HW2a

Alan Wu
MKTG 7760: Probability Models in Marketing

February 11, 2026

1 Question 1

The HW2A dataset has a spreadsheet containing Internet visit data for 2728 Comscore panelists, showing the number of visits each person made to one website (with the disguised name of khakichinos.com). You should ignore the covariate data (we'll discuss/use it in a few weeks). Fit a plain NBD and an NBD with spike at zero to this dataset. Briefly discuss your results.

Before discussing results, we'd like to note a few design decisions for modeling for this dataset:

- For goodness-of-fit tests and model fit, we choose to censor the original data at 10 visits. This reasoning is two fold: the counts of the visits past 10 visits are low and often missing. Furthermore, the list of visits loses continuity after 22 visits. Therefore, we censor at 10 visits.
- The procedure for fitting the NBD on the full raw dataset is similar, but slightly tweaked as compared to a histogram of frequencies. For each observation, we use the gamma formula: $P(X = x) = \frac{\Gamma(r+x)}{\Gamma(r)x!} \cdot (\frac{\alpha}{\alpha+1})^r \cdot (\frac{1}{\alpha+1})^x$ instead of forward recursion. And for the log likelihood, we simply take the logarithm of $P(x = x)$ and sum up over all observations.
- For the goodness of fit tests, we compute probabilities according to forward recursion formulas.

1.1 Plain NBD

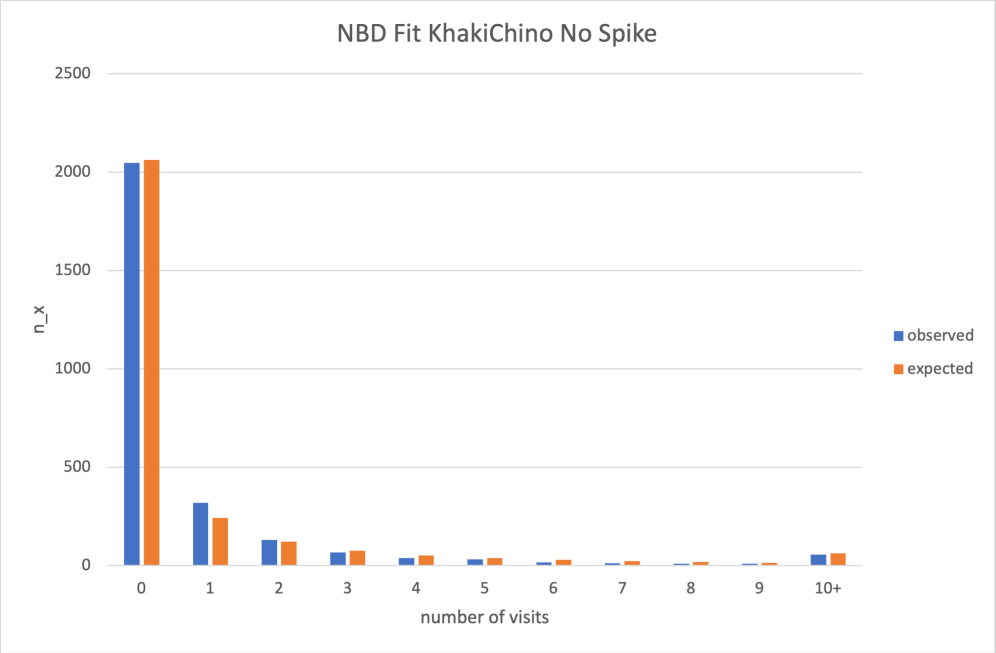


Figure 1: NBD fit of comscore data with no spike

The results of the plain nbd result in $(r, \alpha) = (0.134, 0.141)$. This indicates that the model believes the population of visitors to the website are heterogeneous. When fitting the chi-sq goodness-of-fit test, we use 8 degrees of freedom (11 categories, 2 parameters). The p-value for this test is $p < 0.001$. This indicates that the plain NBD does not explain the data very well, as the observed the data to be generated from this model would be unlikely.

Visually, the model seems to overestimate the 0 category or perhaps overestimates the number of never buyers, and tends to not fit very well across the rest of the categories.

