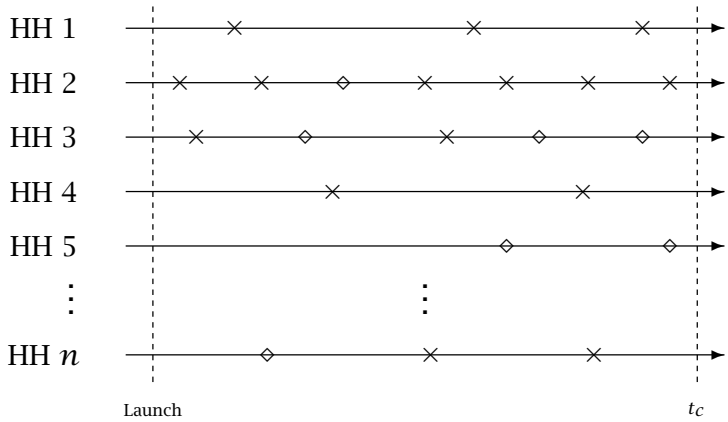


Setting

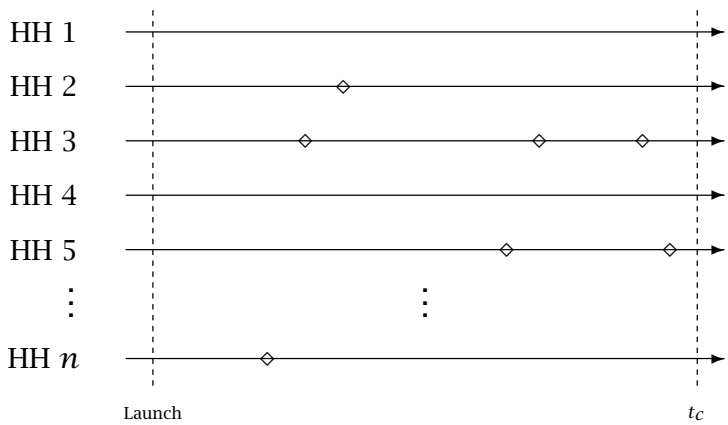
As part of the BehaviorScan® test, we have been monitoring the purchasing behavior of a consumer panel comprising 1499 households.

We summarize these data by determining the cumulative number of panelists that have made a trial, first repeat, second repeat, and so on, purchase by the end of each week.

Purchase Histories: Category



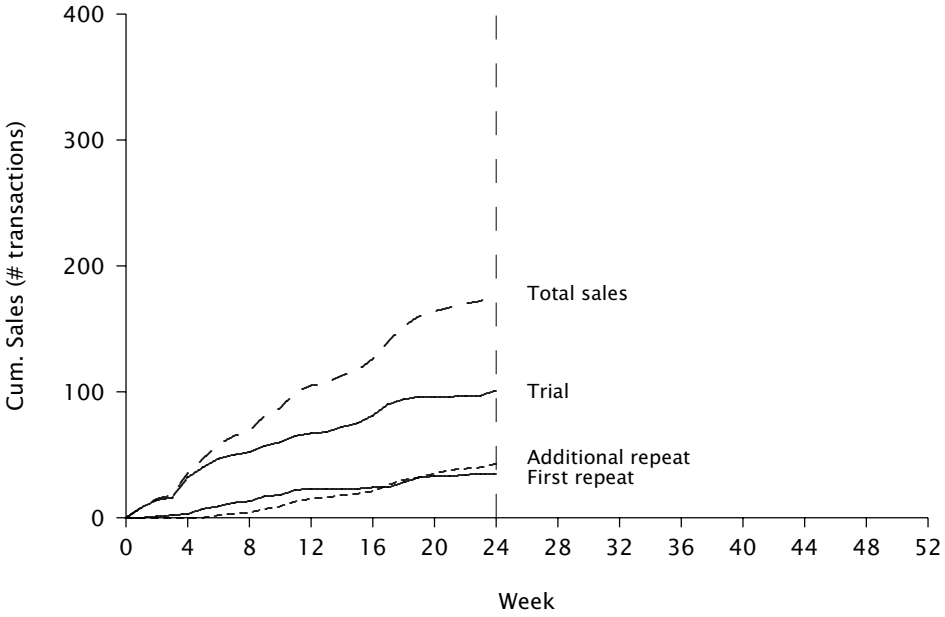
Purchase Histories: New Product Only



Summary of Panel Purchasing

Week	Trial	1st Rpt	2nd Rpt	3rd Rpt	4th Rpt	5th Rpt	6th Rpt	7th Rpt
1	8	0	0	0	0	0	0	0
2	14	1	0	0	0	0	0	0
3	16	2	0	0	0	0	0	0
4	32	3	0	0	0	0	0	0
5	40	7	0	0	0	0	0	0
6	47	9	2	0	0	0	0	0
7	50	12	2	1	0	0	0	0
8	52	13	2	1	1	0	0	0
9	57	17	4	2	1	0	0	0
10	60	18	6	2	1	0	0	0
11	65	22	8	4	1	0	0	0
12	67	23	9	4	1	1	0	0
13	68	23	9	5	1	1	0	0
14	72	23	10	5	2	1	0	0
15	75	23	11	5	2	1	0	0
16	81	24	13	5	2	1	0	0
17	90	24	15	7	2	1	1	0
18	94	28	15	9	4	1	1	0
19	96	32	16	9	5	1	1	0
20	96	33	18	9	5	2	1	0
21	96	33	18	11	5	2	2	0
22	97	34	18	11	5	2	2	1
23	97	35	18	11	5	2	2	2
24	101	35	20	12	5	2	2	2

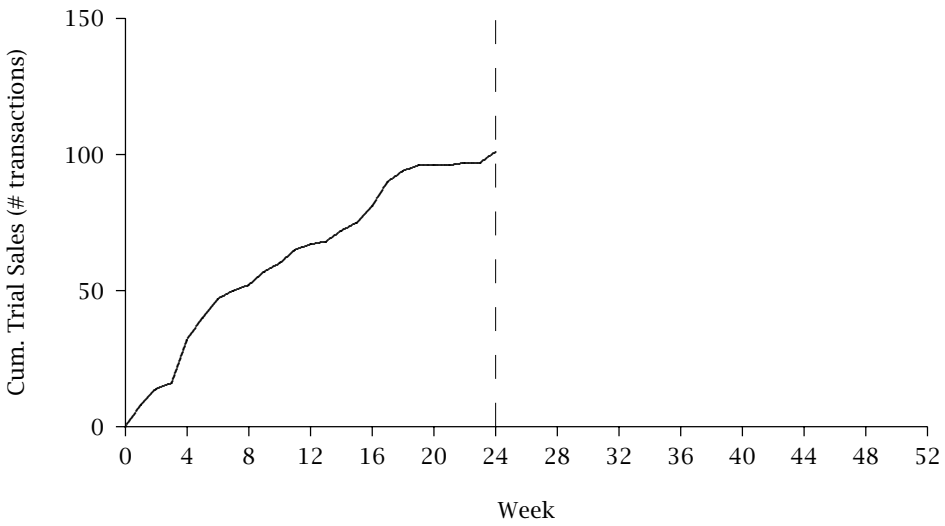
Initial Test-Market Sales



Modelling Objective

Using the purchasing data for the 101 households who had tried Kiwi Bubbles by the end of week 24, we wish to forecast the purchasing behavior of the *whole panel* (i.e., 1499 households) to the end of the year (week 52).

Kiwi Bubbles Trial



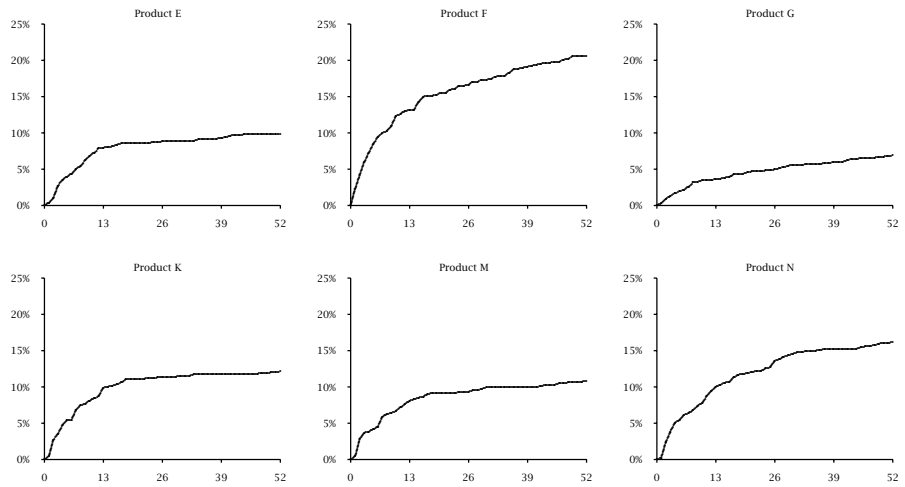
Modelling Trial Sales

We model the cumulative number of triers by time t, $T(t)$, by developing an expression for the probability that a randomly-chosen individual has made a trial purchase by time t.

For a market comprising N individuals (i.e., the size of the panel), we have

$$T(t) = N \cdot P(\text{trial by } t)$$

Illustrative Cumulative Trial Curves



Characterizing P(trial by t)

Cumulative penetration curves (from controlled test-markets) tend to

- increase at a decreasing rate
- towards a penetration limit $< 100\%$

A mathematical expression that captures this is

$$P(\text{trial by } t) = p_0(1 - e^{-\theta_T t})$$

Formal Derivation: Individual-Level Model

- Let T denote the random variable of interest, and t denote a particular realization.
- Assume time-to-trial is distributed exponentially.
- The probability that an individual has tried by time t is given by:

$$F(t) = P(T \leq t) = 1 - e^{-\lambda_T t}$$

- λ_T represents the individual's trial rate.

Formal Derivation: Market-Level Model

Assume two segments of consumers:

Segment	Description	Size	λ_T
1	ever triers	p_0	θ_T
2	never triers	$1 - p_0$	0

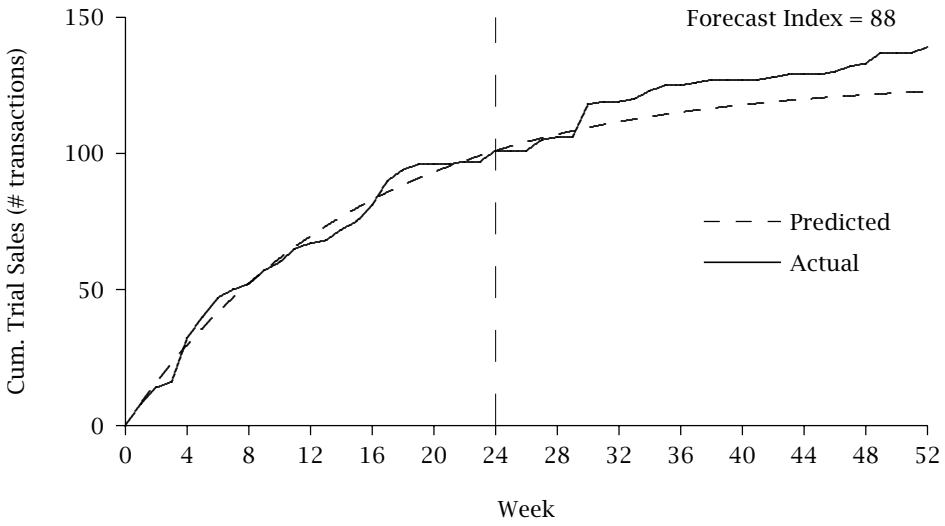
$$
\begin{aligned}
P(\text{trial by } t) = \; & P(T \le t | \text{ever trier}) \times P(\text{ever trier}) + \\
& P(T \le t | \text{never trier}) \times P(\text{never trier}) \\
= \; & p_0 F(t | \lambda_T = \theta_T) + (1 - p_0) F(t | \lambda_T = 0) \\
= \; & p_0 (1 - e^{-\theta_T t})
\end{aligned}
$$

→ the "exponential w/ never triers" model

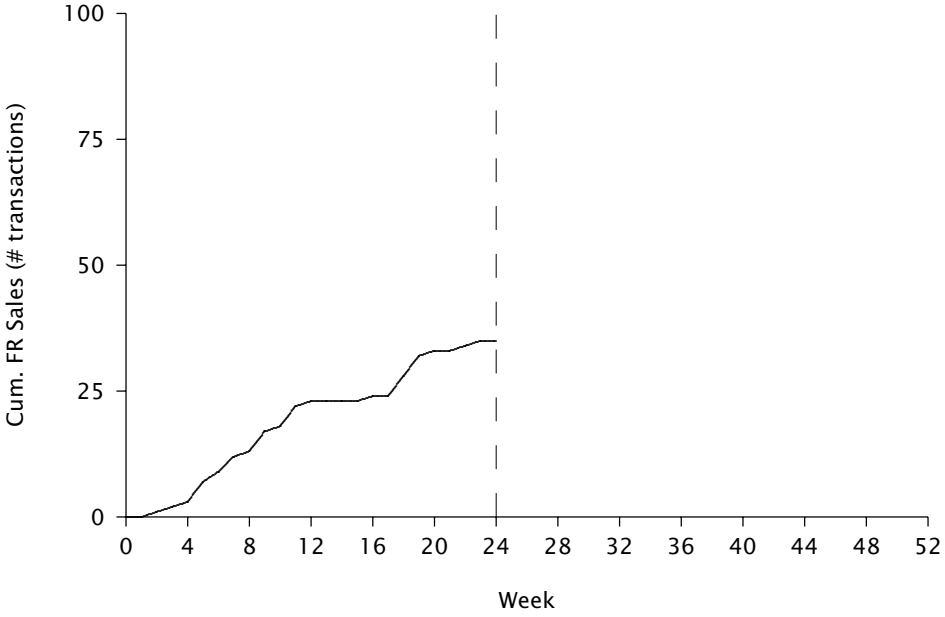
Trial

	A	B	C	D	E
1	p_0	0.0862			
2	theta_T	0.0643			
3	SSE	286.7			
4					
5	Week	P(trial by t)	T(t)	Cum Trl	
6	1	0.0054	8.0	8	0.0
7	2	0.0104	15.6	14	2.5
8	3	0.0151	22.7	16	44.4
9	4	0.0195	29.3	40	7.3
10	5	0.0237	35.5	40	20.1
11	6	0.0276	41.4	47	31.9
12	7	0.0312	46.8	50	10.1
13	8	0.0347	52.0	52	0.0
14	9	0.0379	56.8	57	0.1
15	10	0.0409	61.3	60	1.6
16	11	0.0437	65.5	65	0.3
17	12	0.0463	69.5	67	6.1
18	13	0.0488	73.2	68	26.9
19	14	0.0512	76.7	72	21.9
20	15	0.0533	80.0	75	24.5
21	16	0.0554	83.0	81	4.1
22	17	0.0573	85.9	90	16.9
23	18	0.0591	88.6	94	29.3
24	19	0.0608	91.1	96	23.8
25	20	0.0624	93.5	96	6.3
26	21	0.0639	95.7	96	0.1
27	22	0.0652	97.8	97	0.6
28	23	0.0665	99.8	97	7.6
29	24	0.0678	101.6	101	0.4

Cumulative Trial Forecast



Cumulative First Repeat Sales



Modelling First Repeat

How can an individual have made a first repeat purchase of the new product by the end of week 4?

