MKTG/STAT 4760/7760
Spring 2024

Project 1

We know that count data is a typical form of a multiple varied datasets that we could obtain from various sources. Count data is a dataset that summarizes a process which is repeated over a period of time. The dataset that is used for this project is based on the number of different lines that a customer has in their primary phone plans(more on this to follow). Kindly note that this project document does not follow any sequence/chronological order, though I have tried my level best to prepare this write up in a way that is intuitive and that shows the steps undertaken to reach the conclusions that I have eventually made through analyses.

The Dataset:

The dataset that I have used in this project is based on the number of telephone lines that people have in their primary mobile phone plans. We know that we currently have “Family plans” that could be purchased by families or friends coming together. Though it is the same as having a plan on your mobile but it is operated by some other individual who could be your relative or friend. There is one head in all the family plans who manages the plans on behalf of all the members in the plan.

The dataset can be found on the official website of Mintel (click here for the dataset). Please enable the “count” option and the tabular version in order to obtain the exact figures in the data. The underlying question that was asked after surveying a population of 1878 users, who owned a smartphone and had a mobile network service, was **"How many lines are included in your primary phone plan?" This dataset is based on February 2023 i.e. it is monthly.**

Please note the following about the dataset that is considered: -

1) The dataset is right censored already (obtained the same from the source)
2) The dataset does not have 0s i.e. does not account for people who do not have lines at all. We intend to back out the effective 0s through our model. (More on this to follow)

lines	%	N_x
0		
1	0.34717785	652
2	0.25026624	470
3	0.16719915	314
4	0.14802982	278
5	0.05537806	104
6 or more	0.03194888	60
		1878

Story & Background Set-up

Complementing the belief that the data obtained is dirty and cannot be trusted upon. We are not sure about the origins of data and cannot directly fit our model without cooking up an intuitive story.

I believe that the customers who own a smartphone and have a phone plan choose the number of lines in their plans on a stochastic basis or in an "as if" random manner. **All the customers have varied propensities to go for a certain number of lines in their mobile plans.** Since we assume that the customers decide upon their lines in a random manner, hence we assume that each of them has their **own wheel that has counts from 0 to infinity with some probability of each event, and the outcome of that spin dictates the number of lines they end up choosing.** Moreover, based on this story, we assume that the customers are guided by a Poisson process that in turn guides the count data.

At the individual level, **λ (lambda) is our parameter of interest that denotes the likelihood to spin the numbers on the wheels**. From a distribution point of view, λ is the mean of the Poisson distribution and also the variance observed. However, as discussed, even though **Poisson distribution is equi-dispersed theoretically, due to customers acting via a simple random process, we should expect variance to be more than the mean as it accounts for heterogeneity in the data.**

Since different customers can choose however many lines (theoretically speaking - it can be infinite) based on different lambdas, we have the gamma distribution

as our mixing distribution that accounts for the uncertainty related to the true propensity of each individual/customer.

Putting it together, I have fitted the NBD model over this data and have tried to make it fit better by making logical tweaks to the model. I will keep elaborating on the tweaks as and when they arrive further in my analysis.