1.2 NBD With Spike at 0

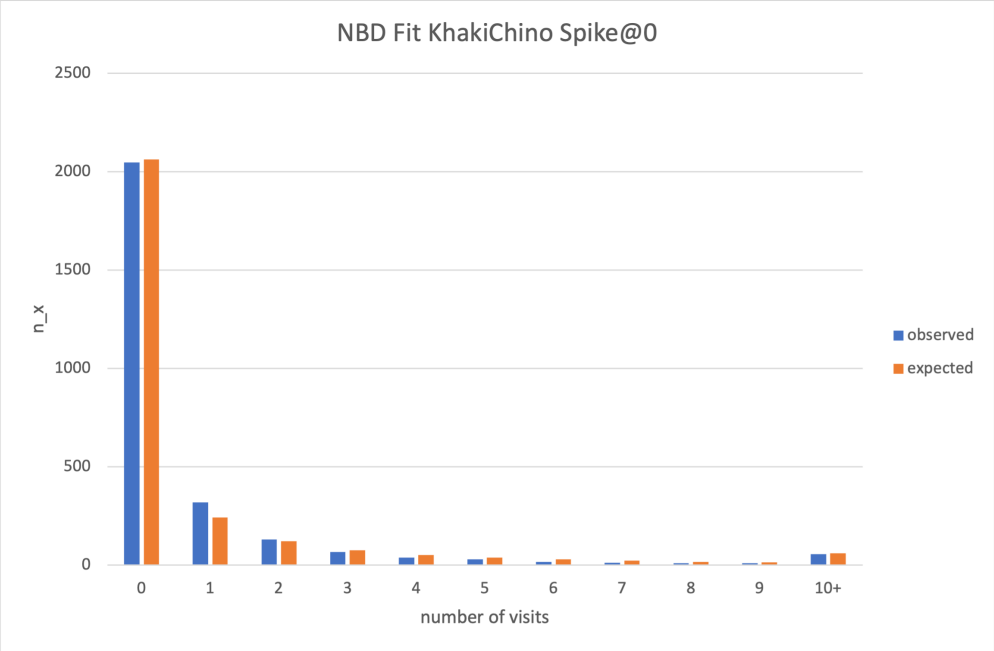


Figure 2: NBD fit of comscore data with spike@0

The results of the nbd with spike at zero result in $(r, \alpha) = (0.134, 0.141)$ and also the spike parameter of 0.0001. This indicates that the model believes the population of visitors to the website are heterogeneous, and also that the spike seems insignificant, as a low spike parameter value indicates a low percentage of hard core never buyers. When fitting the chi-sq goodness-of-fit test, we use 7 degrees of freedom (11 categories, 3 parameters). The p-value for this test is $p < 0.001$. This indicates that the NBD with spike at zero does not explain the data very well, as the observed the data to be generated from this model would be unlikely.

When we compare the two models, we notice that the log likelihood that the spike model converged to is the same if not slightly higher (-2905.6266 vs -2905.6245) compared to the plain nbd model. When we compute the chi-sq LRT test, there yields no number since the difference will be negative. However, this is indicative that adding the spike parameter did not increase the performance of the model.

2 Question 2

The HW2A dataset contains a second sheet showing the distribution of the number of albums purchased across a set of 1574 shopping trips for an online music seller.

2.1 Part A: Estimate a shifted NBD and a truncated NBD. Describe the relevant features and differences between these two models.

Before we discuss the results of the modeling, we can discuss some of the implications of both of the shifted and truncated NBD models.

The shifted model assumes that potential missing categories (in this case those who never purchase albums) do not exist and we simply shift the model to starting at 1. Mechanically, this means we begin the computations of $P(X = x)$ however many count periods later than the original model. For this model, this means that the formula for likelihood at 1 is $(\frac{\alpha}{\alpha+1})^r$, whereas for unshifted model this would be the probability at x = 0. We compute all the remaining probabilities as according to forward recursion formula and can perform goodness of fit in the same way. Specifically, let us define a δ as to how many periods shifted from 0 the model becomes. Therefore, the forward recursion formula is the following:

$$P(x = x) = \begin{cases} (\frac{\alpha}{\alpha+t})^r & \text{for } x = \delta \\ \frac{t(r+x-1-\delta}{(x-\delta)(\alpha+t)} \cdot P(x = x-1) & \text{for } x > \delta \end{cases}$$

In our case, since the shift is only one count, we have the following:

$$P(x = x) = \begin{cases} (\frac{\alpha}{\alpha+t})^r & \text{for } x = 1 \\ \frac{t(r+x-2)}{(x-1)(\alpha+t)} \cdot P(x = x-1) & \text{for } x > 1 \end{cases}$$

since the $\delta = 1$.

The truncated model, on the other hand, assumes the existence of missing categories, but does not fit on the data since it does not exist in the dataset. Instead, we compute probabilities for all missing categories, then normalize the final probabilities by the sum of the probabilities of missing categories. For instance, in this dataset, we normalize the rest of the probabilities by the $1 - P(x = 0)$, since that is missing. This also means once we fit the model, we are able to roll back the values of the missing categories to infer the counts of the missing categories. We cannot do this with the shifted model.

In the parameters (r, α) of the models, we notice that there can be vast differences due to the assumption of missingness.