- she could have made a trial purchase in week 1 and made a second purchase (i.e., her first repeat purchase) somewhere in the intervening three weeks,
- she could have made a trial purchase in week 2 and a second purchase sometime in the following two weeks, or
- she could have made a trial purchase in week 3 and her first repeat purchase sometime in the following week.

	A	B	C	D	E	F	G	H	I	J	K	L	Y	Z
1	**Cumulative First Repeat by Week of Trial**													
2														
3			Week											
4	Trial Week	#HHs	1	2	3	4	5	6	7	8	9	10	23	24
5	1	8	0	1	2	2	3	3	4	4	4	4	5	5
6	2	6		0	0	1	1	1	2	2	2	2	4	4
7	3	2			0	0	1	1	1	1	1	1	1	1
8	4	16				0	2	3	4	5	7	8	11	11
9	5	8					0	1	1	1	2	2	3	3
10	6	7						0	0	0	0	0	1	1
11	7	3							0	0	0	0	0	0
22	18	4											2	2
23	19	2											0	0
24	20	0											0	0
25	21	0											0	0
26	22	1											0	0
27	23	0											0	0
28	24	4												0
29														
30	Cum First Repeat		0	1	2	3	7	9	12	13	17	18	35	35
31	First Repeat		0	1	1	1	4	2	3	1	4	1	1	0

	A	B	C	D	E	F	G	H	I	J	K	L	Y	Z
1	**Cum. First Repeat (as % of Trial) by Week of Trial**													
2														
3			Week											
4	Trial Week	#HHs	1	2	3	4	5	6	7	8	9	10	23	24
5	1	8	0.0	12.5	25.0	25.0	37.5	37.5	50.0	50.0	50.0	50.0	62.5	62.5
6	2	6		0.0	0.0	16.7	16.7	16.7	33.3	33.3	33.3	33.3	66.7	66.7
7	3	2			0.0	0.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
8	4	16				0.0	12.5	18.8	25.0	31.3	43.8	50.0	68.8	68.8
9	5	8					0.0	12.5	12.5	12.5	25.0	25.0	37.5	37.5
10	6	7						0.0	0.0	0.0	0.0	0.0	14.3	14.3
11	7	3							0.0	0.0	0.0	0.0	0.0	0.0
22	18	4											50.0	50.0
23	19	2											0.0	0.0
24	20	0											0.0	0.0
25	21	0											0.0	0.0
26	22	1											0.0	0.0
27	23	0											0.0	0.0
28	24	4												0.0
29														
30	Cum First Repeat		0	1	2	3	7	9	12	13	17	18	35	35
31	First Repeat		0	1	1	1	4	2	3	1	4	1	1	0

=ROUND(SUMPRODUCT($B5:$B28,D5:D28)/100,0)

Modelling First Repeat Sales

$$FR(t) = \sum_{t_0=1}^{t-1} \Big\{ P(\text{first repeat by } t \mid \text{trial at } t_0) \times \left[T(t_0) - T(t_0 - 1) \right] \Big\}$$

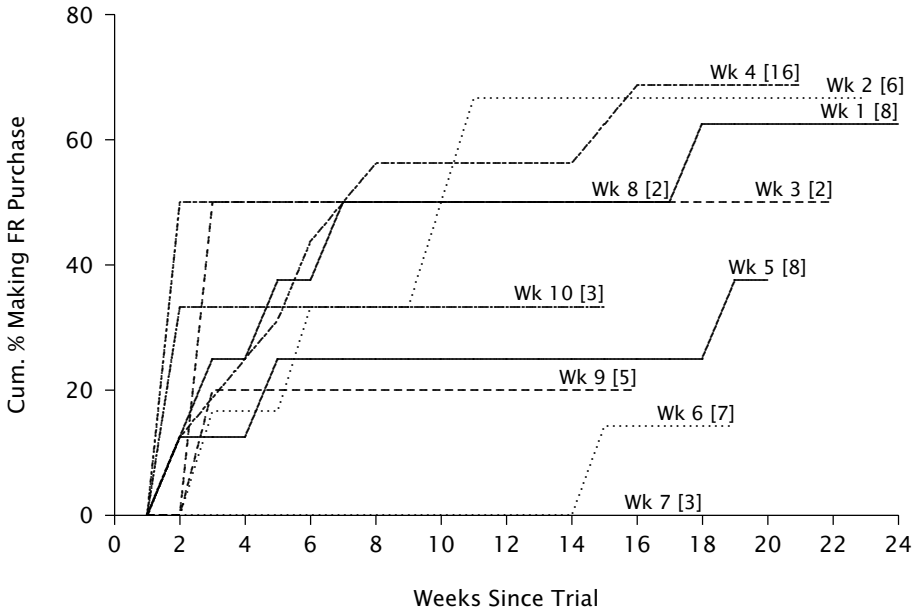
where $T(t_0) - T(t_0 - 1)$ is the number of incremental triers in week t_0.

	A	B	C	D	E	F	G	H	I	J	K	L	Y	Z
1	**Cum. First Repeat (as % of Trial) by Week of Trial**													
2														
3			Week											
4	Trial Week	#HHs	1	2	3	4	5	6	7	8	9	10	23	24
5	1	8	0.0	12.5	25.0	25.0	37.5	37.5	50.0	50.0	50.0	50.0	62.5	62.5
6	2	6		0.0	0.0	16.7	16.7	16.7	33.3	33.3	33.3	33.3	66.7	66.7
7	3	2			0.0	0.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
8	4	16				0.0	12.5	18.8	25.0	31.3	43.8	50.0	68.8	68.8
9	5	8					0.0	12.5	12.5	12.5	25.0	25.0	37.5	37.5
10	6	7						0.0	0.0	0.0	0.0	0.0	14.3	14.3
11	7	3							0.0	0.0	0.0	0.0	0.0	0.0
22	18	4											50.0	50.0
23	19	2											0.0	0.0
24	20	0											0.0	0.0
25	21	0											0.0	0.0
26	22	1											0.0	0.0
27	23	0											0.0	0.0
28	24	4												0.0
29														
30	Cum First Repeat		0	1	2	3	7	9	12	13	17	18	35	35
31	First Repeat		0	1	1	1	4	2	3	1	4	1	1	0

=ROUND(SUMPRODUCT($B5:$B28,D5:D28)/100,0)

	A	B	C	D	E	F	G	H	I	J	K	L	Y	Z
1	**(Shifted) Cum. First Repeat (as % of Trial) by Week of Trial**													
2														
3			Weeks Since Trial											
4	Trial Week	#HHs	0	1	2	3	4	5	6	7	8	9	22	23
5	1	8	0.0	12.5	25.0	25.0	37.5	37.5	50.0	50.0	50.0	50.0	62.5	62.5
6	2	6	0.0	0.0	16.7	16.7	16.7	33.3	33.3	33.3	33.3	50.0	66.7	
7	3	2	0.0	0.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0		
8	4	16	0.0	12.5	18.8	25.0	31.3	43.8	50.0	56.3	56.3	56.3		
9	5	8	0.0	12.5	12.5	12.5	25.0	25.0	25.0	25.0	25.0	25.0		
10	6	7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
11	7	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
22	18	4	0.0	50.0	50.0	50.0	50.0	50.0	50.0					
23	19	2	0.0	0.0	0.0	0.0	0.0	0.0						
24	20	0	0.0	0.0	0.0	0.0	0.0							
25	21	0	0.0	0.0	0.0	0.0								
26	22	1	0.0	0.0	0.0									
27	23	0	0.0	0.0										
28	24	4	0.0											

Empirical Time-to-FR Curves by Time of Trial



Modelling First Repeat

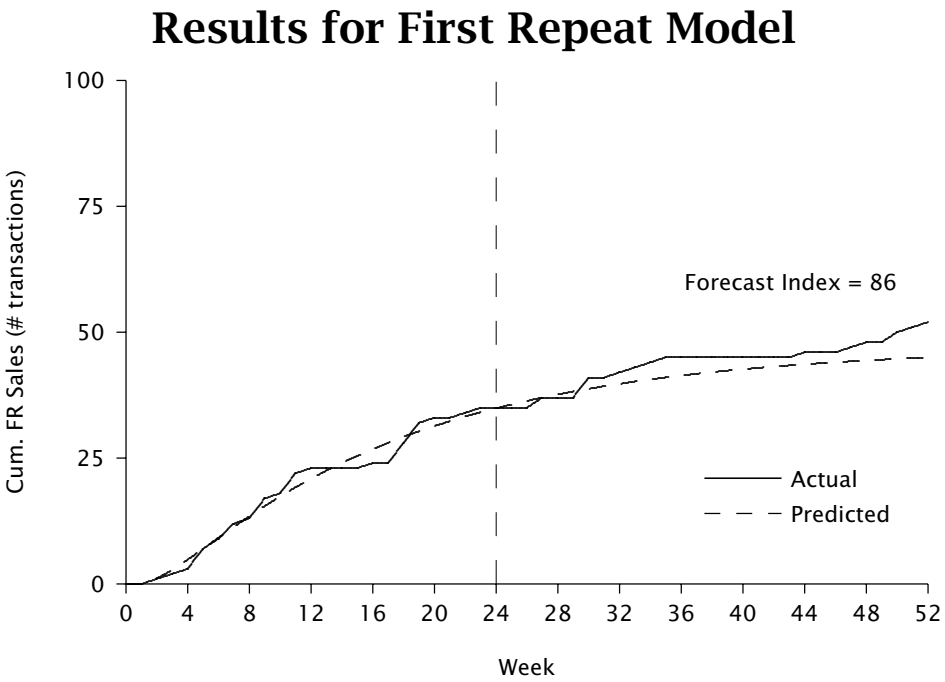
- Let us assume that these empirical time-to-FR curves are realizations of the same underlying curve

$$P(\text{first repeat by } t \mid \text{trial at } t_0) = P(\text{first repeat } t - t_0 \text{ periods after trial})$$