Analysis

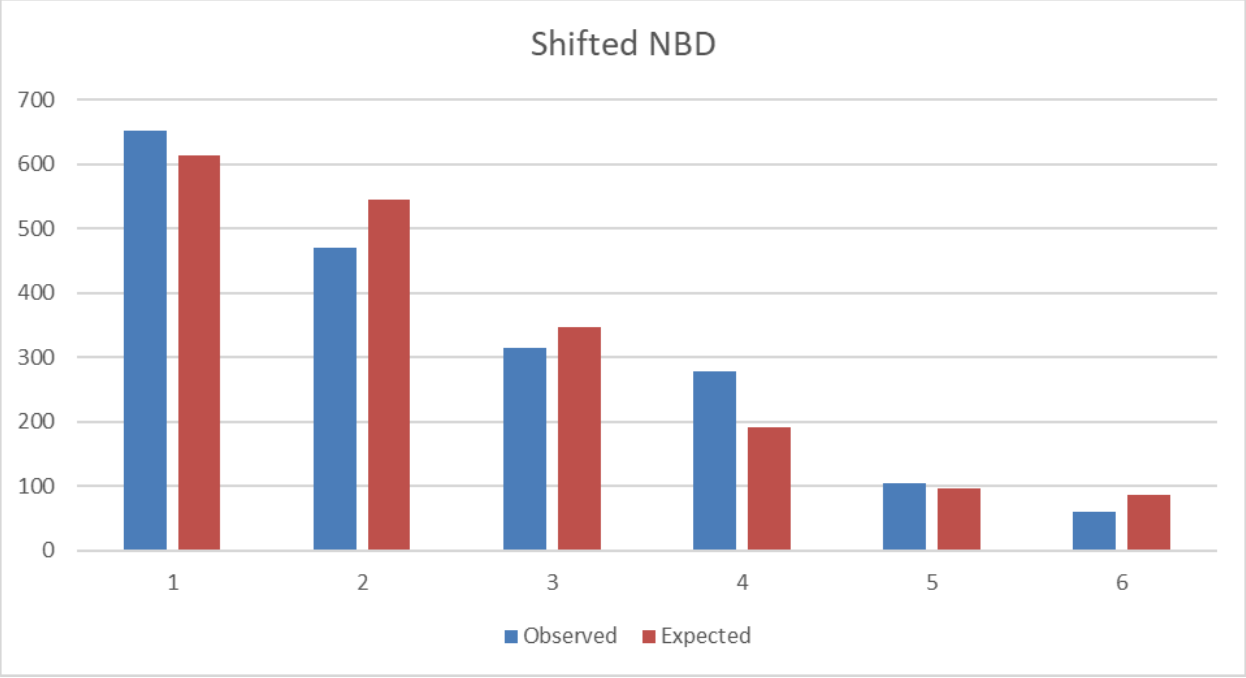
a) Shifted NBD

Since we do not have 0s in our dataset, I decided to fit a "Shifted NBD" model on the data. Kindly note the results for the Shifted NBD model: -

The parameters estimated are as follows: -

r 2.323 (**denotes that the data is very homogeneous**)
alpha 1.617 (**does not matter much, just a unit of time**)
Mean 1.43627776

The plot of the observed and expected numbers based on this model is given below: -

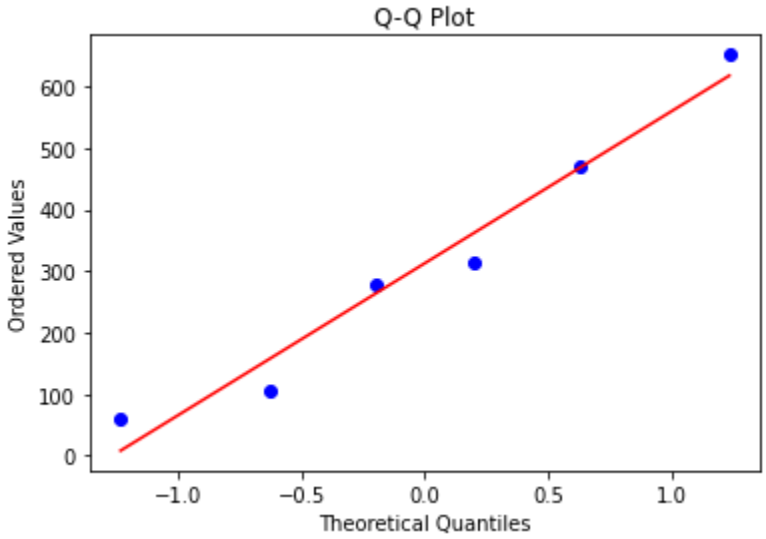


Model Evaluation:

5	**Chi-sq**	64.41951672
6	**df**	3
7	**p-value**	0.0000000000001

Inferences:

1) After plotting the raw data on the Q-Q plot, I realized that the data is normally distributed as all the data points are more or less on the straight line as shown in the figure given below.



I hypothesized that since our data is homogeneous with r > 1, we may have a hump in our data - similar to the bell shaped curve of the normal distribution. The Q-Q plot proved that the data could have been extracted from a normal distribution and testifies to the homogeneity observed via the fitted parameter r

2) The p-value is very low **which denotes that the model fit is not good. Although, the p-value is very different from the estimates of the parameters. It just implies the confidence level of the fit - if we fit this**

model on this data a million times then the likelihood of obtaining these set of parameters is very low.

