Applied Probability Models in Marketing: An Introduction

Peter S. Fader
University of Pennsylvania
faderp@wharton.upenn.edu

Bruce G. S. Hardie
London Business School
bhardie@london.edu

Statistical Modeling Week
October 22, 2002

1

Problem 1: Predicting New Product Trial

(Modeling Timing Data)

Background

Ace Snackfoods, Inc. has developed a new snack product called Krunchy Bits. Before deciding whether or not to “go national” with the new product, the marketing manager for Krunchy Bits has decided to commission a year-long test market using IRI’s BehaviorScan service, with a view to getting a clearer picture of the product’s potential.

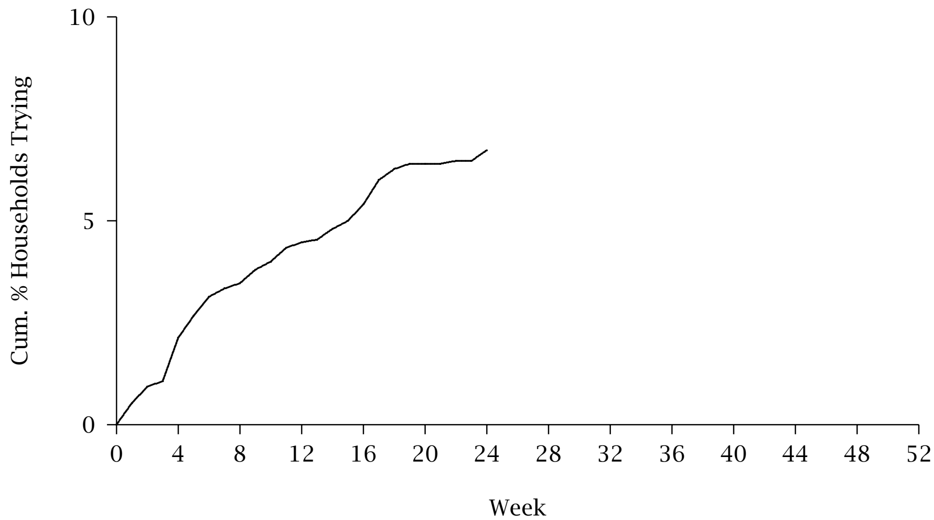
The product has now been under test for 24 weeks. On hand is a dataset documenting the number of households that have made a trial purchase by the end of each week. (The total size of the panel is 1499 households.)

The marketing manager for Krunchy Bits would like a forecast of the product’s year-end performance in the test market. First, she wants a forecast of the percentage of households that will have made a trial purchase by week 52.

Krunchy Bits Cumulative Trial

Week	# Households	Week	# Households
1	8	13	68
2	14	14	72
3	16	15	75
4	32	16	81
5	40	17	90
6	47	18	94
7	50	19	96
8	52	20	96
9	57	21	96
10	60	22	97
11	65	23	97
12	67	24	101

Krunchy Bits Cumulative Trial



5

Approaches to Forecasting Trial

- French curve
- "Curve fitting" — specify a flexible functional form, fit it to the data, and project into the future.
- Probability model

Developing a Model of Trial Purchasing

- Start at the individual-level then aggregate.

  Q: What is the individual-level behavior of interest?

  A: Time (since new product launch) of trial purchase.

- We don't know exactly what is driving the behavior $\Rightarrow$ treat it as a random variable.

The 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:

$$F(t) = P(T \le t) = 1 - e^{-\lambda t}$$

- λ represents the individual's trial rate.

The Market-Level Model

Assume two segments of consumers:

Segment	Description	Size	λ
1	ever triers	p	θ
2	never triers	$1-p$	0

$$\begin{aligned} P(T \le t) &= P(T \le t \mid \text{ever trier}) \times P(\text{ever trier}) + \\ &\quad P(T \le t \mid \text{never trier}) \times P(\text{never trier}) \\ &= pF(t \mid \lambda = \theta) + (1-p)F(t \mid \lambda = 0) \\ &= p(1 - e^{-\theta t}) \end{aligned}$$

→ the "exponential w/ never triers" model

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:

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

- 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 Model Parameters

The log-likelihood function is defined as:

$$
\begin{aligned}
LL(p, \theta|\text{data}) = \; & 8 \times \ln[P(0 < T \leq 1)] \quad + \\
& 6 \times \ln[P(1 < T \leq 2)] \quad + \\
& \ldots \quad + \\
& 4 \times \ln[P(23 < T \leq 24)] + \\
& (1499 - 101) \times \ln[P(T > 24)]
\end{aligned}
$$

The maximum value of the log-likelihood function is $LL = -680.9$, which occurs at $\hat{p} = 0.085$ and $\hat{\theta} = 0.066$.

Forecasting Trial

- $F(t)$ represents the probability that a randomly chosen household has made a trial purchase by time t, where $t = 0$ corresponds to the launch of the new product.
- Let $T(t)$ = cumulative # households that have made a trial purchase by time t:

$$
\begin{aligned}
E[T(t)] &= N \times \hat{F}(t) \\
&= N\hat{p}(1 - e^{-\hat{\theta}t}), \quad t = 1, 2, \ldots
\end{aligned}
$$

  where N is the panel size.
- Use projection factors for market-level estimates.

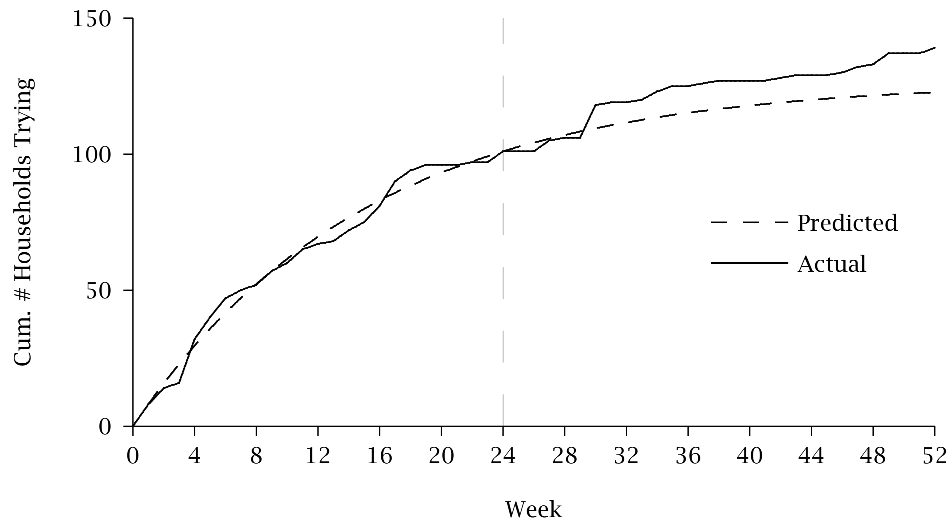
Problem 1 -- Model 1

	A	B	C	D	E	F	G	H
1	Product:	Krunchy Bits			p	0.5		
2	Panelists:	1499			\theta	0.05		
3					LL =	=SUM(F6:F30)		
4		Cum_Trl						
5	Week	# HHs	Incr_Trl		P(T <= t)			E[T(t)]
6	1	8	=B6		=F$1*(1-EXP(-F$2*A6))	=C6*LN(E6)		=B$2*E6
7	2	14	=B7-B6		=F$1*(1-EXP(-F$2*A7))	=C7*LN(E7-E6)		=B$2*E7
8	3	16	=B8-B7		=F$1*(1-EXP(-F$2*A8))	=C8*LN(E8-E7)		=B$2*E8
9	4	32	=B9-B8		=F$1*(1-EXP(-F$2*A9))	=C9*LN(E9-E8)		=B$2*E9
10	5	40	=B10-B9		=F$1*(1-EXP(-F$2*A10))	=C10*LN(E10-E9)		=B$2*E10
11	6	47	=B11-B10		=F$1*(1-EXP(-F$2*A11))	=C11*LN(E11-E10)		=B$2*E11
12	7	50	=B12-B11		=F$1*(1-EXP(-F$2*A12))	=C12*LN(E12-E11)		=B$2*E12
13	8	52	=B13-B12		=F$1*(1-EXP(-F$2*A13))	=C13*LN(E13-E12)		=B$2*E13
14	9	57	=B14-B13		=F$1*(1-EXP(-F$2*A14))	=C14*LN(E14-E13)		=B$2*E14
15	10	60	=B15-B14		=F$1*(1-EXP(-F$2*A15))	=C15*LN(E15-E14)		=B$2*E15
16	11	65	=B16-B15		=F$1*(1-EXP(-F$2*A16))	=C16*LN(E16-E15)		=B$2*E16
17	12	67	=B17-B16		=F$1*(1-EXP(-F$2*A17))	=C17*LN(E17-E16)		=B$2*E17
18	13	68	=B18-B17		=F$1*(1-EXP(-F$2*A18))	=C18*LN(E18-E17)		=B$2*E18
19	14	72	=B19-B18		=F$1*(1-EXP(-F$2*A19))	=C19*LN(E19-E18)		=B$2*E19
20	15	75	=B20-B19		=F$1*(1-EXP(-F$2*A20))	=C20*LN(E20-E19)		=B$2*E20
21	16	81	=B21-B20		=F$1*(1-EXP(-F$2*A21))	=C21*LN(E21-E20)		=B$2*E21
22	17	90	=B22-B21		=F$1*(1-EXP(-F$2*A22))	=C22*LN(E22-E21)		=B$2*E22
23	18	94	=B23-B22		=F$1*(1-EXP(-F$2*A23))	=C23*LN(E23-E22)		=B$2*E23
24	19	96	=B24-B23		=F$1*(1-EXP(-F$2*A24))	=C24*LN(E24-E23)		=B$2*E24
25	20	96	=B25-B24		=F$1*(1-EXP(-F$2*A25))	=C25*LN(E25-E24)		=B$2*E25
26	21	96	=B26-B25		=F$1*(1-EXP(-F$2*A26))	=C26*LN(E26-E25)		=B$2*E26
27	22	97	=B27-B26		=F$1*(1-EXP(-F$2*A27))	=C27*LN(E27-E26)		=B$2*E27
28	23	97	=B28-B27		=F$1*(1-EXP(-F$2*A28))	=C28*LN(E28-E27)		=B$2*E28
29	24	101	=B29-B28		=F$1*(1-EXP(-F$2*A29))	=C29*LN(E29-E28)		=B$2*E29
30	25	101			=F$1*(1-EXP(-F$2*A30))	=(B2-B29)*LN(1-E29)		=B$2*E30
31	26	101			=F$1*(1-EXP(-F$2*A31))			=B$2*E31
32	27	105			=F$1*(1-EXP(-F$2*A32))			=B$2*E32
33	28	106			=F$1*(1-EXP(-F$2*A33))			=B$2*E33
34	29	106			=F$1*(1-EXP(-F$2*A34))			=B$2*E34
35	30	118			=F$1*(1-EXP(-F$2*A35))			=B$2*E35
36	31	119			=F$1*(1-EXP(-F$2*A36))			=B$2*E36
37	32	119			=F$1*(1-EXP(-F$2*A37))			=B$2*E37
38	33	120			=F$1*(1-EXP(-F$2*A38))			=B$2*E38
39	34	123			=F$1*(1-EXP(-F$2*A39))			=B$2*E39
40	35	125			=F$1*(1-EXP(-F$2*A40))			=B$2*E40
41	36	125			=F$1*(1-EXP(-F$2*A41))			=B$2*E41
42	37	126			=F$1*(1-EXP(-F$2*A42))			=B$2*E42
43	38	127			=F$1*(1-EXP(-F$2*A43))			=B$2*E43
44	39	127			=F$1*(1-EXP(-F$2*A44))			=B$2*E44
45	40	127			=F$1*(1-EXP(-F$2*A45))			=B$2*E45
46	41	127			=F$1*(1-EXP(-F$2*A46))			=B$2*E46
47	42	128			=F$1*(1-EXP(-F$2*A47))			=B$2*E47
48	43	129			=F$1*(1-EXP(-F$2*A48))			=B$2*E48
49	44	129			=F$1*(1-EXP(-F$2*A49))			=B$2*E49
50	45	129			=F$1*(1-EXP(-F$2*A50))			=B$2*E50
51	46	130			=F$1*(1-EXP(-F$2*A51))			=B$2*E51
52	47	132			=F$1*(1-EXP(-F$2*A52))			=B$2*E52
53	48	133			=F$1*(1-EXP(-F$2*A53))			=B$2*E53
54	49	137			=F$1*(1-EXP(-F$2*A54))			=B$2*E54
55	50	137			=F$1*(1-EXP(-F$2*A55))			=B$2*E55
56	51	137			=F$1*(1-EXP(-F$2*A56))			=B$2*E56
57	52	139			=F$1*(1-EXP(-F$2*A57))			=B$2*E57

Problem 1 -- Model 1

	A	B	C	D	E	F	G	H
1	Product:	Krunchy Bits			p	0.08456		
2	Panelists:	1499			\theta	0.0664		
3					LL =	-680.9094		
4		Cum_Trl						
5	Week	# HHs	Incr_Trl		P(T <= t)			E[T(t)]
6	1	8	8		0.00543	-41.723		8.14
7	2	14	6		0.01052	-31.691		15.76
8	3	16	2		0.01527	-10.696		22.89
9	4	32	16		0.01972	-86.633		29.57
10	5	40	8		0.02389	-43.848		35.81
11	6	47	7		0.02779	-38.832		41.65
12	7	50	3		0.03143	-16.841		47.12
13	8	52	2		0.03485	-11.360		52.24
14	9	57	5		0.03804	-28.733		57.02
15	10	60	3		0.04103	-17.439		61.50
16	11	65	5		0.04383	-29.397		65.70
17	12	67	2		0.04644	-11.892		69.62
18	13	68	1		0.04889	-6.012		73.29
19	14	72	4		0.05118	-24.314		76.72
20	15	75	3		0.05333	-18.435		79.94
21	16	81	6		0.05533	-37.268		82.95
22	17	90	9		0.05721	-56.500		85.76
23	18	94	4		0.05897	-25.377		88.39
24	19	96	2		0.06061	-12.821		90.86
25	20	96	0		0.06215	0.000		93.16
26	21	96	0		0.06359	0.000		95.32
27	22	97	1		0.06494	-6.610		97.34
28	23	97	0		0.06620	0.000		99.23
29	24	101	4		0.06738	-26.970		101.00
30	25	101			0.06848	-97.518		102.65
31	26	101			0.06951			104.20
32	27	105			0.07048			105.65
33	28	106			0.07139			107.01
34	29	106			0.07223			108.28
35	30	118			0.07302			109.46
36	31	119			0.07377			110.57
37	32	119			0.07446			111.61
38	33	120			0.07511			112.59
39	34	123			0.07572			113.50
40	35	125			0.07628			114.35
41	36	125			0.07682			115.15
42	37	126			0.07731			115.89
43	38	127			0.07778			116.59
44	39	127			0.07821			117.24
45	40	127			0.07862			117.85
46	41	127			0.07900			118.43
47	42	128			0.07936			118.96
48	43	129			0.07969			119.46
49	44	129			0.08001			119.93
50	45	129			0.08030			120.37
51	46	130			0.08057			120.78
52	47	132			0.08083			121.16
53	48	133			0.08107			121.52
54	49	137			0.08129			121.86
55	50	137			0.08150			122.17
56	51	137			0.08170			122.47
57	52	139			0.08188			122.74

Cumulative Trial Forecast



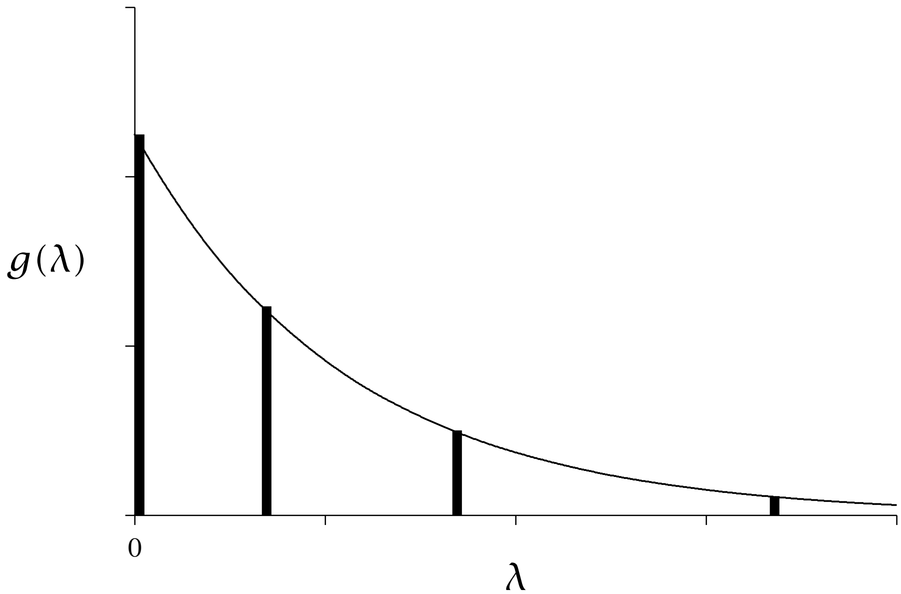
Extending the Basic Model

- The “exponential w/ never triers” model assumes all triers have the same underlying trial rate θ — a bit simplistic.
- Allow for multiple trier “segments” each with a different (latent) trial rate:

$$F(t) = \sum_{s=1}^{S} p_s F(t|\lambda_s), \quad \lambda_1 = 0, \sum_{s=1}^{S} p_s = 1$$

- Replace the discrete distribution with a continuous distribution.

Distribution of Trial Rates



Distribution of Trial Rates

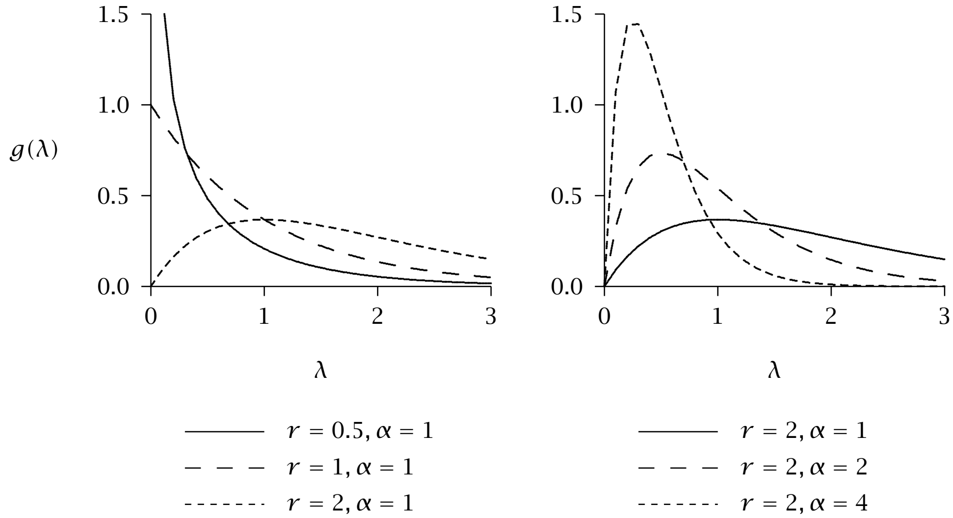
- Assume trial rates are distributed across the population according to a gamma distribution:

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

  where r is the “shape” parameter and α is the “scale” parameter.

- The gamma distribution is a flexible (unimodal) distribution …and is mathematically convenient.

Illustrative Gamma Density Functions



Alternative Market-Level Model

The cumulative distribution of time-to-trial at the market-level is given by:

$$P(T \leq t) = \int_0^\infty P(T \leq t|\lambda)\, g(\lambda)\, d\lambda$$
$$= 1 - \left(\frac{\alpha}{\alpha + t}\right)^r$$

We call this the "exponential-gamma" model.

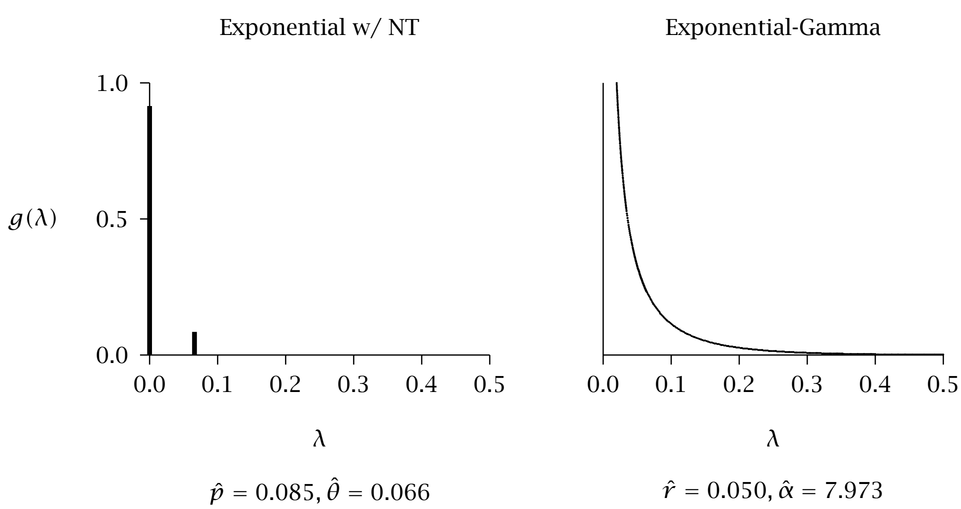
Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned}
LL(r, \alpha|\text{data}) = \; & 8 \times \ln[P(0 < T \le 1)] & + \\
& 6 \times \ln[P(1 < T \le 2)] & + \\
& \quad \ldots & + \\
& 4 \times \ln[P(23 < T \le 24)] & + \\
(1499 - 101) \times & \ln[P(T > 24)]
\end{aligned}$$

The maximum value of the log-likelihood function is $LL = -681.4$, which occurs at $\hat{r} = 0.050$ and $\hat{\alpha} = 7.973$.

Estimated Distribution of λ



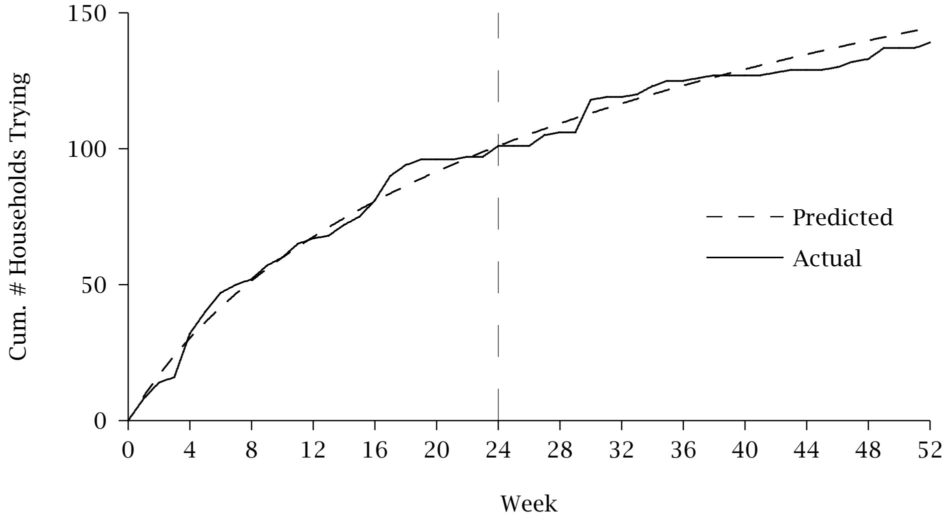
Problem 1 -- Model 2

	A	B	C	D	E	F	G	H
1	Product:	Krunchy Bits			r	1		
2	Panelists:	1499			\alpha	1		
3					LL =	=SUM(F6:F30)		
4		Cum_Trl						
5	Week	# HHs	Incr_Trl		P(T <= t)			E[T(t)]
6	1	8	=B6		=1-(F$2/(F$2+A6))^F$1	=C6*LN(E6)		=B$2*E6
7	2	14	=B7-B6		=1-(F$2/(F$2+A7))^F$1	=C7*LN(E7-E6)		=B$2*E7
8	3	16	=B8-B7		=1-(F$2/(F$2+A8))^F$1	=C8*LN(E8-E7)		=B$2*E8
9	4	32	=B9-B8		=1-(F$2/(F$2+A9))^F$1	=C9*LN(E9-E8)		=B$2*E9
10	5	40	=B10-B9		=1-(F$2/(F$2+A10))^F$1	=C10*LN(E10-E9)		=B$2*E10
11	6	47	=B11-B10		=1-(F$2/(F$2+A11))^F$1	=C11*LN(E11-E10)		=B$2*E11
12	7	50	=B12-B11		=1-(F$2/(F$2+A12))^F$1	=C12*LN(E12-E11)		=B$2*E12
13	8	52	=B13-B12		=1-(F$2/(F$2+A13))^F$1	=C13*LN(E13-E12)		=B$2*E13
14	9	57	=B14-B13		=1-(F$2/(F$2+A14))^F$1	=C14*LN(E14-E13)		=B$2*E14
15	10	60	=B15-B14		=1-(F$2/(F$2+A15))^F$1	=C15*LN(E15-E14)		=B$2*E15
16	11	65	=B16-B15		=1-(F$2/(F$2+A16))^F$1	=C16*LN(E16-E15)		=B$2*E16
17	12	67	=B17-B16		=1-(F$2/(F$2+A17))^F$1	=C17*LN(E17-E16)		=B$2*E17
18	13	68	=B18-B17		=1-(F$2/(F$2+A18))^F$1	=C18*LN(E18-E17)		=B$2*E18
19	14	72	=B19-B18		=1-(F$2/(F$2+A19))^F$1	=C19*LN(E19-E18)		=B$2*E19
20	15	75	=B20-B19		=1-(F$2/(F$2+A20))^F$1	=C20*LN(E20-E19)		=B$2*E20
21	16	81	=B21-B20		=1-(F$2/(F$2+A21))^F$1	=C21*LN(E21-E20)		=B$2*E21
22	17	90	=B22-B21		=1-(F$2/(F$2+A22))^F$1	=C22*LN(E22-E21)		=B$2*E22
23	18	94	=B23-B22		=1-(F$2/(F$2+A23))^F$1	=C23*LN(E23-E22)		=B$2*E23
24	19	96	=B24-B23		=1-(F$2/(F$2+A24))^F$1	=C24*LN(E24-E23)		=B$2*E24
25	20	96	=B25-B24		=1-(F$2/(F$2+A25))^F$1	=C25*LN(E25-E24)		=B$2*E25
26	21	96	=B26-B25		=1-(F$2/(F$2+A26))^F$1	=C26*LN(E26-E25)		=B$2*E26
27	22	97	=B27-B26		=1-(F$2/(F$2+A27))^F$1	=C27*LN(E27-E26)		=B$2*E27
28	23	97	=B28-B27		=1-(F$2/(F$2+A28))^F$1	=C28*LN(E28-E27)		=B$2*E28
29	24	101	=B29-B28		=1-(F$2/(F$2+A29))^F$1	=C29*LN(E29-E28)		=B$2*E29
30	25	101			=1-(F$2/(F$2+A30))^F$1	=(B2-B29)*LN(1-E29)		=B$2*E30
31	26	101			=1-(F$2/(F$2+A31))^F$1			=B$2*E31
32	27	105			=1-(F$2/(F$2+A32))^F$1			=B$2*E32
33	28	106			=1-(F$2/(F$2+A33))^F$1			=B$2*E33
34	29	106			=1-(F$2/(F$2+A34))^F$1			=B$2*E34
35	30	118			=1-(F$2/(F$2+A35))^F$1			=B$2*E35
36	31	119			=1-(F$2/(F$2+A36))^F$1			=B$2*E36
37	32	119			=1-(F$2/(F$2+A37))^F$1			=B$2*E37
38	33	120			=1-(F$2/(F$2+A38))^F$1			=B$2*E38
39	34	123			=1-(F$2/(F$2+A39))^F$1			=B$2*E39
40	35	125			=1-(F$2/(F$2+A40))^F$1			=B$2*E40
41	36	125			=1-(F$2/(F$2+A41))^F$1			=B$2*E41
42	37	126			=1-(F$2/(F$2+A42))^F$1			=B$2*E42
43	38	127			=1-(F$2/(F$2+A43))^F$1			=B$2*E43
44	39	127			=1-(F$2/(F$2+A44))^F$1			=B$2*E44
45	40	127			=1-(F$2/(F$2+A45))^F$1			=B$2*E45
46	41	127			=1-(F$2/(F$2+A46))^F$1			=B$2*E46
47	42	128			=1-(F$2/(F$2+A47))^F$1			=B$2*E47
48	43	129			=1-(F$2/(F$2+A48))^F$1			=B$2*E48
49	44	129			=1-(F$2/(F$2+A49))^F$1			=B$2*E49
50	45	129			=1-(F$2/(F$2+A50))^F$1			=B$2*E50
51	46	130			=1-(F$2/(F$2+A51))^F$1			=B$2*E51
52	47	132			=1-(F$2/(F$2+A52))^F$1			=B$2*E52
53	48	133			=1-(F$2/(F$2+A53))^F$1			=B$2*E53
54	49	137			=1-(F$2/(F$2+A54))^F$1			=B$2*E54
55	50	137			=1-(F$2/(F$2+A55))^F$1			=B$2*E55
56	51	137			=1-(F$2/(F$2+A56))^F$1			=B$2*E56
57	52	139			=1-(F$2/(F$2+A57))^F$1			=B$2*E57

Problem 1 -- Model 2

	A	B	C	D	E	F	G	H
1	Product:	Krunchy Bits			r	0.050245		
2	Panelists:	1499			\alpha	7.973267		
3					LL =	-681.3729		
4		Cum_Trl						
5	Week	# HHs	Incr_Trl		P(T <= t)			E[T(t)]
6	1	8	8		0.00592	-41.036		8.87
7	2	14	6		0.01118	-31.482		16.76
8	3	16	2		0.01592	-10.705		23.86
9	4	32	16		0.02022	-87.175		30.31
10	5	40	8		0.02416	-44.291		36.22
11	6	47	7		0.02780	-39.322		41.67
12	7	50	3		0.03117	-17.078		46.72
13	8	52	2		0.03431	-11.526		51.43
14	9	57	5		0.03725	-29.144		55.84
15	10	60	3		0.04002	-17.672		59.98
16	11	65	5		0.04262	-29.746		63.89
17	12	67	2		0.04509	-12.009		67.59
18	13	68	1		0.04743	-6.057		71.10
19	14	72	4		0.04966	-24.429		74.44
20	15	75	3		0.05178	-18.465		77.62
21	16	81	6		0.05381	-37.205		80.66
22	17	90	9		0.05575	-56.202		83.57
23	18	94	4		0.05761	-25.147		86.36
24	19	96	2		0.05940	-12.654		89.04
25	20	96	0		0.06112	0.000		91.62
26	21	96	0		0.06277	0.000		94.10
27	22	97	1		0.06437	-6.440		96.49
28	23	97	0		0.06591	0.000		98.80
29	24	101	4		0.06740	-26.036		101.04
30	25	101			0.06884	-97.554		103.20
31	26	101			0.07024			105.29
32	27	105			0.07159			107.32
33	28	106			0.07291			109.29
34	29	106			0.07419			111.20
35	30	118			0.07543			113.06
36	31	119			0.07663			114.87
37	32	119			0.07781			116.63
38	33	120			0.07895			118.35
39	34	123			0.08007			120.02
40	35	125			0.08115			121.65
41	36	125			0.08222			123.24
42	37	126			0.08325			124.80
43	38	127			0.08426			126.31
44	39	127			0.08525			127.80
45	40	127			0.08622			129.25
46	41	127			0.08717			130.67
47	42	128			0.08810			132.05
48	43	129			0.08900			133.42
49	44	129			0.08989			134.75
50	45	129			0.09076			136.05
51	46	130			0.09162			137.33
52	47	132			0.09245			138.59
53	48	133			0.09328			139.82
54	49	137			0.09408			141.03
55	50	137			0.09487			142.22
56	51	137			0.09565			143.38
57	52	139			0.09641			144.53

Cumulative Trial Forecast



Further Model Extensions

- Combine a “never triers” term with the “exponential-gamma” model.
- Incorporate the effects of marketing covariates.
- Model repeat sales using a “depth of repeat” formulation, where transitions from one repeat class to the next are modeled using an “exponential-gamma”-type model.