2.1.1 Shifted NBD Model

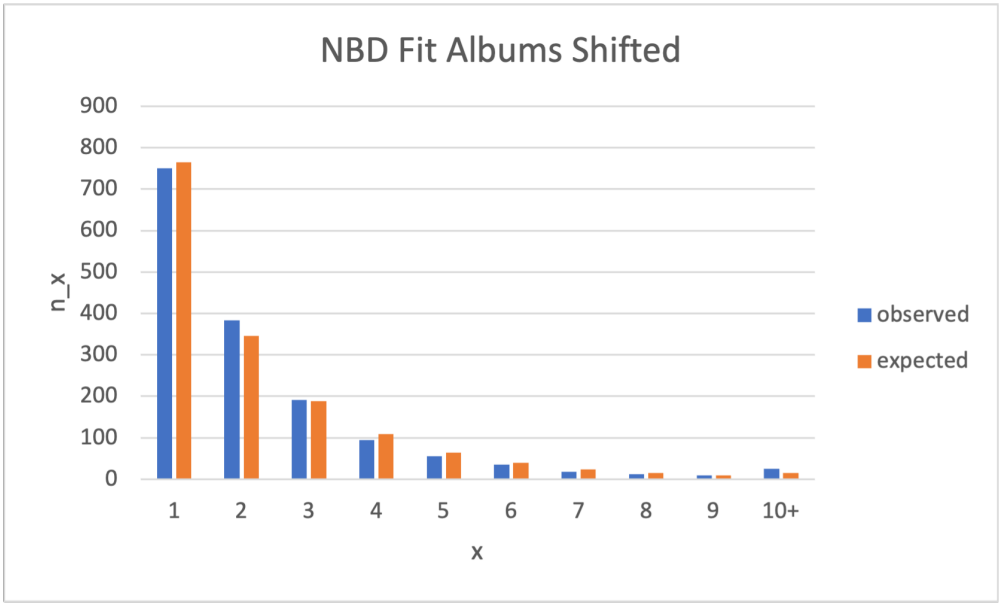


Figure 3: NBD fit of album data shifted model

The results of the shifted result in $(r, \alpha) = (0.705, 0.560)$. This indicates that the model believes the population of album purchasers to observe some heterogeneity. When fitting the chi-sq goodness-of-fit test, we use 7 degrees of freedom (10 categories, 2 parameters). The p-value for this test is $p = 0.020$. This suggests that the model, although capturing some of the shape of the data, still fails to maintain a good fit. This can be visually confirmed as we can see there are overestimates for the 1 purchase category and varying over and underestimates for the rest of the categories. Notably, the model underestimates the 10+ category (25 true vs 14.9 expected).

2.1.2 Truncated NBD Model

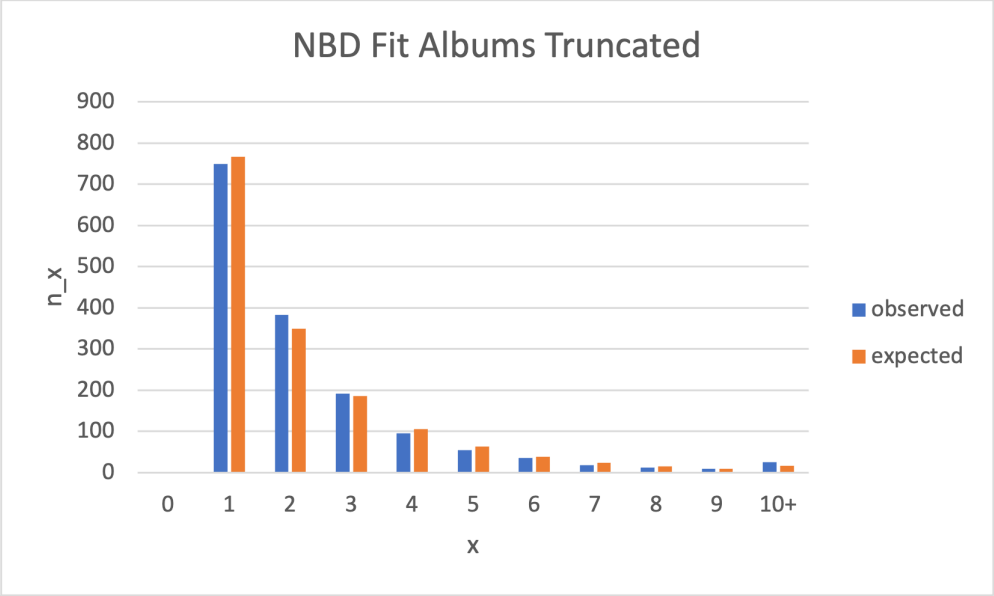


Figure 4: NBD fit of album data truncated model

The results of the truncated result in $(r, \alpha) = (0.330, 0.461)$. This indicates that the model believes the population of album purchasers to be fairly heterogeneous. When fitting the chi-sq goodness-of-fit test, we use 7 degrees of freedom (10 categories, 2 parameters). The p-value for this test is $p = 0.115$. With a p-value of 0.115, the model passes the goodness-of-fit test. This indicates that the NBD model provides a consistent fit for the data, successfully accounting for the population's heterogeneity without significant deviation from the observed purchase frequencies. Visually, the truncated model seems to fit better in most categories but seems to be missing at 1 and 2 purchase categories.

2.1.3 Results Comparison

Notice that the r and α parameter values for the shifted and truncated models differ. The truncated model detects greater heterogeneity because it isolates the active consumer base by excluding non-purchasers. By removing these zeros, the model focuses strictly on the behavioral spread among individuals who have actually made a purchase, which typically reveals a much wider dispersion than the population-wide average suggests. Mathematically, this is reflected in the lower shape parameter $r = 0.330$ compared to the shifted model's $r = 0.705$; in the nbd, a smaller r indicates a more highly skewed Gamma distribution of purchase rates. While the shifted model effectively "smooths" the variance by forcing a start at one, the truncated model's parameters are determined solely by the differences between light and heavy buyers, thereby capturing the high degree of heterogeneity inherent in the active purchasing population.

2.2 Part B: Also fit and describe a "zero-truncated, one-inflated" model.

The procedure for fitting a zero-truncated, one-inflated model is simply an expansion on the truncated model. We first compute the probabilities for each category based on the truncated model and then add a spike parameter at x = 1 to reflect a potential narrative for one-time buyers of albums.