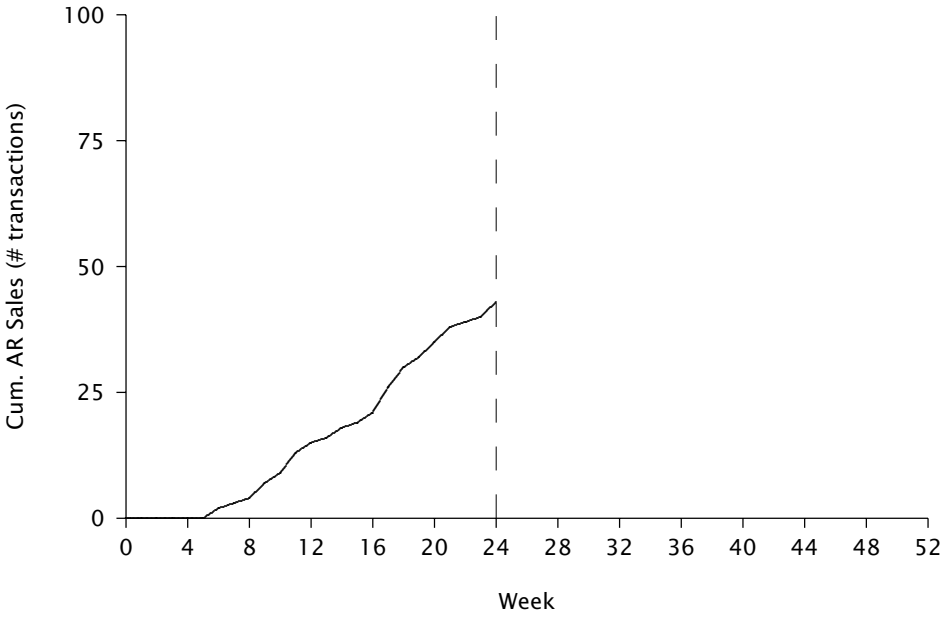
- Specify a mathematical expression for this curve

$$P(\text{first repeat by } t \mid \text{trial at } t_0) = p_1(1 - e^{-\theta_{FR}(t-t_0)})$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	Z	AA
1	p_1	0.3635													
2	theta_FR	0.4614			=SUM(D33:AA33)										
3	SSE	31.6774													
4				Week of First Repeat											
5	Trial Week	Cum Trial	Eligible	1	2	3	4	5	6	7	8	9	10	23	24
6	1	8	8	0.000	0.134	0.219	0.272	0.306	0.327	0.341	0.349	0.354	0.358	0.363	0.363
7	2	14	6		0.000	0.134	0.219	0.272	0.306	0.327	0.341	0.349	0.354	0.363	0.363
8	3	16	2			0.000	0.134	0.219	0.272	0.306	0.327	0.341	0.349	0.363	0.363
9	4	=$B1*(1-EXP(-$B2*(E5-1)))					0.000	0.134	0.219	0.272	0.306	0.327	0.341	0.363	0.363
10	5	40	8			=E7		0.000	0.134	0.219	0.272	0.306	0.327	0.363	0.363
11	6	47	7						0.000	0.134	0.219	0.272	0.306	0.363	0.363
12	7	50	3							0.000	0.134	0.219	0.272	0.363	0.363
23	18	94	4											0.327	0.341
24	19	96	2											0.306	0.327
25	20	96	0											0.272	0.306
26	21	96	0		=SUMPRODUCT($C6:$C29,E6:E29)									0.219	0.272
27	22	97	1											0.134	0.219
28	23	97	0											0.000	0.134
29	24	101	4												0.000
30															
31		Pred Cum FR		0.00	1.07	2.56	3.76	6.67	9.58	12.35	14.50	16.13	17.82	34.40	34.72
32		Act Cum FR		0	1	2	3	7	9	12	13	17	18	35	35
33		squared error		0.0000	0.0056	0.3116	0.5807	0.1087	0.3345	0.1239	2.2605	0.7596	0.0308	0.3595	0.0803
34															
35															
36					=(E31-E32)^2										
37															



Cumulative Additional Repeat Sales



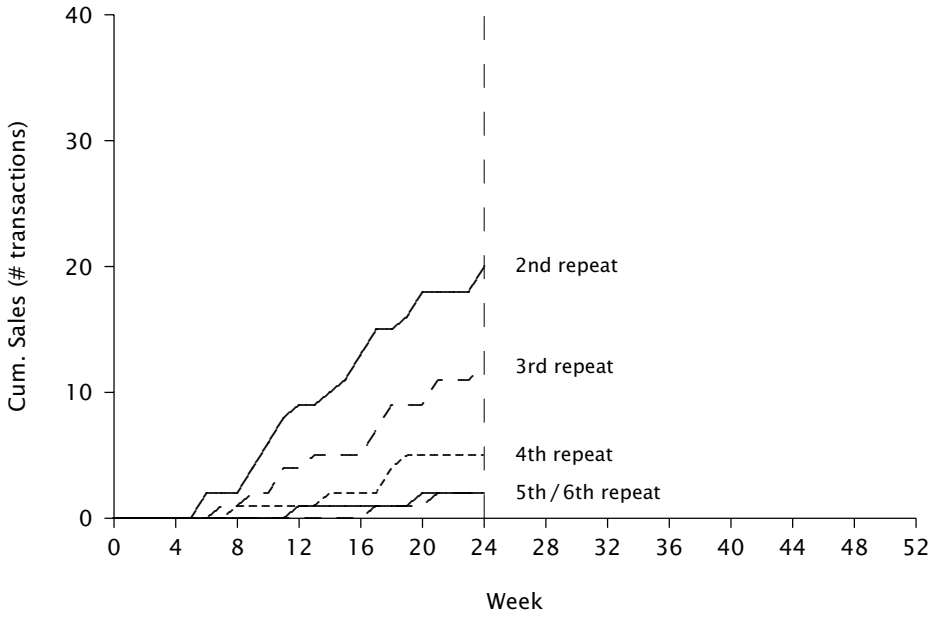
Modelling Additional Repeat

$$AR(t) = \sum_{j \geq 2} R_j(t)$$

where $R_j(t)$ is the cumulative number of individuals who have made at least j repeat purchases by time t.

How do we characterize $R_j(t),\ j = 2, 3, \ldots$?

Cumulative Sales by Depth of Repeat Level



Modelling Second Repeat

How can an individual have made a second repeat purchase by the end of week 5?

- she could have made her 1st repeat purchase in week 2 (which implies her trial purchase occurred in week 1) and made a 3rd purchase of the new product (i.e., her 2nd repeat purchase) somewhere in the intervening three weeks,
- she could have made her 1st repeat purchase in week 3 and a 2nd repeat purchase sometime in the following two weeks, or
- she could have made her 1st repeat purchase in week 4 and her 2nd repeat purchase sometime in the following week.

	A	B	C	D	E	F	G	H	I	J	K	L	Y	Z
1	**Cumulative Second Repeat by Week of First Repeat**													
2														
3			Week											
4	FR Week	#HHs	1	2	3	4	5	6	7	8	9	10	23	24
5	1	0												
6	2	1		0	0	0	0	1	1	1	1	1	1	1
7	3	1			0	0	0	0	0	0	0	0	0	0
8	4	1				0	0	0	0	0	0	0	0	0
9	5	4					0	1	1	1	2	2	3	3
10	6	2						0	0	0	0	0	1	1
11	7	3							0	0	1	3	3	3
22	18	4											2	3
23	19	4											1	2
24	20	1											0	0
25	21	0											0	0
26	22	1											0	0
27	23	1											0	0
28	24	0												0
29														
30	Cum 2nd Repeat		0	0	0	0	0	2	2	2	4	6	18	20
31	2nd Repeat		0	0	0	0	0	2	0	0	2	2	0	2

	A	B	C	D	E	F	G	H	I	J	K	L	Y	Z
1	**Cumulative Second Repeat (as % of FR) by Week of FR**													
2														
3			Week											
4	FR Week	#HHs	1	2	3	4	5	6	7	8	9	10	23	24
5	1	0												
6	2	1		0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
7	3	1			0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	4	1				0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	5	4					0.0	25.0	25.0	25.0	50.0	50.0	75.0	75.0
10	6	2						0.0	0.0	0.0	0.0	0.0	50.0	50.0
11	7	3							0.0	0.0	33.3	100.0	100.0	100.0
22	18	4											50.0	75.0
23	19	4											25.0	50.0
24	20	1											0.0	0.0
25	21	0											0.0	0.0
26	22	1											0.0	0.0
27	23	1											0.0	0.0
28	24	0												0.0
29														
30	Cum 2nd Repeat		0	0	0	0	0	2	2	2	4	6	18	20
31	2nd Repeat		0	0	0	0	0	2	0	0	2	2	0	2

=ROUND(SUMPRODUCT($B5:$B28,D5:D28)/100,0)

Modelling Second Repeat

$$R_2(t) = \sum_{t_1=2}^{t-1} \Big\{ P(\text{second repeat by } t \mid \text{first repeat at } t_1) \times \left[FR(t_1) - FR(t_1 - 1) \right] \Big\}$$

where $FR(t_1) - FR(t_1 - 1)$ is the number of individuals who made their first repeat purchase in week t_1.

More Generally ...

$$R_j(t) = \sum_{t_{j-1}=j}^{t-1} \Big\{ P(j\text{th repeat by } t \mid (j-1)\text{th repeat at } t_{j-1}) \times \left[R_{j-1}(t_{j-1}) - R_{j-1}(t_{j-1} - 1) \right] \Big\}$$

where $R_{j-1}(t_{j-1}) - R_{j-1}(t_{j-1} - 1)$ is the number of individuals who made their $(j-1)$th repeat purchase in week t_{j-1}.

Objective: develop a model for

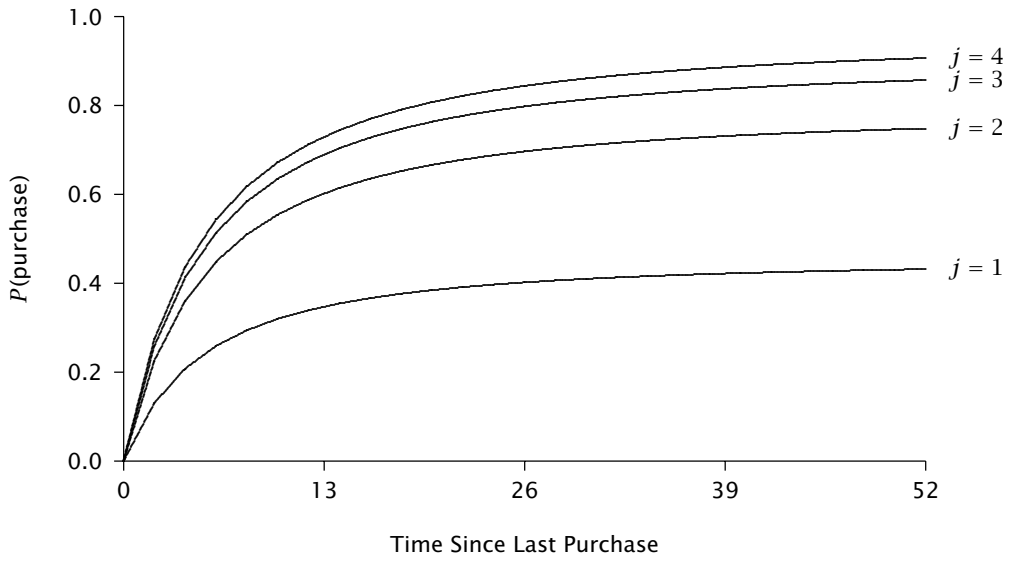
$$P(j\text{th repeat by } t \mid (j-1)\text{th repeat at } t_{j-1})$$

Challenges

- Sparse data for higher orders of repeat ($j = 3, 4, 5$)
- No data for repeat levels we are likely to observe in the forecast period

Are there common patterns across depth-of-repeat levels that we can exploit?

Depth-of-Repeat Curves



Probability of jth Repeat

Following the same logic as for trial and first repeat,

$$P(j\text{th repeat by } t \mid (j-1)\text{th repeat at } t_{j-1}) = p_j(1 - e^{-\theta_{AR}(t-t_{j-1})}), \quad t = t_{j-1} + 1, \ldots$$

Evolution of p_j

The asymptote of the depth-of-repeat curves (i.e., ultimate conversion proportion) increases, at a decreasing rate, as j increases.

> The proportion of consumers who have made their jth repeat purchase within 52 weeks of their $(j-1)$th repeat purchase increases with j.

We model the evolution of the ultimate conversion proportions as

$$p_j = p_\infty(1 - e^{-\gamma j}), \; j \geq 2$$

AR

	A	B	C	D	E	F	G
1	p_infty	0.7816	DoR j	p_j			
2	gamma	1.0014	2	0.6761			
3	theta _AR	0.2309	3	0.7428			
4			4	0.7673			
5	SSE	51.17298	5	0.7763			
6							
7							
8							
9							

=B$1*(1-EXP(-B$2*D2))

='DoR 2'!B4+'DoR 3'!B4+'DoR 4'!B4+'DoR 5'!B4

DoR 2

	A	B	C	D	E	F	G	H	I	J	K	L	M	Z	AA
1	p_j	0.6761		=AR!E2											
2	theta_AR	0.2309													
3	j	2		=AR!B3			=IF($A8<$B$3,0,IF($A8>=F$6,0,$B$1*(1-EXP(-$B$2*(F$6-$A8)))))								
4	SSE_j	18.6012													
5				Week of jth Repeat											
6	Week	Cum j-1	Eligible	1	2	3	4	5	6	7	8	9	10	23	24
7	1	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	2	1	1	0.000	0.000	0.139	0.250	0.338	0.408	0.463	0.507	0.542	0.570	0.671	0.672
9	3	2	1	0.000	0.000	0.000	0.139	0.250	0.338	0.408	0.463	0.507	0.542	0.669	0.671
10	4	3	1	0.000	0.000	0.000	0.000	0.139	0.250	0.338	0.408	0.463	0.507	0.668	0.669
11	5	7	4	0.000	0.000	0.000	0.000	0.000	0.139	0.250	0.338	0.408	0.463	0.666	0.668
12	6	9	2	0.000	0.000	0.000	0.000	0.000	0.000	0.139	0.250	0.338	0.408	0.663	0.666
13	7	12	3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.139	0.250	0.338	0.659	0.663
24	18	28	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.463	0.507
25	19	32	4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.408	0.463
26	20	33	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.338	0.408
27	21	33	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.250	0.338
28	22	34	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.139	0.250
29	23	35	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.139
30	24	35	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31															
32		Pred Cum DoR j		0.00	0.00	0.14	0.39	0.73	1.55	2.49	3.65	4.71	6.11	19.53	20.38
33		Act Cum DoR j		0	0	0	0	0	2	2	2	4	6	18	20
34		squared error		0.0000	0.0000	0.0194	0.1517	0.5292	0.1995	0.2380	2.7154	0.5014	0.0115	2.3365	0.1453