Concepts and Tools Introduced

- Probability models
- (Single-event) timing processes
- Models of new product trial/adoption

Further Reading

Hardie, Bruce G. S., Peter S. Fader, and Michael Wisniewski (1998), "An Empirical Comparison of New Product Trial Forecasting Models," *Journal of Forecasting*, **17** (June–July), 209–29.

Fader, Peter S., Bruce G. S. Hardie, and Robert Zeithammer (2002), "Forecasting New Product Trial in a Controlled Test Market Environment," unpublished working paper. (Available at `http://brucehardie.com/`)

Kalbfleisch, John D. and Ross L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: Wiley.

Introduction to Probability Models

The Logic of Probability Models

- Many researchers attempt to describe/predict behavior using observed variables.
- However, they still use random components in recognition that not all factors are included in the model.
- We treat behavior as if it were “random” (probabilistic, stochastic).
- We propose a model of individual-level behavior which is “summed” across individuals (taking individual differences into account) to obtain a model of aggregate behavior.

Uses of Probability Models

- Understanding market-level behavior patterns
- Prediction
  - To settings (e.g., time periods) beyond the observation period
  - Conditional on past behavior
- Profiling behavioral propensities of individuals
- Benchmarks/norms

Building a Probability Model

(i) Determine the marketing decision problem/ information needed.

(ii) Identify the *observable* individual-level behavior of interest.

- We denote this by x.

(iii) Select a probability distribution that characterizes this individual-level behavior.

- This is denoted by $f(x|\theta)$.
- We view the parameters of this distribution as individual-level *latent traits*.

Building a Probability Model

(iv) Specify a distribution to characterize the distribution of the latent trait variable(s) across the population.

- We denote this by $g(\theta)$.
- This is often called the *mixing distribution*.

(v) Derive the corresponding *aggregate* or *observed* distribution for the behavior of interest:

$$f(x) = \int f(x|\theta)g(\theta)\,d\theta$$

Building a Probability Model

(vi) Estimate the parameters (of the mixing distribution) by fitting the aggregate distribution to the observed data.

(vii) Use the model to solve the marketing decision problem/provide the required information.

Outline

- Problem 1: Predicting New Product Trial
  (Modeling Timing Data)
- Problem 2: Estimating Billboard Exposures
  (Modeling Count Data)
- Problem 3: Test/Roll Decisions in Segmentation-based Direct Marketing
  (Modeling "Choice" Data)
- Problem 4: Characterizing the Purchasing of Hard-Candy
  (Introduction to Finite Mixture Models)
- Problem 5: Who is Visiting khakichinos.com?
  (Incorporating Covariates in Count Models)
- Case Study: Forecasting Repeat Sales at CDNOW

Problem 2:
Estimating Billboard Exposures

(Modeling Count Data)

Background

One advertising medium at the marketer's disposal is the outdoor billboard. The unit of purchase for this medium is usually a "monthly showing," which comprises a specific set of billboards carrying the advertiser's message in a given market.

The effectiveness of a monthly showing is evaluated in terms of three measures: reach, (average) frequency, and gross rating points (GRPs). These measures are determined using data collected from a sample of people in the market.

Respondents record their daily travel on maps. From each respondent's travel map, the total frequency of exposure to the showing over the survey period is counted. An "exposure" is deemed to occur each time the respondent travels by a billboard in the showing, on the street or road closest to that billboard, going towards the billboard's face.

Background

The standard approach to data collection requires each respondent to fill out daily travel maps for *an entire month*. The problem with this is that it is difficult and expensive to get a high proportion of respondents to do this accurately.

B&P Research is interested in developing a means by which it can generate effectiveness measures for a monthly showing from a survey in which respondents fill out travel maps for *only one week*.

Data have been collected from a sample of 250 residents who completed daily travel maps for one week. The sampling process is such that approximately one quarter of the respondents fill out travel maps during each of the four weeks in the target month.

Effectiveness Measures

The effectiveness of a monthly showing is evaluated in terms of three measures:

- Reach: the proportion of the population exposed to the billboard message at least once in the month.
- Average Frequency: the average number of exposures (per month) among those people reached.
- Gross Rating Points (GRPs): the mean number of exposures per 100 people.

Distribution of Billboard Exposures (1 week)

# Exposures	# People	# Exposures	# People
0	48	12	5
1	37	13	3
2	30	14	3
3	24	15	2
4	20	16	2
5	16	17	2
6	13	18	1
7	11	19	1
8	9	20	2
9	7	21	1
10	6	22	1
11	5	23	1

Average # Exposures = 4.456

Modeling Objective

Develop a model that enables us to estimate a billboard showing's reach, average frequency, and GRPs for the month using the one-week data.

Modeling Issues

- Modeling the exposures to showing in a week.
- Estimating summary statistics of the exposure distribution for a longer period of time (i.e., one month).

Modeling One Week Exposures

- Let the random variable X denote the number of exposures to the showing in a week.
- At the individual-level, X is assumed to be Poisson distributed with (exposure) rate parameter λ:

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

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

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

Modeling One Week Exposures

- The distribution of exposures at the population-level is given by:

$$\begin{aligned} P(X = x) &= \int_0^\infty P(X = x|\lambda)\, g(\lambda)\, d\lambda \\ &= \frac{\Gamma(r + x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha + 1}\right)^r \left(\frac{1}{\alpha + 1}\right)^x \end{aligned}$$

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

- The mean of the NBD is given by $E(X) = r/\alpha$.

Computing NBD Probabilities

- Note that

$$\frac{P(X = x)}{P(X = x - 1)} = \frac{r + x - 1}{x(\alpha + 1)}$$

- We can therefore compute NBD probabilities using the following *forward recursion* formula:

$$P(X = x) = \begin{cases} \left(\frac{\alpha}{\alpha + 1}\right)^r & x = 0 \\ \frac{r + x - 1}{x(\alpha + 1)} \times P(X = x - 1) & x \ge 1 \end{cases}$$

Estimating Model Parameters

The log-likelihood function is defined as:

$$\begin{aligned} LL(r, \alpha | \text{data}) = & \; 48 \times \ln[P(X = 0)] \; + \\ & \; 37 \times \ln[P(X = 1)] \; + \\ & \; 30 \times \ln[P(X = 2)] \; + \\ & \; \ldots \; + \\ & \; 1 \times \ln[P(X = 23)] \end{aligned}$$

The maximum value of the log-likelihood function is $LL = -649.7$, which occurs at $\hat{r} = 0.969$ and $\hat{\alpha} = 0.218$.

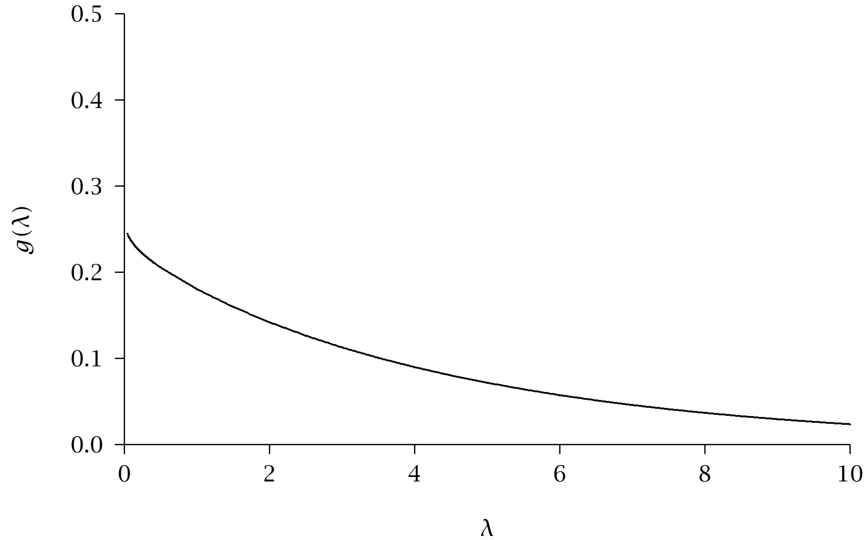
Problem 2 -- Parameter Estimation

	A	B	C	D
1	r	1		
2	\alpha	1	LL=	=SUM(D5:D28)
3				
4	x	f_x	P(X=x)	
5	0	48	=(B2/(B2+1))^B1	=B5*LN(C5)
6	1	37	=C5*(B$1+A6-1)/(A6*(B$2+1))	=B6*LN(C6)
7	2	30	=C6*(B$1+A7-1)/(A7*(B$2+1))	=B7*LN(C7)
8	3	24	=C7*(B$1+A8-1)/(A8*(B$2+1))	=B8*LN(C8)
9	4	20	=C8*(B$1+A9-1)/(A9*(B$2+1))	=B9*LN(C9)
10	5	16	=C9*(B$1+A10-1)/(A10*(B$2+1))	=B10*LN(C10)
11	6	13	=C10*(B$1+A11-1)/(A11*(B$2+1))	=B11*LN(C11)
12	7	11	=C11*(B$1+A12-1)/(A12*(B$2+1))	=B12*LN(C12)
13	8	9	=C12*(B$1+A13-1)/(A13*(B$2+1))	=B13*LN(C13)
14	9	7	=C13*(B$1+A14-1)/(A14*(B$2+1))	=B14*LN(C14)
15	10	6	=C14*(B$1+A15-1)/(A15*(B$2+1))	=B15*LN(C15)
16	11	5	=C15*(B$1+A16-1)/(A16*(B$2+1))	=B16*LN(C16)
17	12	5	=C16*(B$1+A17-1)/(A17*(B$2+1))	=B17*LN(C17)
18	13	3	=C17*(B$1+A18-1)/(A18*(B$2+1))	=B18*LN(C18)
19	14	3	=C18*(B$1+A19-1)/(A19*(B$2+1))	=B19*LN(C19)
20	15	2	=C19*(B$1+A20-1)/(A20*(B$2+1))	=B20*LN(C20)
21	16	2	=C20*(B$1+A21-1)/(A21*(B$2+1))	=B21*LN(C21)
22	17	2	=C21*(B$1+A22-1)/(A22*(B$2+1))	=B22*LN(C22)
23	18	1	=C22*(B$1+A23-1)/(A23*(B$2+1))	=B23*LN(C23)
24	19	1	=C23*(B$1+A24-1)/(A24*(B$2+1))	=B24*LN(C24)
25	20	2	=C24*(B$1+A25-1)/(A25*(B$2+1))	=B25*LN(C25)
26	21	1	=C25*(B$1+A26-1)/(A26*(B$2+1))	=B26*LN(C26)
27	22	1	=C26*(B$1+A27-1)/(A27*(B$2+1))	=B27*LN(C27)
28	23	1	=C27*(B$1+A28-1)/(A28*(B$2+1))	=B28*LN(C28)

Problem 2 -- Parameter Estimation

	A	B	C	D
1	r	0.96926		
2	\alpha	0.21752	LL=	-649.6888
3				
4	x	f_x	P(X=x)	
5	0	48	0.18837	-80.128
6	1	37	0.14996	-70.203
7	2	30	0.12128	-63.291
8	3	24	0.09859	-55.603
9	4	20	0.08035	-50.427
10	5	16	0.06559	-43.589
11	6	13	0.05360	-38.041
12	7	11	0.04383	-34.402
13	8	9	0.03586	-29.953
14	9	7	0.02935	-24.699
15	10	6	0.02403	-22.370
16	11	5	0.01969	-19.639
17	12	5	0.01613	-20.636
18	13	3	0.01321	-12.979
19	14	3	0.01083	-13.576
20	15	2	0.00888	-9.449
21	16	2	0.00728	-9.846
22	17	2	0.00597	-10.243
23	18	1	0.00489	-5.320
24	19	1	0.00401	-5.519
25	20	2	0.00329	-11.434
26	21	1	0.00270	-5.915
27	22	1	0.00221	-6.113
28	23	1	0.00182	-6.312

Estimated Distribution of λ



$\hat{r} = 0.969, \hat{\alpha} = 0.218$

NBD for a Non-Unit Time Period

- Let $X(t)$ be the number of exposures occurring in an observation period of length t time units.
- If, for a unit time period, the distribution of exposures *at the individual-level* is distributed Poisson with rate parameter λ, then $X(t)$ has a Poisson distribution with rate parameter λt:

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

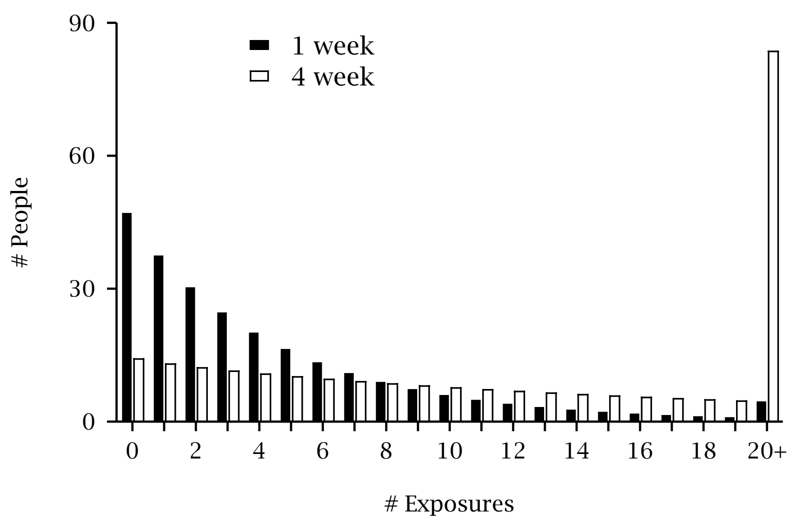
NBD for a Non-Unit Time Period

- The distribution of exposures at the population-level is given by:

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

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

Exposure Distributions: 1 week vs. 4 week



Effectiveness of Monthly Showing

- For $t = 4$, we have:
  - $P(X(t) = 0) = 0.056$, and
  - $E[X(t)] = 17.82$
- It follows that:
  - Reach $= 1 - P(X(t) = 0)$
    $= 94.4\%$
  - Frequency $= E[X(t)]/(1 - P(X(t) = 0))$
    $= 18.9$
  - GRPs $= 100 \times E[X(t)]$
    $= 1782$

Concepts and Tools Introduced

- Counting processes
- The NBD model
- Extrapolating an observed histogram over time
- Using models to estimate "exposure distributions" for media vehicles

Problem 2 -- Solution

	A	B
1	r	='Parameter Estimation'!B1
2	\alpha	='Parameter Estimation'!B2
3	t	4
4		
5	P(X(t)=0)	=(B2/(B2+B3))^B1
6	E[X(t)]	=B1*B3/B2
7		
8	Reach	=1-B5
9	Frequency	=B6/B8
10	GRPs	=100*B6

Problem 2 -- Solution

	A	B
1	r	0.96926
2	\alpha	0.21752
3	t	4
4		
5	P(X(t)=0)	0.056
6	E[X(t)]	17.82
7		
8	Reach	94.4%
9	Frequency	18.9
10	GRPs	1782

Further Reading

Greene, Jerome D. (1982), *Consumer Behavior Models for Non-Statisticians*, New York: Praeger.

Morrison, Donald G. and David C. Schmittlein (1988), "Generalizing the NBD Model for Customer Purchases: What Are the Implications and Is It Worth the Effort?" *Journal of Business and Economic Statistics*, **6** (April), 145–59.

Ehrenberg, A. S. C. (1988), *Repeat-Buying*, 2nd edn., London: Charles Griffin & Company, Ltd. (Available online at <`http://www.empgens.com/ehrenberg.html#repeat`>.)

Problem 3:
Test/Roll Decisions in Segmentation-based Direct Marketing

(Modeling "Choice" Data)

The "Segmentation" Approach

1. Divide the customer list into a set of (homogeneous) segments.
2. Test customer response by mailing to a random sample of each segment.
3. Rollout to segments with a response rate (RR) above some cut-off point,

$$\text{e.g., RR} > \frac{\text{cost of each mailing}}{\text{unit margin}}$$

Ben's Knick Knacks, Inc.

- A consumer durable product (unit margin = $161.50, mailing cost per 10,000 = $3343)
- 126 segments formed from customer database on the basis of past purchase history information
- Test mailing to 3.24% of database

Ben's Knick Knacks, Inc.

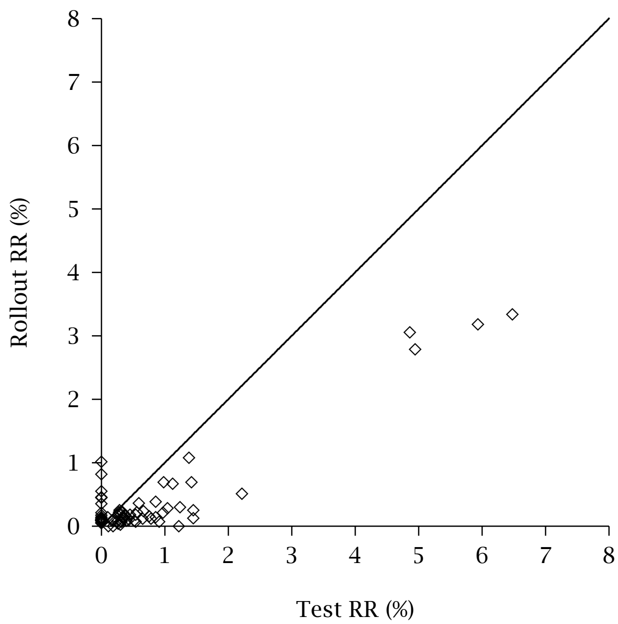
Standard approach:

- Rollout to all segments with

$$\text{Test RR} > \frac{3343/10,000}{161.50} = 0.00207$$

- 51 segments pass this hurdle

Test vs. Actual Response Rate



Modeling Objective

Develop a model that leverages the whole data set to make better informed decisions.

Model Development

Notation:

$$N_s = \text{size of segment } s \ (s = 1, \ldots, S)$$
$$m_s = \# \text{ members of segment } s \text{ tested}$$
$$X_s = \# \text{ responses to test in segment } s$$

Assume: All members of segment s have the same (unknown) response probability $p_s \Rightarrow X_s$ is a binomial random variable

$$P(X_s = x_s | m_s, p_s) = \binom{m_s}{x_s} p_s^{x_s} (1 - p_s)^{m_s - x_s}$$

Distribution of Response Probabilities

- Heterogeneity in p_s is captured using a beta distribution:

$$g(p_s) = \frac{1}{B(\alpha, \beta)} p_s^{\alpha-1}(1 - p_s)^{\beta-1}$$

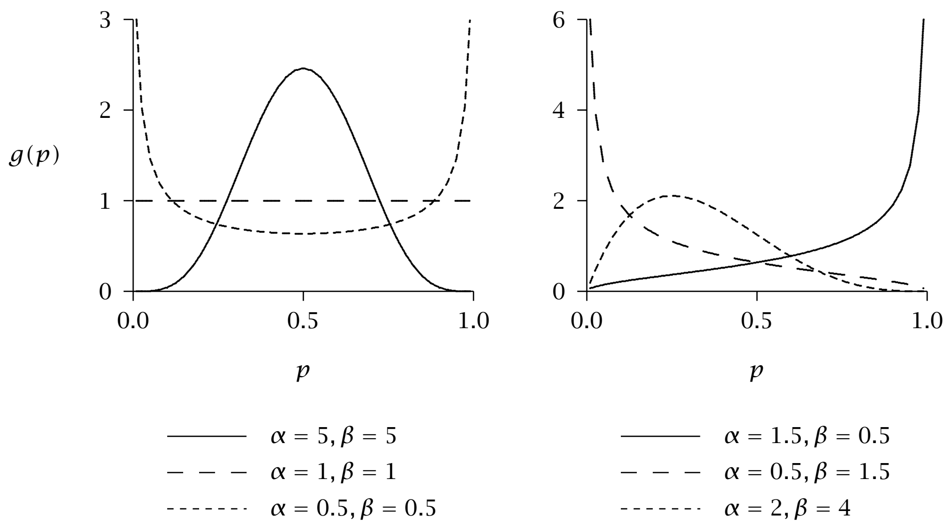
- The beta function, $B(\alpha, \beta)$, can be expressed as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

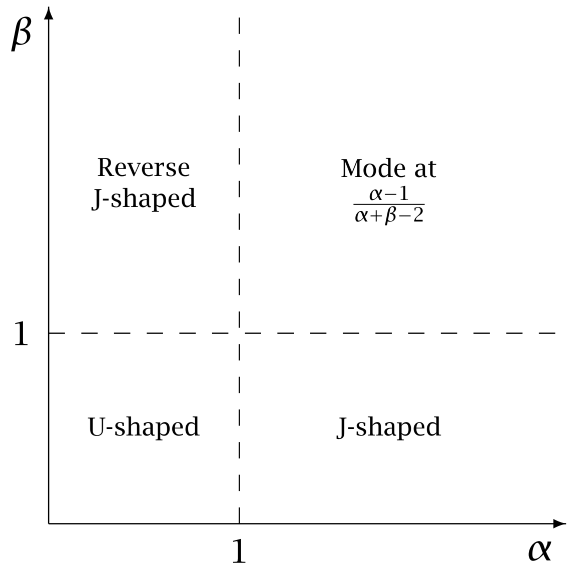
- The mean of the beta distribution is given by

$$E(p_s) = \frac{\alpha}{\alpha + \beta}$$

Illustrative Beta Density Functions



Shape of the Beta Density



The Beta Binomial Model

The aggregate distribution of responses to a mailing of size m_s is given by

$$\begin{aligned} P(X_s = x_s | m_s) &= \int_0^1 P(X_s = x_s | m_s, p_s)\, g(p_s)\, dp_s \\ &= \binom{m_s}{x_s} \frac{B(\alpha + x_s, \beta + m_s - x_s)}{B(\alpha, \beta)} \end{aligned}$$

Estimating Model Parameters

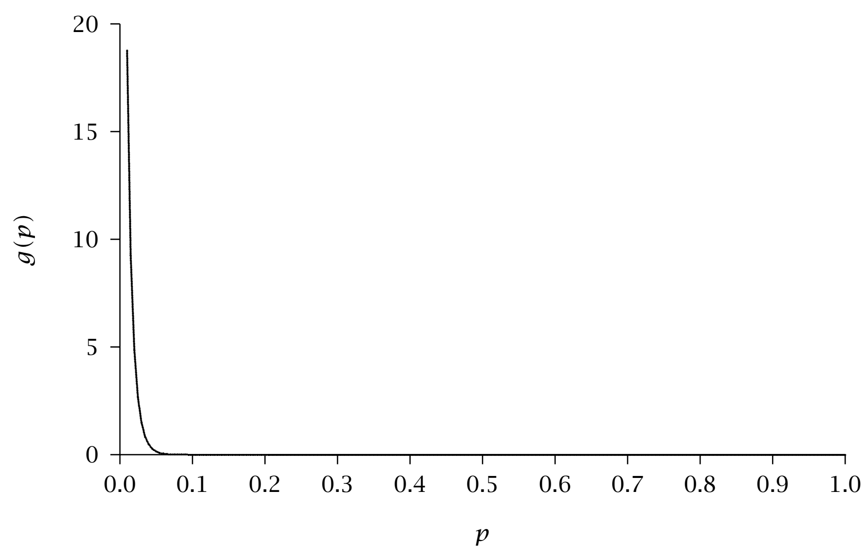
The log-likelihood function is defined as:

$$LL(\alpha, \beta|\text{data}) = \sum_{s=1}^{126} \ln[P(X_s = x_s|m_s)]$$

$$= \sum_{s=1}^{126} \ln\Bigg[\frac{m_s!}{(m_s - x_s)!\,x_s!} \underbrace{\frac{\Gamma(\alpha + x_s)\Gamma(\beta + m_s - x_s)}{\Gamma(\alpha + \beta + m_s)}}_{B(\alpha+x_s,\beta+m_s-x_s)} \underbrace{\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{1/B(\alpha,\beta)}\Bigg]$$

The maximum value of the log-likelihood function is $LL = -200.5$, which occurs at $\hat{\alpha} = 0.439$ and $\hat{\beta} = 95.411$.