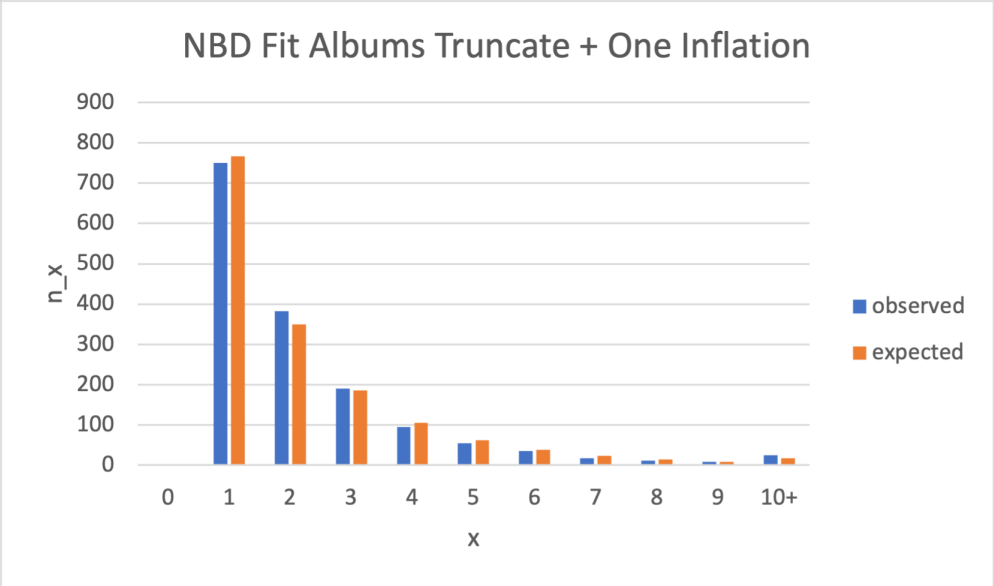


Figure 5: NBD fit of album data zero-truncated, one-inflated model

The results of the zero-truncated, one-inflated result in $(r, \alpha) = (0.330, 0.461)$ and spike parameter of 0.00001. This indicates that the model believes the population of album purchasers to be fairly heterogeneous. When fitting the chi-sq goodness-of-fit test, we use 6 degrees of freedom (10 categories, 3 parameters). The p-value for this test is $p = 0.070$. The performance is similar to the original truncated model, meaning there is an absence of a one-time buyer only segment and that they do not behave differently than normal buyers of other counts.

2.3 Part C: Finally, fit an sBG model to the data. What would its "story" be in this context (compared to the way we used it in Session 1)?

The sBG model suggests that each shopping trip is governed by a constant probability p that the customer will stop purchasing albums in a given sitting. A shopper who purchases 9 albums is someone who 'survived' 8 opportunities to stop and finally decided to quit on the 9th. Crucially, the model assumes that this stopping probability varies across the 1,574 trips according to a Beta distribution, accounting for the fact that some shoppers are naturally more inclined to buy in bulk than others.

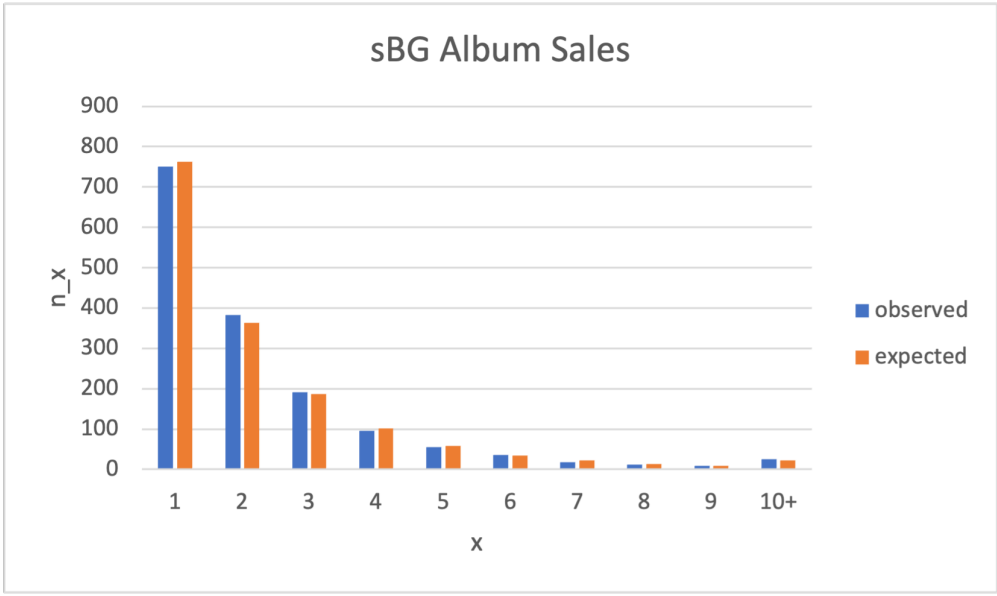


Figure 6: sBG fit of album data

The sBG model yielded parameter estimates of $(\alpha, \beta) = (5.911, 6.282)$. In contrast to models with parameters below 1, these relatively high and balanced values indicate a homogeneous population. Because α and β are similar and significantly greater than 1, the underlying Beta distribution of p (the stopping probability) is unimodal and roughly symmetric, centered near 0.5. This "bell-shaped" distribution suggests that most shoppers share a similar, moderate propensity to stop after each purchase, rather than being split into extreme "power buyers" and "one-time" shoppers.For the Chi-squared goodness-of-fit test, we utilized 7 degrees of freedom (10 categories, 2 parameters). The resulting p-value of $p = 0.855$ is exceptionally high, suggesting that the model provides an excellent fit for the observed data.