3) Complementary to my story
   I feel that these results complement my story well. In my opinion, practically speaking now, there are only a certain number of lines that people could opt for. Even the telecom operators roll out family plans that do not surpass 8 or 10 lines per plan. **Hence, people would go for the number of lines lying in the range [1,4] - that is where we see our expected numbers going. Hence, the majority of the population would be fairly homogeneous due to this and we would see a plot with an interior mode kind of hump.**

   Diving deeper into the details of my story, I contend that most people do not want to go through the hassle of joining family plans as being the head of the plan implies that you would have to manage the plans of other members / lines too.
   That is why we see "too many" ones in model figures.
   Besides, there could be couples as well who would enroll in a plan with 2 lines in order to reduce the cost. Similarly, couples may have a kid or 2 kids (generally) who all go for 3 and 4 lines respectively to manage things well and avail the cost benefits.
   Hence, I posited that majorly people would have the number of lines in the range of [1,4].

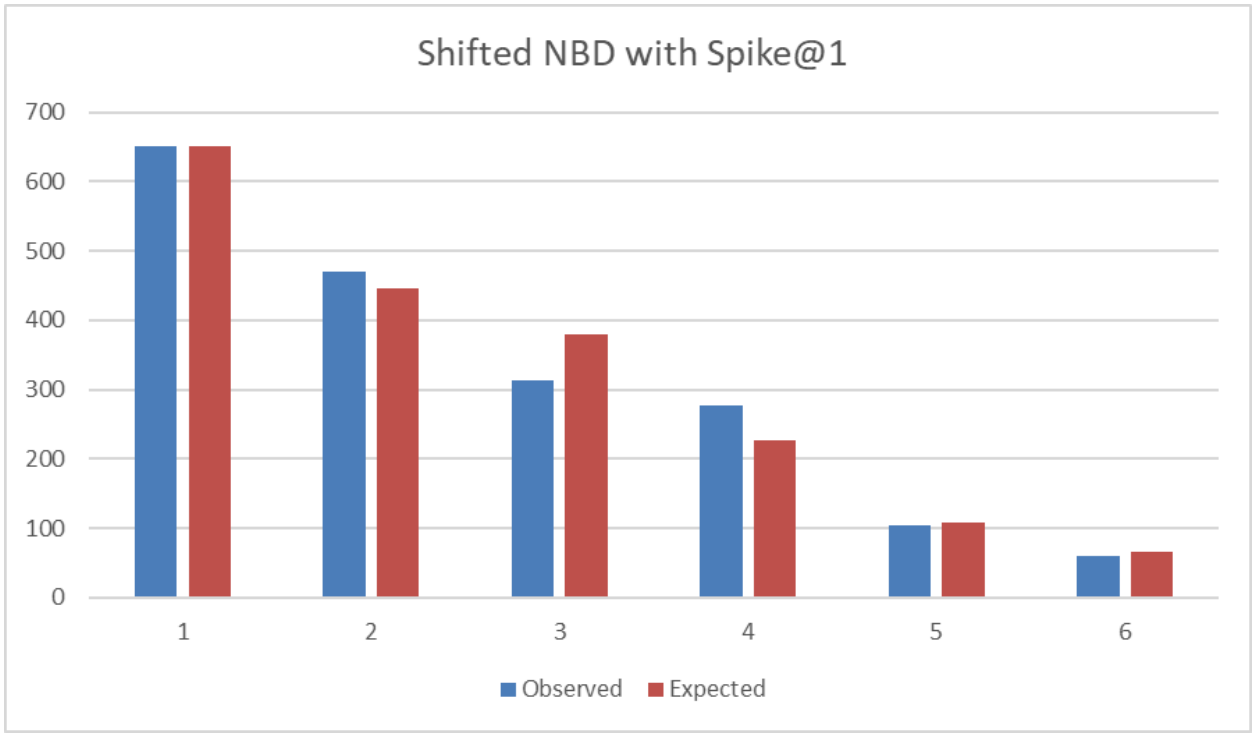
b) **Shifted NBD with Spike@1**

In order to improve my model fit, I decided to add a spike @ 1 which would essentially **segment the customers into "Hard core one liners" and people who contemplate buying more lines, but fortunately or unfortunately, have to settle at one after spinning their respective "poisson wheels."**

I believe that many phone users would not want to go through the hassle of managing multiple lines and would be glad to go for a single line that serves their purpose. Due to this, there would be "hard core one liners" who I intend to segment out by adding a spike at one.