Estimated Distribution of *p*



$\hat{\alpha} = 0.439, \hat{\beta} = 95.411, \bar{p} = 0.0046$

Problem 3 -- Model (a)

	A	B	C	D	E	F
1	\alpha	1		B(\alpha,\beta)	=EXP(GAMMALN(B1)+GAMMALN(B2)-GAMMALN(B1+B2))	
2	\beta	1				
3					LL =	=SUM(F6:F131)
4						
5	Segment	m_s	x_s		P(X=x|m)	
6	1	34	0		=COMBIN(B6,C6)*EXP(GAMMALN(B$1+C6)+GAMMALN(B$2+B6-C6)-GAMMALN(B$1+B$2+B6))/E$1	=LN(E6)
7	2	102	1		=COMBIN(B7,C7)*EXP(GAMMALN(B$1+C7)+GAMMALN(B$2+B7-C7)-GAMMALN(B$1+B$2+B7))/E$1	=LN(E7)
8	3	53	0		=COMBIN(B8,C8)*EXP(GAMMALN(B$1+C8)+GAMMALN(B$2+B8-C8)-GAMMALN(B$1+B$2+B8))/E$1	=LN(E8)
9	4	145	2		=COMBIN(B9,C9)*EXP(GAMMALN(B$1+C9)+GAMMALN(B$2+B9-C9)-GAMMALN(B$1+B$2+B9))/E$1	=LN(E9)
10	5	1254	62		=COMBIN(B10,C10)*EXP(GAMMALN(B$1+C10)+GAMMALN(B$2+B10-C10)-GAMMALN(B$1+B$2+B10))/E$1	=LN(E10)
11	6	144	7		=COMBIN(B11,C11)*EXP(GAMMALN(B$1+C11)+GAMMALN(B$2+B11-C11)-GAMMALN(B$1+B$2+B11))/E$1	=LN(E11)
12	7	1235	80		=COMBIN(B12,C12)*EXP(GAMMALN(B$1+C12)+GAMMALN(B$2+B12-C12)-GAMMALN(B$1+B$2+B12))/E$1	=LN(E12)
13	8	573	34		=COMBIN(B13,C13)*EXP(GAMMALN(B$1+C13)+GAMMALN(B$2+B13-C13)-GAMMALN(B$1+B$2+B13))/E$1	=LN(E13)
14	9	1083	24		=COMBIN(B14,C14)*EXP(GAMMALN(B$1+C14)+GAMMALN(B$2+B14-C14)-GAMMALN(B$1+B$2+B14))/E$1	=LN(E14)
15	10	352	5		=COMBIN(B15,C15)*EXP(GAMMALN(B$1+C15)+GAMMALN(B$2+B15-C15)-GAMMALN(B$1+B$2+B15))/E$1	=LN(E15)
16	11	817	7		=COMBIN(B16,C16)*EXP(GAMMALN(B$1+C16)+GAMMALN(B$2+B16-C16)-GAMMALN(B$1+B$2+B16))/E$1	=LN(E16)
17	12	118	0		=COMBIN(B17,C17)*EXP(GAMMALN(B$1+C17)+GAMMALN(B$2+B17-C17)-GAMMALN(B$1+B$2+B17))/E$1	=LN(E17)
18	13	1049	3		=COMBIN(B18,C18)*EXP(GAMMALN(B$1+C18)+GAMMALN(B$2+B18-C18)-GAMMALN(B$1+B$2+B18))/E$1	=LN(E18)
19	14	452	3		=COMBIN(B19,C19)*EXP(GAMMALN(B$1+C19)+GAMMALN(B$2+B19-C19)-GAMMALN(B$1+B$2+B19))/E$1	=LN(E19)
20	15	338	2		=COMBIN(B20,C20)*EXP(GAMMALN(B$1+C20)+GAMMALN(B$2+B20-C20)-GAMMALN(B$1+B$2+B20))/E$1	=LN(E20)
21	16	168	0		=COMBIN(B21,C21)*EXP(GAMMALN(B$1+C21)+GAMMALN(B$2+B21-C21)-GAMMALN(B$1+B$2+B21))/E$1	=LN(E21)
22	17	242	3		=COMBIN(B22,C22)*EXP(GAMMALN(B$1+C22)+GAMMALN(B$2+B22-C22)-GAMMALN(B$1+B$2+B22))/E$1	=LN(E22)
23	18	185	1		=COMBIN(B23,C23)*EXP(GAMMALN(B$1+C23)+GAMMALN(B$2+B23-C23)-GAMMALN(B$1+B$2+B23))/E$1	=LN(E23)
24	19	116	0		=COMBIN(B24,C24)*EXP(GAMMALN(B$1+C24)+GAMMALN(B$2+B24-C24)-GAMMALN(B$1+B$2+B24))/E$1	=LN(E24)
25	20	69	1		=COMBIN(B25,C25)*EXP(GAMMALN(B$1+C25)+GAMMALN(B$2+B25-C25)-GAMMALN(B$1+B$2+B25))/E$1	=LN(E25)
26	21	193	1		=COMBIN(B26,C26)*EXP(GAMMALN(B$1+C26)+GAMMALN(B$2+B26-C26)-GAMMALN(B$1+B$2+B26))/E$1	=LN(E26)
27	22	82	1		=COMBIN(B27,C27)*EXP(GAMMALN(B$1+C27)+GAMMALN(B$2+B27-C27)-GAMMALN(B$1+B$2+B27))/E$1	=LN(E27)
28	23	265	1		=COMBIN(B28,C28)*EXP(GAMMALN(B$1+C28)+GAMMALN(B$2+B28-C28)-GAMMALN(B$1+B$2+B28))/E$1	=LN(E28)
29	24	171	0		=COMBIN(B29,C29)*EXP(GAMMALN(B$1+C29)+GAMMALN(B$2+B29-C29)-GAMMALN(B$1+B$2+B29))/E$1	=LN(E29)
30	25	1554	7		=COMBIN(B30,C30)*EXP(GAMMALN(B$1+C30)+GAMMALN(B$2+B30-C30)-GAMMALN(B$1+B$2+B30))/E$1	=LN(E30)
31	26	1339	4		=COMBIN(B31,C31)*EXP(GAMMALN(B$1+C31)+GAMMALN(B$2+B31-C31)-GAMMALN(B$1+B$2+B31))/E$1	=LN(E31)
32	27	1167	4		=COMBIN(B32,C32)*EXP(GAMMALN(B$1+C32)+GAMMALN(B$2+B32-C32)-GAMMALN(B$1+B$2+B32))/E$1	=LN(E32)
33	28	621	2		=COMBIN(B33,C33)*EXP(GAMMALN(B$1+C33)+GAMMALN(B$2+B33-C33)-GAMMALN(B$1+B$2+B33))/E$1	=LN(E33)
34	29	1013	1		=COMBIN(B34,C34)*EXP(GAMMALN(B$1+C34)+GAMMALN(B$2+B34-C34)-GAMMALN(B$1+B$2+B34))/E$1	=LN(E34)
35	30	544	1		=COMBIN(B35,C35)*EXP(GAMMALN(B$1+C35)+GAMMALN(B$2+B35-C35)-GAMMALN(B$1+B$2+B35))/E$1	=LN(E35)
36	31	731	1		=COMBIN(B36,C36)*EXP(GAMMALN(B$1+C36)+GAMMALN(B$2+B36-C36)-GAMMALN(B$1+B$2+B36))/E$1	=LN(E36)
37	32	326	0		=COMBIN(B37,C37)*EXP(GAMMALN(B$1+C37)+GAMMALN(B$2+B37-C37)-GAMMALN(B$1+B$2+B37))/E$1	=LN(E37)
38	33	772	1		=COMBIN(B38,C38)*EXP(GAMMALN(B$1+C38)+GAMMALN(B$2+B38-C38)-GAMMALN(B$1+B$2+B38))/E$1	=LN(E38)
39	34	335	1		=COMBIN(B39,C39)*EXP(GAMMALN(B$1+C39)+GAMMALN(B$2+B39-C39)-GAMMALN(B$1+B$2+B39))/E$1	=LN(E39)
40	35	235	0		=COMBIN(B40,C40)*EXP(GAMMALN(B$1+C40)+GAMMALN(B$2+B40-C40)-GAMMALN(B$1+B$2+B40))/E$1	=LN(E40)
41	36	218	0		=COMBIN(B41,C41)*EXP(GAMMALN(B$1+C41)+GAMMALN(B$2+B41-C41)-GAMMALN(B$1+B$2+B41))/E$1	=LN(E41)
42	37	221	0		=COMBIN(B42,C42)*EXP(GAMMALN(B$1+C42)+GAMMALN(B$2+B42-C42)-GAMMALN(B$1+B$2+B42))/E$1	=LN(E42)
43	38	103	1		=COMBIN(B43,C43)*EXP(GAMMALN(B$1+C43)+GAMMALN(B$2+B43-C43)-GAMMALN(B$1+B$2+B43))/E$1	=LN(E43)
44	39	170	0		=COMBIN(B44,C44)*EXP(GAMMALN(B$1+C44)+GAMMALN(B$2+B44-C44)-GAMMALN(B$1+B$2+B44))/E$1	=LN(E44)
45	40	45	0		=COMBIN(B45,C45)*EXP(GAMMALN(B$1+C45)+GAMMALN(B$2+B45-C45)-GAMMALN(B$1+B$2+B45))/E$1	=LN(E45)

Problem 3 -- Model

	A	B	C	D	E	F	G	H	I
1	\alpha	0.439	B(\alpha,\beta)		0.273				
2	\beta	95.411							
3					LL =	-200.548		cutoff	0.00207
4									
5	Segment	m_s	x_s		P(X=x|m)			E[p_s|x_s]	Roll?
6	1	34	0		0.87448	-0.134		0.00338	Y
7	2	102	1		0.16556	-1.798		0.00727	Y
8	3	53	0		0.82334	-0.194		0.00295	Y
9	4	145	2		0.07694	-2.565		0.01013	Y
10	5	1254	62		0.00015	-8.793		0.04626	Y
11	6	144	7		0.00301	-5.805		0.03101	Y
12	7	1235	80		0.00003	-10.403		0.06044	Y
13	8	573	34		0.00014	-8.869		0.05149	Y
14	9	1083	24		0.00362	-5.622		0.02073	Y
15	10	352	5		0.03010	-3.503		0.01214	Y
16	11	817	7		0.02810	-3.572		0.00815	Y
17	12	118	0		0.70182	-0.354		0.00205	N
18	13	1049	3		0.06653	-2.710		0.00300	Y
19	14	452	3		0.06735	-2.698		0.00628	Y
20	15	338	2		0.09913	-2.311		0.00562	Y
21	16	168	0		0.63981	-0.447		0.00166	N
22	17	242	3		0.05465	-2.907		0.01018	Y
23	18	185	1		0.18091	-1.710		0.00512	Y
24	19	116	0		0.70473	-0.350		0.00207	Y
25	20	69	1		0.14588	-1.925		0.00873	Y
26	21	193	1		0.18122	-1.708		0.00498	Y
27	22	82	1		0.15531	-1.862		0.00809	Y
28	23	265	1		0.18042	-1.712		0.00399	Y
29	24	171	0		0.63664	-0.452		0.00164	N
30	25	1554	7		0.03089	-3.477		0.00451	Y
31	26	1339	4		0.05107	-2.975		0.00309	Y
32	27	1167	4		0.05197	-2.957		0.00352	Y
33	28	621	2		0.09808	-2.322		0.00340	Y
34	29	1013	1		0.13667	-1.990		0.00130	N
35	30	544	1		0.16210	-1.820		0.00225	Y
36	31	731	1		0.15052	-1.894		0.00174	N
37	32	326	0		0.52048	-0.653		0.00104	N
38	33	772	1		0.14826	-1.909		0.00166	N
39	34	335	1		0.17658	-1.734		0.00334	Y
40	35	235	0		0.57918	-0.546		0.00133	N
41	36	218	0		0.59277	-0.523		0.00140	N
42	37	221	0		0.59030	-0.527		0.00139	N
43	38	103	1		0.16596	-1.796		0.00724	Y
44	39	170	0		0.63769	-0.450		0.00165	N
45	40	45	0		0.84365	-0.170		0.00312	Y
46	41	237	0		0.57764	-0.549		0.00132	N
47	42	86	0		0.75377	-0.283		0.00241	Y
48	43	297	1		0.17887	-1.721		0.00366	Y
49	44	415	0		0.47847	-0.737		0.00086	N
50	45	187	0		0.62053	-0.477		0.00155	N
51	46	248	0		0.56944	-0.563		0.00128	N

Applying the Model

What is our best guess of p_s given a response of x_s to a test mailing of size m_s?

Intuitively, we would expect each segment's true mean to be a weighted-average of

- the population average (our best guess with no data), and
- the observed response rate.

$$E(p_s|x_s, m_s) \approx \omega \frac{\alpha}{\alpha + \beta} + (1 - \omega) \frac{x_s}{m_s}$$

Bayes Theorem

- The *prior distribution* $g(p)$ captures the possible values p can take on, prior to collecting any information about the specific individual.
- The *posterior distribution* $g(p|x)$ is the conditional distribution of p, given the observed data x. It represents our updated opinion about the possible values p can take on, now that we have some information x about the specific individual.
- According to Bayes theorem:

$$g(p|x) = \frac{f(x|p)g(p)}{\int f(x|p)g(p)\,dp}$$

Bayes Theorem

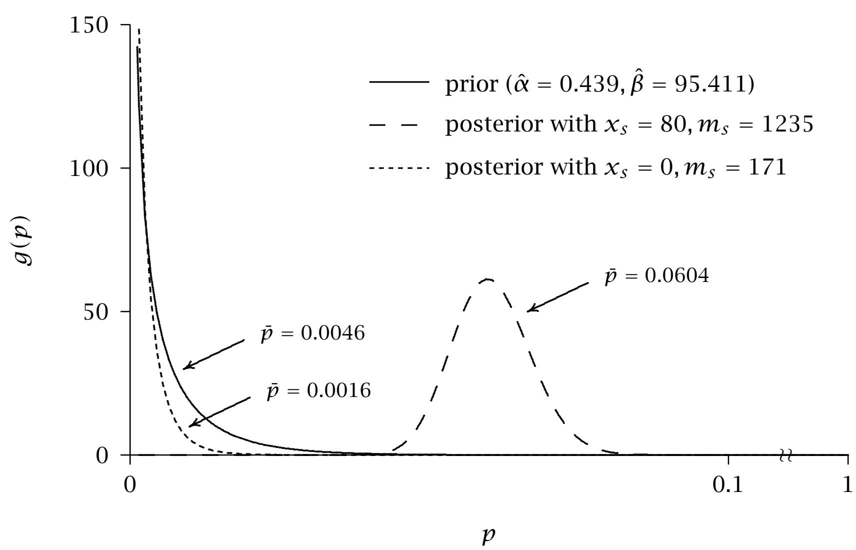
For the beta-binomial model, we have:

$$g(p_s | X_s = x_s, m_s) = \frac{\overbrace{P(X_s = x_s | m_s, p_s)}^{\text{binomial}} \overbrace{g(p_s)}^{\text{beta}}}{\underbrace{\int_0^1 P(X_s = x_s | m_s, p_s)\, g(p_s)\, dp_s}_{\text{beta-binomial}}}$$

$$= \frac{1}{B(\alpha + x_s, \beta + m_s - x_s)} p_s^{\alpha + x_s - 1} (1 - p_s)^{\beta + m_s - x_s - 1}$$

which is a beta distribution with parameters $\alpha + x_s$ and $\beta + m_s - x_s$.

Distribution of p



Applying the Model

Recall that the mean of the beta distribution is $\alpha/(\alpha+\beta)$. Therefore

$$E(p_s|X_s = x_s, m_s) = \frac{\alpha + x_s}{\alpha + \beta + m_s}$$

which can be written as

$$\left(\frac{\alpha+\beta}{\alpha+\beta+m_s}\right)\frac{\alpha}{\alpha+\beta} + \left(\frac{m_s}{\alpha+\beta+m_s}\right)\frac{x_s}{m_s}$$

- a weighted average of the test RR (x_s/m_s) and the population mean ($\alpha/(\alpha+\beta)$).
- "Regressing the test RR to the mean"

Model-Based Decision Rule

- Rollout to segments with:

$$E(p_s|X_s = x_s, m_s) > \frac{3343/10,000}{161.5} = 0.00207$$

- 66 segments pass this hurdle
- To test this model, we compare model predictions with managers' actions. (We also examine the performance of the "standard" approach.)

Problem 3 -- Model (b)

<table>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th><th>F</th><th>G</th><th>H</th><th>I</th></tr>
<tr><td>1</td><td>\alpha</td><td>0.439</td><td></td><td>B(\alpha,\beta)</td><td>0.2733</td><td></td><td></td><td></td><td></td></tr>
<tr><td>2</td><td>\beta</td><td>95.411</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td></td><td></td><td></td><td></td><td>LL =</td><td>-200.548</td><td></td><td>cutoff</td><td>=(3343/10000)/161.5</td></tr>
<tr><td>4</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>Segment</td><td>m_s</td><td>x_s</td><td></td><td>P(X=x|m)</td><td></td><td></td><td>E[p_s|x_s]</td><td>Roll?</td></tr>
<tr><td>6</td><td>1</td><td>34</td><td>0</td><td></td><td>0.8745</td><td>-0.1341</td><td></td><td>=(B$1+C6)/(B$1+B$2+B6)</td><td>=IF(H6>=I$3,"Y","N")</td></tr>
<tr><td>7</td><td>2</td><td>102</td><td>1</td><td></td><td>0.1656</td><td>-1.7984</td><td></td><td>=(B$1+C7)/(B$1+B$2+B7)</td><td>=IF(H7>=I$3,"Y","N")</td></tr>
<tr><td>8</td><td>3</td><td>53</td><td>0</td><td></td><td>0.8233</td><td>-0.1944</td><td></td><td>=(B$1+C8)/(B$1+B$2+B8)</td><td>=IF(H8>=I$3,"Y","N")</td></tr>
<tr><td>9</td><td>4</td><td>145</td><td>2</td><td></td><td>0.0769</td><td>-2.5647</td><td></td><td>=(B$1+C9)/(B$1+B$2+B9)</td><td>=IF(H9>=I$3,"Y","N")</td></tr>
<tr><td>10</td><td>5</td><td>1254</td><td>62</td><td></td><td>0.0002</td><td>-8.7933</td><td></td><td>=(B$1+C10)/(B$1+B$2+B10)</td><td>=IF(H10>=I$3,"Y","N")</td></tr>
<tr><td>11</td><td>6</td><td>144</td><td>7</td><td></td><td>0.003</td><td>-5.8046</td><td></td><td>=(B$1+C11)/(B$1+B$2+B11)</td><td>=IF(H11>=I$3,"Y","N")</td></tr>
<tr><td>12</td><td>7</td><td>1235</td><td>80</td><td></td><td>0</td><td>-10.4029</td><td></td><td>=(B$1+C12)/(B$1+B$2+B12)</td><td>=IF(H12>=I$3,"Y","N")</td></tr>
<tr><td>13</td><td>8</td><td>573</td><td>34</td><td></td><td>0.0001</td><td>-8.8693</td><td></td><td>=(B$1+C13)/(B$1+B$2+B13)</td><td>=IF(H13>=I$3,"Y","N")</td></tr>
<tr><td>14</td><td>9</td><td>1083</td><td>24</td><td></td><td>0.0036</td><td>-5.6216</td><td></td><td>=(B$1+C14)/(B$1+B$2+B14)</td><td>=IF(H14>=I$3,"Y","N")</td></tr>
<tr><td>15</td><td>10</td><td>352</td><td>5</td><td></td><td>0.0301</td><td>-3.5032</td><td></td><td>=(B$1+C15)/(B$1+B$2+B15)</td><td>=IF(H15>=I$3,"Y","N")</td></tr>
<tr><td>16</td><td>11</td><td>817</td><td>7</td><td></td><td>0.0281</td><td>-3.5719</td><td></td><td>=(B$1+C16)/(B$1+B$2+B16)</td><td>=IF(H16>=I$3,"Y","N")</td></tr>
<tr><td>17</td><td>12</td><td>118</td><td>0</td><td></td><td>0.7018</td><td>-0.3541</td><td></td><td>=(B$1+C17)/(B$1+B$2+B17)</td><td>=IF(H17>=I$3,"Y","N")</td></tr>
<tr><td>18</td><td>13</td><td>1049</td><td>3</td><td></td><td>0.0665</td><td>-2.7102</td><td></td><td>=(B$1+C18)/(B$1+B$2+B18)</td><td>=IF(H18>=I$3,"Y","N")</td></tr>
<tr><td>19</td><td>14</td><td>452</td><td>3</td><td></td><td>0.0674</td><td>-2.6978</td><td></td><td>=(B$1+C19)/(B$1+B$2+B19)</td><td>=IF(H19>=I$3,"Y","N")</td></tr>
<tr><td>20</td><td>15</td><td>338</td><td>2</td><td></td><td>0.0991</td><td>-2.3113</td><td></td><td>=(B$1+C20)/(B$1+B$2+B20)</td><td>=IF(H20>=I$3,"Y","N")</td></tr>
<tr><td>21</td><td>16</td><td>168</td><td>0</td><td></td><td>0.6398</td><td>-0.4466</td><td></td><td>=(B$1+C21)/(B$1+B$2+B21)</td><td>=IF(H21>=I$3,"Y","N")</td></tr>
<tr><td>22</td><td>17</td><td>242</td><td>3</td><td></td><td>0.0547</td><td>-2.9067</td><td></td><td>=(B$1+C22)/(B$1+B$2+B22)</td><td>=IF(H22>=I$3,"Y","N")</td></tr>
<tr><td>23</td><td>18</td><td>185</td><td>1</td><td></td><td>0.1809</td><td>-1.7098</td><td></td><td>=(B$1+C23)/(B$1+B$2+B23)</td><td>=IF(H23>=I$3,"Y","N")</td></tr>
<tr><td>24</td><td>19</td><td>116</td><td>0</td><td></td><td>0.7047</td><td>-0.3499</td><td></td><td>=(B$1+C24)/(B$1+B$2+B24)</td><td>=IF(H24>=I$3,"Y","N")</td></tr>
<tr><td>25</td><td>20</td><td>69</td><td>1</td><td></td><td>0.1459</td><td>-1.925</td><td></td><td>=(B$1+C25)/(B$1+B$2+B25)</td><td>=IF(H25>=I$3,"Y","N")</td></tr>
<tr><td>26</td><td>21</td><td>193</td><td>1</td><td></td><td>0.1812</td><td>-1.708</td><td></td><td>=(B$1+C26)/(B$1+B$2+B26)</td><td>=IF(H26>=I$3,"Y","N")</td></tr>
<tr><td>27</td><td>22</td><td>82</td><td>1</td><td></td><td>0.1553</td><td>-1.8623</td><td></td><td>=(B$1+C27)/(B$1+B$2+B27)</td><td>=IF(H27>=I$3,"Y","N")</td></tr>
<tr><td>28</td><td>23</td><td>265</td><td>1</td><td></td><td>0.1804</td><td>-1.7125</td><td></td><td>=(B$1+C28)/(B$1+B$2+B28)</td><td>=IF(H28>=I$3,"Y","N")</td></tr>
<tr><td>29</td><td>24</td><td>171</td><td>0</td><td></td><td>0.6366</td><td>-0.4516</td><td></td><td>=(B$1+C29)/(B$1+B$2+B29)</td><td>=IF(H29>=I$3,"Y","N")</td></tr>
<tr><td>30</td><td>25</td><td>1554</td><td>7</td><td></td><td>0.0309</td><td>-3.4774</td><td></td><td>=(B$1+C30)/(B$1+B$2+B30)</td><td>=IF(H30>=I$3,"Y","N")</td></tr>
<tr><td>31</td><td>26</td><td>1339</td><td>4</td><td></td><td>0.0511</td><td>-2.9745</td><td></td><td>=(B$1+C31)/(B$1+B$2+B31)</td><td>=IF(H31>=I$3,"Y","N")</td></tr>
<tr><td>32</td><td>27</td><td>1167</td><td>4</td><td></td><td>0.052</td><td>-2.9572</td><td></td><td>=(B$1+C32)/(B$1+B$2+B32)</td><td>=IF(H32>=I$3,"Y","N")</td></tr>
<tr><td>33</td><td>28</td><td>621</td><td>2</td><td></td><td>0.0981</td><td>-2.3219</td><td></td><td>=(B$1+C33)/(B$1+B$2+B33)</td><td>=IF(H33>=I$3,"Y","N")</td></tr>
<tr><td>34</td><td>29</td><td>1013</td><td>1</td><td></td><td>0.1367</td><td>-1.9902</td><td></td><td>=(B$1+C34)/(B$1+B$2+B34)</td><td>=IF(H34>=I$3,"Y","N")</td></tr>
<tr><td>35</td><td>30</td><td>544</td><td>1</td><td></td><td>0.1621</td><td>-1.8195</td><td></td><td>=(B$1+C35)/(B$1+B$2+B35)</td><td>=IF(H35>=I$3,"Y","N")</td></tr>
<tr><td>36</td><td>31</td><td>731</td><td>1</td><td></td><td>0.1505</td><td>-1.8936</td><td></td><td>=(B$1+C36)/(B$1+B$2+B36)</td><td>=IF(H36>=I$3,"Y","N")</td></tr>
</table>

Results

	Standard	Manager	Model
# Segments (Rule)	51		66
# Segments (Act.)	46	71	53
Contacts	682,392	858,728	732,675
Responses	4,463	4,804	4,582
Profit	$492,651	$488,773	$495,060

Use of model results in a profit increase of $6287; 126,053 fewer contacts, saved for another offering.

Concepts and Tools Introduced

- “Choice” processes
- The Beta Binomial model
- “Regression-to-the-mean” and the use of models to capture such an effect
- Bayes theorem (and “empirical Bayes” methods)
- Using “empirical Bayes” methods in the development of targeted marketing campaigns

Further Reading

Colombo, Richard and Donald G. Morrison (1988), "Blacklisting Social Science Departments with Poor Ph.D. Submission Rates," *Management Science*, **34** (June), 696–706.

Morwitz, Vicki G. and David C. Schmittlein (1998), "Testing New Direct Marketing Offerings: The Interplay of Management Judgment and Statistical Models," *Management Science*, **44** (May), 610–28.

Sabavala, Darius J. and Donald G. Morrison (1977), "A Model of TV Show Loyalty," *Journal of Advertising Research*, **17** (December), 35–43.

Further Applications and Tools/ Modeling Issues

Recap

- The preceding three problems introduce simple models for three behavioral processes:
  - Timing ⟶ "when"
  - Counting ⟶ "how many"
  - "Choice" ⟶ "whether/which"
- Each of these simple models has multiple applications.
- More complex behavioral phenomena can be captured by combining models from each of these processes.

Further Applications: Timing Models

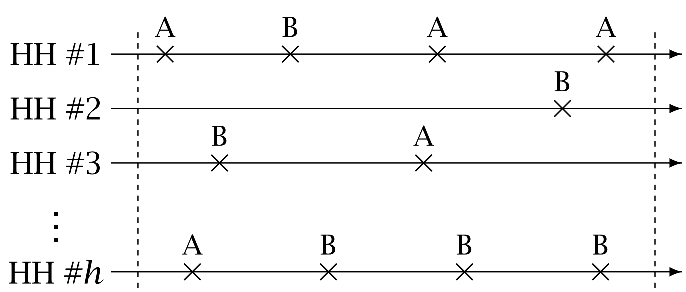
- Repeat purchasing of new products
- Response times:
  - Coupon redemptions
  - Survey response
  - Direct mail (response, returns, repeat sales)
- Customer retention/attrition
- Other durations:
  - Salesforce job tenure
  - Length of web site browsing session

Further Applications: Count Models

- Repeat purchasing
- Customer concentration ("80/20" rules)
- Salesforce productivity/allocation
- Number of page views during a web site browsing session

Further Applications: "Choice" Models

- Brand choice



- Media exposure
- Multibrand choice (BB → Dirichlet Multinomial)
- Taste tests (discrimination tests)
- "Click-through" behavior

Integrated Models

- Counting + Timing
  - catalog purchases (purchasing | “alive” & “death” process)
  - “stickiness” (# visits & duration/visit)
- Counting + Counting
  - purchase volume (# transactions & units/transaction)
  - page views/month (# visits & pages/visit)
- Counting + Choice
  - brand purchasing (category purchasing & brand choice)
  - “conversion” behavior (# visits & buy/not-buy)

77

Further Issues

Relaxing usual assumptions:

- Non-exponential purchasing (greater regularity) $\longrightarrow$ non-Poisson counts
- Non-gamma/beta heterogeneity (e.g., “hard-core” non-buyers, “hard-core” loyals)
- Nonstationarity — latent traits vary over time

The basic models are quite robust to these departures.