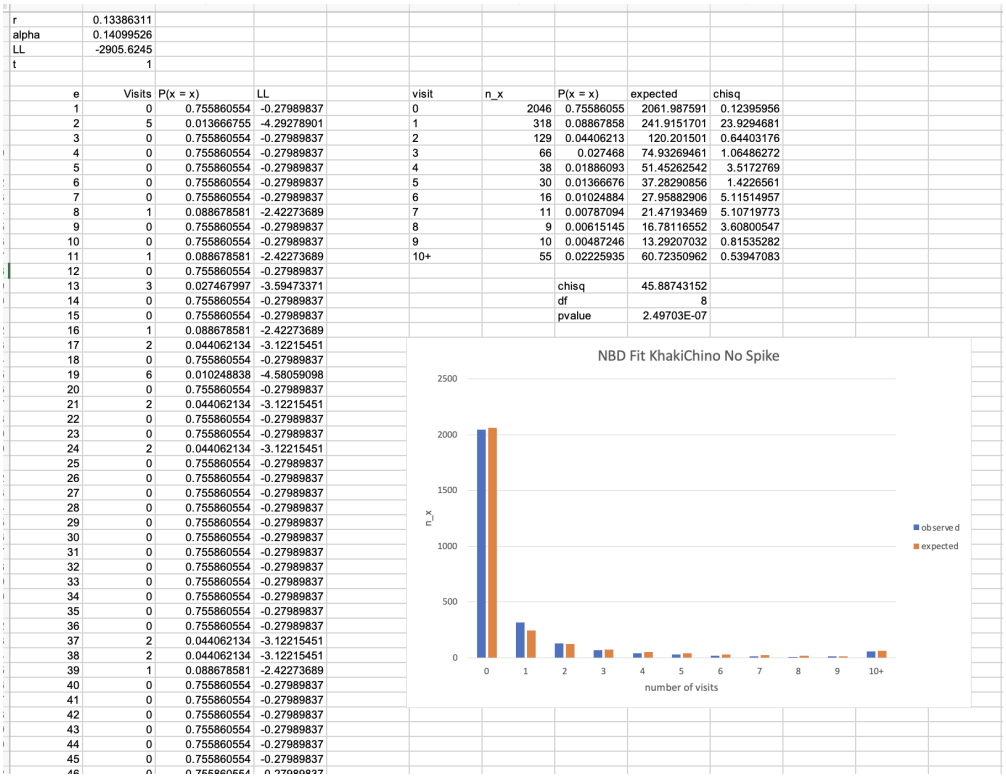
2.4 Part D: Which of these 4 models would you choose and why?

To make a decision for the best model, we compare each based on their statistical fit and the underlying "story" they tell about consumer behavior. While the sBG model provides a statistically superior fit in this instance with a p-value of 0.855, its narrative—that consumers engage in a sequential "coin-flip" process where they decide to stop after each individual album—can feel implausible in a real-world retail setting. In contrast, we often appreciate models that view behavior as counts over a time period, such as the NBD model. This perspective suggests that consumers possess a latent, long-term "purchase rate," which is a more intuitive reflection of a person's stable interest in music. Rather than a series of impulsive stopping decisions, the NBD narrative allows us to see a shopper who buys ten albums not as someone who "survived" nine trials, but as a "heavy buyer" with a high consumption velocity. Ultimately, while the sBG captures the mechanics of a single shopping trip with high precision, the NBD provides a more managerially useful story by linking observed counts to the enduring traits of the customer.

Furthermore, the choice between truncated and shifted nbd models is ultimately a managerial decision driven by specific business goals. For a more general application, the Zero-Truncated NBD model is our preferred choice over the Shifted NBD. This is because it is logically plausible that many potential customers had a purchase rate that resulted in zero albums during this period. Within this subset, we choose the zero truncated nbd over the zero truncated, one-inflated version. Given that the spike parameter π was negligible and there was no measurable improvement in the Chi-squared goodness-of-fit or the p-value, the simpler model is preferred for its parsimony and its ability to accurately capture the distribution without unnecessary complexity.