Results

r	15.75378511	**Mean**	1.778271	
alpha	8.859044849			
spike@1	0.198534127			



Chi-sq	24.22561469
df	2
p-value	0.00001
LRT	36.69693595
df	1
p-value	1E-09

- Denotes even higher homogeneity with the increased value of r
- The mean is quite similar to the one in shifted NBD

- The p-value for the Chi-square Likelihood ratio Test is very low which is what we root for - signifies that the model with the spike is significantly different from the plain shifted NBD model.

Inferences

1) We observe that the model fit has significantly improved as shown by the LRT test. Though the p-value has increased manifold, it is still not something that we would consider ideal/good.

2) **Heterogeneity with the addition of Spike:**
Arguably, the increase in the value of r to such a number can be attributed to the addition of a spike. I believe, adding a spike, chopped off a segment of the population (**"hard core one liners"** in this case) that further declined heterogeneity.

3) Though the estimated parameters have increased, however, we observe that the mean is consistent in both the fitted models that does indicate the robustness of the NBD model.

c) Truncated NBD model

Adding some bells and whistles to my analysis, I was curious to account for the zero counts in my model. Since, we do not have zero counts in our model (perhaps they were lost by our assistant while compiling the data), we would try to run a truncated NBD model in order to back out the effective 0s!

I am well aware of the fact that the Shifted NBD model is very different from the truncated model. In the truncated model, we assume that we have the 0s unlike the Shifted NBD where we do not bother about the zero counts. The generative stories will be very different, in fact the story that I shall be delineating now could potentially violate the one told earlier. However, I did want to bring a new perspective into my analysis. The base story would not change which is why I am able to embark on this path of fitting a truncated NBD model after having fitted the Shifted NBD.

Story wise, the NBD model would just help us back out the potential size of the market for these telecom plans. **One could also argue that the effective 0s backed out from the NBD model would correspond to customers who were very unlucky to have**

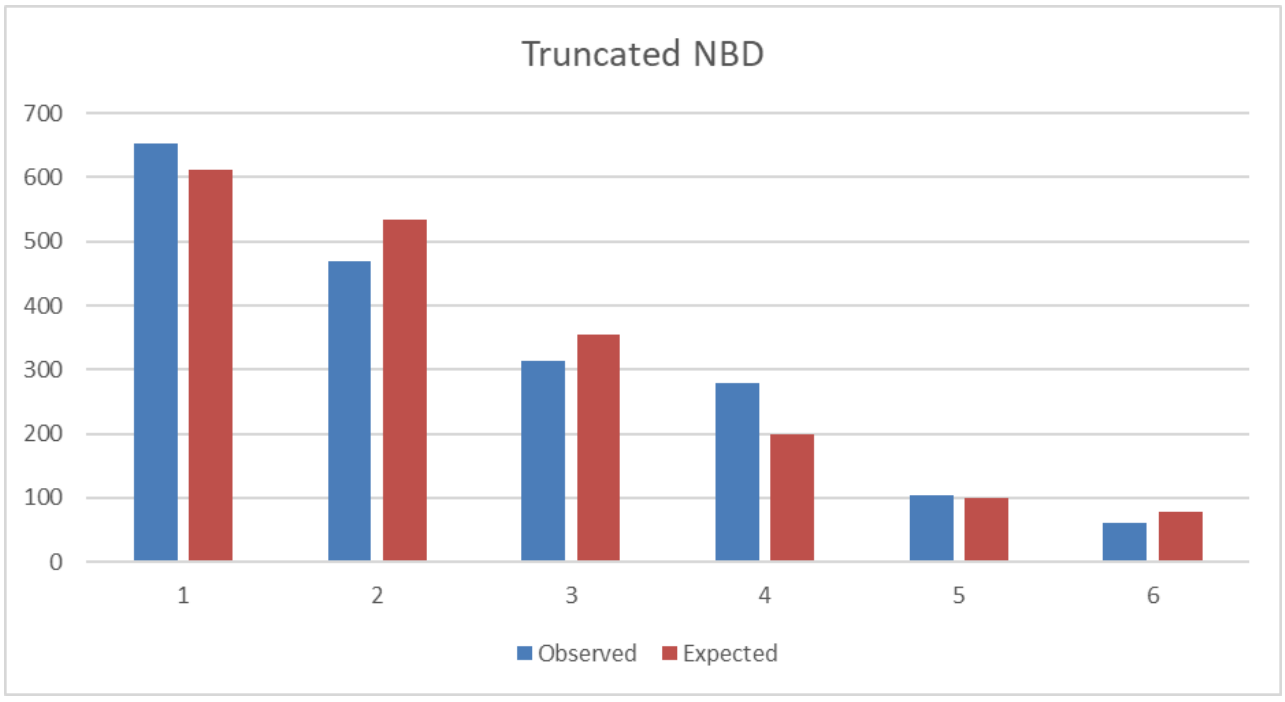
got 0 in their wheel spins. <u>The 0s estimated here would not involve hard-core customers.</u>