Extensions

- Latent class/finite mixture models
- Introducing covariate effects
- Hierarchical Bayes methods

Problem 4: Characterizing the Purchasing of Hard-Candy

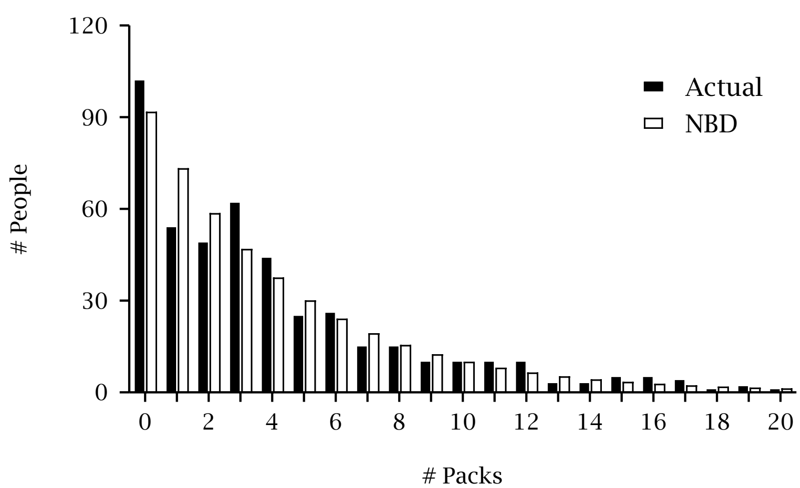
(Introduction to Finite Mixture Models)

Distribution of Hard-Candy Purchases

# Packs	# People	# Packs	# People
0	102	11	10
1	54	12	10
2	49	13	3
3	62	14	3
4	44	15	5
5	25	16	5
6	26	17	4
7	15	18	1
8	15	19	2
9	10	20	1
10	10		

Source: Dillon and Kumar (1994)

Fit of NBD



$$\hat{r} = 0.998, \hat{\alpha} = 0.250$$

Problem 4 -- NBD

	A	B	C	D	E	F	G	H	I	J
1	r	0.997657								
2	alpha	0.249962								
3	LL	=SUM(D6:D26)								
4										
5	# Packs	Observed	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
6	0	102	=(B2/(B2+1))^B1	=B6*LN(C6)	=B$27*C6		=A6	=B6	=E6	=(H6-I6)^2/I6
7	1	54	=(B$1+A7-1)/(A7*(B$2+1))*C6	=B7*LN(C7)	=B$27*C7		=A7	=B7	=E7	=(H7-I7)^2/I7
8	2	49	=(B$1+A8-1)/(A8*(B$2+1))*C7	=B8*LN(C8)	=B$27*C8		=A8	=B8	=E8	=(H8-I8)^2/I8
9	3	62	=(B$1+A9-1)/(A9*(B$2+1))*C8	=B9*LN(C9)	=B$27*C9		=A9	=B9	=E9	=(H9-I9)^2/I9
10	4	44	=(B$1+A10-1)/(A10*(B$2+1))*C9	=B10*LN(C10)	=B$27*C10		=A10	=B10	=E10	=(H10-I10)^2/I10
11	5	25	=(B$1+A11-1)/(A11*(B$2+1))*C10	=B11*LN(C11)	=B$27*C11		=A11	=B11	=E11	=(H11-I11)^2/I11
12	6	26	=(B$1+A12-1)/(A12*(B$2+1))*C11	=B12*LN(C12)	=B$27*C12		=A12	=B12	=E12	=(H12-I12)^2/I12
13	7	15	=(B$1+A13-1)/(A13*(B$2+1))*C12	=B13*LN(C13)	=B$27*C13		=A13	=B13	=E13	=(H13-I13)^2/I13
14	8	15	=(B$1+A14-1)/(A14*(B$2+1))*C13	=B14*LN(C14)	=B$27*C14		=A14	=B14	=E14	=(H14-I14)^2/I14
15	9	10	=(B$1+A15-1)/(A15*(B$2+1))*C14	=B15*LN(C15)	=B$27*C15		=A15	=B15	=E15	=(H15-I15)^2/I15
16	10	10	=(B$1+A16-1)/(A16*(B$2+1))*C15	=B16*LN(C16)	=B$27*C16		=A16	=B16	=E16	=(H16-I16)^2/I16
17	11	10	=(B$1+A17-1)/(A17*(B$2+1))*C16	=B17*LN(C17)	=B$27*C17		=A17	=B17	=E17	=(H17-I17)^2/I17
18	12	10	=(B$1+A18-1)/(A18*(B$2+1))*C17	=B18*LN(C18)	=B$27*C18		=A18	=B18	=E18	=(H18-I18)^2/I18
19	13	3	=(B$1+A19-1)/(A19*(B$2+1))*C18	=B19*LN(C19)	=B$27*C19		=A19	=B19	=E19	=(H19-I19)^2/I19
20	14	3	=(B$1+A20-1)/(A20*(B$2+1))*C19	=B20*LN(C20)	=B$27*C20		=A20	=B20	=E20	=(H20-I20)^2/I20
21	15	5	=(B$1+A21-1)/(A21*(B$2+1))*C20	=B21*LN(C21)	=B$27*C21		15+	=SUM(B21:B26)	=SUM(E21:E26)	=(H21-I21)^2/I21
22	16	5	=(B$1+A22-1)/(A22*(B$2+1))*C21	=B22*LN(C22)	=B$27*C22					=SUM(J6:J21)
23	17	4	=(B$1+A23-1)/(A23*(B$2+1))*C22	=B23*LN(C23)	=B$27*C23					
24	18	1	=(B$1+A24-1)/(A24*(B$2+1))*C23	=B24*LN(C24)	=B$27*C24				# params	2
25	19	2	=(B$1+A25-1)/(A25*(B$2+1))*C24	=B25*LN(C25)	=B$27*C25				df	=16-J24-1
26	20	1	=(B$1+A26-1)/(A26*(B$2+1))*C25	=B26*LN(C26)	=B$27*C26					
27		=SUM(B6:B26)							p-value	=CHIDIST(J22,J25)

Problem 4 -- NBD

	A	B	C	D	E	F	G	H	I	J
1	r	0.998								
2	alpha	0.250								
3	LL	-1140.02								
4										
5	# Packs	Observed	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
6	0	102	0.20073	-163.79	91.5		0	102	91.5	1.20
7	1	54	0.16021	-98.89	73.1		1	54	73.1	4.97
8	2	49	0.12802	-100.72	58.4		2	49	58.4	1.51
9	3	62	0.10234	-141.32	46.7		3	62	46.7	5.04
10	4	44	0.08183	-110.14	37.3		4	44	37.3	1.20
11	5	25	0.06543	-68.17	29.8		5	25	29.8	0.78
12	6	26	0.05233	-76.71	23.9		6	26	23.9	0.19
13	7	15	0.04185	-47.60	19.1		7	15	19.1	0.87
14	8	15	0.03347	-50.96	15.3		8	15	15.3	0.00
15	9	10	0.02677	-36.20	12.2		9	10	12.2	0.40
16	10	10	0.02141	-38.44	9.8		10	10	9.8	0.01
17	11	10	0.01713	-40.67	7.8		11	10	7.8	0.61
18	12	10	0.01370	-42.90	6.2		12	10	6.2	2.25
19	13	3	0.01096	-13.54	5.0		13	3	5.0	0.80
20	14	3	0.00876	-14.21	4.0		14	3	4.0	0.25
21	15	5	0.00701	-24.80	3.2		15+	18	11.8	3.27
22	16	5	0.00561	-25.92	2.6					23.35
23	17	4	0.00449	-21.63	2.0					
24	18	1	0.00359	-5.63	1.6				# params	2
25	19	2	0.00287	-11.71	1.3				df	13
26	20	1	0.00230	-6.08	1.0					
27		456							p-value	0.038

The Zero-Inflated NBD Model

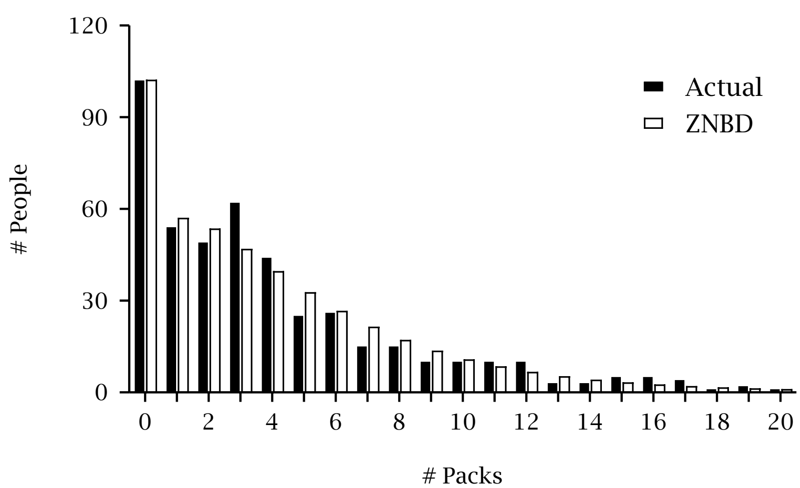
Because of the "excessive" number of zeros, let us consider the zero-inflated NBD (ZNBD) model:

- a proportion π of the population never buy hard-candy
- the visiting behavior of the "ever buyers" can be characterized by the NBD model

$$P(X = x) = \delta_{x=0}\pi + (1 - \pi) \times \frac{\Gamma(r + x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha + 1}\right)^r \left(\frac{1}{\alpha + 1}\right)^x$$

This is sometimes called the NBD with hard-core non-buyers model.

Fit of ZNBD



$$\hat{\pi} = 0.113, \hat{r} = 1.504, \hat{\alpha} = 0.334$$

Problem 4 -- ZNBD

	A	B	C	D	E	F	G	H	I	J	K
1	r	0.997657									
2	alpha	0.249962									
3	pi	0.113103									
4	LL	=SUM(E8:E28)									
5											
6			P(X=x)								
7	# Packs	Observed	NBD	ZNBD	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
8	0	102	=(B2/(B2+1))^B1	=(A8=0)*B$3+(1-B$3)*C8	=B8*LN(D8)	=B$29*D8		=A8	=B8	=F8	=(I8-J8)^2/J8
9	1	54	=(B$1+A9-1)/(A9*(B$2+1))*C8	=(A9=0)*B$3+(1-B$3)*C9	=B9*LN(D9)	=B$29*D9		=A9	=B9	=F9	=(I9-J9)^2/J9
10	2	49	=(B$1+A10-1)/(A10*(B$2+1))*C9	=(A10=0)*B$3+(1-B$3)*C10	=B10*LN(D10)	=B$29*D10		=A10	=B10	=F10	=(I10-J10)^2/J10
11	3	62	=(B$1+A11-1)/(A11*(B$2+1))*C10	=(A11=0)*B$3+(1-B$3)*C11	=B11*LN(D11)	=B$29*D11		=A11	=B11	=F11	=(I11-J11)^2/J11
12	4	44	=(B$1+A12-1)/(A12*(B$2+1))*C11	=(A12=0)*B$3+(1-B$3)*C12	=B12*LN(D12)	=B$29*D12		=A12	=B12	=F12	=(I12-J12)^2/J12
13	5	25	=(B$1+A13-1)/(A13*(B$2+1))*C12	=(A13=0)*B$3+(1-B$3)*C13	=B13*LN(D13)	=B$29*D13		=A13	=B13	=F13	=(I13-J13)^2/J13
14	6	26	=(B$1+A14-1)/(A14*(B$2+1))*C13	=(A14=0)*B$3+(1-B$3)*C14	=B14*LN(D14)	=B$29*D14		=A14	=B14	=F14	=(I14-J14)^2/J14
15	7	15	=(B$1+A15-1)/(A15*(B$2+1))*C14	=(A15=0)*B$3+(1-B$3)*C15	=B15*LN(D15)	=B$29*D15		=A15	=B15	=F15	=(I15-J15)^2/J15
16	8	15	=(B$1+A16-1)/(A16*(B$2+1))*C15	=(A16=0)*B$3+(1-B$3)*C16	=B16*LN(D16)	=B$29*D16		=A16	=B16	=F16	=(I16-J16)^2/J16
17	9	10	=(B$1+A17-1)/(A17*(B$2+1))*C16	=(A17=0)*B$3+(1-B$3)*C17	=B17*LN(D17)	=B$29*D17		=A17	=B17	=F17	=(I17-J17)^2/J17
18	10	10	=(B$1+A18-1)/(A18*(B$2+1))*C17	=(A18=0)*B$3+(1-B$3)*C18	=B18*LN(D18)	=B$29*D18		=A18	=B18	=F18	=(I18-J18)^2/J18
19	11	10	=(B$1+A19-1)/(A19*(B$2+1))*C18	=(A19=0)*B$3+(1-B$3)*C19	=B19*LN(D19)	=B$29*D19		=A19	=B19	=F19	=(I19-J19)^2/J19
20	12	10	=(B$1+A20-1)/(A20*(B$2+1))*C19	=(A20=0)*B$3+(1-B$3)*C20	=B20*LN(D20)	=B$29*D20		=A20	=B20	=F20	=(I20-J20)^2/J20
21	13	3	=(B$1+A21-1)/(A21*(B$2+1))*C20	=(A21=0)*B$3+(1-B$3)*C21	=B21*LN(D21)	=B$29*D21		=A21	=B21	=F21	=(I21-J21)^2/J21
22	14	3	=(B$1+A22-1)/(A22*(B$2+1))*C21	=(A22=0)*B$3+(1-B$3)*C22	=B22*LN(D22)	=B$29*D22		=A22	=B22	=F22	=(I22-J22)^2/J22
23	15	5	=(B$1+A23-1)/(A23*(B$2+1))*C22	=(A23=0)*B$3+(1-B$3)*C23	=B23*LN(D23)	=B$29*D23		15+	=SUM(B23:B28)	=SUM(F23:F28)	=(I23-J23)^2/J23
24	16	5	=(B$1+A24-1)/(A24*(B$2+1))*C23	=(A24=0)*B$3+(1-B$3)*C24	=B24*LN(D24)	=B$29*D24					=SUM(K8:K23)
25	17	4	=(B$1+A25-1)/(A25*(B$2+1))*C24	=(A25=0)*B$3+(1-B$3)*C25	=B25*LN(D25)	=B$29*D25					
26	18	1	=(B$1+A26-1)/(A26*(B$2+1))*C25	=(A26=0)*B$3+(1-B$3)*C26	=B26*LN(D26)	=B$29*D26				# params	3
27	19	2	=(B$1+A27-1)/(A27*(B$2+1))*C26	=(A27=0)*B$3+(1-B$3)*C27	=B27*LN(D27)	=B$29*D27				df	=16-K26-1
28	20	1	=(B$1+A28-1)/(A28*(B$2+1))*C27	=(A28=0)*B$3+(1-B$3)*C28	=B28*LN(D28)	=B$29*D28					
29		=SUM(B8:B28)								p-value	=CHIDIST(K24,K27)

Problem 4 -- ZNBD

	A	B	C	D	E	F	G	H	I	J	K
1	r	1.504									
2	alpha	0.334									
3	pi	0.113									
4	LL	-1136.17									
5											
6			P(X=x)								
7	# Packs	Observed	NBD	ZNBD	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
8	0	102	0.12468	0.22368	-152.75	102.0		0	102	102.0	0.00
9	1	54	0.14054	0.12465	-112.44	56.8		1	54	56.8	0.14
10	2	49	0.13188	0.11697	-105.15	53.3		2	49	53.3	0.35
11	3	62	0.11545	0.10239	-141.29	46.7		3	62	46.7	5.02
12	4	44	0.09743	0.08641	-107.74	39.4		4	44	39.4	0.54
13	5	25	0.08039	0.07130	-66.02	32.5		5	25	32.5	1.74
14	6	26	0.06531	0.05793	-74.06	26.4		6	26	26.4	0.01
15	7	15	0.05248	0.04654	-46.01	21.2		7	15	21.2	1.82
16	8	15	0.04181	0.03708	-49.42	16.9		8	15	16.9	0.22
17	9	10	0.03309	0.02935	-35.28	13.4		9	10	13.4	0.86
18	10	10	0.02605	0.02311	-37.68	10.5		10	10	10.5	0.03
19	11	10	0.02042	0.01811	-40.11	8.3		11	10	8.3	0.37
20	12	10	0.01595	0.01415	-42.58	6.5		12	10	6.5	1.95
21	13	3	0.01242	0.01101	-13.53	5.0		13	3	5.0	0.81
22	14	3	0.00964	0.00855	-14.28	3.9		14	3	3.9	0.21
23	15	5	0.00747	0.00663	-25.08	3.0		15+	18	10.4	5.48
24	16	5	0.00578	0.00512	-26.37	2.3					19.54
25	17	4	0.00446	0.00395	-22.13	1.8					
26	18	1	0.00343	0.00305	-5.79	1.4				# params	3
27	19	2	0.00264	0.00234	-12.11	1.1				df	12
28	20	1	0.00203	0.00180	-6.32	0.8					
29		456								p-value	0.076

What is Wrong With the NBD Model?

The assumptions underlying the model could be wrong on two accounts:

i. at the individual-level, the number of purchases is not Poisson distributed

ii. purchase rates (λ) are not gamma-distributed

Relaxing the Gamma Assumption

- Replace the continuous distribution with a discrete distribution by allowing for multiple (discrete) segments each with a different (latent) buying rate:

$$P(X = x) = \sum_{s=1}^{S} \pi_s P(X = x|\lambda_s), \quad \sum_{s=1}^{S} \pi_s = 1$$

- This is called a finite mixture model.
- We often reparameterize the mixing proportions for computational convenience:

$$\pi_s = \frac{\exp(\theta_s)}{\sum_{s'=1}^{S} \exp(\theta_{s'})}, \quad \theta_S = 0.$$

	A	B	C	D	E	F	G	H	I	J
1	lambda	3.9912280686								
2	LL	=SUM(D5:D25)								
3										
4	# Packs	Observed	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
5	0	102	=POISSON(A5,B$1,FALSE)	=B5*LN(C5)	=B$26*C5		=A5	=B5	=E5	=(H5-I5)^2/I5
6	1	54	=POISSON(A6,B$1,FALSE)	=B6*LN(C6)	=B$26*C6		=A6	=B6	=E6	=(H6-I6)^2/I6
7	2	49	=POISSON(A7,B$1,FALSE)	=B7*LN(C7)	=B$26*C7		=A7	=B7	=E7	=(H7-I7)^2/I7
8	3	62	=POISSON(A8,B$1,FALSE)	=B8*LN(C8)	=B$26*C8		=A8	=B8	=E8	=(H8-I8)^2/I8
9	4	44	=POISSON(A9,B$1,FALSE)	=B9*LN(C9)	=B$26*C9		=A9	=B9	=E9	=(H9-I9)^2/I9
10	5	25	=POISSON(A10,B$1,FALSE)	=B10*LN(C10)	=B$26*C10		=A10	=B10	=E10	=(H10-I10)^2/I10
11	6	26	=POISSON(A11,B$1,FALSE)	=B11*LN(C11)	=B$26*C11		=A11	=B11	=E11	=(H11-I11)^2/I11
12	7	15	=POISSON(A12,B$1,FALSE)	=B12*LN(C12)	=B$26*C12		=A12	=B12	=E12	=(H12-I12)^2/I12
13	8	15	=POISSON(A13,B$1,FALSE)	=B13*LN(C13)	=B$26*C13		=A13	=B13	=E13	=(H13-I13)^2/I13
14	9	10	=POISSON(A14,B$1,FALSE)	=B14*LN(C14)	=B$26*C14		=A14	=B14	=E14	=(H14-I14)^2/I14
15	10	10	=POISSON(A15,B$1,FALSE)	=B15*LN(C15)	=B$26*C15		=A15	=B15	=E15	=(H15-I15)^2/I15
16	11	10	=POISSON(A16,B$1,FALSE)	=B16*LN(C16)	=B$26*C16		=A16	=B16	=E16	=(H16-I16)^2/I16
17	12	10	=POISSON(A17,B$1,FALSE)	=B17*LN(C17)	=B$26*C17		=A17	=B17	=E17	=(H17-I17)^2/I17
18	13	3	=POISSON(A18,B$1,FALSE)	=B18*LN(C18)	=B$26*C18		=A18	=B18	=E18	=(H18-I18)^2/I18
19	14	3	=POISSON(A19,B$1,FALSE)	=B19*LN(C19)	=B$26*C19		=A19	=B19	=E19	=(H19-I19)^2/I19
20	15	5	=POISSON(A20,B$1,FALSE)	=B20*LN(C20)	=B$26*C20		15+	=SUM(B20:B25)	=SUM(E20:E25)	=(H20-I20)^2/I20
21	16	5	=POISSON(A21,B$1,FALSE)	=B21*LN(C21)	=B$26*C21					=SUM(J5:J20)
22	17	4	=POISSON(A22,B$1,FALSE)	=B22*LN(C22)	=B$26*C22					
23	18	1	=POISSON(A23,B$1,FALSE)	=B23*LN(C23)	=B$26*C23				# params	1
24	19	2	=POISSON(A24,B$1,FALSE)	=B24*LN(C24)	=B$26*C24				df	=16-J23-1
25	20	1	=POISSON(A25,B$1,FALSE)	=B25*LN(C25)	=B$26*C25					
26		=SUM(B5:B25)							p-value	=CHIDIST(J21,J24)

Problem 4 -- Poisson

	A	B	C	D	E	F	G	H	I	J
1	lambda	3.991								
2	LL	-1545.00								
3										
4	# Packs	Observed	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
5	0	102	0.01848	-407.11	8.4		0	102	8.4	1039.25
6	1	54	0.07375	-140.78	33.6		1	54	33.6	12.34
7	2	49	0.14717	-93.89	67.1		2	49	67.1	4.89
8	3	62	0.19579	-101.10	89.3		3	62	89.3	8.34
9	4	44	0.19536	-71.85	89.1		4	44	89.1	22.82
10	5	25	0.15595	-46.46	71.1		5	25	71.1	29.90
11	6	26	0.10374	-58.91	47.3		6	26	47.3	9.59
12	7	15	0.05915	-42.42	27.0		7	15	27.0	5.31
13	8	15	0.02951	-52.85	13.5		8	15	13.5	0.18
14	9	10	0.01309	-43.36	6.0		9	10	6.0	2.72
15	10	10	0.00522	-52.55	2.4		10	10	2.4	24.37
16	11	10	0.00190	-62.68	0.9		11	10	0.9	96.58
17	12	10	0.00063	-73.69	0.3		12	10	0.3	328.19
18	13	3	0.00019	-25.65	0.1		13	3	0.1	96.07
19	14	3	0.00006	-29.42	0.0		14	3	0.0	351.76
20	15	5	0.00001	-55.65	0.0		15+	18	0.0	36512.67
21	16	5	0.00000	-62.59	0.0					38544.98
22	17	4	0.00000	-55.87	0.0					
23	18	1	0.00000	-15.47	0.0				# params	1
24	19	2	0.00000	-34.07	0.0				df	14
25	20	1	0.00000	-18.64	0.0					
26		456							p-value	0.000

Problem 4 -- 2seg Poisson

	A	B	C	D	E	F	G
1	lambda_1	1.8021538					
2	lambda_2	9.1206784					
3	pi	0.7008857					
4	LL	=SUM(F8:F28)					
5	BIC	=-2*B4+L26*LN(B29)					
6							
7	# Packs	Observed	Seg1	Seg2	P(X=x)	LL	Expected
8	0	102	=POISSON(A8,B$1,FALSE)	=POISSON(A8,B$2,FALSE)	=B$3*C8+(1-B$3)*D8	=B8*LN(E8)	=B$29*E8
9	1	54	=POISSON(A9,B$1,FALSE)	=POISSON(A9,B$2,FALSE)	=B$3*C9+(1-B$3)*D9	=B9*LN(E9)	=B$29*E9
10	2	49	=POISSON(A10,B$1,FALSE)	=POISSON(A10,B$2,FALSE)	=B$3*C10+(1-B$3)*D10	=B10*LN(E10)	=B$29*E10
11	3	62	=POISSON(A11,B$1,FALSE)	=POISSON(A11,B$2,FALSE)	=B$3*C11+(1-B$3)*D11	=B11*LN(E11)	=B$29*E11
12	4	44	=POISSON(A12,B$1,FALSE)	=POISSON(A12,B$2,FALSE)	=B$3*C12+(1-B$3)*D12	=B12*LN(E12)	=B$29*E12
13	5	25	=POISSON(A13,B$1,FALSE)	=POISSON(A13,B$2,FALSE)	=B$3*C13+(1-B$3)*D13	=B13*LN(E13)	=B$29*E13
14	6	26	=POISSON(A14,B$1,FALSE)	=POISSON(A14,B$2,FALSE)	=B$3*C14+(1-B$3)*D14	=B14*LN(E14)	=B$29*E14
15	7	15	=POISSON(A15,B$1,FALSE)	=POISSON(A15,B$2,FALSE)	=B$3*C15+(1-B$3)*D15	=B15*LN(E15)	=B$29*E15
16	8	15	=POISSON(A16,B$1,FALSE)	=POISSON(A16,B$2,FALSE)	=B$3*C16+(1-B$3)*D16	=B16*LN(E16)	=B$29*E16
17	9	10	=POISSON(A17,B$1,FALSE)	=POISSON(A17,B$2,FALSE)	=B$3*C17+(1-B$3)*D17	=B17*LN(E17)	=B$29*E17
18	10	10	=POISSON(A18,B$1,FALSE)	=POISSON(A18,B$2,FALSE)	=B$3*C18+(1-B$3)*D18	=B18*LN(E18)	=B$29*E18
19	11	10	=POISSON(A19,B$1,FALSE)	=POISSON(A19,B$2,FALSE)	=B$3*C19+(1-B$3)*D19	=B19*LN(E19)	=B$29*E19
20	12	10	=POISSON(A20,B$1,FALSE)	=POISSON(A20,B$2,FALSE)	=B$3*C20+(1-B$3)*D20	=B20*LN(E20)	=B$29*E20
21	13	3	=POISSON(A21,B$1,FALSE)	=POISSON(A21,B$2,FALSE)	=B$3*C21+(1-B$3)*D21	=B21*LN(E21)	=B$29*E21
22	14	3	=POISSON(A22,B$1,FALSE)	=POISSON(A22,B$2,FALSE)	=B$3*C22+(1-B$3)*D22	=B22*LN(E22)	=B$29*E22
23	15	5	=POISSON(A23,B$1,FALSE)	=POISSON(A23,B$2,FALSE)	=B$3*C23+(1-B$3)*D23	=B23*LN(E23)	=B$29*E23
24	16	5	=POISSON(A24,B$1,FALSE)	=POISSON(A24,B$2,FALSE)	=B$3*C24+(1-B$3)*D24	=B24*LN(E24)	=B$29*E24
25	17	4	=POISSON(A25,B$1,FALSE)	=POISSON(A25,B$2,FALSE)	=B$3*C25+(1-B$3)*D25	=B25*LN(E25)	=B$29*E25
26	18	1	=POISSON(A26,B$1,FALSE)	=POISSON(A26,B$2,FALSE)	=B$3*C26+(1-B$3)*D26	=B26*LN(E26)	=B$29*E26
27	19	2	=POISSON(A27,B$1,FALSE)	=POISSON(A27,B$2,FALSE)	=B$3*C27+(1-B$3)*D27	=B27*LN(E27)	=B$29*E27
28	20	1	=POISSON(A28,B$1,FALSE)	=POISSON(A28,B$2,FALSE)	=B$3*C28+(1-B$3)*D28	=B28*LN(E28)	=B$29*E28
29		=SUM(B8:B28)					

Problem 4 -- 2seg Poisson

	A	B	C	D	E	F	G	H	I	J	K	L
1	lambda_1	1.802										
2	lambda_2	9.121										
3	pi	0.701										
4	LL	-1188.83										
5	BIC	2396.03										
6												
7	# Packs	Observed	Seg1	Seg2	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
8	0	102	0.16494	0.00011	0.11564	-220.04	52.7		0	102	52.7	46.03
9	1	54	0.29725	0.00100	0.20864	-84.63	95.1		1	54	95.1	17.79
10	2	49	0.26785	0.00455	0.18909	-81.61	86.2		2	49	86.2	16.07
11	3	62	0.16090	0.01383	0.11691	-133.07	53.3		3	62	53.3	1.42
12	4	44	0.07249	0.03154	0.06024	-123.61	27.5		4	44	27.5	9.95
13	5	25	0.02613	0.05753	0.03552	-83.44	16.2		5	25	16.2	4.78
14	6	26	0.00785	0.08745	0.03166	-89.77	14.4		6	26	14.4	9.26
15	7	15	0.00202	0.11395	0.03550	-50.07	16.2		7	15	16.2	0.09
16	8	15	0.00046	0.12991	0.03918	-48.60	17.9		8	15	17.9	0.46
17	9	10	0.00009	0.13165	0.03944	-32.33	18.0		9	10	18.0	3.55
18	10	10	0.00002	0.12007	0.03593	-33.26	16.4		10	10	16.4	2.49
19	11	10	0.00000	0.09956	0.02978	-35.14	13.6		11	10	13.6	0.94
20	12	10	0.00000	0.07567	0.02263	-37.88	10.3		12	10	10.3	0.01
21	13	3	0.00000	0.05309	0.01588	-12.43	7.2		13	3	7.2	2.48
22	14	3	0.00000	0.03459	0.01035	-13.71	4.7		14	3	4.7	0.63
23	15	5	0.00000	0.02103	0.00629	-25.34	2.9		15+	18	6.1	22.94
24	16	5	0.00000	0.01199	0.00359	-28.15	1.6					138.88
25	17	4	0.00000	0.00643	0.00192	-25.01	0.9					
26	18	1	0.00000	0.00326	0.00097	-6.93	0.4				# params	3
27	19	2	0.00000	0.00156	0.00047	-15.33	0.2				df	12
28	20	1	0.00000	0.00071	0.00021	-8.45	0.1					
29		456									p-value	0.000