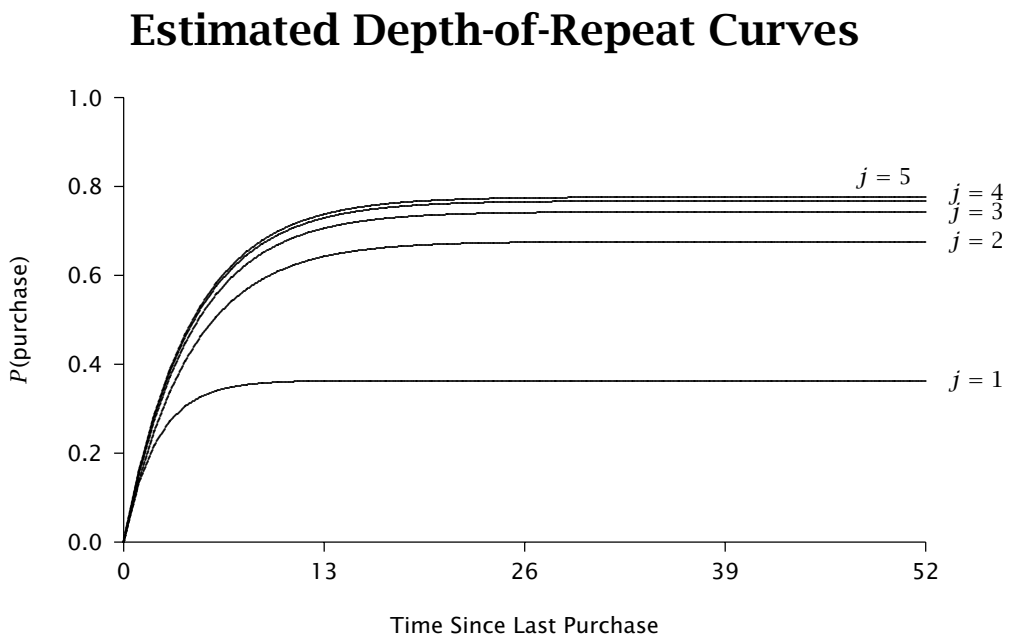
DoR 3

	A	B	C	D	E	F	G	H	I	J	K	L	M	Z	AA
1	p_j	0.7428													
2	theta_AR	0.2309													
3	j	3													
4	SSE_j	17.7239													
5				Week of jth Repeat											
6	Week	Cum j-1	Eligible	1	2	3	4	5	6	7	8	9	10	23	24
7	1	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	2	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	3	0	0	0.000	0.000	0.000	0.153	0.275	0.371	0.448	0.509	0.557	0.595	0.736	0.737
10	4	0	0	0.000	0.000	0.000	0.000	0.153	0.275	0.371	0.448	0.509	0.557	0.734	0.736
11	5	0	0	0.000	0.000	0.000	0.000	0.000	0.153	0.275	0.371	0.448	0.509	0.731	0.734
12	6	2	2	0.000	0.000	0.000	0.000	0.000	0.000	0.153	0.275	0.371	0.448	0.728	0.731
13	7	2	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.153	0.275	0.371	0.724	0.728
24	18	15	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.509	0.557
25	19	16	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.448	0.509
26	20	18	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.371	0.448
27	21	18	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.275	0.371
28	22	18	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.153	0.275
29	23	18	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.153
30	24	20	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31															
32		Pred Cum DoR j		0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.55	0.74	1.20	11.14	11.60
33		Act Cum DoR j		0	0	0	0	0	0	1	1	2	2	11	12
34		squared error		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4811	0.2029	1.5811	0.6365	0.0204	0.1581

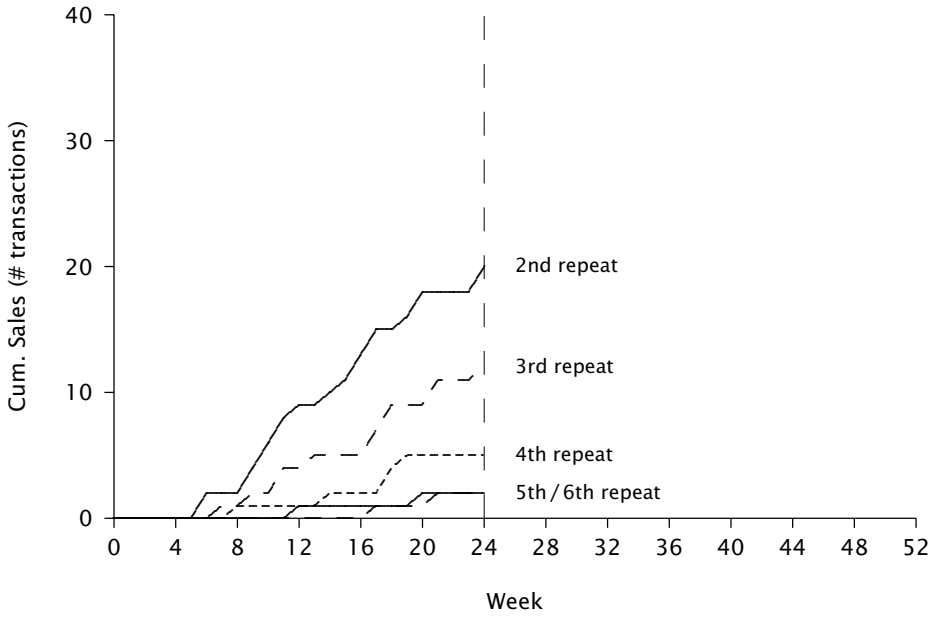
DoR 4

	A	B	C	D	E	F	G	H	I	J	K	L	M	Z	AA
1	p_j	0.7673													
2	theta_AR	0.2309													
3	j	4													
4	SSE_j	10.5477													
5				Week of jth Repeat											
6	Week	Cum j-1	Eligible	1	2	3	4	5	6	7	8	9	10	23	24
7	1	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	2	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	3	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	4	0	0	0.000	0.000	0.000	0.000	0.158	0.284	0.384	0.463	0.526	0.575	0.758	0.760
11	5	0	0	0.000	0.000	0.000	0.000	0.000	0.158	0.284	0.384	0.463	0.526	0.755	0.758
12	6	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.158	0.284	0.384	0.463	0.752	0.755
13	7	1	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.158	0.284	0.384	0.748	0.752
24	18	9	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.526	0.575
25	19	9	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.463	0.526
26	20	9	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.384	0.463
27	21	11	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.284	0.384
28	22	11	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.158	0.284
29	23	11	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.158
30	24	12	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31															
32		Pred Cum DoR j		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.28	0.54	6.38	6.81
33		Act Cum DoR j		0	0	0	0	0	0	0	1	1	1	5	5
34		squared error		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7086	0.5129	0.2100	1.9171	3.2711

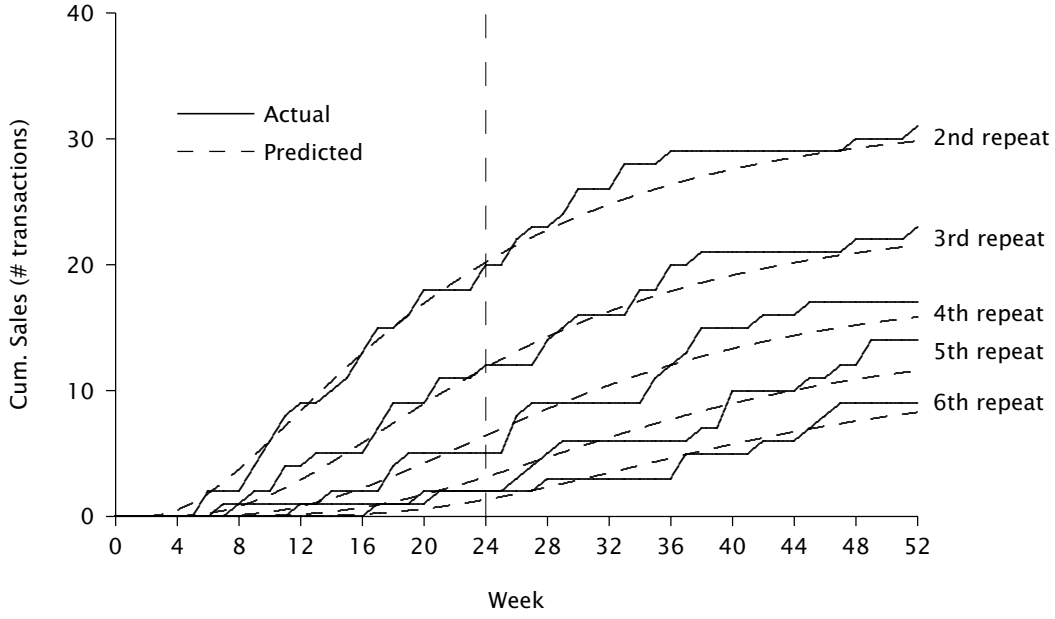
	A	B	C	D	E	F	G	H	I	J	K	L	M	Z	AA
1	p_j	0.7763													
2	theta_AR	0.2309													
3	j	5													
4	SSE_j	4.3002													
5				Week of jth Repeat											
6	Week	Cum j-1	Eligible	1	2	3	4	5	6	7	8	9	10	23	24
7	1	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	2	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	3	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	4	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
11	5	0	0	0.000	0.000	0.000	0.000	0.000	0.160	0.287	0.388	0.468	0.532	0.764	0.767
12	6	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.160	0.287	0.388	0.468	0.761	0.764
13	7	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.160	0.287	0.388	0.757	0.761
24	18	4	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.532	0.582
25	19	5	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.468	0.532
26	20	5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.388	0.468
27	21	5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.287	0.388
28	22	5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.160	0.287
29	23	5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.160
30	24	5	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31															
32		Pred Cum DoR j		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.29	2.96	3.15
33		Act Cum DoR j		0	0	0	0	0	0	0	0	0	0	2	2
34		squared error		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0256	0.0825	0.9269	1.3277



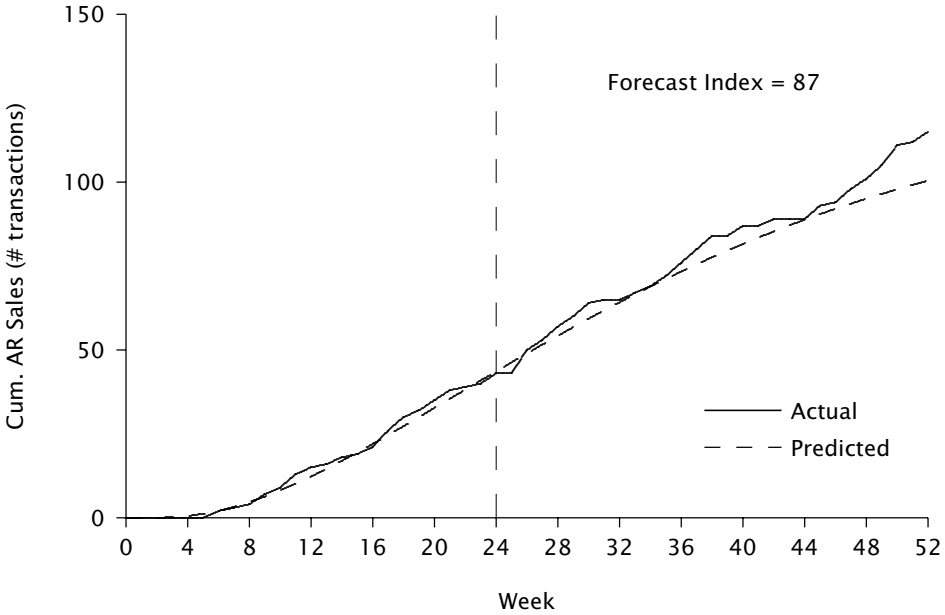
Cumulative Sales by Depth of Repeat Level