3 Appendix

3.1 Total Excel Tables

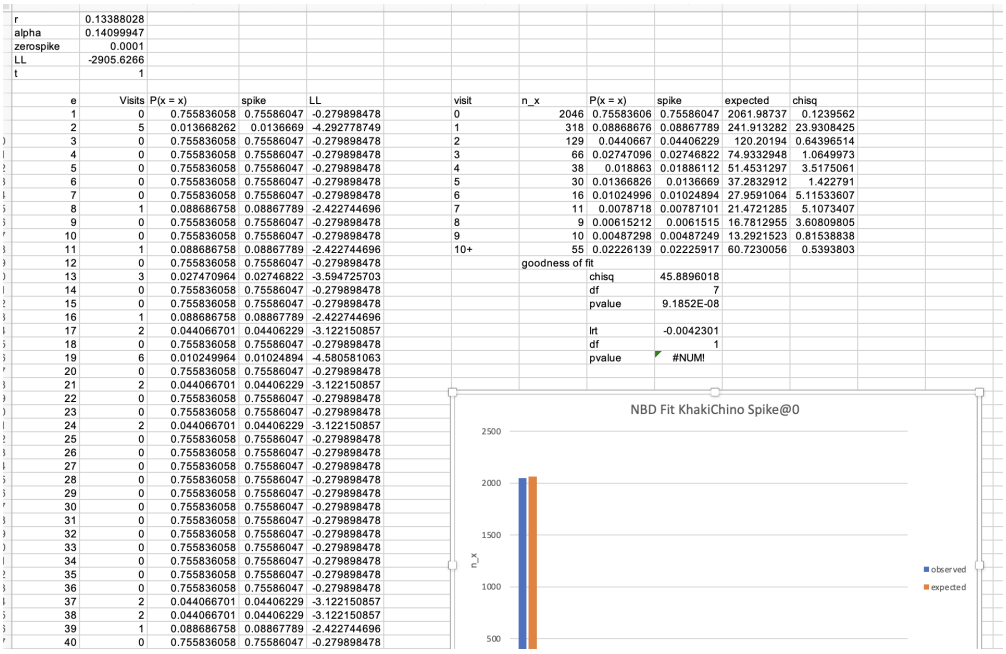


r	0.13386311
alpha	0.14099526
LL	-2905.6245
t	1

e	Visits	P(x = x)	LL
1	0	0.755860554	-0.27989837
2	5	0.013666755	-4.29278901
3	0	0.755860554	-0.27989837
4	0	0.755860554	-0.27989837
5	0	0.755860554	-0.27989837
6	0	0.755860554	-0.27989837
7	0	0.755860554	-0.27989837
8	1	0.088678581	-2.42273689
9	0	0.755860554	-0.27989837
10	0	0.755860554	-0.27989837
11	1	0.088678581	-2.42273689
12	0	0.755860554	-0.27989837
13	3	0.027467997	-3.59473371
14	0	0.755860554	-0.27989837
15	0	0.755860554	-0.27989837
16	1	0.088678581	-2.42273689
17	2	0.044062134	-3.12215451
18	0	0.755860554	-0.27989837
19	6	0.010248838	-4.58059098
20	0	0.755860554	-0.27989837
21	2	0.044062134	-3.12215451
22	0	0.755860554	-0.27989837
23	0	0.755860554	-0.27989837
24	2	0.044062134	-3.12215451
25	0	0.755860554	-0.27989837
26	0	0.755860554	-0.27989837
27	0	0.755860554	-0.27989837
28	0	0.755860554	-0.27989837
29	0	0.755860554	-0.27989837
30	0	0.755860554	-0.27989837
31	0	0.755860554	-0.27989837
32	0	0.755860554	-0.27989837
33	0	0.755860554	-0.27989837
34	0	0.755860554	-0.27989837
35	0	0.755860554	-0.27989837
36	0	0.755860554	-0.27989837
37	2	0.044062134	-3.12215451
38	2	0.044062134	-3.12215451
39	1	0.088678581	-2.42273689
40	0	0.755860554	-0.27989837
41	0	0.755860554	-0.27989837
42	0	0.755860554	-0.27989837
43	0	0.755860554	-0.27989837
44	0	0.755860554	-0.27989837
45	0	0.755860554	-0.27989837

visit	n_x	P(x = x)	expected	chisq
0	2046	0.75586055	2061.987591	0.12395956
1	318	0.08867858	241.9151701	23.9294681
2	129	0.04406213	120.201501	0.64403176
3	66	0.027468	74.93269461	1.06486272
4	38	0.01886093	51.45262542	3.5172769
5	30	0.01366676	37.28290856	1.4226561
6	16	0.01024884	27.95882906	5.11514957
7	11	0.00787094	21.47193469	5.10719773
8	9	0.00615145	16.78116552	3.60800547
9	10	0.00487246	13.29207032	0.81535282
10+	55	0.02225935	60.72350962	0.53947083
		chisq	45.88743152	
		df	8	
		pvalue	2.49703E-07	

Figure 7: nbd model on comscore data no spike



r	0.13388028
alpha	0.14099947
zerospike	0.0001
LL	-2905.6266
t	1

e	Visits	P(x = x)	spike	LL
1	0	0.755836058	0.75586047	-0.279898478
2	5	0.013668262	0.0136669	-4.292778749
3	0	0.755836058	0.75586047	-0.279898478
4	0	0.755836058	0.75586047	-0.279898478
5	0	0.755836058	0.75586047	-0.279898478
6	0	0.755836058	0.75586047	-0.279898478
7	0	0.755836058	0.75586047	-0.279898478
8	1	0.088686758	0.08867789	-2.422744696
9	0	0.755836058	0.75586047	-0.279898478
10	0	0.755836058	0.75586047	-0.279898478
11	1	0.088686758	0.08867789	-2.422744696
12	0	0.755836058	0.75586047	-0.279898478
13	3	0.027470964	0.02746822	-3.594725703
14	0	0.755836058	0.75586047	-0.279898478
15	0	0.755836058	0.75586047	-0.279898478
16	1	0.088686758	0.08867789	-2.422744696
17	2	0.044066701	0.04406229	-3.122150857
18	0	0.755836058	0.75586047	-0.279898478
19	6	0.010249964	0.01024894	-4.580581063
20	0	0.755836058	0.75586047	-0.279898478
21	2	0.044066701	0.04406229	-3.122150857
22	0	0.755836058	0.75586047	-0.279898478
23	0	0.755836058	0.75586047	-0.279898478
24	2	0.044066701	0.04406229	-3.122150857
25	0	0.755836058	0.75586047	-0.279898478
26	0	0.755836058	0.75586047	-0.279898478
27	0	0.755836058	0.75586047	-0.279898478
28	0	0.755836058	0.75586047	-0.279898478
29	0	0.755836058	0.75586047	-0.279898478
30	0	0.755836058	0.75586047	-0.279898478
31	0	0.755836058	0.75586047	-0.279898478
32	0	0.755836058	0.75586047	-0.279898478
33	0	0.755836058	0.75586047	-0.279898478
34	0	0.755836058	0.75586047	-0.279898478
35	0	0.755836058	0.75586047	-0.279898478
36	0	0.755836058	0.75586047	-0.279898478
37	2	0.044066701	0.04406229	-3.122150857
38	2	0.044066701	0.04406229	-3.122150857
39	1	0.088686758	0.08867789	-2.422744696
40	0	0.755836058	0.75586047	-0.279898478