	A	B	C	D	E	F	G	H
1	lambda_1	3.483317						
2	lambda_2	11.21581						
3	lambda_3	0.290543						
4	theta_1	0.674427	=EXP(B4)					
5	theta_2	-0.43042	=EXP(B5)					
6	theta_3	0	=EXP(B6)					
7	LL	=SUM(G11:G31)						
8	BIC	=-2*B7+M29*LN(B32)						
9			=C4/SUM(C4:C6)	=C5/SUM(C4:C6)	=C6/SUM(C4:C6)			
10	# Packs	Observed	Seg1	Seg2	Seg3	P(X=x)	LL	Expected
11	0	102	=POISSON(A11,B$1,FALSE)	=POISSON(A11,B$2,FALSE)	=POISSON(A11,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C11:E11)	=B11*LN(F11)	=B$32*F11
12	1	54	=POISSON(A12,B$1,FALSE)	=POISSON(A12,B$2,FALSE)	=POISSON(A12,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C12:E12)	=B12*LN(F12)	=B$32*F12
13	2	49	=POISSON(A13,B$1,FALSE)	=POISSON(A13,B$2,FALSE)	=POISSON(A13,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C13:E13)	=B13*LN(F13)	=B$32*F13
14	3	62	=POISSON(A14,B$1,FALSE)	=POISSON(A14,B$2,FALSE)	=POISSON(A14,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C14:E14)	=B14*LN(F14)	=B$32*F14
15	4	44	=POISSON(A15,B$1,FALSE)	=POISSON(A15,B$2,FALSE)	=POISSON(A15,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C15:E15)	=B15*LN(F15)	=B$32*F15
16	5	25	=POISSON(A16,B$1,FALSE)	=POISSON(A16,B$2,FALSE)	=POISSON(A16,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C16:E16)	=B16*LN(F16)	=B$32*F16
17	6	26	=POISSON(A17,B$1,FALSE)	=POISSON(A17,B$2,FALSE)	=POISSON(A17,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C17:E17)	=B17*LN(F17)	=B$32*F17
18	7	15	=POISSON(A18,B$1,FALSE)	=POISSON(A18,B$2,FALSE)	=POISSON(A18,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C18:E18)	=B18*LN(F18)	=B$32*F18
19	8	15	=POISSON(A19,B$1,FALSE)	=POISSON(A19,B$2,FALSE)	=POISSON(A19,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C19:E19)	=B19*LN(F19)	=B$32*F19
20	9	10	=POISSON(A20,B$1,FALSE)	=POISSON(A20,B$2,FALSE)	=POISSON(A20,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C20:E20)	=B20*LN(F20)	=B$32*F20
21	10	10	=POISSON(A21,B$1,FALSE)	=POISSON(A21,B$2,FALSE)	=POISSON(A21,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C21:E21)	=B21*LN(F21)	=B$32*F21
22	11	10	=POISSON(A22,B$1,FALSE)	=POISSON(A22,B$2,FALSE)	=POISSON(A22,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C22:E22)	=B22*LN(F22)	=B$32*F22
23	12	10	=POISSON(A23,B$1,FALSE)	=POISSON(A23,B$2,FALSE)	=POISSON(A23,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C23:E23)	=B23*LN(F23)	=B$32*F23
24	13	3	=POISSON(A24,B$1,FALSE)	=POISSON(A24,B$2,FALSE)	=POISSON(A24,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C24:E24)	=B24*LN(F24)	=B$32*F24
25	14	3	=POISSON(A25,B$1,FALSE)	=POISSON(A25,B$2,FALSE)	=POISSON(A25,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C25:E25)	=B25*LN(F25)	=B$32*F25
26	15	5	=POISSON(A26,B$1,FALSE)	=POISSON(A26,B$2,FALSE)	=POISSON(A26,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C26:E26)	=B26*LN(F26)	=B$32*F26
27	16	5	=POISSON(A27,B$1,FALSE)	=POISSON(A27,B$2,FALSE)	=POISSON(A27,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C27:E27)	=B27*LN(F27)	=B$32*F27
28	17	4	=POISSON(A28,B$1,FALSE)	=POISSON(A28,B$2,FALSE)	=POISSON(A28,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C28:E28)	=B28*LN(F28)	=B$32*F28
29	18	1	=POISSON(A29,B$1,FALSE)	=POISSON(A29,B$2,FALSE)	=POISSON(A29,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C29:E29)	=B29*LN(F29)	=B$32*F29
30	19	2	=POISSON(A30,B$1,FALSE)	=POISSON(A30,B$2,FALSE)	=POISSON(A30,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C30:E30)	=B30*LN(F30)	=B$32*F30
31	20	1	=POISSON(A31,B$1,FALSE)	=POISSON(A31,B$2,FALSE)	=POISSON(A31,B$3,FALSE)	=SUMPRODUCT(C$9:E$9,C31:E31)	=B31*LN(F31)	=B$32*F31
32		=SUM(B11:B31)						

Problem 4 -- 3seg Poisson

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	lambda_1	3.483											
2	lambda_2	11.216											
3	lambda_3	0.291											
4	theta_1	0.674	1.962908										
5	theta_2	-0.430	0.650233										
6	theta_3	0	1										
7	LL	-1132.04											
8	BIC	2294.70											
9			0.54327	0.17996	0.27677								
10	# Packs	Observed	Seg1	Seg2	Seg3	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
11	0	102	0.03071	0.00001	0.74786	0.22367	-152.76	102.0		0	102	102.0	0.00
12	1	54	0.10696	0.00015	0.21728	0.11827	-115.28	53.9		1	54	53.9	0.00
13	2	49	0.18628	0.00085	0.03157	0.11009	-108.12	50.2		2	49	50.2	0.03
14	3	62	0.21629	0.00317	0.00306	0.11892	-132.02	54.2		3	62	54.2	1.11
15	4	44	0.18835	0.00887	0.00022	0.10399	-99.59	47.4		4	44	47.4	0.25
16	5	25	0.13122	0.01991	0.00001	0.07487	-64.80	34.1		5	25	34.1	2.45
17	6	26	0.07618	0.03721	0.00000	0.04808	-78.91	21.9		6	26	21.9	0.76
18	7	15	0.03791	0.05962	0.00000	0.03132	-51.95	14.3		7	15	14.3	0.04
19	8	15	0.01651	0.08359	0.00000	0.02401	-55.94	10.9		8	15	10.9	1.50
20	9	10	0.00639	0.10417	0.00000	0.02222	-38.07	10.1		9	10	10.1	0.00
21	10	10	0.00223	0.11684	0.00000	0.02224	-38.06	10.1		10	10	10.1	0.00
22	11	10	0.00070	0.11913	0.00000	0.02182	-38.25	10.0		11	10	10.0	0.00
23	12	10	0.00020	0.11134	0.00000	0.02015	-39.05	9.2		12	10	9.2	0.07
24	13	3	0.00005	0.09606	0.00000	0.01732	-12.17	7.9		13	3	7.9	3.04
25	14	3	0.00001	0.07696	0.00000	0.01386	-12.84	6.3		14	3	6.3	1.74
26	15	5	0.00000	0.05754	0.00000	0.01036	-22.85	4.7		15+	18	12.8	2.08
27	16	5	0.00000	0.04034	0.00000	0.00726	-24.63	3.3					13.07
28	17	4	0.00000	0.02661	0.00000	0.00479	-21.37	2.2					
29	18	1	0.00000	0.01658	0.00000	0.00298	-5.81	1.4				# params	5
30	19	2	0.00000	0.00979	0.00000	0.00176	-12.68	0.8				df	10
31	20	1	0.00000	0.00549	0.00000	0.00099	-6.92	0.5					
32		456										p-value	0.220

Parameter Estimates

	Seg 1	Seg 2	Seg 3	LL
λ	3.991			−1545.00
λ_s	1.802	9.121		−1188.83
π_s	0.701	0.299		
λ_s	0.291	3.483	11.216	−1132.04
π_s	0.277	0.543	0.180	

How Many Segments?

- Controlling for the extra parameters, is an $S + 1$ segment model better than an S segment model?
- We can't use the likelihood ratio test because its properties are violated
- It is standard practice to use "information-theoretic" model selection criteria
- A common measure is the Bayesian information criterion:

$$\text{BIC} = -2LL + p\ln(N)$$

  where p is the number of parameters and N is the sample size
- Rule: choose S to minimize BIC

Problem 4 -- 4seg Poisson

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	lambda_1	3.002												
2	lambda_2	0.205												
3	lambda_3	7.418												
4	lambda_4	12.873												
5	theta_1	1.598	4.943453											
6	theta_2	0.876	2.401291											
7	theta_3	0.398	1.488702											
8	theta_4	0	1											
9	LL	-1130.07												
10	BIC	2303.00												
11			0.50272	0.24420	0.15139	0.10169								
12	# Packs	Observed	Seg1	Seg2	Seg3	Seg4	P(X=x)	LL	Expected		# Packs	Observed	Expected	(O-E)^2/E
13	0	102	0.04969	0.81487	0.00060	0.00000	0.22406	-152.58	102.2		0	102	102.2	0.00
14	1	54	0.14917	0.16683	0.00445	0.00003	0.11641	-116.14	53.1		1	54	53.1	0.02
15	2	49	0.22390	0.01708	0.01652	0.00021	0.11925	-104.20	54.4		2	49	54.4	0.53
16	3	62	0.22404	0.00117	0.04084	0.00091	0.11919	-131.88	54.4		3	62	54.4	1.08
17	4	44	0.16814	0.00006	0.07574	0.00294	0.09631	-102.97	43.9		4	44	43.9	0.00
18	5	25	0.10095	0.00000	0.11237	0.00756	0.06853	-67.01	31.2		5	25	31.2	1.25
19	6	26	0.05051	0.00000	0.13892	0.01622	0.04807	-78.91	21.9		6	26	21.9	0.76
20	7	15	0.02166	0.00000	0.14722	0.02984	0.03621	-49.78	16.5		7	15	16.5	0.14
21	8	15	0.00813	0.00000	0.13652	0.04801	0.02964	-52.78	13.5		8	15	13.5	0.16
22	9	10	0.00271	0.00000	0.11252	0.06867	0.02538	-36.74	11.6		9	10	11.6	0.21
23	10	10	0.00081	0.00000	0.08347	0.08839	0.02203	-38.15	10.0		10	10	10.0	0.00
24	11	10	0.00022	0.00000	0.05629	0.10344	0.01915	-39.55	8.7		11	10	8.7	0.18
25	12	10	0.00006	0.00000	0.03480	0.11096	0.01658	-41.00	7.6		12	10	7.6	0.79
26	13	3	0.00001	0.00000	0.01986	0.10987	0.01419	-12.77	6.5		13	3	6.5	1.86
27	14	3	0.00000	0.00000	0.01052	0.10102	0.01187	-13.30	5.4		14	3	5.4	1.07
28	15	5	0.00000	0.00000	0.00520	0.08669	0.00960	-23.23	4.4		15+	18	14.1	1.11
29	16	5	0.00000	0.00000	0.00241	0.06975	0.00746	-24.49	3.4					9.17
30	17	4	0.00000	0.00000	0.00105	0.05281	0.00553	-20.79	2.5					
31	18	1	0.00000	0.00000	0.00043	0.03777	0.00391	-5.55	1.8				# params	7
32	19	2	0.00000	0.00000	0.00017	0.02559	0.00263	-11.88	1.2				df	8
33	20	1	0.00000	0.00000	0.00006	0.01647	0.00168	-6.39	0.8					
34		456											p-value	0.329

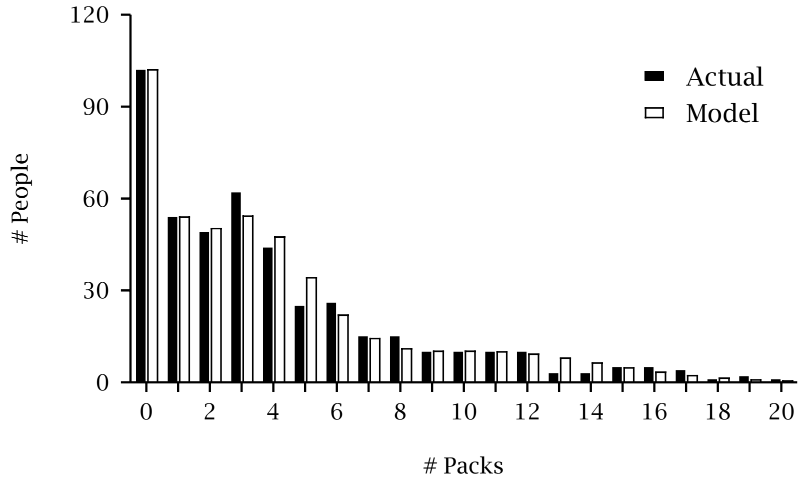
Summary of Model Fit

Model	LL	# params	BIC	χ^2 p-value
NBD	−1140.02	2	2292.29	0.04
ZNBD	−1136.17	3	2290.70	0.08
Poisson	−1545.00	1	3096.12	0.00
2 seg Poisson	−1188.83	3	2396.03	0.00
3 seg Poisson	−1132.04	5	2294.70	0.22
4 seg Poisson	−1130.07	7	2303.00	0.33

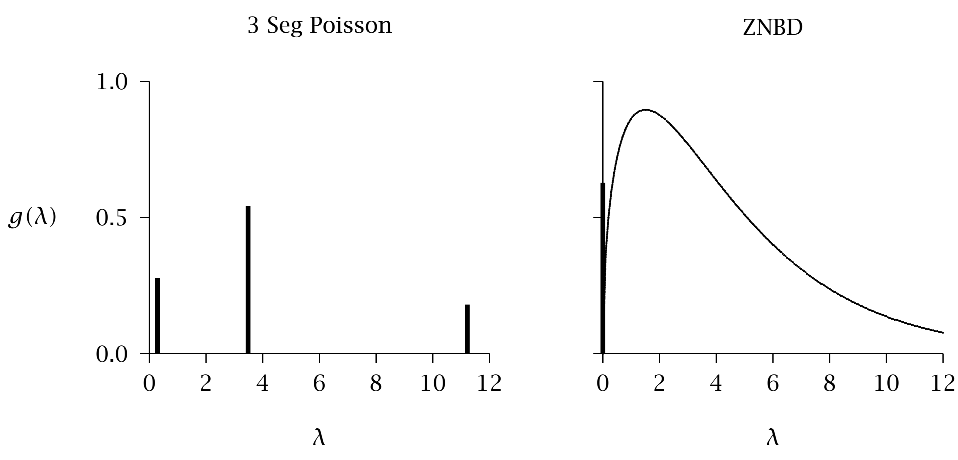
LatentGOLD Results

	Seg 1	Seg 2	Seg 3	Seg 4	LL
mean	3.991				−1545.00
class size	1.000				
mean	1.801	9.115			−1188.83
class size	0.700	0.300			
mean	3.483	0.291	11.210		−1132.04
class size	0.542	0.277	0.181		
mean	2.976	0.202	7.247	12.787	−1130.07
class size	0.500	0.243	0.156	0.106	

Fit of 3 Segment Poisson



Implied Heterogeneity Distribution



$\hat{\pi} = 0.113, \hat{r} = 1.504, \hat{\alpha} = 0.334$

Classification Using Bayes Theorem

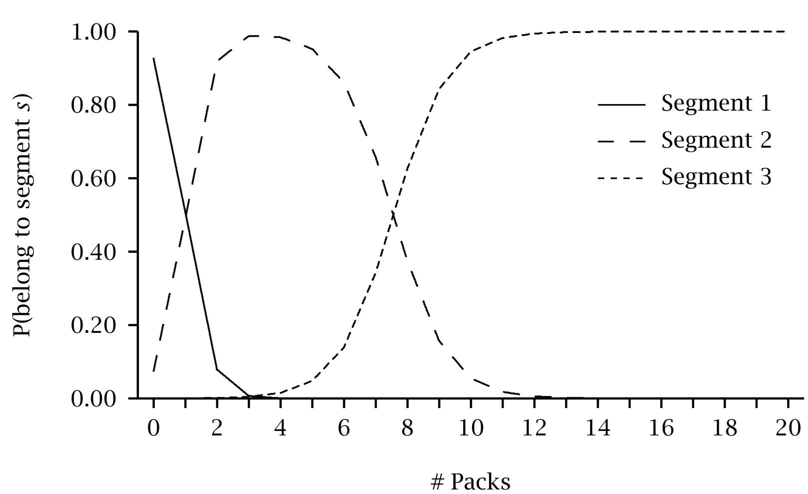
To which "segment" of the mixing distribution does each observation x belong?

- The π_s can be interpreted as the prior probability of λ_s
- By Bayes theorem,

$$P(s \mid X = x) = \frac{P(X = x \mid \lambda_s)\pi_s}{\sum_{s'=1}^{S} P(X = x \mid \lambda_{s'})\pi_{s'}},$$

which can be interpreted as the posterior probability of λ_s

Posterior Probabilities



Conditional Expectations

What is the expected purchase quantity over the next month for a customer who purchased seven packs last week?

$$
\begin{aligned}
E[X(4)] &= E[X(4)|\text{seg } 1]\, P(\text{seg } 1|X=7) \\
&\quad + E[X(4)|\text{seg } 2]\, P(\text{seg } 2|X=7) \\
&\quad + E[X(4)|\text{seg } 3]\, P(\text{seg } 3|X=7) \\
&= (4 \times 0.291) \times 0.0000 \\
&\quad + (4 \times 3.483) \times 0.6575 \\
&\quad + (4 \times 11.216) \times 0.3425 \\
&= 24.5
\end{aligned}
$$

… or 13.9 with “hard assignment” to segment 2.

Concepts and Tools Introduced

- Finite mixture models
- Discrete vs. continuous mixing distributions
- Probability models for classification

Further Reading

Dillon, Wiliam R. and Ajith Kumar (1994), “Latent Structure and Other Mixture Models in Marketing: An Integrative Survey and Overview,” in Richard P. Bagozzi (ed.), *Advanced Methods of Marketing Research*, Oxford: Blackwell.

McLachlan, Geoffrey and David Peel (2000), *Finite Mixture Models*, New York: John Wiley & Sons.

Wedel, Michel and Wagner A. Kamakura (1998), *Market Segmentation: Conceptual and Methodological Foundations*, Boston, MA: Kluwer Academic Publishers.

Problem 5: Who is Visiting khakichinos.com?

(Incorporating Covariates in Count Models)

Background

Khaki Chinos, Inc. is an established clothing catalog company with an online presence at khakichinos.com. While the company is able to track the online *purchasing* behavior of its customers, it has no real idea as to the pattern of *visiting* behaviors by the broader Internet population.

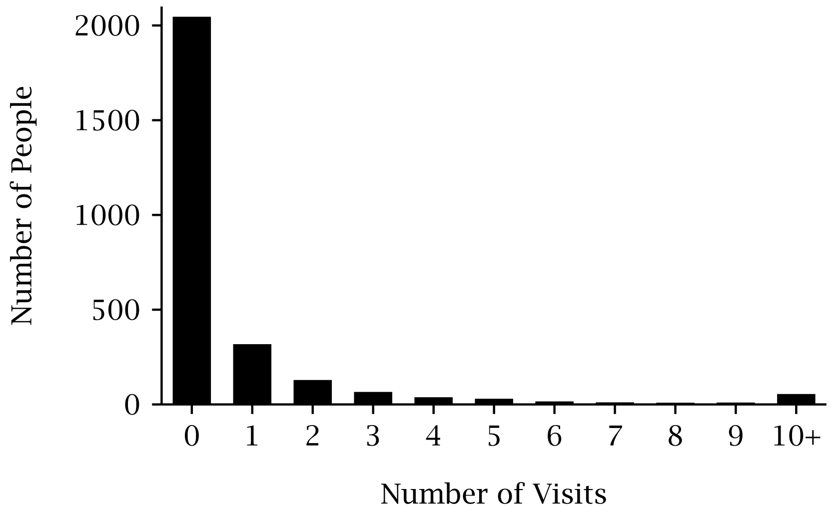
In order to gain an understanding of the aggregate visiting patterns, some Media Metrix panel data has been purchased. For a sample of 2728 people who visited an online apparel site at least once during the second-half of 2000, the dataset reports how many visits each person made to the khakichinos.com web site, along with some demographic information.

Management would like to know whether frequency of visiting the web site is related to demographic characteristics.

Raw Data

ID	# Visits	ln(Income)	Sex	ln(Age)	HH Size
1	0	11.38	1	3.87	2
2	5	9.77	1	4.04	1
3	0	11.08	0	3.33	2
4	0	10.92	1	3.95	3
5	0	10.92	1	2.83	3
6	0	10.92	0	2.94	3
7	0	11.19	0	3.66	2
8	1	11.74	0	4.08	2
9	0	10.02	0	4.25	1
	…				

Distribution of Visits



Modeling Count Data

Recall the NBD:

- At the individual-level, $Y \sim \text{Poisson}(\lambda)$
- λ is distributed across the population according to a gamma distribution with parameters r and α

$$P(Y = y) = \frac{\Gamma(r + y)}{\Gamma(r)y!} \left(\frac{\alpha}{\alpha + 1}\right)^r \left(\frac{1}{\alpha + 1}\right)^y$$

Observed vs. Unobserved Heterogeneity

Unobserved Heterogeneity:

- People differ in their mean (visiting) rate λ
- To account for heterogeneity in λ, we assume it is distributed across the population according to some (parametric) distribution
- But there is no attempt to *explain* how people differ in their mean rates

Observed Heterogeneity:

- We observe how people differ on a set of observable independent (explanatory) variables
- We explicitly link an individual's λ to her observable characteristics

The Poisson Regression Model

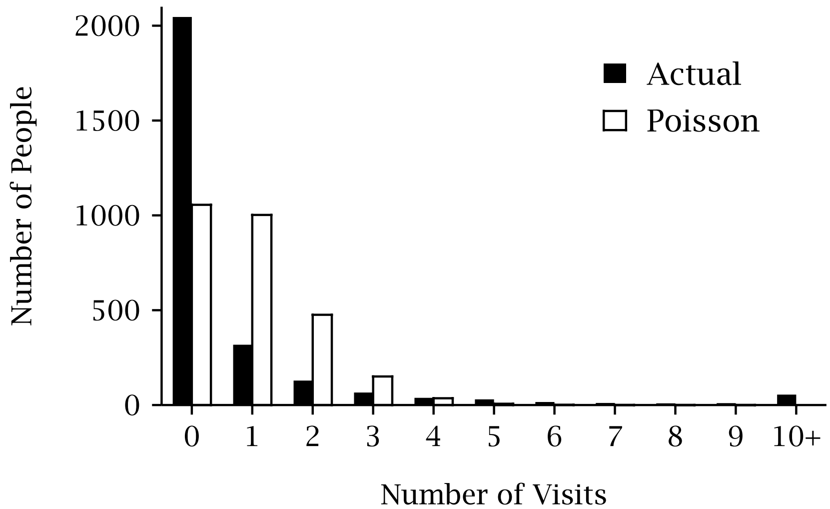
- Let the random variable Y_i denote the number of times individual i visits the site in a unit time period
- At the individual-level, Y_i is assumed to be distributed Poisson with mean λ_i:

$$P(Y_i = y|\lambda_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}$$

- An individual's mean is related to her observable characteristics through the function

$$\lambda_i = \lambda_0 \exp(\boldsymbol{\beta}'\mathbf{x}_i)$$

Fit of Poisson



$\hat{\lambda} = 0.949, LL = -6378.6$

Poisson Regression Results

Variable	Coefficient
λ_0	0.0439
Income	0.0938
Sex	0.0043
Age	0.5882
HH Size	−0.0359
LL	−6291.5
LL_{Poiss}	−6378.6
LR (df = 4)	174.2

Problem 5 -- Poisson reg

	A	B	C	D	E	F	G	H	I	J	K
1	\lambda_0	0.04387			LL =	=SUM(K9:K2736)					
2	B_inc	0.09385									
3	B_sex	0.00426									
4	B_age	0.58825									
5	B_size	-0.0359									
6				=TRANSPOSE(B2:B5)	=TRANSPOSE(B2:B5)	=TRANSPOSE(B2:B5)	=TRANSPOSE(B2:B5)				
7											
8	ID	Total		Income	Sex	Age	Size		lambda	P(Y=y)	ln[P(Y=y)]
9	1	0		11.3793940723457	1	3.87120101090789	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D9:G9))	=POISSON(B9,I9,FALSE)	=LN(J9)
10	2	5		9.76995615991161	1	4.04305126783455	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D10:G10))	=POISSON(B10,I10,FALSE)	=LN(J10)
11	3	0		11.0821425488778	0	3.3322045101752	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D11:G11))	=POISSON(B11,I11,FALSE)	=LN(J11)
12	4	0		10.9150884642146	1	3.95124371858143	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D12:G12))	=POISSON(B12,I12,FALSE)	=LN(J12)
13	5	0		10.9150884642146	1	2.83321334405622	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D13:G13))	=POISSON(B13,I13,FALSE)	=LN(J13)
14	6	0		10.9150884642146	0	2.94443897916644	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D14:G14))	=POISSON(B14,I14,FALSE)	=LN(J14)
15	7	0		11.1913418408428	0	3.66356164612965	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D15:G15))	=POISSON(B15,I15,FALSE)	=LN(J15)
16	8	1		11.7360690162844	0	4.07753744390572	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D16:G16))	=POISSON(B16,I16,FALSE)	=LN(J16)
17	9	0		10.0212705881925	0	4.24849524204936	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D17:G17))	=POISSON(B17,I17,FALSE)	=LN(J17)
18	10	0		10.9150884642146	0	3.85014760171006	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D18:G18))	=POISSON(B18,I18,FALSE)	=LN(J18)
19	11	1		10.7684849900227	0	3.93182563272433	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D19:G19))	=POISSON(B19,I19,FALSE)	=LN(J19)
20	12	0		10.9150884642146	0	3.98898404656427	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D20:G20))	=POISSON(B20,I20,FALSE)	=LN(J20)
21	13	3		10.5320962119585	0	3.63758615972639	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D21:G21))	=POISSON(B21,I21,FALSE)	=LN(J21)
22	14	0		10.9150884642146	0	3.61091791264422	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D22:G22))	=POISSON(B22,I22,FALSE)	=LN(J22)
23	15	0		10.2219412836547	1	3.58351893845611	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D23:G23))	=POISSON(B23,I23,FALSE)	=LN(J23)
24	16	1		10.7684849900227	1	3.25809653802148	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D24:G24))	=POISSON(B24,I24,FALSE)	=LN(J24)
25	17	2		12.2060726455302	0	3.66356164612965	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D25:G25))	=POISSON(B25,I25,FALSE)	=LN(J25)
26	18	0		10.7684849900227	0	3.95124371858143	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D26:G26))	=POISSON(B26,I26,FALSE)	=LN(J26)
27	19	6		11.1913418408428	1	3.3322045101752	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D27:G27))	=POISSON(B27,I27,FALSE)	=LN(J27)
28	20	0		10.3889953683178	1	3.58351893845611	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D28:G28))	=POISSON(B28,I28,FALSE)	=LN(J28)
29	21	2		10.7684849900227	1	3.3322045101752	4		=B$1*EXP(SUMPRODUCT(D$6:G$6,D29:G29))	=POISSON(B29,I29,FALSE)	=LN(J29)
30	22	0		11.1913418408428	1	3.46573590279973	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D30:G30))	=POISSON(B30,I30,FALSE)	=LN(J30)
31	23	0		11.1913418408428	1	3.43398720448515	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D31:G31))	=POISSON(B31,I31,FALSE)	=LN(J31)
32	24	2		11.7360690162844	1	3.80666248977032	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D32:G32))	=POISSON(B32,I32,FALSE)	=LN(J32)
33	25	0		11.3793940723457	0	4.27666611901606	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D33:G33))	=POISSON(B33,I33,FALSE)	=LN(J33)
34	26	0		10.3889953683178	0	4.21950770517611	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D34:G34))	=POISSON(B34,I34,FALSE)	=LN(J34)
35	27	0		10.6572593549125	1	3.49650756146648	4		=B$1*EXP(SUMPRODUCT(D$6:G$6,D35:G35))	=POISSON(B35,I35,FALSE)	=LN(J35)
36	28	0		12.0725412529057	0	3.95124371858143	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D36:G36))	=POISSON(B36,I36,FALSE)	=LN(J36)
37	29	0		10.9150884642146	1	3.80666248977032	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D37:G37))	=POISSON(B37,I37,FALSE)	=LN(J37)
38	30	0		10.9150884642146	0	3.52636052461616	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D38:G38))	=POISSON(B38,I38,FALSE)	=LN(J38)
39	31	0		11.1913418408428	1	3.36729582998647	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D39:G39))	=POISSON(B39,I39,FALSE)	=LN(J39)
40	32	0		10.2219412836547	1	3.13549421592915	4		=B$1*EXP(SUMPRODUCT(D$6:G$6,D40:G40))	=POISSON(B40,I40,FALSE)	=LN(J40)
41	33	0		11.3793940723457	0	3.3322045101752	4		=B$1*EXP(SUMPRODUCT(D$6:G$6,D41:G41))	=POISSON(B41,I41,FALSE)	=LN(J41)
42	34	0		9.07680897935166	1	3.40119738166216	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D42:G42))	=POISSON(B42,I42,FALSE)	=LN(J42)
43	35	0		10.0212705881925	1	3.52636052461616	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D43:G43))	=POISSON(B43,I43,FALSE)	=LN(J43)
44	36	0		11.0821425488778	0	4.06044301054642	4		=B$1*EXP(SUMPRODUCT(D$6:G$6,D44:G44))	=POISSON(B44,I44,FALSE)	=LN(J44)
45	37	2		10.2219412836547	1	3.68887945411394	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D45:G45))	=POISSON(B45,I45,FALSE)	=LN(J45)
46	38	2		12.0725412529057	1	3.68887945411394	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D46:G46))	=POISSON(B46,I46,FALSE)	=LN(J46)
47	39	1		11.0821425488778	0	4.17438726989564	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D47:G47))	=POISSON(B47,I47,FALSE)	=LN(J47)
48	40	0		9.52879410309472	1	2.70805020110221	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D48:G48))	=POISSON(B48,I48,FALSE)	=LN(J48)
49	41	0		11.0821425488778	1	3.80666248977032	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D49:G49))	=POISSON(B49,I49,FALSE)	=LN(J49)
50	42	0		11.3793940723457	1	4.12713438504509	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D50:G50))	=POISSON(B50,I50,FALSE)	=LN(J50)
51	43	0		11.3793940723457	0	4.17438726989564	3		=B$1*EXP(SUMPRODUCT(D$6:G$6,D51:G51))	=POISSON(B51,I51,FALSE)	=LN(J51)
52	44	0		10.5320962119585	1	3.55534806148941	6		=B$1*EXP(SUMPRODUCT(D$6:G$6,D52:G52))	=POISSON(B52,I52,FALSE)	=LN(J52)
53	45	0		10.7684849900227	0	3.2188758248682	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D53:G53))	=POISSON(B53,I53,FALSE)	=LN(J53)
54	46	0		11.3793940723457	1	3.36729582998647	2		=B$1*EXP(SUMPRODUCT(D$6:G$6,D54:G54))	=POISSON(B54,I54,FALSE)	=LN(J54)
55	47	0		11.7360690162844	0	3.04452243772342	4		=B$1*EXP(SUMPRODUCT(D$6:G$6,D55:G55))	=POISSON(B55,I55,FALSE)	=LN(J55)
56	48	0		10.7684849900227	0	3.52636052461616	1		=B$1*EXP(SUMPRODUCT(D$6:G$6,D56:G56))	=POISSON(B56,I56,FALSE)	=LN(J56)