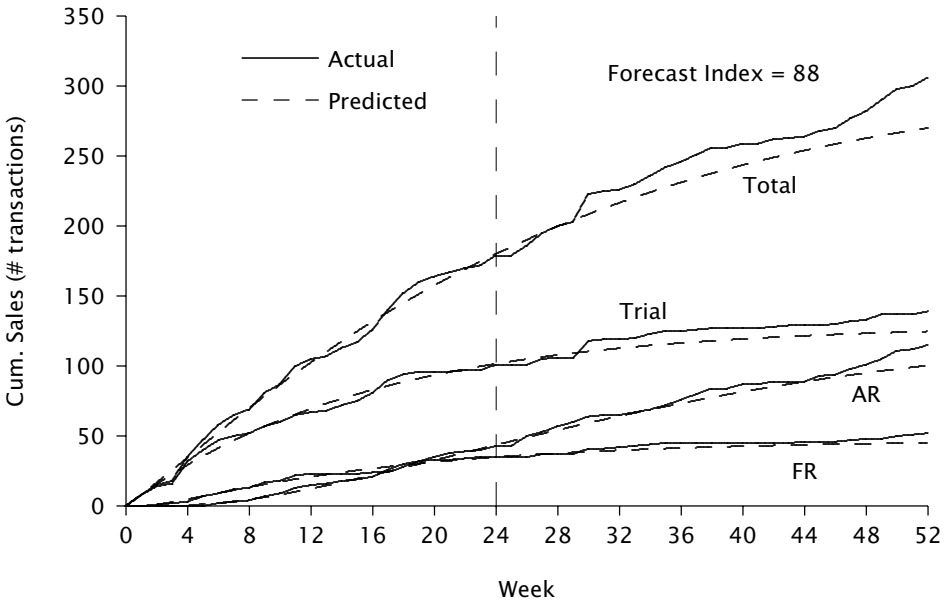
Results by Depth of Repeat Level



Results for Additional Repeat Model



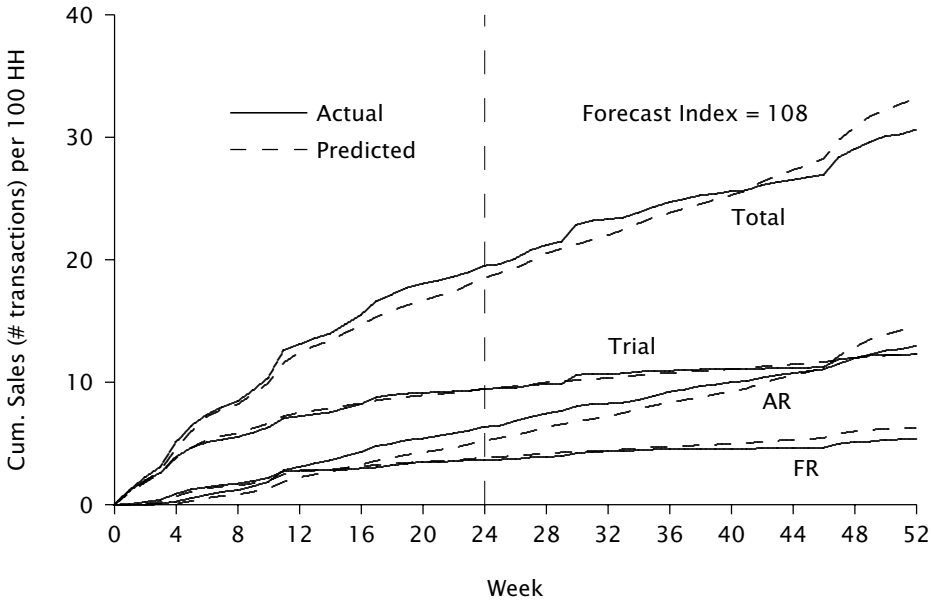
Creating an Overall Sales Forecast



Extending the Basic Model

- The trial model assumes all triers have the same underlying trial rate θ_T — a bit simplistic.
- Allow for individual differences in (latent) trial rates across the population.
- We incorporate the effects of marketing mix covariates by assuming that the probability of an individual buying in week t, *given* she has yet to make a trial purchase, is a function of marketing activity in week t.
- Similar modifications to first repeat model, etc.

Overall Sales Forecast for the Extended Model



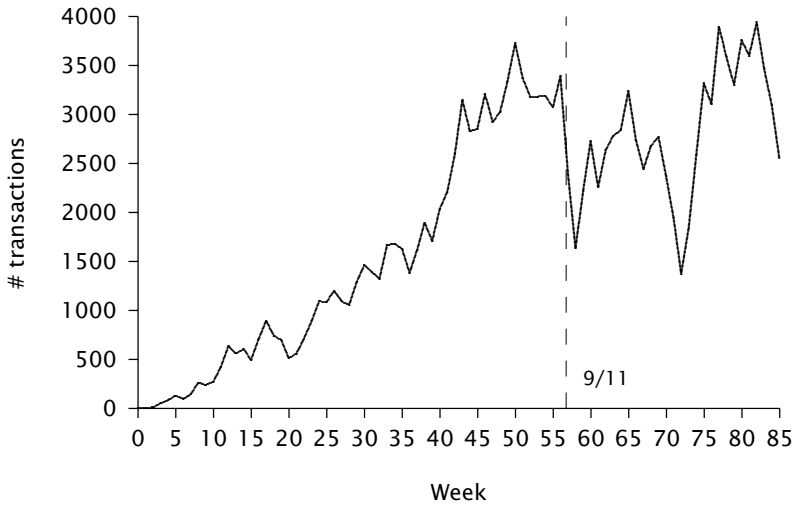
Implementation Issues

- Projection from panel to market
- Projection from market to region/country
- Adjustments for distribution build, seasonality, etc.

Application:

Using a Depth-of-Repeat Model to Determine the Impact of 9/11 on Online Travel Sales

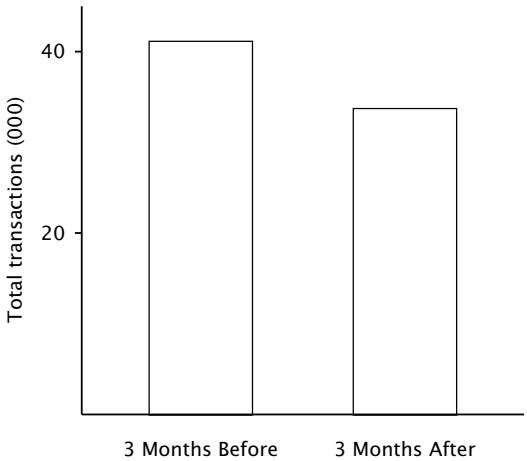
Setting



Systematic sample from a well-known online travel provider, drawn from its inception through March 2002 (26 weeks after 9/11).

Impact of 9/11 on Sales

Total Transactions — Pre and Post 9/11

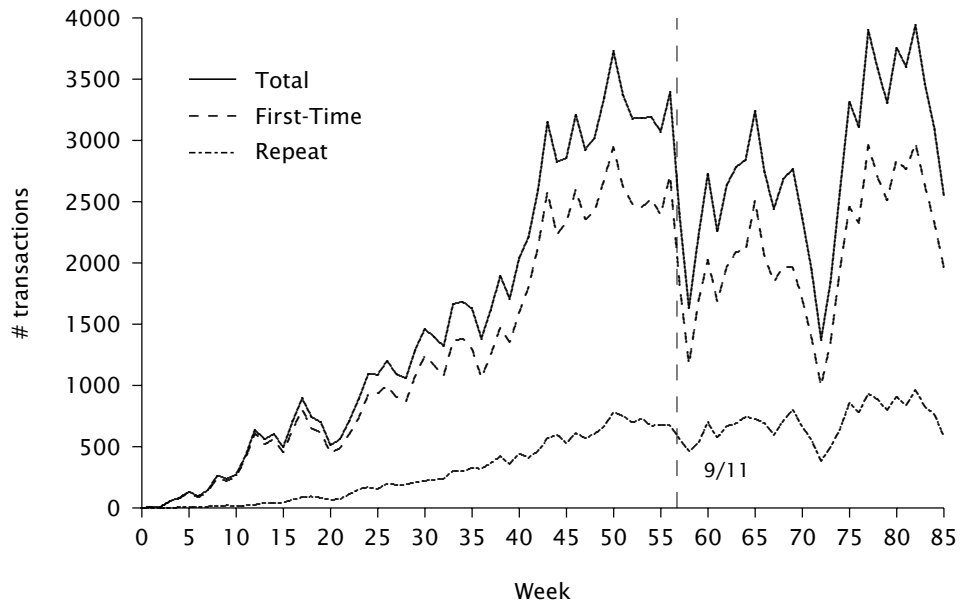


Aggregate measures (3 months before vs. 3 months) after show a decline of 23% from 9/11

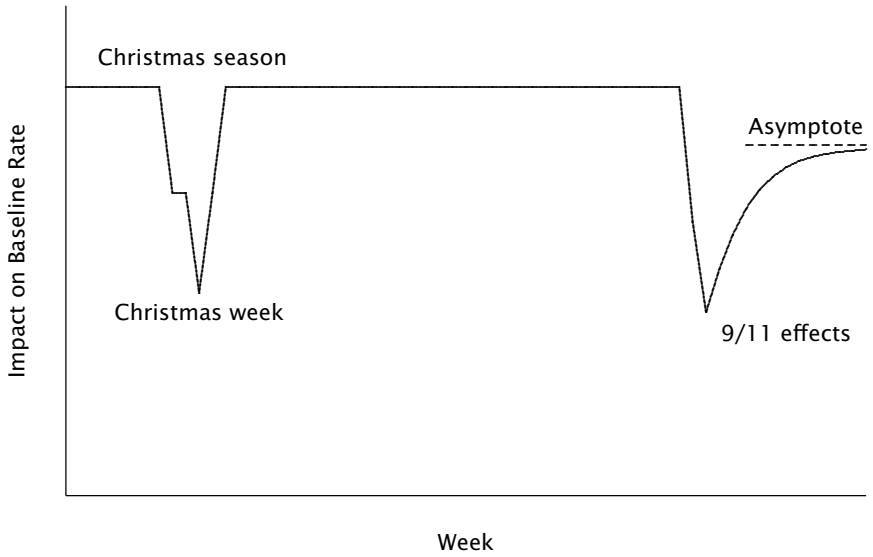
Measuring the True Effects

- Need to ask “what would the sales have been had the event of interest not happened?”
- Need to recognize that total transactions are the sum of first-time and subsequent (repeat) transactions

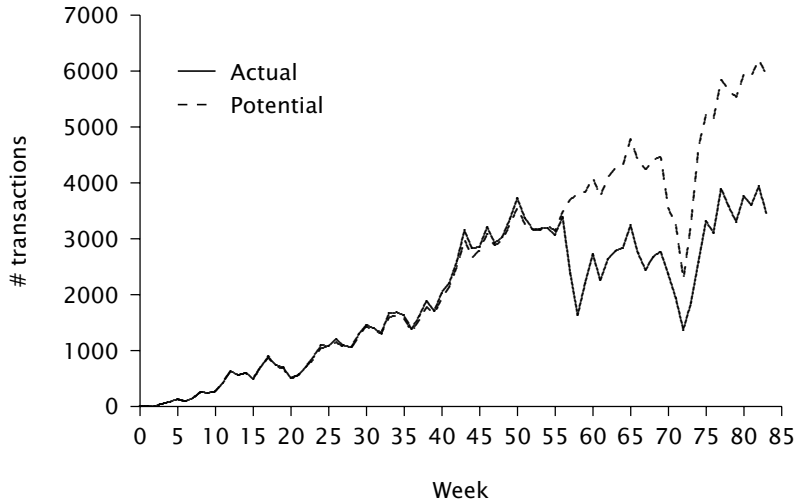
First-Time vs. Repeat Transactions



Hypothesized Time-Varying Covariate Effect



Complete Effect of 9/11



- An estimated loss of 39% of potential transactions
- 83% of this effect comes from lost potential "adopters" (first-time and repeat transactions)

Conclusions

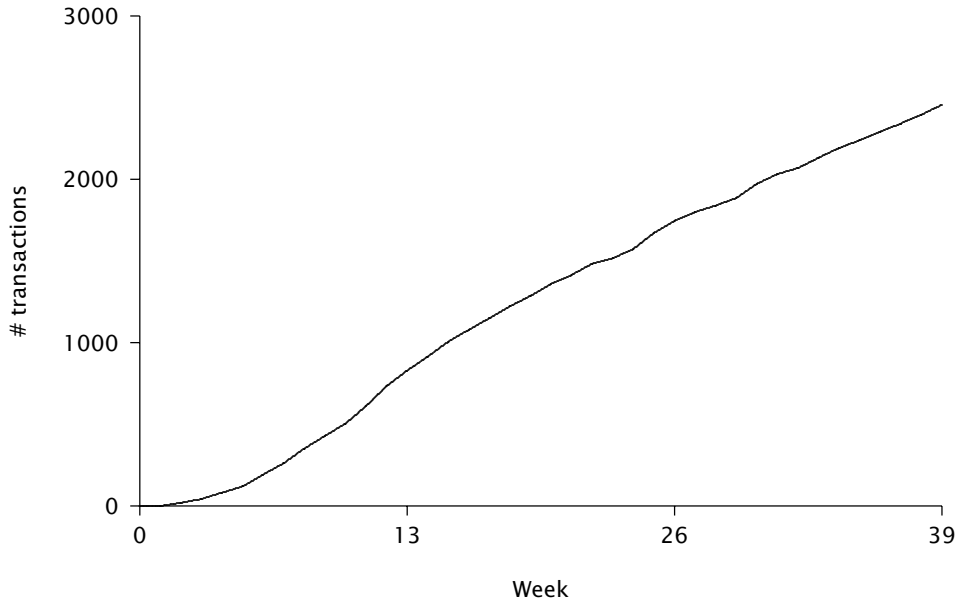
- Decompose new product sales into separate trial and repeat components.
- For each component, tell a behaviorally plausible “story” at the individual customer level, then aggregate up to the market level.
- While the parameters vary across these components, the same basic structural model can often be used for each of them.
- The different stages of “additional repeat” can be linked together and projected in a very parsimonious manner.
- Explanatory variables may be useful additions, but they are often not essential to developing an accurate sales forecasting model.