visit	n_x	P(x = x)	spike	expected	chisq
0	2046	0.75583606	0.75586047	2061.98737	0.1239562
1	318	0.08868676	0.08867789	241.913282	23.9308425
2	129	0.0440667	0.04406229	120.20194	0.64396514
3	66	0.02747096	0.02746822	74.9332948	1.0649973
4	38	0.018863	0.01886112	51.4531297	3.5175061
5	30	0.01366826	0.0136669	37.2832912	1.422791
6	16	0.01024996	0.01024894	27.9591064	5.11533607
7	11	0.0078718	0.00787101	21.4721285	5.1073407
8	9	0.00615212	0.0061515	16.7812955	3.60809805
9	10	0.00487298	0.00487249	13.2921523	0.81538838
10+	55	0.02226139	0.02225917	60.7230056	0.5393803

goodness of fit		
	chisq	45.8896018
	df	7
	pvalue	9.1852E-08
	lrt	-0.0042301
	df	1
	pvalue	#NUM!

Figure 8: nbd model on comscore data spike@0

alpha	0.55993136				
LL	-2384.18903				
t	1				
x	n_x	P(X = x)	LL	expected	chisq
0					
1	750	0.48577768	-541.503157	764.614068	0.27931865
2	383	0.2194445	-580.879223	345.40564	4.09181482
3	191	0.11990364	-405.123773	188.728326	0.02734355
4	95	0.06929816	-253.587012	109.075299	1.8163052
5	55	0.04114405	-175.487178	64.7607337	1.47113716
6	36	0.02481771	-133.063118	39.0630777	0.24018704
7	18	0.01512643	-75.4436145	23.8089964	1.41729784
8	12	0.00928776	-56.1486952	14.6189333	0.46917321
9	9	0.00573416	-46.451819	9.02557495	7.2469E-05
10+	25	0.00946592	-116.501443	14.8993519	6.84748514
	chisq	16.6601351			
	df	7			
	pvalue	0.01972415			

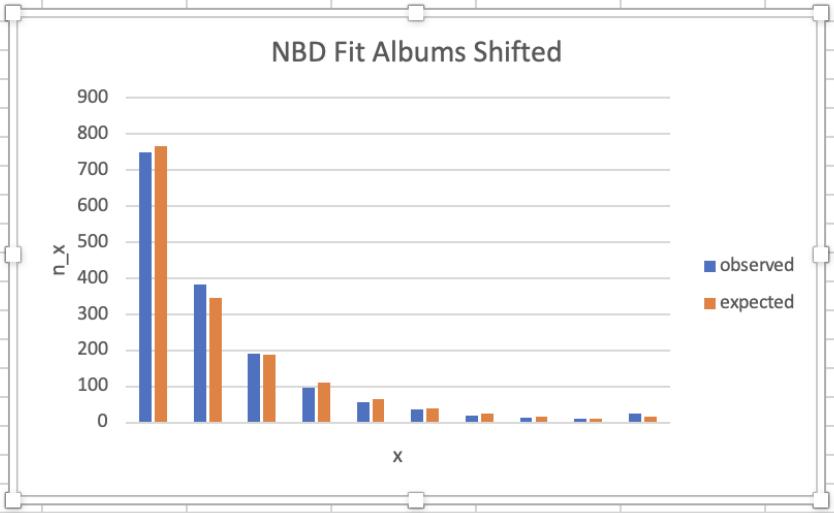


Figure 9: shifted nbd model on albums

r	0.33013426						
alpha	0.46085358						
LL	-2381.9801						
t	1						
x	n_x	P(X = x)	rescale	LL	expected	chisq	
0		0.68326362					
1	750	0.15440886	0.4874996	-538.84935	767.324376	0.39114357	
2	383	0.07029606	0.22193871	-576.55058	349.331533	3.24495659	
3	191	0.03737524	0.1180011	-408.17871	185.733735	0.14931887	
4	95	0.02129997	0.06724827	-256.43958	105.848779	1.111926	
5	55	0.0126271	0.03986629	-177.22233	62.7495364	0.95706387	
6	36	0.00767863	0.02424298	-133.90661	38.1584504	0.12209374	
7	18	0.00475327	0.01500701	-75.58628	23.6210349	1.33762274	
8	12	0.00298131	0.0094126	-55.988477	14.815427	0.53502537	
9	9	0.00188891	0.00596365	-46.098651	9.38678726	0.01593776	
10+	25	0.00342702	0.01081978	-113.15947	17.0303414	3.72954701	
						chisq	11.59463553
	Best N	3395.43236				df	7
						pvalue	0.114702501