Problem 5 -- Poisson reg

	A	B	C	D	E	F	G	H	I	J
1	\lambda_0	0.0439			LL =	-6291.497				
2	B_inc	0.0938								
3	B_sex	0.0043								
4	B_age	0.5882								
5	B_size	-0.0359								
6				0.0938	0.0043	0.5882	-0.0359			
7										
8	ID	Total		Income	Sex	Age	HH Size		lambda	P(Y=y)
9	1	0		11.38	1	3.87	2		1.16317	0.31249
10	2	5		9.77	1	4.04	1		1.14695	0.00525
11	3	0		11.08	0	3.33	2		0.82031	0.44029
12	4	0		10.92	1	3.95	3		1.12609	0.32430
13	5	0		10.92	1	2.83	3		0.58338	0.55801
14	6	0		10.92	0	2.94	3		0.62017	0.53785
15	7	0		11.19	0	3.66	2		1.00712	0.36527
16	8	1		11.74	0	4.08	2		1.35220	0.34977
17	9	0		10.02	0	4.25	1		1.31954	0.26726
18	10	0		10.92	0	3.85	3		1.05656	0.34765
19	11	1		10.77	0	3.93	2		1.13340	0.36488
20	12	0		10.92	0	3.99	2		1.18839	0.30471
21	13	3		10.53	0	3.64	2		0.93235	0.05317
22	14	0		10.92	0	3.61	1		0.98621	0.37299
23	15	0		10.22	1	3.58	3		0.84992	0.42745
24	16	1		10.77	1	3.26	3		0.73879	0.35291
25	17	2		12.21	0	3.66	2		1.10774	0.20266
26	18	0		10.77	0	3.95	2		1.14642	0.31777
27	19	6		11.19	1	3.33	2		0.83230	0.00020
28	20	0		10.39	1	3.58	2		0.89492	0.40864
29	21	2		10.77	1	3.33	4		0.74449	0.13163
30	22	0		11.19	1	3.47	2		0.90031	0.40644
31	23	0		11.19	1	3.43	2		0.88365	0.41327
32	24	2		11.74	1	3.81	2		1.15796	0.21060
33	25	0		11.38	0	4.28	2		1.47020	0.22988
34	26	0		10.39	0	4.22	2		1.29542	0.27378
35	27	0		10.66	1	3.50	4		0.81152	0.44418
36	28	0		12.07	0	3.95	2		1.29566	0.27372
37	29	0		10.92	1	3.81	3		1.03428	0.35548
38	30	0		10.92	0	3.53	3		0.87333	0.41756
39	31	0		11.19	1	3.37	2		0.84966	0.42756
40	32	0		10.22	1	3.14	4		0.62998	0.53260
41	33	0		11.38	0	3.33	4		0.78506	0.45609
42	34	0		9.08	1	3.40	1		0.73675	0.47867
43	35	0		10.02	1	3.53	1		0.86654	0.42040
44	36	0		11.08	0	4.06	4		1.17175	0.30982
45	37	2		10.22	1	3.69	2		0.93733	0.17206
46	38	2		12.07	1	3.69	2		1.11510	0.20385
47	39	1		11.08	0	4.17	1		1.39549	0.34568
48	40	0		9.53	1	2.71	3		0.47585	0.62136
49	41	0		11.08	1	3.81	3		1.05062	0.34972
50	42	0		11.38	1	4.13	3		1.30446	0.27132
51	43	0		11.38	0	4.17	3		1.33553	0.26302
52	44	0		10.53	1	3.56	6		0.77275	0.46174

Comparing Expected Visit Behavior

	Person A	Person B
Income	59,874	98,716
Sex	M	F
Age	55	33
HH Size	4	2

Who is less likely to have visited the web site?

$$\begin{aligned}\lambda_A &= 0.0439 \times \exp\big(0.0938 \times \ln(59,874) + 0.0043 \times 0 \\ &\quad + 0.5882 \times \ln(55) - 0.0359 \times 4\big) \\ &= 1.127 \\ \lambda_B &= 0.0439 \times \exp\big(0.0938 \times \ln(98,716) + 0.0043 \times 1 \\ &\quad + 0.5882 \times \ln(33) - 0.0359 \times 2\big) \\ &= 0.944\end{aligned}$$

Is $\boldsymbol{\beta}$ Different from 0?

Consider two models, A and B:

> If we can arrive at model B by placing *k* constraints on the parameters of model A, we say that model B is *nested* within model A.

The Poisson model is nested within the Poisson regression model by imposing the constraint $\boldsymbol{\beta} = \mathbf{0}$.

We use the *likelihood ratio test* to determine whether model A, which has more parameters, fits the data better than model B.

The Likelihood Ratio Test

- The null hypothesis is that model A is not different from model B
- Compute the test statistic

$$LR = -2(LL_B - LL_A)$$

- Reject null hypothesis if $LR > \chi^2_{.05,k}$

Computing Standard Errors

- Excel
  - indirectly via a series of likelihood ratio tests
- General modeling environments (e.g., MATLAB, Gauss)
  - easily computed from the Hessian matrix (computed directly or as a by-product of optimization)
- Advanced statistics packages (e.g., Limdep, S-Plus, LatentGOLD)
  - they come for free

S-Plus Poisson Regression Results

Coefficients:

	Value	Std. Error	t value
(Intercept)	-3.126238804	0.40578080	-7.7042552
Income	0.093828021	0.03436347	2.7304580
Sex	0.004259338	0.04089411	0.1041553
Age	0.588249213	0.05472896	10.7484079
HH Size	-0.035907406	0.01528397	-2.3493511

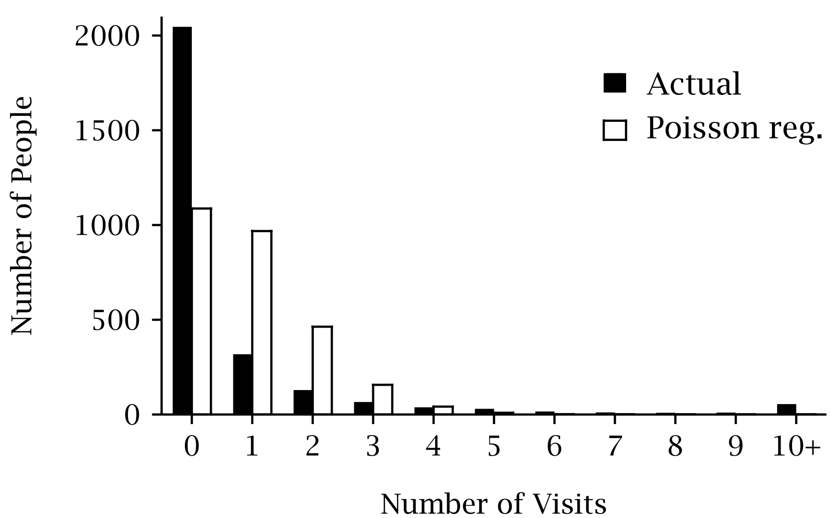
Limdep Poisson Regression Results

Variable	Coefficient	Standard Error	b/St.Er.
Constant	-3.122103284	.40565119	-7.697
INCOME	.9305546493E-01	.34332533E-01	2.710
SEX	.4312514407E-02	.40904265E-01	.105
AGE	.5893014445	.54790230E-01	10.756
HH SIZE	-.3577795361E-01	.15287122E-01	-2.340

LatentGOLD Poisson Regression Results

	Class1	Wald	p-value
Intercept	.	.	.
Dependent Variable	(beta)		
	Class1	Wald	p-value
TOTAL			
	-3.1262	59.3196	1.3e-14
Predictors			
INCOME			
	0.0938	7.4513	0.0064
SEX			
	0.0043	0.0108	0.92
AGE			
	0.5882	115.4660	6.2e-27
HH SIZE			
	-0.0359	5.5165	0.019

Fit of Poisson Regression



The ZIP Regression Model

Because of the “excessive” number of zeros, let us consider the zero-inflated Poisson (ZIP) regression model:

- a proportion π of those people who go to online apparel sites will never visit khakichinos.com
- the visiting behavior of the “ever visitors” can be characterized by the Poisson regression model

$$P(Y_i = y) = \delta_{y=0}\pi + (1-\pi) \times \frac{[\lambda_0 \exp(\boldsymbol{\beta}'\mathbf{x}_i)]^y e^{-\lambda_0 \exp(\boldsymbol{\beta}'\mathbf{x}_i)}}{y!}$$

ZIP Regression Results

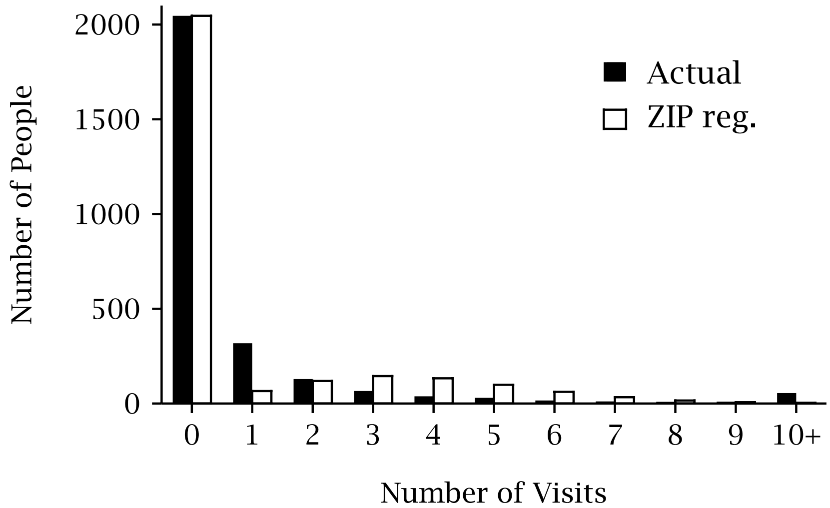
Variable	Coefficient
λ_0	6.6231
Income	−0.0891
Sex	−0.1327
Age	0.1141
HH Size	0.0196
π	0.7433
LL	−4297.5
$LL_{\text{Poiss reg}}$	−6291.5
LR (df = 1)	3988.0

	A	B	C	D	E	F	G	H	I	J	K
1	\lambda_0	6.6231			LL =	=SUM(K10:K2737)					
2	pi	0.7433									
3	B_inc	-0.0891									
4	B_sex	-0.1327									
5	B_age	0.1141									
6	B_size	0.0196									
7				=TRANSPOSE(B3:B6)	=TRANSPOSE(B3:B6)	=TRANSPOSE(B3:B6)	=TRANSPOSE(B3:B6)				
8											
9	ID	Total		Income	Sex	Age	Size		lambda	P(Y=y)	ln[P(Y=y)]
10	1	0		11.3793940723457	1	3.87120101090789	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D10:G10))	=IF(B10=0,B$2,0)+(1-B$2)*POISSON(B10,I10,FALSE)	=LN(J10)
11	2	5		9.76995615991161	1	4.04305126783455	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D11:G11))	=IF(B11=0,B$2,0)+(1-B$2)*POISSON(B11,I11,FALSE)	=LN(J11)
12	3	0		11.0821425488778	0	3.3322045101752	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D12:G12))	=IF(B12=0,B$2,0)+(1-B$2)*POISSON(B12,I12,FALSE)	=LN(J12)
13	4	0		10.9150884642146	1	3.95124371858143	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D13:G13))	=IF(B13=0,B$2,0)+(1-B$2)*POISSON(B13,I13,FALSE)	=LN(J13)
14	5	0		10.9150884642146	1	2.83321334405622	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D14:G14))	=IF(B14=0,B$2,0)+(1-B$2)*POISSON(B14,I14,FALSE)	=LN(J14)
15	6	0		10.9150884642146	0	2.94443897916644	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D15:G15))	=IF(B15=0,B$2,0)+(1-B$2)*POISSON(B15,I15,FALSE)	=LN(J15)
16	7	0		11.1913418408428	0	3.66356164612965	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D16:G16))	=IF(B16=0,B$2,0)+(1-B$2)*POISSON(B16,I16,FALSE)	=LN(J16)
17	8	1		11.7360690162844	0	4.07753744390572	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D17:G17))	=IF(B17=0,B$2,0)+(1-B$2)*POISSON(B17,I17,FALSE)	=LN(J17)
18	9	0		10.0212705881925	0	4.24849524204936	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D18:G18))	=IF(B18=0,B$2,0)+(1-B$2)*POISSON(B18,I18,FALSE)	=LN(J18)
19	10	0		10.9150884642146	0	3.85014760171006	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D19:G19))	=IF(B19=0,B$2,0)+(1-B$2)*POISSON(B19,I19,FALSE)	=LN(J19)
20	11	1		10.7684849900227	0	3.93182563272433	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D20:G20))	=IF(B20=0,B$2,0)+(1-B$2)*POISSON(B20,I20,FALSE)	=LN(J20)
21	12	0		10.9150884642146	0	3.98898404656427	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D21:G21))	=IF(B21=0,B$2,0)+(1-B$2)*POISSON(B21,I21,FALSE)	=LN(J21)
22	13	3		10.5320962119585	0	3.63758615972639	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D22:G22))	=IF(B22=0,B$2,0)+(1-B$2)*POISSON(B22,I22,FALSE)	=LN(J22)
23	14	0		10.9150884642146	0	3.61091791264422	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D23:G23))	=IF(B23=0,B$2,0)+(1-B$2)*POISSON(B23,I23,FALSE)	=LN(J23)
24	15	0		10.2219412836547	1	3.58351893845611	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D24:G24))	=IF(B24=0,B$2,0)+(1-B$2)*POISSON(B24,I24,FALSE)	=LN(J24)
25	16	1		10.7684849900227	1	3.25809653802148	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D25:G25))	=IF(B25=0,B$2,0)+(1-B$2)*POISSON(B25,I25,FALSE)	=LN(J25)
26	17	2		12.2060726455302	0	3.66356164612965	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D26:G26))	=IF(B26=0,B$2,0)+(1-B$2)*POISSON(B26,I26,FALSE)	=LN(J26)
27	18	0		10.7684849900227	0	3.95124371858143	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D27:G27))	=IF(B27=0,B$2,0)+(1-B$2)*POISSON(B27,I27,FALSE)	=LN(J27)
28	19	6		11.1913418408428	1	3.3322045101752	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D28:G28))	=IF(B28=0,B$2,0)+(1-B$2)*POISSON(B28,I28,FALSE)	=LN(J28)
29	20	0		10.3889953683178	1	3.58351893845611	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D29:G29))	=IF(B29=0,B$2,0)+(1-B$2)*POISSON(B29,I29,FALSE)	=LN(J29)
30	21	2		10.7684849900227	1	3.3322045101752	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D30:G30))	=IF(B30=0,B$2,0)+(1-B$2)*POISSON(B30,I30,FALSE)	=LN(J30)
31	22	0		11.1913418408428	1	3.46573590279973	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D31:G31))	=IF(B31=0,B$2,0)+(1-B$2)*POISSON(B31,I31,FALSE)	=LN(J31)
32	23	0		11.1913418408428	1	3.43398720448515	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D32:G32))	=IF(B32=0,B$2,0)+(1-B$2)*POISSON(B32,I32,FALSE)	=LN(J32)
33	24	2		11.7360690162844	1	3.80666248977032	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D33:G33))	=IF(B33=0,B$2,0)+(1-B$2)*POISSON(B33,I33,FALSE)	=LN(J33)
34	25	0		11.3793940723457	0	4.27666611901606	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D34:G34))	=IF(B34=0,B$2,0)+(1-B$2)*POISSON(B34,I34,FALSE)	=LN(J34)
35	26	0		10.3889953683178	0	4.21950770517611	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D35:G35))	=IF(B35=0,B$2,0)+(1-B$2)*POISSON(B35,I35,FALSE)	=LN(J35)
36	27	0		10.6572593549125	1	3.49650756146648	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D36:G36))	=IF(B36=0,B$2,0)+(1-B$2)*POISSON(B36,I36,FALSE)	=LN(J36)
37	28	0		12.0725412529057	0	3.95124371858143	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D37:G37))	=IF(B37=0,B$2,0)+(1-B$2)*POISSON(B37,I37,FALSE)	=LN(J37)
38	29	0		10.9150884642146	1	3.80666248977032	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D38:G38))	=IF(B38=0,B$2,0)+(1-B$2)*POISSON(B38,I38,FALSE)	=LN(J38)
39	30	0		10.9150884642146	0	3.52636052461616	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D39:G39))	=IF(B39=0,B$2,0)+(1-B$2)*POISSON(B39,I39,FALSE)	=LN(J39)
40	31	0		11.1913418408428	1	3.36729582998647	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D40:G40))	=IF(B40=0,B$2,0)+(1-B$2)*POISSON(B40,I40,FALSE)	=LN(J40)
41	32	0		10.2219412836547	1	3.13549421592915	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D41:G41))	=IF(B41=0,B$2,0)+(1-B$2)*POISSON(B41,I41,FALSE)	=LN(J41)
42	33	0		11.3793940723457	0	3.3322045101752	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D42:G42))	=IF(B42=0,B$2,0)+(1-B$2)*POISSON(B42,I42,FALSE)	=LN(J42)
43	34	0		9.07680897935166	1	3.40119738166216	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D43:G43))	=IF(B43=0,B$2,0)+(1-B$2)*POISSON(B43,I43,FALSE)	=LN(J43)
44	35	0		10.0212705881925	1	3.52636052461616	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D44:G44))	=IF(B44=0,B$2,0)+(1-B$2)*POISSON(B44,I44,FALSE)	=LN(J44)
45	36	0		11.0821425488778	0	4.06044301054642	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D45:G45))	=IF(B45=0,B$2,0)+(1-B$2)*POISSON(B45,I45,FALSE)	=LN(J45)
46	37	2		10.2219412836547	1	3.68887945411394	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D46:G46))	=IF(B46=0,B$2,0)+(1-B$2)*POISSON(B46,I46,FALSE)	=LN(J46)
47	38	2		12.0725412529057	1	3.68887945411394	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D47:G47))	=IF(B47=0,B$2,0)+(1-B$2)*POISSON(B47,I47,FALSE)	=LN(J47)
48	39	1		11.0821425488778	0	4.17438726989564	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D48:G48))	=IF(B48=0,B$2,0)+(1-B$2)*POISSON(B48,I48,FALSE)	=LN(J48)
49	40	0		9.52879410309472	1	2.70805020110221	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D49:G49))	=IF(B49=0,B$2,0)+(1-B$2)*POISSON(B49,I49,FALSE)	=LN(J49)
50	41	0		11.0821425488778	1	3.80666248977032	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D50:G50))	=IF(B50=0,B$2,0)+(1-B$2)*POISSON(B50,I50,FALSE)	=LN(J50)
51	42	0		11.3793940723457	1	4.12713438504509	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D51:G51))	=IF(B51=0,B$2,0)+(1-B$2)*POISSON(B51,I51,FALSE)	=LN(J51)
52	43	0		11.3793940723457	0	4.17438726989564	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D52:G52))	=IF(B52=0,B$2,0)+(1-B$2)*POISSON(B52,I52,FALSE)	=LN(J52)
53	44	0		10.5320962119585	1	3.55534806148941	6		=B$1*EXP(SUMPRODUCT(D$7:G$7,D53:G53))	=IF(B53=0,B$2,0)+(1-B$2)*POISSON(B53,I53,FALSE)	=LN(J53)
54	45	0		10.7684849900227	0	3.2188758248682	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D54:G54))	=IF(B54=0,B$2,0)+(1-B$2)*POISSON(B54,I54,FALSE)	=LN(J54)
55	46	0		11.3793940723457	1	3.36729582998647	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D55:G55))	=IF(B55=0,B$2,0)+(1-B$2)*POISSON(B55,I55,FALSE)	=LN(J55)
56	47	0		11.7360690162844	0	3.04452243772342	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D56:G56))	=IF(B56=0,B$2,0)+(1-B$2)*POISSON(B56,I56,FALSE)	=LN(J56)
57	48	0		10.7684849900227	0	3.52636052461616	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D57:G57))	=IF(B57=0,B$2,0)+(1-B$2)*POISSON(B57,I57,FALSE)	=LN(J57)
58	49	0		10.3889953683178	1	2.83321334405622	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D58:G58))	=IF(B58=0,B$2,0)+(1-B$2)*POISSON(B58,I58,FALSE)	=LN(J58)
59	50	0		10.3889953683178	1	2.63905732961526	3		=B$1*EXP(SUMPRODUCT(D$7:G$7,D59:G59))	=IF(B59=0,B$2,0)+(1-B$2)*POISSON(B59,I59,FALSE)	=LN(J59)
60	51	0		11.0821425488778	0	3.73766961828337	5		=B$1*EXP(SUMPRODUCT(D$7:G$7,D60:G60))	=IF(B60=0,B$2,0)+(1-B$2)*POISSON(B60,I60,FALSE)	=LN(J60)
61	52	6		9.76995615991161	0	3.29583686600433	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D61:G61))	=IF(B61=0,B$2,0)+(1-B$2)*POISSON(B61,I61,FALSE)	=LN(J61)
62	53	16		9.76995615991161	0	3.13549421592915	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D62:G62))	=IF(B62=0,B$2,0)+(1-B$2)*POISSON(B62,I62,FALSE)	=LN(J62)
63	54	1		10.2219412836547	1	3.46573590279973	1		=B$1*EXP(SUMPRODUCT(D$7:G$7,D63:G63))	=IF(B63=0,B$2,0)+(1-B$2)*POISSON(B63,I63,FALSE)	=LN(J63)
64	55	0		11.3793940723457	1	3.55534806148941	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D64:G64))	=IF(B64=0,B$2,0)+(1-B$2)*POISSON(B64,I64,FALSE)	=LN(J64)
65	56	0		11.7360690162844	0	3.93182563272433	4		=B$1*EXP(SUMPRODUCT(D$7:G$7,D65:G65))	=IF(B65=0,B$2,0)+(1-B$2)*POISSON(B65,I65,FALSE)	=LN(J65)
66	57	0		10.3889953683178	1	3.73766961828337	2		=B$1*EXP(SUMPRODUCT(D$7:G$7,D66:G66))	=IF(B66=0,B$2,0)+(1-B$2)*POISSON(B66,I66,FALSE)	=LN(J66)

Problem 5 -- ZIP reg

	A	B	C	D	E	F	G	H	I	J
1	\lambda_0	6.6231			LL =	-4297.472				
2	pi	0.7433								
3	B_inc	-0.0891								
4	B_sex	-0.1327								
5	B_age	0.1141								
6	B_size	0.0196								
7				-0.0891	-0.1327	0.1141	0.0196			
8										
9	ID	Total		Income	Sex	Age	HH Size		lambda	P(Y=y)
10	1	0		11.38	1	3.87	2		3.40193	0.75184
11	2	5		9.77	1	4.04	1		3.92698	0.03936
12	3	0		11.08	0	3.33	2		3.75094	0.74932
13	4	0		10.92	1	3.95	3		3.64889	0.74996
14	5	0		10.92	1	2.83	3		3.21182	0.75363
15	6	0		10.92	0	2.94	3		3.71435	0.74954
16	7	0		11.19	0	3.66	2		3.85775	0.74871
17	8	1		11.74	0	4.08	2		3.85266	0.02099
18	9	0		10.02	0	4.25	1		4.48880	0.74617
19	10	0		10.92	0	3.85	3		4.11879	0.74746
20	11	1		10.77	0	3.93	2		4.13048	0.01705
21	12	0		10.92	0	3.99	2		4.10353	0.74752
22	13	3		10.53	0	3.64	2		4.07915	0.04914
23	14	0		10.92	0	3.61	1		3.85413	0.74872
24	15	0		10.22	1	3.58	3		3.72197	0.74949
25	16	1		10.77	1	3.26	3		3.41574	0.02881
26	17	2		12.21	0	3.66	2		3.52410	0.04699
27	18	0		10.77	0	3.95	2		4.13964	0.74737
28	19	6		11.19	1	3.33	2		3.25307	0.01633
29	20	0		10.39	1	3.58	2		3.59593	0.75033
30	21	2		10.77	1	3.33	4		3.51278	0.04722
31	22	0		11.19	1	3.47	2		3.30302	0.75272
32	23	0		11.19	1	3.43	2		3.29107	0.75284
33	24	2		11.74	1	3.81	2		3.27128	0.05214
34	25	0		11.38	0	4.28	2		4.06854	0.74767
35	26	0		10.39	0	4.22	2		4.41520	0.74639
36	27	0		10.66	1	3.50	4		3.61493	0.75019
37	28	0		12.07	0	3.95	2		3.68532	0.74973
38	29	0		10.92	1	3.81	3		3.58919	0.75038
39	30	0		10.92	0	3.53	3		3.96938	0.74813
40	31	0		11.19	1	3.37	2		3.26612	0.75308
41	32	0		10.22	1	3.14	4		3.60630	0.75025
42	33	0		11.38	0	3.33	4		3.79855	0.74904
43	34	0		9.08	1	3.40	1		3.88226	0.74857
44	35	0		10.02	1	3.53	1		3.62011	0.75016
45	36	0		11.08	0	4.06	4		4.23856	0.74699
46	37	2		10.22	1	3.69	2		3.69403	0.04356
47	38	2		12.07	1	3.69	2		3.13223	0.05493
48	39	1		11.08	0	4.17	1		4.04934	0.01812
49	40	0		9.53	1	2.71	3		3.58278	0.75042
50	41	0		11.08	1	3.81	3		3.53613	0.75076
51	42	0		11.38	1	4.13	3		3.57193	0.75050

Fit of ZIP Regression



NBD Regression

The explanatory variables may not fully capture the differences among individuals

To capture the remaining (unobserved) component of differences among individuals, let λ_0 vary across the population according to a gamma distribution with parameters r and α:

$$P(Y_i = y) = \frac{\Gamma(r+y)}{\Gamma(r)y!}\left(\frac{\alpha}{\alpha + \exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^r \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\alpha + \exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^y$$

- Known as the “Negbin II” model in most textbooks
- Collapses to the NBD when $\boldsymbol{\beta} = \mathbf{0}$