Hence, I hypothesize that the counts in effective 0s would be more or less the same as these indicate the potential buyers who just happened to be unfortunate.

Results

r	5.836424	**Mean**	2.994554
alpha	2.92618		

chi	50.26013
df	3
p-value	7.03E-11
effective 0s	411.7767



Inferences & Spike

- Again we observe that r is very high denoting homogeneity.
- The fit of the model is still not up to the mark
- The effective 0s are decently outputted, though the number could have been different had the model fitted better. Note that these are potential buyers and do not include “hard core nevers”
- The figure 411 for effective zeroes denote those who do contemplate to buy lines but do not.

d) Inference on Propensity

I intended to check the propensity of the customers to buy more than 3 lines - as mentioned earlier, the homogeneity baked in the data was indicative of the fact that people do not go over and above 3-4 lines per plan.
I leveraged the GAMMA distribution in order to check for the propensity of the customers to buy more than 4 lines.

Formula used: 1-GAMMA.DIST(A26,B1,1/B2,TRUE)

According to the model, the proportion of people who are willing to buy more than 4 lines is just 2%. This makes sense in the light of reasons discussed previously.

Propensity to buy more than 4 lines

Propensity	Cumultive prob	
4	0.981189808	**0.01881**

e) Refraining from other models

I refrained from fitting an sBG model or a beta binomial model as this is not a dataset where the customers are contemplating one of the two things (coin flipping process) or a choice dataset (with an upper bound) - this is a typical count dataset where customers contemplate bing a certain number of lines in their smartphone plans.

Limitations

1) I feel that the dataset on the purchase of the number of lines in a phone plan is a repeated process but over a very long period of time. People do not bring modifications to their plans in a short span. For example, if a customer has opted for some plan then he would stick to it for a considerable duration. **Put it simply, this process is not as frequently repeated as buying a coffee creamer or seeing a billboard.**
2) The dataset was right censored right from its origin which did not allow me to draw out the parameters of the NBD model via Means 7 Zeroes estimation or Method of moments. **This is because we cannot compute the mean or the variance for an observed right censored data directly.**

3) It is worth noting that, as we discussed in the class, since we have only targetted the buyers of the plan with a smartphone and a mobile plan) and for the reason discussed above that the time frame is relatively short, we were bound to get a model with high homogeneity.

   This is very similar to the colgate analysis we did in the lecture. If we just consider the buyers and go for a data that is as short a duration as a day, our estimates were relatively homogeneous. As we increase the time, more heterogeneity is bound to kick in!

Further Implementations

1) As further steps, it would be nice to look into covariates in order to see how different segments behave and compare the results across each of the segments.

2) Purchase of lines would be impacted by “social contagions” as well. For example, if there are good deals, people do come together and buy family plans, or perhaps children in the family buying new phones may change the plans held by customers previously. Hence it would be nice to account for these social changes that are not guided on an individual level.