Part 2
Forecasting Individual-Level Repeat-Buying

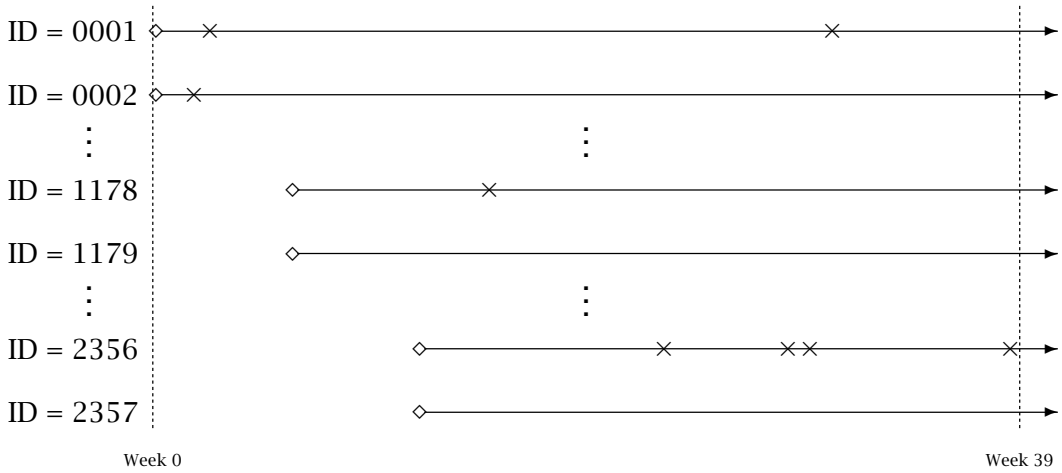
Setting

- New customers at CDNOW, 1/97–3/97
- Systematic sample (1/10) drawn from panel of 23,570 new customers
- 39-week calibration period
- 39-week forecasting (holdout) period

Cumulative Repeat Transactions



Purchase Histories



Raw Data

	A	B	C
1	ID	x	T
2	0001	2	38.86
3	0002	1	38.86
4	0003	0	38.86
5	0004	0	38.86
6	0005	0	38.86
7	0006	7	38.86
8	0007	1	38.86
9	0008	0	38.86
10	0009	2	38.86
11	0010	0	38.86
12	0011	5	38.86
13	0012	0	38.86
14	0013	0	38.86
15	0014	0	38.86
16	0015	0	38.86
17	0016	0	38.86
18	0017	10	38.86
19	0018	1	38.86
20	0019	3	38.71
1178	1177	0	32.71
1179	1178	1	32.71
1180	1179	0	32.71
1181	1180	0	32.71
2356	2355	0	27.00
2357	2356	4	27.00
2358	2357	0	27.00

Modelling Objective

Given this customer database, we wish to determine the level of transactions that should be expected in next period (e.g., 39 weeks) by those on the customer list, both individually and collectively.

Modelling the Purchasing Process

- A customer purchases "randomly" with an average transaction rate λ
- Transaction rates vary across customers

Modelling the Purchasing Process

- Let the random variable $X(t)$ denote the number of transactions in a period of length t time units.
- At the individual-level, $X(t)$ is assumed to be Poisson distributed with (exposure) rate parameter λt:

$$P(X(t) = x|\lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

- Transaction rates (λ) are distributed across the population according to a gamma distribution:

$$g(\lambda) = \frac{\alpha^r \lambda^{r-1} e^{-\alpha\lambda}}{\Gamma(r)}$$

Modelling the Purchasing Process

- The distribution of transactions for a randomly-chosen individual is given by:

$$\begin{aligned} P(X(t) = x) &= \int_0^\infty P(X(t) = x|\lambda)\, g(\lambda)\, d\lambda \\ &= \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha+t}\right)^r \left(\frac{t}{\alpha+t}\right)^x \end{aligned}$$

  This is called the Negative Binomial Distribution, or NBD model.

- The mean of the NBD is given by $E[X(t)] = rt/\alpha$.

Estimating Model Parameters

We estimate the model parameters using the method of *maximum likelihood*.

- The likelihood function is defined as the probability of observing all of the data points
- This probability is computed using the model and is viewed as a function of the model parameters:

$$L(\text{parameters}) = p(\text{data}|\text{parameters})$$

- For any given set of parameters, $L(\cdot)$ tells us the probability of obtaining the actual data
- For a given dataset, the maximum likelihood estimates of the model parameters are those values that maximize $L(\cdot)$

Estimating NBD Model Parameters

The likelihood function is defined as:

$$\begin{aligned} L(r, \alpha|\text{data}) &= P(X(T_1) = x_1|r, \alpha) \\ &\times P(X(T_2) = x_2|r, \alpha) \\ &\cdots \\ &\times P(X(T_{2357}) = x_{2357}|r, \alpha) \end{aligned}$$

The log-likelihood function is defined as:

$$LL(r, \alpha|\text{data}) = \sum_{i=1}^{2357} \ln\left[P(X(T_i) = x_i|r, \alpha)\right]$$

NBD Estimation

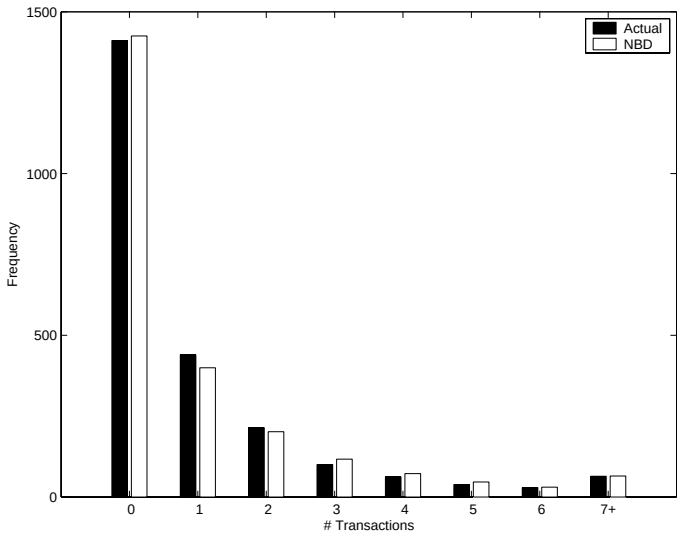
	A	B	C	D	E
1	r	0.3848			
2	alpha	12.0720			
3	LL	-3193.06			
4					
5	ID	x	T	P(X(T)=x)	ln(.)
6	0001	2	38.86	0.08913	-2.4177
7	0002	1	38.86	0.16871	-1.7796
8	0003	0	38.86	0.57471	-0.5539
9	0004	0	38.86	0.57471	-0.5539
10	0005	0	38.86	0.57471	-0.5539
11	0006	7	38.86	0.01113	-4.4984
12					6
13					9
14					7
15	0010	0	38.86	0.57471	-0.5539
2359	2354	5	27.00	0.01576	-4.1503
2360	2355	0	27.00	0.63641	-0.4519
2361	2356	4	27.00	0.02601	-3.6494
2362	2357	0	27.00	0.63641	-0.4519

=SUM(E6:E2362)

=LN(D6)

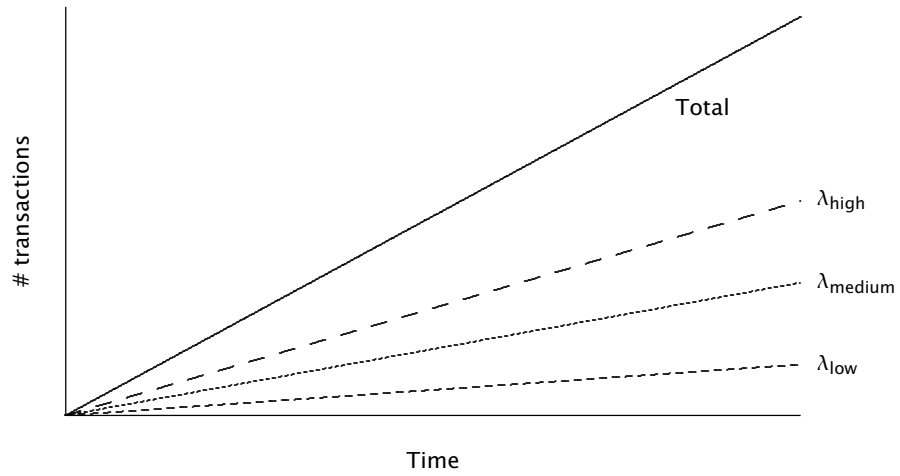
=EXP(GAMMALN(B1+B6)-GAMMALN(B1))/
FACT(B6)* (B2/(B2+C6))^B1*
(C6/(B2+C6))^B6

Frequency of Repeat Transactions



Computing Cumulative Repeat Transactions

Assuming Poisson purchasing, an individual's cumulative repeat transactions at time t equals λt.



Computing Cumulative Repeat Transactions

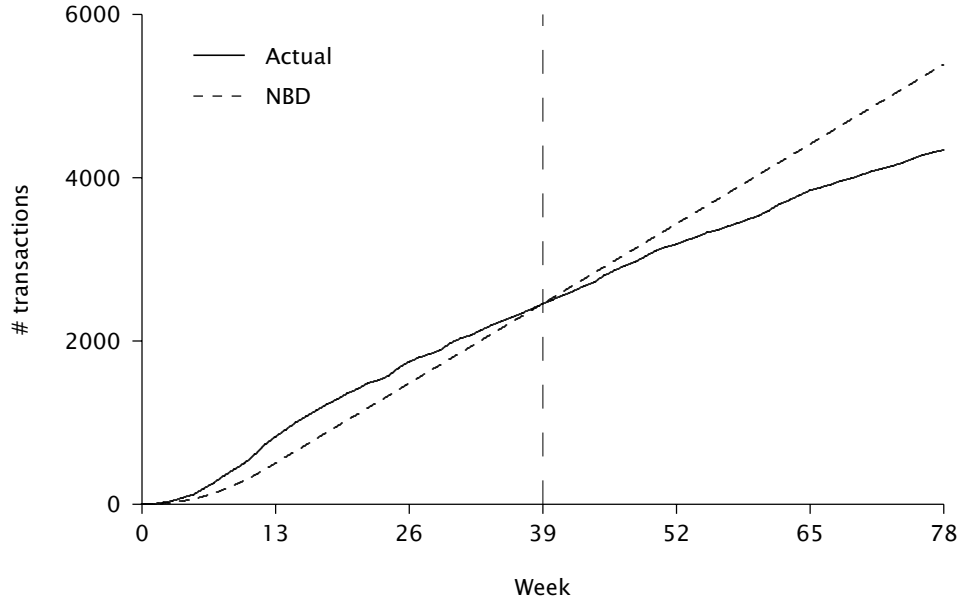
Total Repeat Transactions by t

$$= \sum_{s=1}^{84} \delta_{(t>\frac{s}{7})} n_s E[X(t-\tfrac{s}{7})]$$

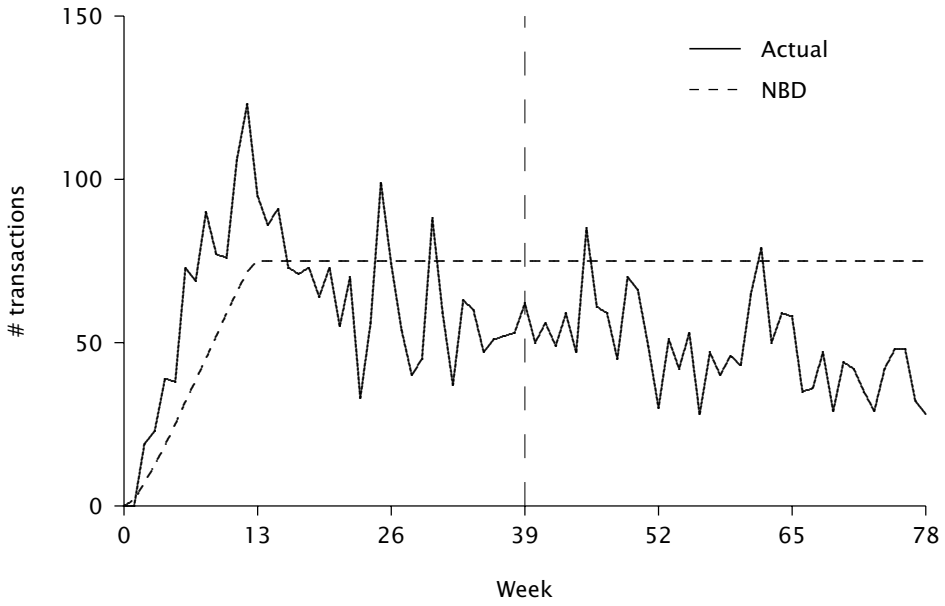
where

- n_s is the number of customers who made their first purchase at CDNOW on day s of 1997 (and therefore have $t - \frac{s}{7}$ weeks within which to make repeat purchases)
- $\delta_{(t>\frac{s}{7})} = 1$ if $t > \frac{s}{7}$, 0 otherwise.