	A	B	C	D	E	F	G	H	I	J	K
1	r	0.1388			LL =	=SUM(K10:K2737)					
2	alpha	8.1979									
3	B_inc	0.0734									
4	B_sex	-0.0093									
5	B_age	0.9022									
6	B_size	-0.0243									
7				=TRANSPOSE(B3:B6)	=TRANSPOSE(B3:B6)	=TRANSPOSE(B3:B6)	=TRANSPOSE(B3:B6)				
8											
9	ID	Total		Income	Sex	Age	Size		exp(BX)	P(Y=y)	ln[P(Y=y)]
10	1	0		11.3793940723457	1	3.87120101090789	2		=EXP(SUMPRODUCT(D$7:G$7,D10:G10))	=EXP(GAMMALN(B$1+B10)-GAMMALN(B$1))/FACT(B10)*(B$2/(B$2+I10))^B$1*(I10/(B$2+I10))^B10	=LN(J10)
11	2	5		9.76995615991161	1	4.04305126783455	1		=EXP(SUMPRODUCT(D$7:G$7,D11:G11))	=EXP(GAMMALN(B$1+B11)-GAMMALN(B$1))/FACT(B11)*(B$2/(B$2+I11))^B$1*(I11/(B$2+I11))^B11	=LN(J11)
12	3	0		11.0821425488778	0	3.3322045101752	2		=EXP(SUMPRODUCT(D$7:G$7,D12:G12))	=EXP(GAMMALN(B$1+B12)-GAMMALN(B$1))/FACT(B12)*(B$2/(B$2+I12))^B$1*(I12/(B$2+I12))^B12	=LN(J12)
13	4	0		10.9150884642146	1	3.95124371858143	3		=EXP(SUMPRODUCT(D$7:G$7,D13:G13))	=EXP(GAMMALN(B$1+B13)-GAMMALN(B$1))/FACT(B13)*(B$2/(B$2+I13))^B$1*(I13/(B$2+I13))^B13	=LN(J13)
14	5	0		10.9150884642146	1	2.83321334405622	3		=EXP(SUMPRODUCT(D$7:G$7,D14:G14))	=EXP(GAMMALN(B$1+B14)-GAMMALN(B$1))/FACT(B14)*(B$2/(B$2+I14))^B$1*(I14/(B$2+I14))^B14	=LN(J14)
15	6	0		10.9150884642146	0	2.94443897916644	3		=EXP(SUMPRODUCT(D$7:G$7,D15:G15))	=EXP(GAMMALN(B$1+B15)-GAMMALN(B$1))/FACT(B15)*(B$2/(B$2+I15))^B$1*(I15/(B$2+I15))^B15	=LN(J15)
16	7	0		11.1913418408428	0	3.66356164612965	2		=EXP(SUMPRODUCT(D$7:G$7,D16:G16))	=EXP(GAMMALN(B$1+B16)-GAMMALN(B$1))/FACT(B16)*(B$2/(B$2+I16))^B$1*(I16/(B$2+I16))^B16	=LN(J16)
17	8	1		11.7360690162844	0	4.07753744390572	2		=EXP(SUMPRODUCT(D$7:G$7,D17:G17))	=EXP(GAMMALN(B$1+B17)-GAMMALN(B$1))/FACT(B17)*(B$2/(B$2+I17))^B$1*(I17/(B$2+I17))^B17	=LN(J17)
18	9	0		10.0212705881925	0	4.24849524204936	1		=EXP(SUMPRODUCT(D$7:G$7,D18:G18))	=EXP(GAMMALN(B$1+B18)-GAMMALN(B$1))/FACT(B18)*(B$2/(B$2+I18))^B$1*(I18/(B$2+I18))^B18	=LN(J18)
19	10	0		10.9150884642146	0	3.85014760171006	3		=EXP(SUMPRODUCT(D$7:G$7,D19:G19))	=EXP(GAMMALN(B$1+B19)-GAMMALN(B$1))/FACT(B19)*(B$2/(B$2+I19))^B$1*(I19/(B$2+I19))^B19	=LN(J19)
20	11	1		10.7684849900227	0	3.93182563272433	2		=EXP(SUMPRODUCT(D$7:G$7,D20:G20))	=EXP(GAMMALN(B$1+B20)-GAMMALN(B$1))/FACT(B20)*(B$2/(B$2+I20))^B$1*(I20/(B$2+I20))^B20	=LN(J20)
21	12	0		10.9150884642146	0	3.98898404656427	2		=EXP(SUMPRODUCT(D$7:G$7,D21:G21))	=EXP(GAMMALN(B$1+B21)-GAMMALN(B$1))/FACT(B21)*(B$2/(B$2+I21))^B$1*(I21/(B$2+I21))^B21	=LN(J21)
22	13	3		10.5320962119585	0	3.63758615972639	2		=EXP(SUMPRODUCT(D$7:G$7,D22:G22))	=EXP(GAMMALN(B$1+B22)-GAMMALN(B$1))/FACT(B22)*(B$2/(B$2+I22))^B$1*(I22/(B$2+I22))^B22	=LN(J22)
23	14	0		10.9150884642146	0	3.61091791264422	1		=EXP(SUMPRODUCT(D$7:G$7,D23:G23))	=EXP(GAMMALN(B$1+B23)-GAMMALN(B$1))/FACT(B23)*(B$2/(B$2+I23))^B$1*(I23/(B$2+I23))^B23	=LN(J23)
24	15	0		10.2219412836547	1	3.58351893845611	3		=EXP(SUMPRODUCT(D$7:G$7,D24:G24))	=EXP(GAMMALN(B$1+B24)-GAMMALN(B$1))/FACT(B24)*(B$2/(B$2+I24))^B$1*(I24/(B$2+I24))^B24	=LN(J24)
25	16	1		10.7684849900227	1	3.25809653802148	3		=EXP(SUMPRODUCT(D$7:G$7,D25:G25))	=EXP(GAMMALN(B$1+B25)-GAMMALN(B$1))/FACT(B25)*(B$2/(B$2+I25))^B$1*(I25/(B$2+I25))^B25	=LN(J25)
26	17	2		12.2060726455302	0	3.66356164612965	2		=EXP(SUMPRODUCT(D$7:G$7,D26:G26))	=EXP(GAMMALN(B$1+B26)-GAMMALN(B$1))/FACT(B26)*(B$2/(B$2+I26))^B$1*(I26/(B$2+I26))^B26	=LN(J26)
27	18	0		10.7684849900227	0	3.95124371858143	2		=EXP(SUMPRODUCT(D$7:G$7,D27:G27))	=EXP(GAMMALN(B$1+B27)-GAMMALN(B$1))/FACT(B27)*(B$2/(B$2+I27))^B$1*(I27/(B$2+I27))^B27	=LN(J27)
28	19	6		11.1913418408428	1	3.3322045101752	2		=EXP(SUMPRODUCT(D$7:G$7,D28:G28))	=EXP(GAMMALN(B$1+B28)-GAMMALN(B$1))/FACT(B28)*(B$2/(B$2+I28))^B$1*(I28/(B$2+I28))^B28	=LN(J28)
29	20	0		10.3889953683178	1	3.58351893845611	2		=EXP(SUMPRODUCT(D$7:G$7,D29:G29))	=EXP(GAMMALN(B$1+B29)-GAMMALN(B$1))/FACT(B29)*(B$2/(B$2+I29))^B$1*(I29/(B$2+I29))^B29	=LN(J29)
30	21	2		10.7684849900227	1	3.3322045101752	4		=EXP(SUMPRODUCT(D$7:G$7,D30:G30))	=EXP(GAMMALN(B$1+B30)-GAMMALN(B$1))/FACT(B30)*(B$2/(B$2+I30))^B$1*(I30/(B$2+I30))^B30	=LN(J30)
31	22	0		11.1913418408428	1	3.46573590279973	2		=EXP(SUMPRODUCT(D$7:G$7,D31:G31))	=EXP(GAMMALN(B$1+B31)-GAMMALN(B$1))/FACT(B31)*(B$2/(B$2+I31))^B$1*(I31/(B$2+I31))^B31	=LN(J31)
32	23	0		11.1913418408428	1	3.43398720448515	2		=EXP(SUMPRODUCT(D$7:G$7,D32:G32))	=EXP(GAMMALN(B$1+B32)-GAMMALN(B$1))/FACT(B32)*(B$2/(B$2+I32))^B$1*(I32/(B$2+I32))^B32	=LN(J32)
33	24	2		11.7360690162844	1	3.80666248977032	2		=EXP(SUMPRODUCT(D$7:G$7,D33:G33))	=EXP(GAMMALN(B$1+B33)-GAMMALN(B$1))/FACT(B33)*(B$2/(B$2+I33))^B$1*(I33/(B$2+I33))^B33	=LN(J33)
34	25	0		11.3793940723457	0	4.27666611901606	2		=EXP(SUMPRODUCT(D$7:G$7,D34:G34))	=EXP(GAMMALN(B$1+B34)-GAMMALN(B$1))/FACT(B34)*(B$2/(B$2+I34))^B$1*(I34/(B$2+I34))^B34	=LN(J34)
35	26	0		10.3889953683178	0	4.21950770517611	2		=EXP(SUMPRODUCT(D$7:G$7,D35:G35))	=EXP(GAMMALN(B$1+B35)-GAMMALN(B$1))/FACT(B35)*(B$2/(B$2+I35))^B$1*(I35/(B$2+I35))^B35	=LN(J35)
36	27	0		10.6572593549125	1	3.49650756146648	4		=EXP(SUMPRODUCT(D$7:G$7,D36:G36))	=EXP(GAMMALN(B$1+B36)-GAMMALN(B$1))/FACT(B36)*(B$2/(B$2+I36))^B$1*(I36/(B$2+I36))^B36	=LN(J36)
37	28	0		12.0725412529057	0	3.95124371858143	2		=EXP(SUMPRODUCT(D$7:G$7,D37:G37))	=EXP(GAMMALN(B$1+B37)-GAMMALN(B$1))/FACT(B37)*(B$2/(B$2+I37))^B$1*(I37/(B$2+I37))^B37	=LN(J37)
38	29	0		10.9150884642146	1	3.80666248977032	3		=EXP(SUMPRODUCT(D$7:G$7,D38:G38))	=EXP(GAMMALN(B$1+B38)-GAMMALN(B$1))/FACT(B38)*(B$2/(B$2+I38))^B$1*(I38/(B$2+I38))^B38	=LN(J38)
39	30	0		10.9150884642146	0	3.52636052461616	3		=EXP(SUMPRODUCT(D$7:G$7,D39:G39))	=EXP(GAMMALN(B$1+B39)-GAMMALN(B$1))/FACT(B39)*(B$2/(B$2+I39))^B$1*(I39/(B$2+I39))^B39	=LN(J39)
40	31	0		11.1913418408428	1	3.36729582998647	2		=EXP(SUMPRODUCT(D$7:G$7,D40:G40))	=EXP(GAMMALN(B$1+B40)-GAMMALN(B$1))/FACT(B40)*(B$2/(B$2+I40))^B$1*(I40/(B$2+I40))^B40	=LN(J40)
41	32	0		10.2219412836547	1	3.13549421592915	4		=EXP(SUMPRODUCT(D$7:G$7,D41:G41))	=EXP(GAMMALN(B$1+B41)-GAMMALN(B$1))/FACT(B41)*(B$2/(B$2+I41))^B$1*(I41/(B$2+I41))^B41	=LN(J41)
42	33	0		11.3793940723457	0	3.3322045101752	4		=EXP(SUMPRODUCT(D$7:G$7,D42:G42))	=EXP(GAMMALN(B$1+B42)-GAMMALN(B$1))/FACT(B42)*(B$2/(B$2+I42))^B$1*(I42/(B$2+I42))^B42	=LN(J42)
43	34	0		9.07680897935166	1	3.40119738166216	1		=EXP(SUMPRODUCT(D$7:G$7,D43:G43))	=EXP(GAMMALN(B$1+B43)-GAMMALN(B$1))/FACT(B43)*(B$2/(B$2+I43))^B$1*(I43/(B$2+I43))^B43	=LN(J43)
44	35	0		10.0212705881925	1	3.52636052461616	1		=EXP(SUMPRODUCT(D$7:G$7,D44:G44))	=EXP(GAMMALN(B$1+B44)-GAMMALN(B$1))/FACT(B44)*(B$2/(B$2+I44))^B$1*(I44/(B$2+I44))^B44	=LN(J44)
45	36	0		11.0821425488778	0	4.06044301054642	4		=EXP(SUMPRODUCT(D$7:G$7,D45:G45))	=EXP(GAMMALN(B$1+B45)-GAMMALN(B$1))/FACT(B45)*(B$2/(B$2+I45))^B$1*(I45/(B$2+I45))^B45	=LN(J45)
46	37	2		10.2219412836547	1	3.68887945411394	2		=EXP(SUMPRODUCT(D$7:G$7,D46:G46))	=EXP(GAMMALN(B$1+B46)-GAMMALN(B$1))/FACT(B46)*(B$2/(B$2+I46))^B$1*(I46/(B$2+I46))^B46	=LN(J46)
47	38	2		12.0725412529057	1	3.68887945411394	2		=EXP(SUMPRODUCT(D$7:G$7,D47:G47))	=EXP(GAMMALN(B$1+B47)-GAMMALN(B$1))/FACT(B47)*(B$2/(B$2+I47))^B$1*(I47/(B$2+I47))^B47	=LN(J47)
48	39	1		11.0821425488778	0	4.17438726989564	1		=EXP(SUMPRODUCT(D$7:G$7,D48:G48))	=EXP(GAMMALN(B$1+B48)-GAMMALN(B$1))/FACT(B48)*(B$2/(B$2+I48))^B$1*(I48/(B$2+I48))^B48	=LN(J48)
49	40	0		9.52879410309472	1	2.70805020110221	3		=EXP(SUMPRODUCT(D$7:G$7,D49:G49))	=EXP(GAMMALN(B$1+B49)-GAMMALN(B$1))/FACT(B49)*(B$2/(B$2+I49))^B$1*(I49/(B$2+I49))^B49	=LN(J49)
50	41	0		11.0821425488778	1	3.80666248977032	3		=EXP(SUMPRODUCT(D$7:G$7,D50:G50))	=EXP(GAMMALN(B$1+B50)-GAMMALN(B$1))/FACT(B50)*(B$2/(B$2+I50))^B$1*(I50/(B$2+I50))^B50	=LN(J50)
51	42	0		11.3793940723457	1	4.12713438504509	3		=EXP(SUMPRODUCT(D$7:G$7,D51:G51))	=EXP(GAMMALN(B$1+B51)-GAMMALN(B$1))/FACT(B51)*(B$2/(B$2+I51))^B$1*(I51/(B$2+I51))^B51	=LN(J51)
52	43	0		11.3793940723457	0	4.17438726989564	3		=EXP(SUMPRODUCT(D$7:G$7,D52:G52))	=EXP(GAMMALN(B$1+B52)-GAMMALN(B$1))/FACT(B52)*(B$2/(B$2+I52))^B$1*(I52/(B$2+I52))^B52	=LN(J52)
53	44	0		10.5320962119585	1	3.55534806148941	6		=EXP(SUMPRODUCT(D$7:G$7,D53:G53))	=EXP(GAMMALN(B$1+B53)-GAMMALN(B$1))/FACT(B53)*(B$2/(B$2+I53))^B$1*(I53/(B$2+I53))^B53	=LN(J53)
54	45	0		10.7684849900227	0	3.2188758248682	1		=EXP(SUMPRODUCT(D$7:G$7,D54:G54))	=EXP(GAMMALN(B$1+B54)-GAMMALN(B$1))/FACT(B54)*(B$2/(B$2+I54))^B$1*(I54/(B$2+I54))^B54	=LN(J54)
55	46	0		11.3793940723457	1	3.36729582998647	2		=EXP(SUMPRODUCT(D$7:G$7,D55:G55))	=EXP(GAMMALN(B$1+B55)-GAMMALN(B$1))/FACT(B55)*(B$2/(B$2+I55))^B$1*(I55/(B$2+I55))^B55	=LN(J55)
56	47	0		11.7360690162844	0	3.04452243772342	4		=EXP(SUMPRODUCT(D$7:G$7,D56:G56))	=EXP(GAMMALN(B$1+B56)-GAMMALN(B$1))/FACT(B56)*(B$2/(B$2+I56))^B$1*(I56/(B$2+I56))^B56	=LN(J56)
57	48	0		10.7684849900227	0	3.52636052461616	1		=EXP(SUMPRODUCT(D$7:G$7,D57:G57))	=EXP(GAMMALN(B$1+B57)-GAMMALN(B$1))/FACT(B57)*(B$2/(B$2+I57))^B$1*(I57/(B$2+I57))^B57	=LN(J57)
58	49	0		10.3889953683178	1	2.83321334405622	3		=EXP(SUMPRODUCT(D$7:G$7,D58:G58))	=EXP(GAMMALN(B$1+B58)-GAMMALN(B$1))/FACT(B58)*(B$2/(B$2+I58))^B$1*(I58/(B$2+I58))^B58	=LN(J58)
59	50	0		10.3889953683178	1	2.63905732961526	3		=EXP(SUMPRODUCT(D$7:G$7,D59:G59))	=EXP(GAMMALN(B$1+B59)-GAMMALN(B$1))/FACT(B59)*(B$2/(B$2+I59))^B$1*(I59/(B$2+I59))^B59	=LN(J59)
60	51	0		11.0821425488778	0	3.73766961828337	5		=EXP(SUMPRODUCT(D$7:G$7,D60:G60))	=EXP(GAMMALN(B$1+B60)-GAMMALN(B$1))/FACT(B60)*(B$2/(B$2+I60))^B$1*(I60/(B$2+I60))^B60	=LN(J60)
61	52	6		9.76995615991161	0	3.29583686600433	1		=EXP(SUMPRODUCT(D$7:G$7,D61:G61))	=EXP(GAMMALN(B$1+B61)-GAMMALN(B$1))/FACT(B61)*(B$2/(B$2+I61))^B$1*(I61/(B$2+I61))^B61	=LN(J61)
62	53	16		9.76995615991161	0	3.13549421592915	1		=EXP(SUMPRODUCT(D$7:G$7,D62:G62))	=EXP(GAMMALN(B$1+B62)-GAMMALN(B$1))/FACT(B62)*(B$2/(B$2+I62))^B$1*(I62/(B$2+I62))^B62	=LN(J62)
63	54	1		10.2219412836547	1	3.46573590279973	1		=EXP(SUMPRODUCT(D$7:G$7,D63:G63))	=EXP(GAMMALN(B$1+B63)-GAMMALN(B$1))/FACT(B63)*(B$2/(B$2+I63))^B$1*(I63/(B$2+I63))^B63	=LN(J63)
64	55	0		11.3793940723457	1	3.55534806148941	2		=EXP(SUMPRODUCT(D$7:G$7,D64:G64))	=EXP(GAMMALN(B$1+B64)-GAMMALN(B$1))/FACT(B64)*(B$2/(B$2+I64))^B$1*(I64/(B$2+I64))^B64	=LN(J64)
65	56	0		11.7360690162844	0	3.93182563272433	4		=EXP(SUMPRODUCT(D$7:G$7,D65:G65))	=EXP(GAMMALN(B$1+B65)-GAMMALN(B$1))/FACT(B65)*(B$2/(B$2+I65))^B$1*(I65/(B$2+I65))^B65	=LN(J65)
66	57	0		10.3889953683178	1	3.73766961828337	2		=EXP(SUMPRODUCT(D$7:G$7,D66:G66))	=EXP(GAMMALN(B$1+B66)-GAMMALN(B$1))/FACT(B66)*(B$2/(B$2+I66))^B$1*(I66/(B$2+I66))^B66	=LN(J66)
67	58	1		11.1913418408428	1	3.49650756146648	2		=EXP(SUMPRODUCT(D$7:G$7,D67:G67))	=EXP(GAMMALN(B$1+B67)-GAMMALN(B$1))/FACT(B67)*(B$2/(B$2+I67))^B$1*(I67/(B$2+I67))^B67	=LN(J67)
68	59	0		10.2219412836547	0	3.61091791264422	4		=EXP(SUMPRODUCT(D$7:G$7,D68:G68))	=EXP(GAMMALN(B$1+B68)-GAMMALN(B$1))/FACT(B68)*(B$2/(B$2+I68))^B$1*(I68/(B$2+I68))^B68	=LN(J68)
69	60	0		10.9150884642146	0	3.89182029811063	1		=EXP(SUMPRODUCT(D$7:G$7,D69:G69))	=EXP(GAMMALN(B$1+B69)-GAMMALN(B$1))/FACT(B69)*(B$2/(B$2+I69))^B$1*(I69/(B$2+I69))^B69	=LN(J69)
70	61	0		10.9150884642146	1	3.46573590279973	3		=EXP(SUMPRODUCT(D$7:G$7,D70:G70))	=EXP(GAMMALN(B$1+B70)-GAMMALN(B$1))/FACT(B70)*(B$2/(B$2+I70))^B$1*(I70/(B$2+I70))^B70	=LN(J70)
71	62	0		10.3889953683178	1	3.36729582998647	2		=EXP(SUMPRODUCT(D$7:G$7,D71:G71))	=EXP(GAMMALN(B$1+B71)-GAMMALN(B$1))/FACT(B71)*(B$2/(B$2+I71))^B$1*(I71/(B$2+I71))^B71	=LN(J71)
72	63	0		10.3889953683178	0	3.46573590279973	2		=EXP(SUMPRODUCT(D$7:G$7,D72:G72))	=EXP(GAMMALN(B$1+B72)-GAMMALN(B$1))/FACT(B72)*(B$2/(B$2+I72))^B$1*(I72/(B$2+I72))^B72	=LN(J72)
73	64	0		10.7684849900227	1	3.78418963391826	2		=EXP(SUMPRODUCT(D$7:G$7,D73:G73))	=EXP(GAMMALN(B$1+B73)-GAMMALN(B$1))/FACT(B73)*(B$2/(B$2+I73))^B$1*(I73/(B$2+I73))^B73	=LN(J73)
74	65	0		9.76995615991161	1	3.71357206670431	2		=EXP(SUMPRODUCT(D$7:G$7,D74:G74))	=EXP(GAMMALN(B$1+B74)-GAMMALN(B$1))/FACT(B74)*(B$2/(B$2+I74))^B$1*(I74/(B$2+I74))^B74	=LN(J74)
75	66	0		10.0212705881925	1	3.17805383034795	2		=EXP(SUMPRODUCT(D$7:G$7,D75:G75))	=EXP(GAMMALN(B$1+B75)-GAMMALN(B$1))/FACT(B75)*(B$2/(B$2+I75))^B$1*(I75/(B$2+I75))^B75	=LN(J75)
76	67	0		11.3793940723457	0	4.00733318523247	2		=EXP(SUMPRODUCT(D$7:G$7,D76:G76))	=EXP(GAMMALN(B$1+B76)-GAMMALN(B$1))/FACT(B76)*(B$2/(B$2+I76))^B$1*(I76/(B$2+I76))^B76	=LN(J76)

Problem 5 -- NBD reg

	A	B	C	D	E	F	G	H	I	J
1	r	0.1388			LL =	-2888.966				
2	alpha	8.1979								
3	B_inc	0.0734								
4	B_sex	-0.0093								
5	B_age	0.9022								
6	B_size	-0.0243								
7				0.0734	-0.0093	0.9022	-0.0243			
8										
9	ID	Total		Income	Sex	Age	HH Size		exp(BX)	P(Y=y)
10	1	0		11.38	1	3.87	2		71.51161	0.72936
11	2	5		9.77	1	4.04	1		76.02589	0.01587
12	3	0		11.08	0	3.33	2		43.42559	0.77467
13	4	0		10.92	1	3.95	3		72.50603	0.72810
14	5	0		10.92	1	2.83	3		26.44384	0.81876
15	6	0		10.92	0	2.94	3		29.50734	0.80919
16	7	0		11.19	0	3.66	2		59.02749	0.74680
17	8	1		11.74	0	4.08	2		89.25195	0.09014
18	9	0		10.02	0	4.25	1		94.07931	0.70456
19	10	0		10.92	0	3.85	3		66.80224	0.73555
20	11	1		10.77	0	3.93	2		72.89216	0.09075
21	12	0		10.92	0	3.99	2		77.57994	0.72197
22	13	3		10.53	0	3.64	2		54.93643	0.02795
23	14	0		10.92	0	3.61	1		56.51751	0.75075
24	15	0		10.22	1	3.58	3		49.45389	0.76289
25	16	1		10.77	1	3.26	3		38.38151	0.08984
26	17	2		12.21	0	3.66	2		63.59217	0.04587
27	18	0		10.77	0	3.95	2		74.18036	0.72603
28	19	6		11.19	1	3.33	2		43.37107	0.00859
29	20	0		10.39	1	3.58	2		51.29650	0.75957
30	21	2		10.77	1	3.33	4		40.04943	0.04257
31	22	0		11.19	1	3.47	2		48.92360	0.76387
32	23	0		11.19	1	3.43	2		47.54218	0.76647
33	24	2		11.74	1	3.81	2		69.25654	0.04625
34	25	0		11.38	0	4.28	2		104.05587	0.69552
35	26	0		10.39	0	4.22	2		91.89630	0.70667
36	27	0		10.66	1	3.50	4		46.07074	0.76932
37	28	0		12.07	0	3.95	2		81.63227	0.71736
38	29	0		10.92	1	3.81	3		63.63946	0.73996
39	30	0		10.92	0	3.53	3		49.88038	0.76211
40	31	0		11.19	1	3.37	2		44.76608	0.77192
41	32	0		10.22	1	3.14	4		32.21815	0.80143
42	33	0		11.38	0	3.33	4		42.27648	0.77710
43	34	0		9.08	1	3.40	1		40.49327	0.78098
44	35	0		10.02	1	3.53	1		48.58828	0.76449
45	36	0		11.08	0	4.06	4		79.79024	0.71943
46	37	2		10.22	1	3.69	2		55.72407	0.04515
47	38	2		12.07	1	3.69	2		63.83215	0.04589
48	39	1		11.08	0	4.17	1		95.12148	0.08988
49	40	0		9.53	1	2.71	3		21.33489	0.83709
50	41	0		11.08	1	3.81	3		64.42466	0.73884
51	42	0		11.38	1	4.13	3		87.92064	0.71066

NBD Regression Results

Variable	Coefficient
r	0.1388
α	8.1979
Income	0.0734
Sex	−0.0093
Age	0.9022
HH Size	−0.0243
LL	−2889.0

S-Plus NBD Regression Results

```
Coefficients:
                    Value Std. Error    t value
(Intercept) -4.047149702 1.10159557 -3.6738979
     Income  0.074549233 0.09555222  0.7801936
        Sex -0.005240835 0.11592793 -0.0452077
        Age  0.889862966 0.14072030  6.3236289
    HH Size -0.025094493 0.04187696 -0.5992435

              Theta:  0.13878
          Std. Err.:  0.00726
```

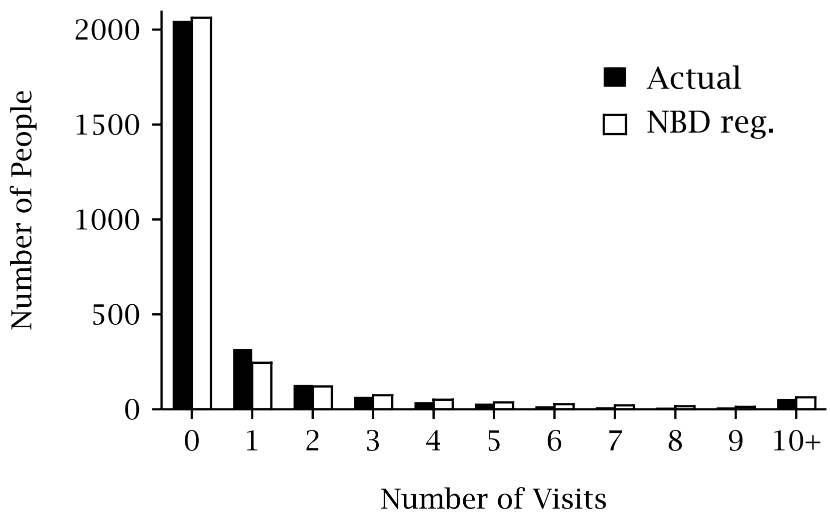
Limdep NBD Regression Results

```
+---------+--------------+---------------+--------+
|Variable | Coefficient  | Standard Error |b/St.Er.|
+---------+--------------+---------------+--------+
 Constant -4.077239653      1.0451741       -3.901
 INCOME    .7237686001E-01  .76663437E-01     .944
 SEX      -.9009160129E-02  .11425700        -.079
 AGE       .9045111135      .17741724        5.098
 HH SIZE  -.2406546843E-01  .38695426E-01    -.622
          Overdispersion parameter
 Alpha     7.206708844      .33334006       21.620
```

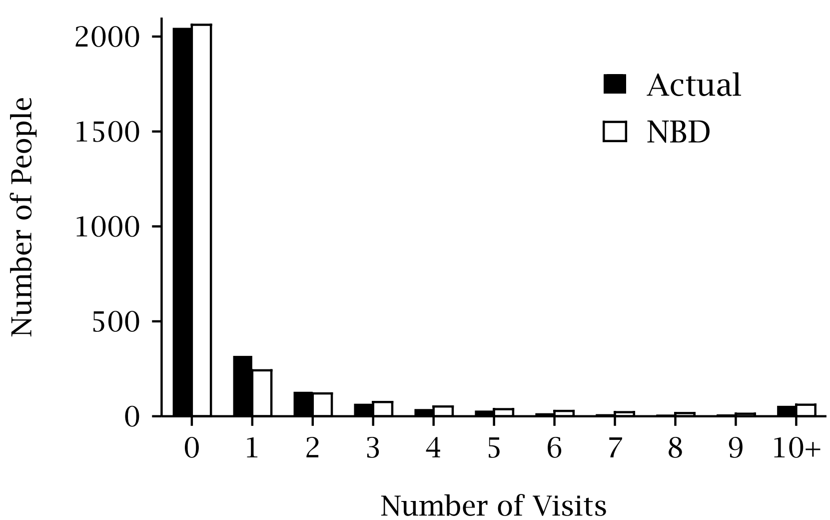

Summary of Regression Results

Variable	Poisson	ZIP	NBD
λ_0	0.0439	6.6231	
r			0.1388
α			8.1979
Income	0.0938	−0.0891	0.0734
Sex	0.0043	−0.1327	−0.0093
Age	0.5882	0.1141	0.9022
HH Size	−0.0359	0.0196	−0.0243
π		0.7433	
LL	−6291.5	−4297.5	−2889.0

Fit of NBD Regression



Fit of NBD



$$\hat{r} = 0.134, \hat{\alpha} = 0.141, LL = -2905.6$$

Concepts and Tools Introduced

- Incorporating covariate effects in count models
- Poisson (and NBD) regression models
- The value of covariates is frequently over-emphasized

Further Reading

Cameron, A. Colin and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.

Wedel, Michel and Wagner A. Kamakura (1998), *Market Segmentation: Conceptual and Methodological Foundations*, Boston, MA: Kluwer Academic Publishers.

Winkelmann, Rainer (1997), *Econometric Analysis of Count Data*, 2nd, revised and enlarged edition, Berlin: Springer.