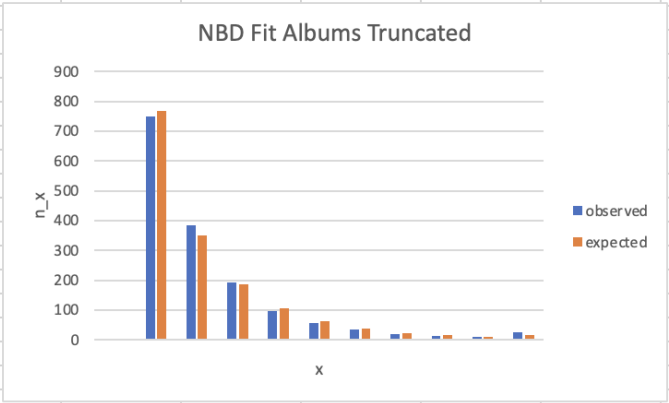


Figure 10: truncated nbd model on albums

r	0.330195165							
alpha	0.460874872							
spike@1	1E-05							
LL	-2381.98041							
t	1							
x	n_x	P(X = x)	rescale	spike	LL	expected	chisq	
0		0.683222752						
1	750	0.154425854	0.487490356	0.487495481	-538.855693	767.3178877	0.390853959	
2	383	0.070305995	0.22194143	0.22193921	-576.549725	349.3323167	3.244798279	
3	191	0.037380954	0.118003911	0.118002731	-408.176074	185.7362988	0.149171435	
4	95	0.021303309	0.067250122	0.06724945	-256.437912	105.8506338	1.112286708	
5	55	0.012629074	0.039867364	0.039866966	-177.221397	62.75060394	0.957311286	
6	36	0.007679808	0.024243559	0.024243316	-133.906116	38.15897949	0.122151915	
7	18	0.004753969	0.015007295	0.015007145	-75.58612	23.62124545	1.337711022	
8	12	0.002981734	0.009412716	0.009412622	-55.9884453	14.81546627	0.53503887	
9	9	0.001889159	0.005963684	0.005963624	-46.0986926	9.386743965	0.015934268	
10+	25	0.003427392	0.010819564	0.010819456	-113.160232	17.02982395	3.730144626	
							chisq	11.59540237
	Best N	3394.791188					df	6
							pvalue	0.071627995

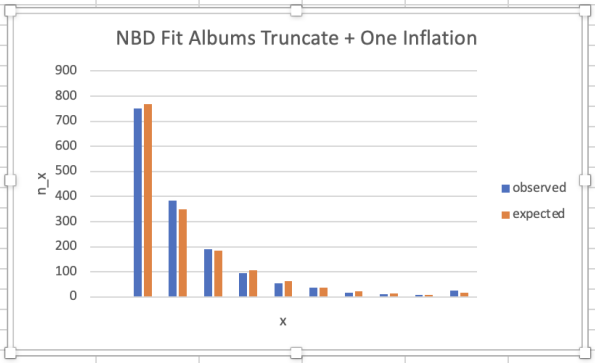


Figure 11: zero-truncated, one inflated nbd model on albums

alpha	5.91116285					
beta	6.28259818					
LL	-2378.01564					
x	n_x	P(x = x)	LL	pred	% error	chisq
1	750	0.48476945	-543.061394	763.027117	1.74%	0.2224112
2	383	0.23083726	-561.494204	363.337851	5.13%	1.06402377
3	191	0.11843901	-407.47121	186.423004	2.40%	0.11237288
4	95	0.06456484	-260.308103	101.625057	6.97%	0.4318953
5	55	0.0370099	-181.311342	58.2535811	5.92%	0.18171913
6	36	0.02213349	-137.183885	34.8381117	3.23%	0.03875022
7	18	0.01372576	-77.1926518	21.6043519	20.02%	0.60133036
8	12	0.00878348	-56.8185895	13.8251994	15.21%	0.24096237
9	9	0.0057774	-46.3842125	9.09362888	1.04%	0.00096401
10+	25	0.0139594	-106.790052	21.9720976	12.11%	0.41726524
			chisq	3.31169449		
			df	7		
			pvalue	0.85474959		

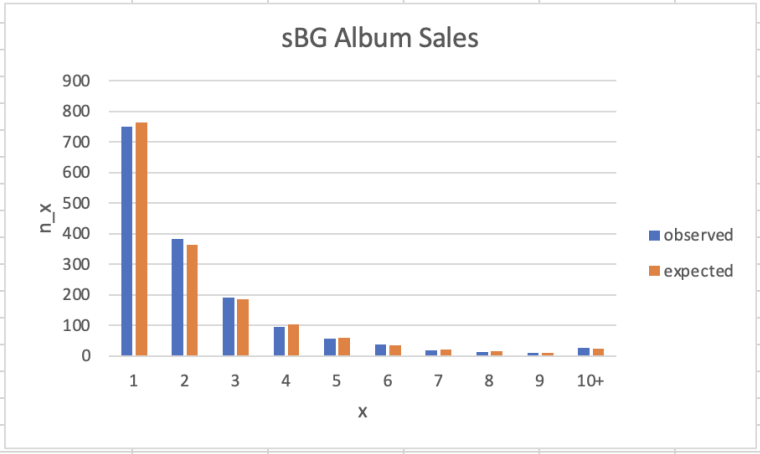


Figure 12: sbg model on albums