Tracking Cumulative Repeat Transactions



Tracking Weekly Repeat Transactions

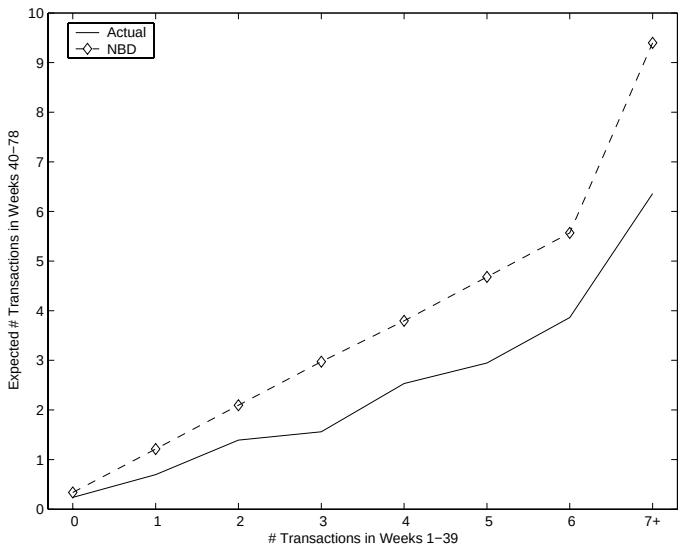


Conditional Expectations

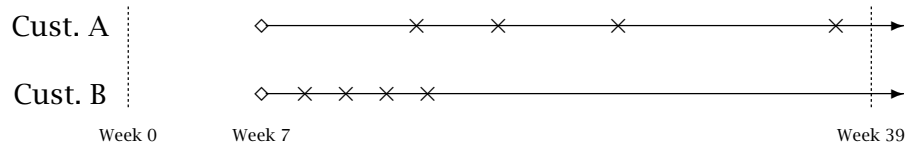
- We are interested in computing $E[Y(t)|\text{data}]$, the expected number of transactions in an adjacent period $(T, T+t]$, conditional on the observed purchase history.
- For the NBD, a straight-forward application of Bayes' theorem gives us

$$E\left[Y(t)|X(T)=x\right] = \left(\frac{r+x}{\alpha+T}\right) t$$

Conditional Expectations



Conditional Expectations

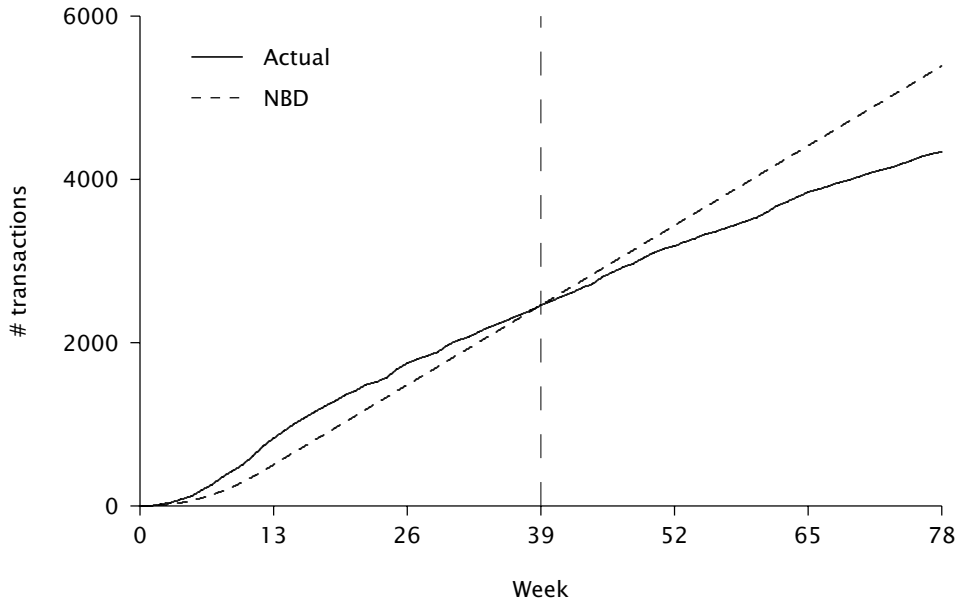


According to the NBD model:

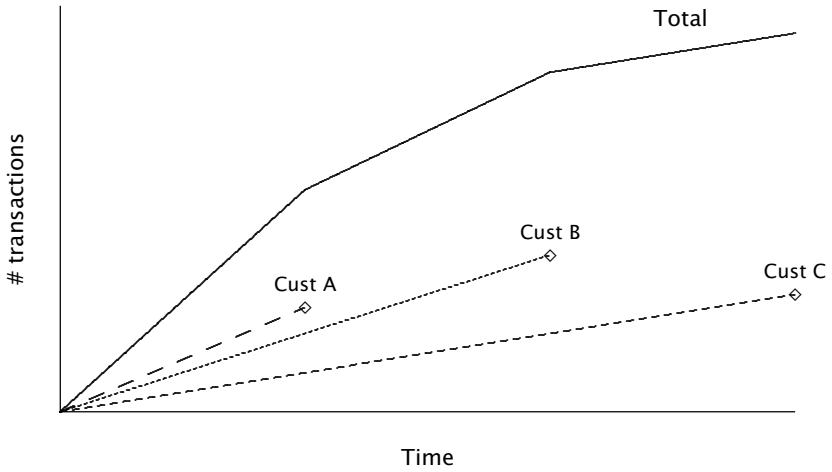
Cust. A: $E[Y(39)|X(32) = 4] = 3.88$

Cust. B: $E[Y(39)|X(32) = 4] = ?$

Tracking Cumulative Repeat Transactions



Model Development: Conceptual Overview



Modelling the Purchasing Process

Purchase Process:

- While active, a customer purchases "randomly" with an average transaction rate λ
- Transaction rates vary across customers

Dropout Process:

- After any transaction, a customer tosses a coin

  heads → become inactive
  tails → remain active

- P(heads) varies across customers

The BG/NBD Model

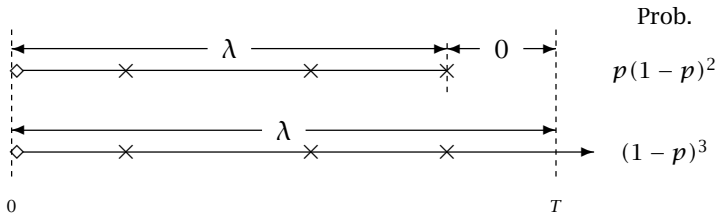
Purchase Process:

- While active, # transactions made by a customer follows a Poisson process with transaction rate λ.
- Heterogeneity in transaction rates across customers is distributed gamma(r, α).

Dropout Process:

- After any transaction, a customer becomes inactive with probability p.
- Heterogeneity in dropout probabilities across customers is distributed beta(a, b).

Model Illustration: Within Customer



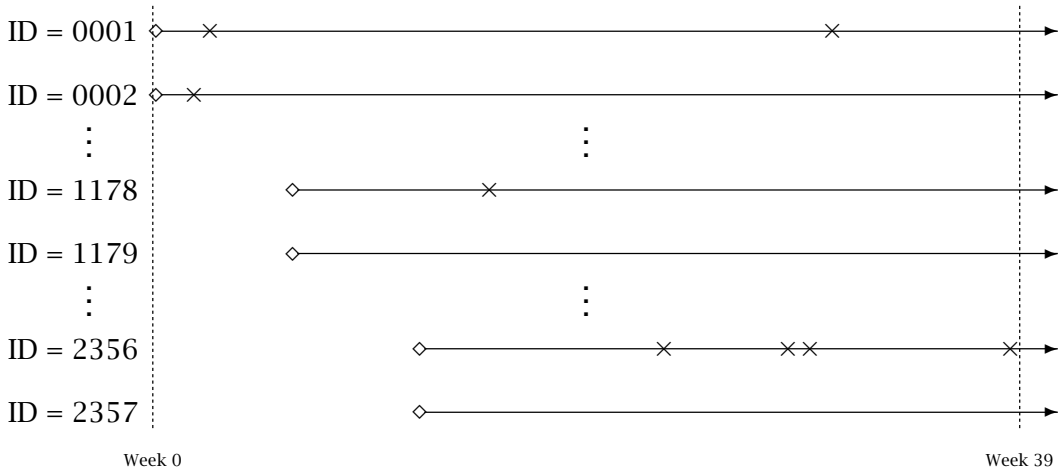
- More generally,

$$L(\lambda, p \mid t_1, t_2, \ldots, t_x, T) = (1-p)^x \lambda^x e^{-\lambda T} + \delta_{x>0}\, p(1-p)^{x-1} \lambda^x e^{-\lambda t_x}$$

where $\delta_{x>0} = 1$ if $x > 0$, 0 otherwise

- No need for the complete purchase history; "recency" and "frequency" are sufficient summary statistics

Purchase Histories



Raw Data

	A	B	C	D
1	ID	x	t_x	T
2	0001	2	30.43	38.86
3	0002	1	1.71	38.86
4	0003	0	0.00	38.86
5	0004	0	0.00	38.86
6	0005	0	0.00	38.86
7	0006	7	29.43	38.86
8	0007	1	5.00	38.86
9	0008	0	0.00	38.86
10	0009	2	35.71	38.86
11	0010	0	0.00	38.86
12	0011	5	24.43	38.86
13	0012	0	0.00	38.86
14	0013	0	0.00	38.86
15	0014	0	0.00	38.86
16	0015	0	0.00	38.86
17	0016	0	0.00	38.86
18	0017	10	34.14	38.86
19	0018	1	4.86	38.86
20	0019	3	28.29	38.71
1178	1177	0	0.00	32.71
1179	1178	1	8.86	32.71
1180	1179	0	0.00	32.71
1181	1180	0	0.00	32.71
2356	2355	0	0.00	27.00
2357	2356	4	26.57	27.00
2358	2357	0	0.00	27.00

Model Likelihood Function

For a randomly-chosen customer with x transactions in the period $(0, T]$, the last occurring at t_x:

$$L(r, \alpha, a, b \mid X = x, t_x, T) = \mathsf{A}_1 \cdot \mathsf{A}_2 \cdot (\mathsf{A}_3 + \delta_{x>0}\,\mathsf{A}_4)$$

where

$$\mathsf{A}_1 = \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)}$$

$$\mathsf{A}_2 = \frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)}$$

$$\mathsf{A}_3 = \left(\frac{1}{\alpha+T}\right)^{r+x}$$

$$\mathsf{A}_4 = \left(\frac{a}{b+x-1}\right)\left(\frac{1}{\alpha+t_x}\right)^{r+x}$$

BGNBD Estimation

	A	B	C	D	E	F	G	H	I
1	r	0.243							
2	alpha	4.414							
3	a	0.793							
4	b	2.426							
5	LL	-9582.4							
6									
7	ID	x	t_x	T	ln(.)	ln(A_1)	ln(A_2)	ln(A_3)	ln(A_4)
8	0001	2	30.43	38.86	-9.4596	-0.8390	-0.4910	-8.4489	-9.4265
9	0002	1	1.71	38.86	-4.4711	-1.0562	-0.2828	-4.6814	-3.3709
10	[illegible]	[illegible]	0.00	38.86	-0.5538	0.3602	0.0000	-0.9140	0.0000
11	0004	0	0.00	38.86	-0.5538	0.3602	0.0000	-0.9140	0.0000
12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
14	0007	1	5.00	38.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
15	0008	0	0.00	38.86	-0.5538	0.3602	0.0000	-0.9140	0.0000
16	0009	2	35.71	38.86	-9.5367	-0.8390	-0.4910	-8.4489	-9.7432
17	0010	0	0.00	38.86	-0.5538	0.3602	0.0000	-0.9140	0.0000
2362	2355	0	0.00	27.00	-0.4761	0.3602	0.0000	-0.8363	0.0000
2363	2356	4	26.57	27.00	-14.1284	1.1450	-0.7922	-14.6252	-16.4902
2364	2357	0	0.00	27.00	-0.4761	0.3602	0.0000	-0.8363	0.0000

=GAMMALN(B$1+B8)-GAMMALN(B$1)+B$1*LN(B$2)

=IF(B8>0,LN(B$3)-LN(B$4+B8-1)-(B$1+B8)*LN(B$2+C8),0)

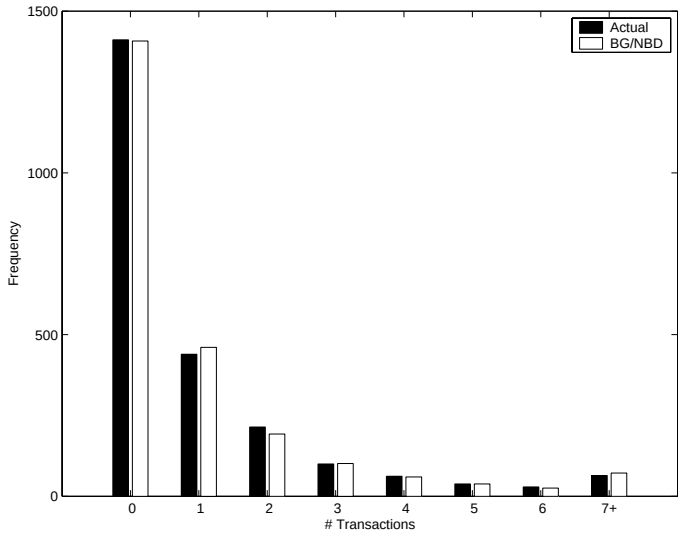
=-(B$1+B8)*LN(B$2+D8)

=SUM(E8:E2364)

=F8+G8+LN(EXP(H8)+(B8>0)*EXP(I8))

=GAMMALN(B$3+B$4)+GAMMALN(B$4+B8)-GAMMALN(B$4)-GAMMALN(B$3+B$4+B8)

Frequency of Repeat Transactions



Computing Expected # Transactions

Unconditional: $E(X(t) \mid r, \alpha, a, b)$

$$\frac{a+b-1}{a-1}\left[1-\left(\frac{\alpha}{\alpha+t}\right)^{r}\,{}_2F_1(r,b;a+b-1;\tfrac{t}{\alpha+t})\right]$$

Conditional: $E(Y(t) \mid X=x, t_x, T, r, \alpha, a, b)$

$$\frac{\frac{a+b+x-1}{a-1}\left[1-\left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x}\,{}_2F_1(r+x,b+x;a+b+x-1;\tfrac{t}{\alpha+T+t})\right]}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}}$$