Introducing Covariates: The General Case

- Select a probability distribution that characterizes the individual-level behavior of interest:

$$f(y|\theta_i)$$

- Make the individual-level latent trait(s) a function of (time-invariant) covariates:

$$\theta_i = s(\theta_0, \mathbf{x}_i)$$

- Specify a mixing distribution to capture the heterogeneity in θ_i not "explained" by $\mathbf{x}_i$
- Derive the corresponding aggregate distribution

$$f(y|\mathbf{x}_i) = \int f(y|\theta_0, \mathbf{x}_i) g(\theta_0)\, d\theta_0$$

Covariates in Timing Models

- If the covariates are time-invariant, we can make λ a direct function of covariates:

$$F(t) = 1 - e^{-\lambda_0 \exp(\boldsymbol{\beta}'\mathbf{x}_i)t}$$

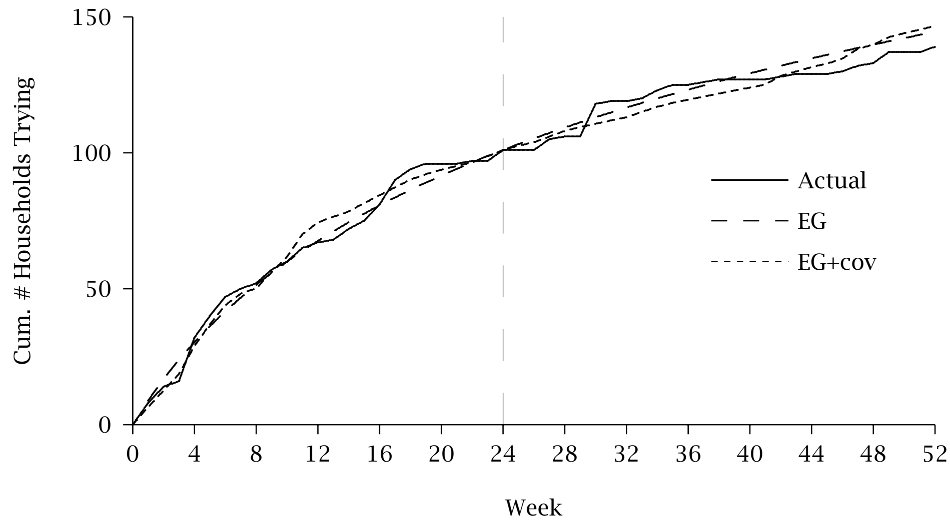
- If the covariates are time-varying (i.e., $\mathbf{x}_{it}$), we incorporate their effects via the hazard rate function

$$F(t) = 1 - e^{-\lambda_0 A(t)}$$

  where $A(t) = \sum_{j=1}^{t} \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})$

- Known as "proportional hazards regression"

Comparing EG with EG+cov



Covariates in “Choice” Models

Two options for binary choice:

- The beta-logistic model
  - a generalization of the beta-binomial model in which the mean is made a function of (time-invariant) covariates
  - covariate effects not introduced at the level of the individual
- Finite mixture of binary logits:

$$P(Y = 1) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\exp(\boldsymbol{\beta}'\mathbf{x}_i) + 1}$$

with some elements of $\boldsymbol{\beta}$ varying across segments

Case Study:
Forecasting Repeat Sales at CDNOW

Some Important Questions

- What is the expected lifetime value of our customer base?
- How effective are our promotional campaigns at generating incremental sales volume?
- What type of capacity do we need to maintain?

At this stage of the Internet's evolution, accurate sales forecasts are as much of an oxymoron as "military intelligence".

Buchanan and Lukaszewski (1997)
Measuring the Impact of Your Web Site

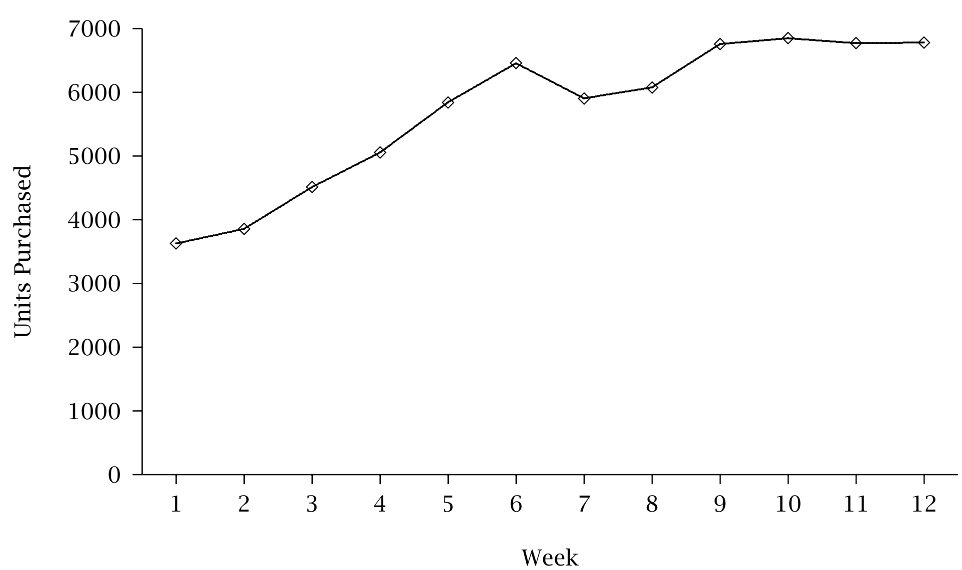
Case Study

- CDNOW — a leading on-line music retailer
- Focus on a single cohort of new customers
- Data covering initial ("trial") and repeat purchases for a 3-month period (1/97–3/97)
  - nearly 70,000 units (CDs) purchased by over 23,000 people

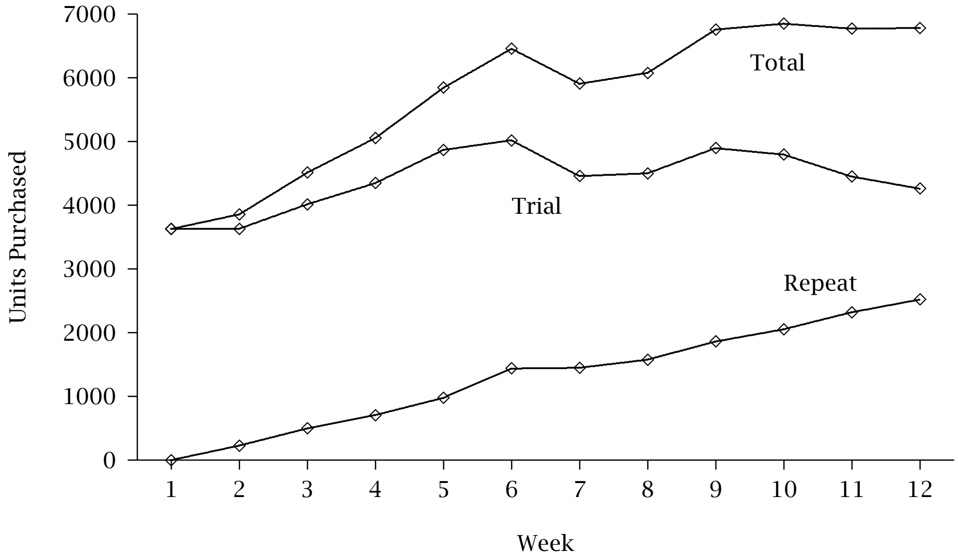
Objectives

- Develop a medium-term forecast of unit purchases by this cohort
- Achieve this using an easily implementable model (i.e., Excel-based)

Quarter 1 Sales

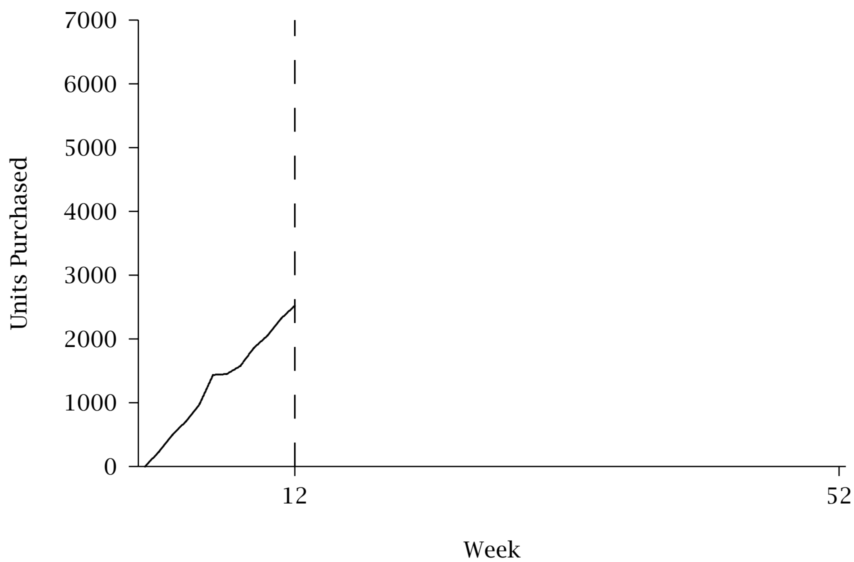


Trial/Repeat Decomposition



137

Forecasting Weekly Repeat Sales



Modeling Repeat Sales

- Depth of repeat decomposition
  - model interpurchase times and quantity (given purchase occasion)
- Counting process
  - individual-level time series of counts
- Or …a different kind of counting process
  - series of cohort-level cross-sections

Raw Data: Repeat Sales for Cohort

Units	Week											
Purchased	1	2	3	4	5	6	7	8	9	10	11	12
0	0	1478	3033	4763	6608	8616	10829	12716	14698	16774	18881	20902
1	0	50	70	113	138	205	181	266	278	341	351	314
2	0	21	45	62	80	105	120	137	149	183	211	221
3	0	8	31	36	51	73	64	99	116	110	118	159
4	0	8	10	30	30	53	43	52	49	66	76	73
5	0	0	13	17	19	21	24	34	30	41	35	38
6	0	4	2	5	10	11	22	15	26	26	29	25
7	0	0	2	2	11	12	14	12	10	11	7	12
8	0	0	3	2	4	10	6	5	9	13	9	17
9	0	1	2	0	3	3	9	2	2	3	9	6
10+	0	4	5	8	8	17	11	9	14	11	18	14
Total units	3627	3857	4512	5054	5843	6456	5906	6077	6757	6848	6770	6781
Trial units	3627	3630	4015	4347	4867	5019	4457	4501	4896	4795	4450	4261
Incr. triers	1574	1642	1822	1924	2164	2197	2024	2034	2198	2165	2037	1789
Cum. triers	1574	3216	5038	6962	9126	11323	13347	15381	17579	19744	21781	23570

Week 3 Purchasing

Consider the relative frequency distribution:

0	1	2	3	4	5	6	7	8	9	10+
0.943	0.022	0.014	0.010	0.003	0.004	0.001	0.001	0.001	0.001	0.002

A mix of purchases by those who "tried" in week 1 or week 2:

$$P(X_3 = x) = \tfrac{1574}{1574+1642} \times P(R_{3|1} = x) + \tfrac{1642}{1574+1642} \times P(R_{3|2} = x)$$

Model Development

- Model each week's purchasing using a finite mixture model with *known* mixing weights:

$$P(X_w = x) = \frac{1}{\sum_{i=1}^{w-1} n_i} \left[\sum_{i=1}^{w-1} n_i P(R_{w|i} = x) \right]$$

  where n_i = # people "trying" in week i.

- Need to develop submodel for $P(R_{w|i} = x)$.

Distribution of Purchases By Week 1 Triers

Units Purchased	Week 1	2	3	4	5	6	7	8	9	10	11	12
0	0	1478	1485	1487	1506	1500	1520	1524	1522	1530	1523	1531
1	0	50	29	24	21	23	17	18	19	21	17	9
2	0	21	25	26	20	16	13	12	10	12	8	16
3	0	8	16	11	10	19	6	8	10	7	8	6
4	0	8	4	11	7	7	5	5	6	1	7	8
5	0	0	9	5	4	4	9	2	3	1	2	3
6	0	4	0	3	3	1	2	1	1	1	4	0
7	0	0	1	2	1	0	0	2	1	1	0	1
8	0	0	1	2	1	2	0	1	0	0	1	0
9	0	1	1	0	0	1	1	0	0	0	2	0
10+	0	4	3	3	1	1	1	1	2	0	2	0

Distribution of Purchases By Week 2 Triers

Units Purchased	Week 1	2	3	4	5	6	7	8	9	10	11	12
0	0	0	1548	1542	1574	1554	1578	1581	1583	1578	1577	1599
1	0	0	41	41	20	39	24	25	20	21	23	12
2	0	0	20	18	12	10	12	12	15	18	16	14
3	0	0	15	17	15	12	11	9	12	7	14	3
4	0	0	6	12	3	10	6	8	6	4	4	3
5	0	0	4	8	4	6	1	4	4	6	3	3
6	0	0	2	1	6	2	3	2	1	3	4	5
7	0	0	1	0	4	2	3	0	0	2	0	1
8	0	0	2	0	1	3	0	0	0	1	0	1
9	0	0	1	0	0	2	2	0	0	1	0	1
10+	0	0	2	3	3	2	2	1	1	1	1	0

Modeling $R_{w|i}$

Decompose purchasing into two components:

- Do I consider the possibility of making a repeat purchase this week?
- If so, how many CDs — if any — do I buy?

Purchase Quantity Submodel

- For a customer who considers purchasing in week w, assume the number of units purchased is distributed geometric:

$$P(R_w = x \mid q) = q(1-q)^x, \; x = 0, 1, 2, \ldots$$

- To capture heterogeneity, assume $q \sim \text{beta}(\alpha, \beta)$
- This leads to the *beta-geometric* distribution:

$$\begin{aligned} P(R_w = x) &= \int_0^1 P(R_w = x \mid q) g(q) dq \\ &= \frac{B(\alpha+1, \beta+x)}{B(\alpha, \beta)} \end{aligned}$$

Purchase Quantity Submodel

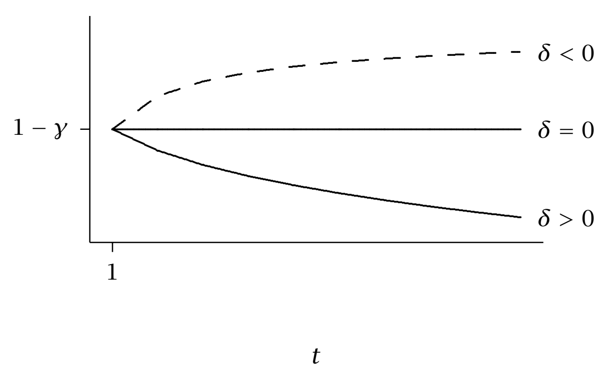
- Beta-geometric probabilities can be computed without having to evaluate the beta function!
- Use forward recursion:

$$P(R_w = x) = \begin{cases} \dfrac{\alpha}{\alpha+\beta} & x = 0 \\ \dfrac{\beta+x-1}{\alpha+\beta+x} P(R_w = x-1) & x \geq 1 \end{cases}$$

"Purchase Consideration" Submodel

The probability that a customer is "out of the market" t weeks after her first purchase, π_t, can vary over time. We capture this as follows:

$$\pi_t = 1 - \gamma t^{\delta}$$



Modeling $R_{w|i}$

Combining the two components:

$$P(R_{w|i} = x) = \begin{cases} \pi_{w|i} + (1 - \pi_{w|i})\, P(R_w = 0) & x = 0 \\ (1 - \pi_{w|i})\, P(R_w = x) & x \geq 1 \end{cases}$$

where $\pi_{w|i} = 1 - \gamma(w - i)^{\delta}$.

Furthermore:

$$E(R_{w|i}) = \gamma(w - i)^{\delta} \frac{\beta}{\alpha - 1}$$

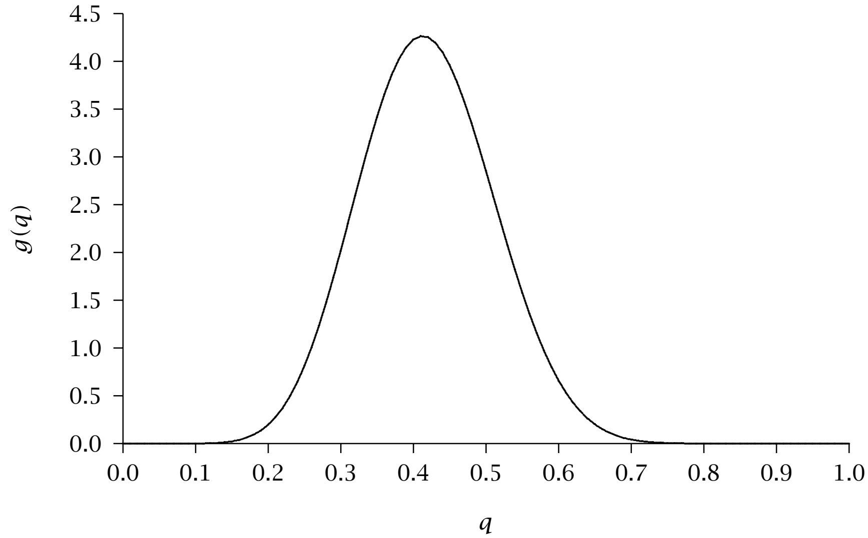
Predicting Unit Sales

Let N_w = total number of units purchased by the (eligible) cohort members in week w:

$$E(N_w) = \sum_{i=1}^{w-1} n_i E(R_{w|i}) \qquad w \leq 12$$

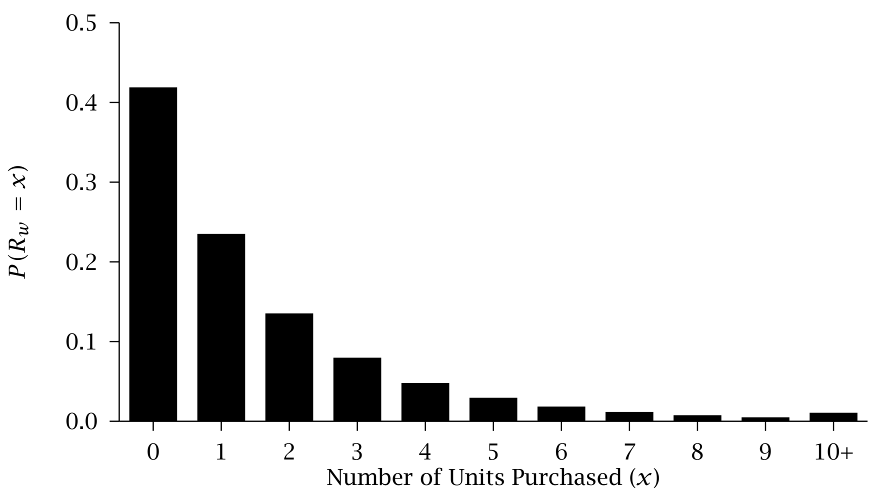
$$E(N_w) = \sum_{i=1}^{12} n_i E(R_{w|i}) \qquad w > 12$$

Estimated Distribution of q



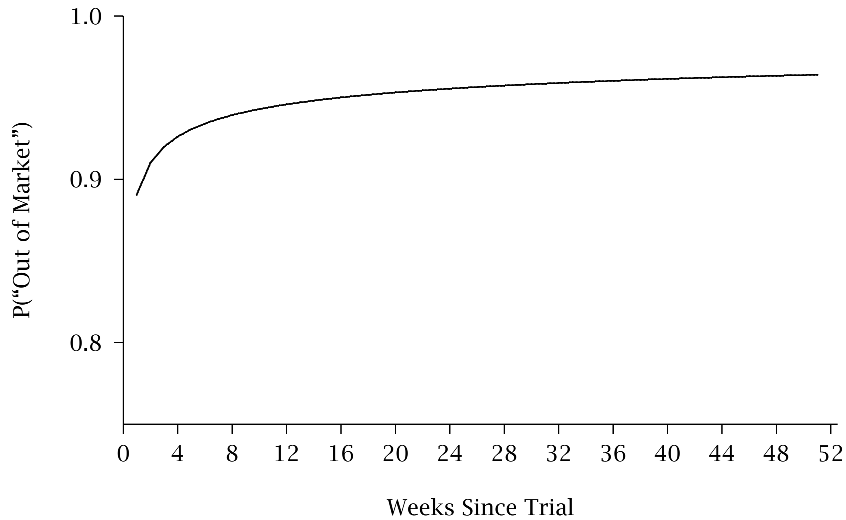
$\hat{\alpha} = 11.82, \hat{\beta} = 16.40$

Estimated Distribution of Purchase Quantity



$\hat{\alpha} = 11.82, \hat{\beta} = 16.40$

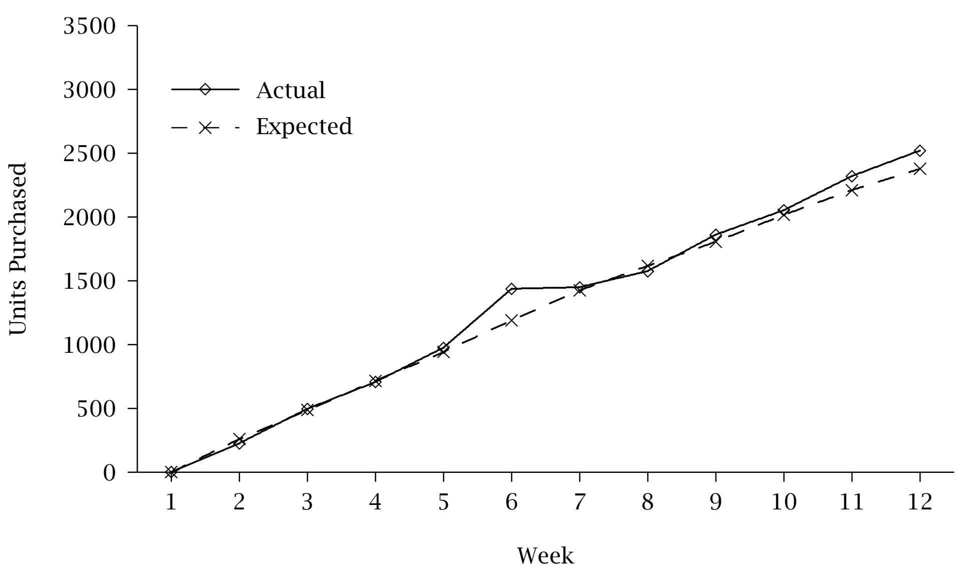
Predicted "Purchase Consideration" Curve



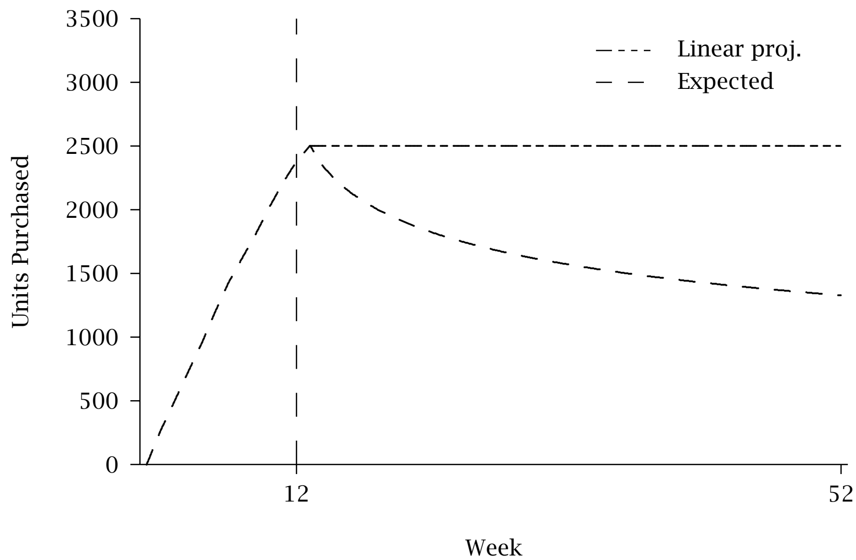
$$\hat{\gamma} = 0.109, \hat{\delta} = -0.283$$

153

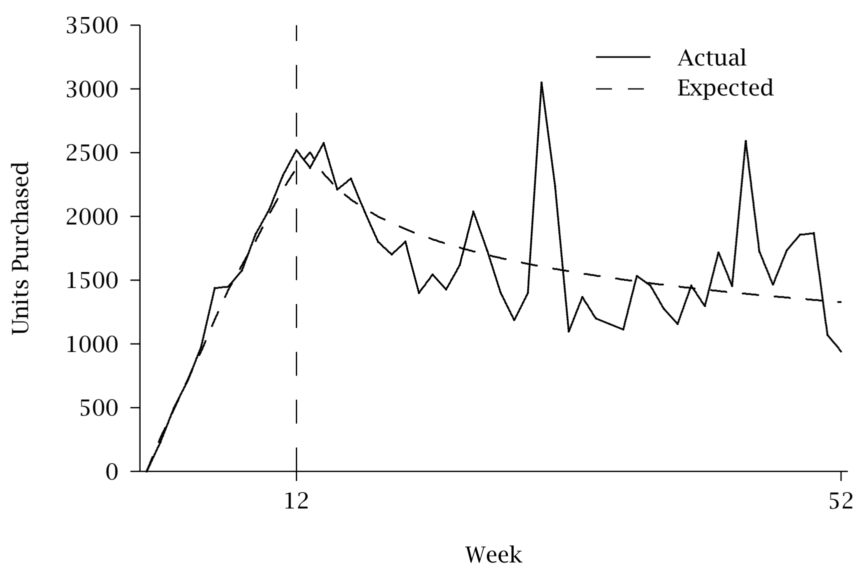
Model Fit



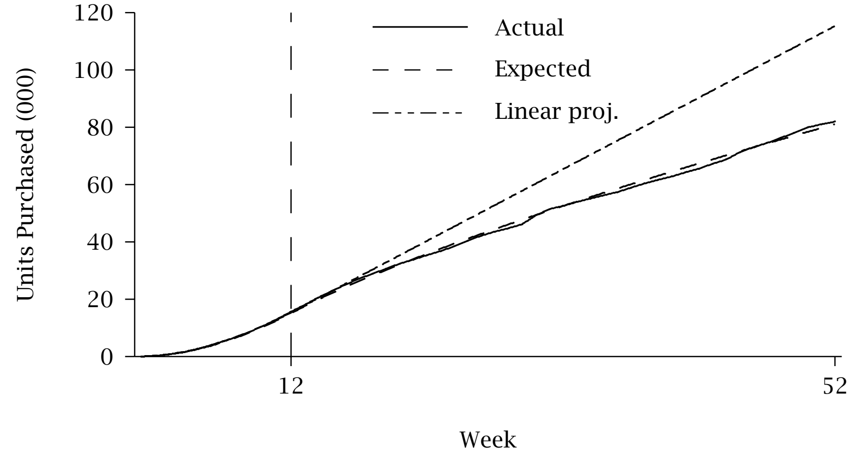
Weekly Repeat Sales Forecast(s)



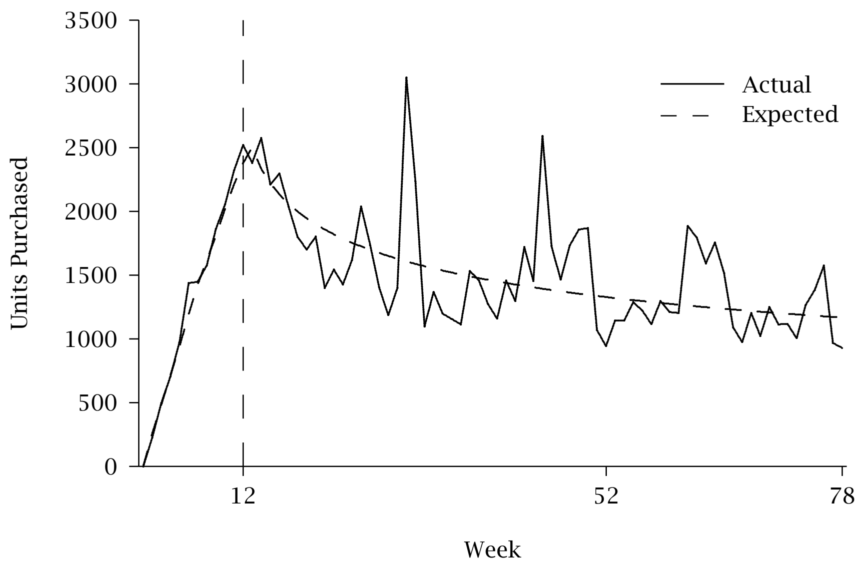
Weekly Repeat Sales Forecast



Cumulative Repeat Sales Forecast



Weekly Repeat Sales Forecast



Future Research

- Stability of parameters across cohorts
- Compare performance relative to more complex models (e.g., Fader and Hardie 1999; Fader, Hardie, and Huang 2002)
  - fit and forecasting
  - parameter validity
  - cost/benefit
- Develop a model for the arrival of "triers"

Conclusions

- Purchasing behavior on the Internet is predictable
- Lessons learned from bricks-and-mortar (e.g., CPG) are more transferable than many imagine
- Rich data don't necessarily require complex analyses — powerful models can be built in Excel

- The Excel spreadsheets associated with this tutorial can be found at:

  `http://brucehardie.com/talks.html`

- A note on how to build the complete CDNOW model and generate the associated sales forecast in an Excel spreadsheet (along with a copy of this spreadsheet) can be found at:

  `http://brucehardie.com/pmnotes.html`