The Gaussian Hypergeometric Function

$$ {}_2F_1(a,b;c;z) = \sum_{j=0}^{\infty} u_j $$

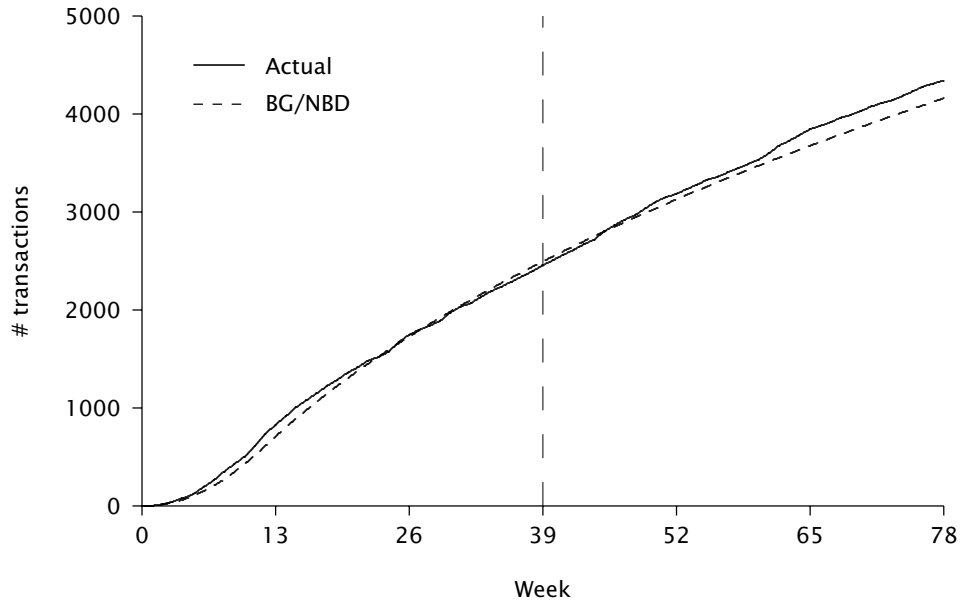
$$ \text{where } u_j = \frac{\Gamma(a+j)\Gamma(b+j)}{\Gamma(c+j)} \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)} \frac{z^j}{j!} $$

Easy to compute, albeit tedious in Excel, given the recursion

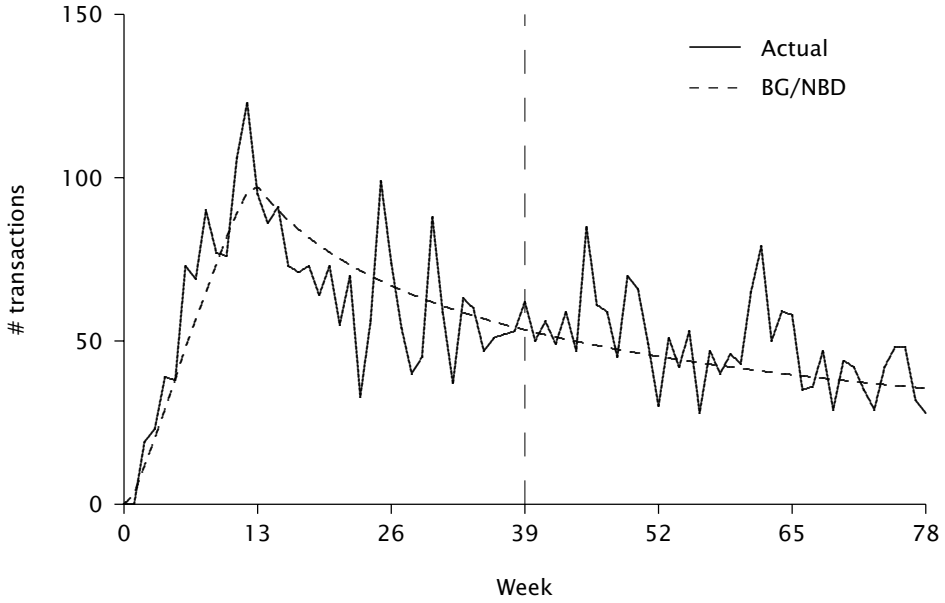
$$ \frac{u_j}{u_{j-1}} = \frac{(a+j-1)(b+j-1)}{(c+j-1)j} z\,, \quad j = 1,2,3,\ldots $$

where $u_0 = 1$

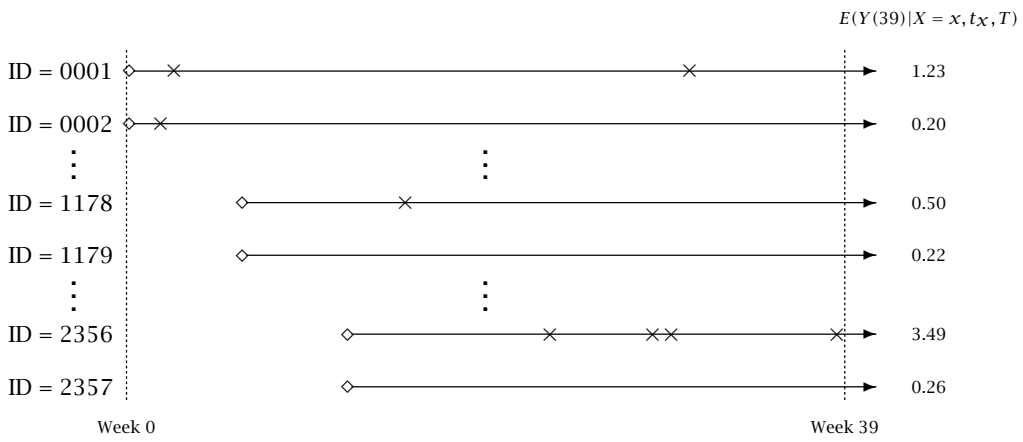
Tracking Cumulative Repeat Transactions



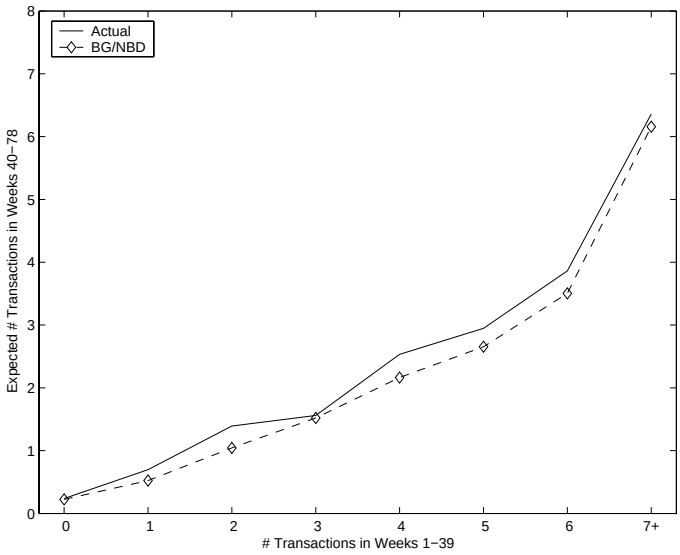
Tracking Weekly Repeat Transactions



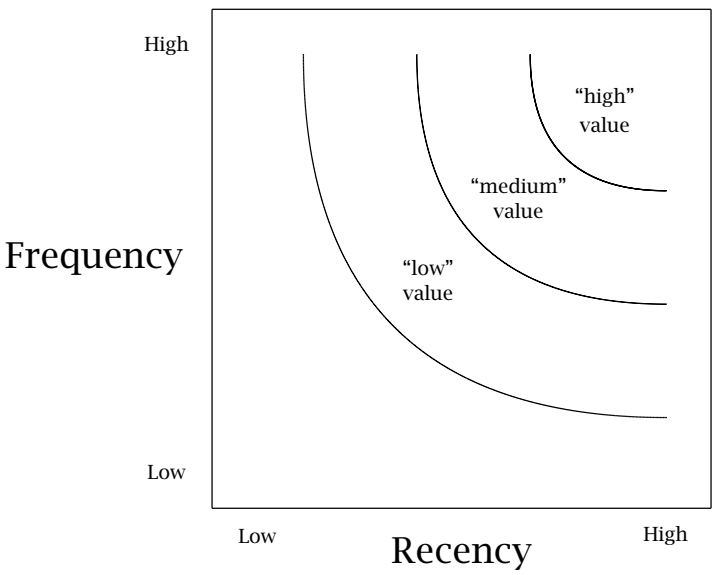
Conditional Expectations



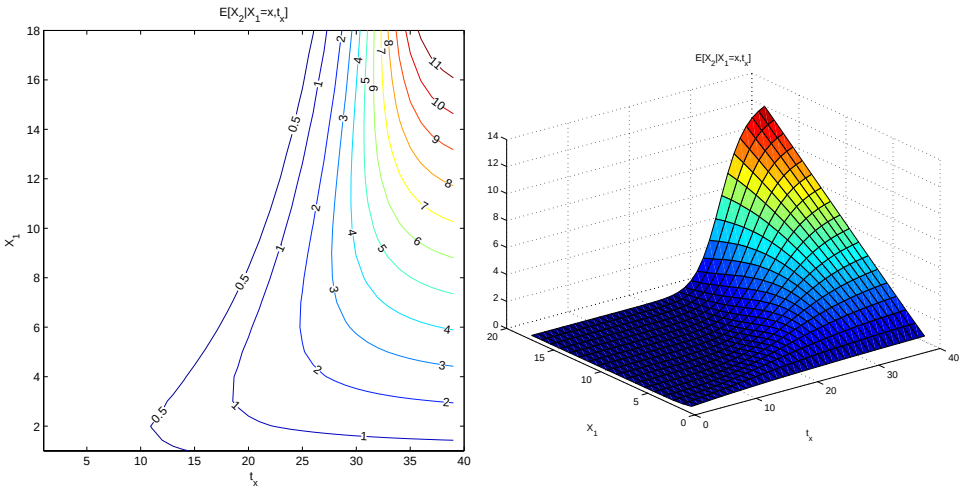
Conditional Expectations



Conditional Expectations



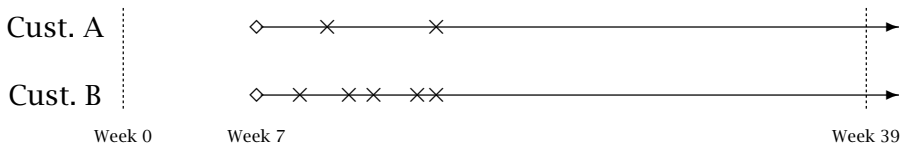
Conditional Expectations



The "Increasing Frequency" Paradox

For people with "low" recency, higher frequency is a bad thing.

Who is a better customer?



Conclusions

- Static models (NBD) are inadequate for customer base analysis
- Key is to capture the idea that customers become inactive
- Recency and frequency seem to provide enough information to capture the whole process (no need for the complete purchase histories)
- Usefulness of “iso-value” contours to understand tradeoffs between recency and frequency

Part 3
Links to the Broader Literature

"Simple" versus "Correct" Models

- The depth-of-repeat model is easy to implement and yields good forecasts of aggregate purchasing behavior
- Using separate submodels for trial, first repeat, second repeat, and so on, that fail to consider an individual's complete purchasing history may yield biased insights into underlying customer behavior.
- Apparent dynamics may not be ...

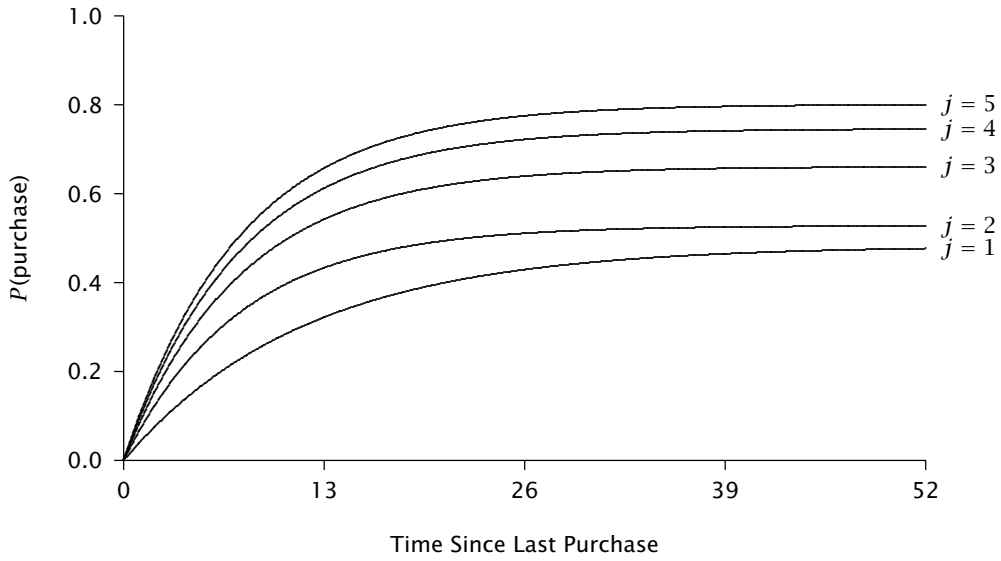
"Simple" versus "Correct" Models

Illustration:

- Let us fit the basic DoR model to 24 weeks of purchasing data generated for 1500 people using an NBD model with $r = 0.5$ and $\alpha = 40$
- An excellent fit to the data, with parameter estimates

p_1	θ_{FR}	p_∞	γ	θ_{AR}
0.483	0.085	0.900	0.442	0.132

Implied Depth-of-Repeat Curves



"Simple" versus "Correct" Models

- The "correct" way to model such a multiple-event timing process is to first condition on the individual, incorporate any effects of nonstationarity, and then bring in cross-sectional heterogeneity.
- Fader et al.'s (2004) "dynamic changepoint" model
  - a generalization of the BG/NBD model
  - difficult to implement, requiring data on the timing of each individual's purchases and specialist modelling environments
  - minor improvements in forecast accuracy may not outweigh the incremental cost of implementation

Models for Customer Base Analysis
Classifying Purchasing Processes (SMC 1987)