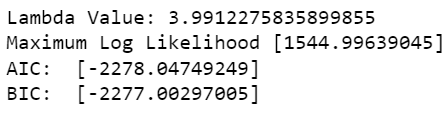
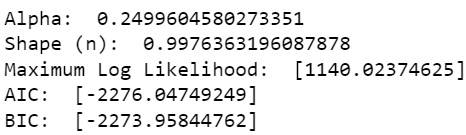
**PART - 1**

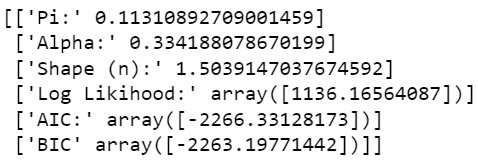
Q1a. The lambda value for the Poisson model after optimization is 3.9912275835899855, the maximum Log likelihood function value for the Poisson model is obtained as 1544.99639045 AIC: -2278.04749249, BIC: -2277.00297005



Q1b. The Alpha value for the NBD model is 0.2499604580273351, Shape parameter value for NBD model is 0.9976363196087878 and the Log likelihood function value for NBD model after optimization is 1140.02374625 AIC: -2276.04749249, BIC: -2273.95844762

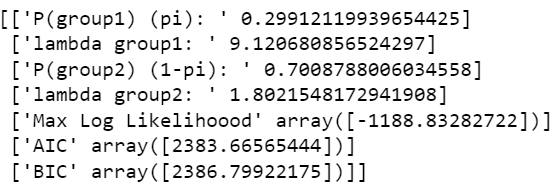


Q1c. The Alpha value for the Zero-inflated NBD model 0.334188078670199, shape parameter value for Zero-inflated NBD model 1.5039147037674592, pie value for Zero-inflated NBD model 0.11310892709001459, Maximum Log likelihood value for Zero inflated NBD model after optimization 1136.16564087, AIC: -2266.33128173, BIC -2263.19771442



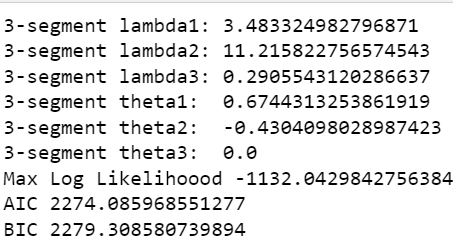
Q1d. 2 - Segment

λ1 value for two-segment 9.120680856524297, λ2 value for two-segment 1.8021548172941908, π 1 value for two-segment 0.29912119939654425, π2 value for two-segment 0.7008788006034558, Maximum Log likelihood value for 2-segment finite mixture model after optimization -1188.83282722 AIC - 2383.66565444, BIC - 2386.79922175



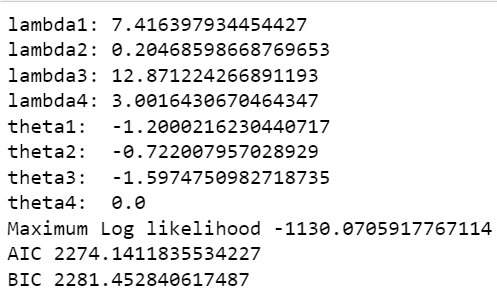
3 - Segment

λ1 value for three-segment 3.483324982796871, λ2 value for three-segment 11.215822756574543, λ3 value for three-segment 0.2905543120286637, θ1 value for three-segment 0.6744313253861919, θ2 value for three-segment -0.4304098028987423, θ3 value for three-segment 0.0, Maximum Log likelihood value for 3-segment finite mixture model after optimization -1132.0429842756384, AIC 2274.085968551277 BIC 2279.308580739894

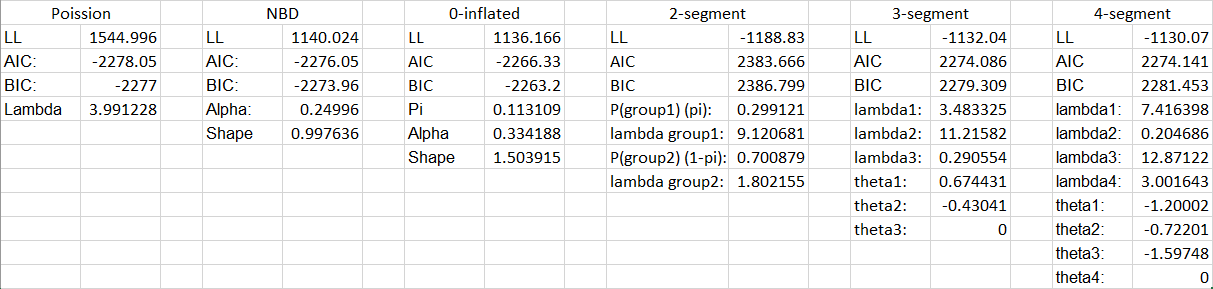


4 - Segment

λ1 value for three-segment 7.416397934454427, λ2 value for three-segment 0.20468598668769653, λ3 value for three-segment 12.871224266891193, λ4 value for four-segment 3.0016430670464347, θ1 value for three-segment -1.2000216230440717, θ2 value for three-segment -0.722007957028929, θ3 value for three-segment -1.5974750982718735, θ4 value for four-segment 0.0, Maximum Log likelihood value for 4-segment finite mixture model after optimization -1130.0705917767114



Q2.

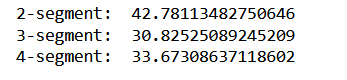


Of the models developed, the “best” one would be the 0-inflated model. With the lowest AIC and BIC and a log likelihood right in the middle of all of the other models, it has a “sweet spot” between enough parameters to adequately estimate the model, while not overburdening it with superfluous computations.

The differences between the models result in how good they are at estimating parameters of the model. Models like the poission account for no variability in parameter estimated, while the 0-inflated NBD accounts for the most. The 2-4 segment models may be beneficial for identifying customer segments, but in this case show that customers are more likely to be entirely heterogeneous (NBD-based models) then to be split into 2-4 homogenous groups.

Q3a

Given that the customer has purchased 5 packs in the last week, the following are the expected values for purchases in the next 8 weeks:



Which notes that, at the very least, the customer is likely to purchase more than 30 packs. It’s interesting to note that the 3 and 4 segment model give higher values then multiplying 5\*8, implying that the customer is likely to purchase less then 5 packs a week in the future.

Q3b

The following is the likelihood of the total amount of purchases the customer will make in the next 8 weeks, given they have made 9 purchases this week, model dependent:



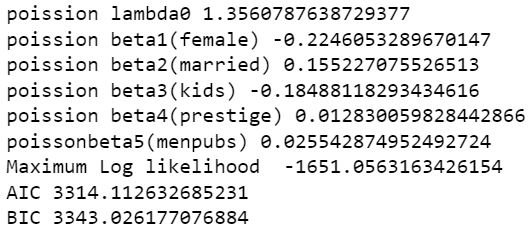
It’s interesting to see that this model is far more spread. A naive guess (8\*9 – 72 purchases) is about the same as the 2-segment model, but the 3 segment guesses higher and the 4-segment guesses lower. The manager will have to use their domain knowledge to decipher which one is more likely to be true.

**PART - 2**

Q1. The optimized log likelihood value for Poisson Regression is -1651.0563163426154

The other associated parameters for Poisson Regression are λ0 = 1.3560787638729377, β1 = - 0.2246053289670147, β2 = 0.155227075526513, β3 = -0.18488118293434616, β4= 0.012830059828442866 and β5= 0.025542874952492724, AIC 3314.112632685231, BIC 3343.026177076884

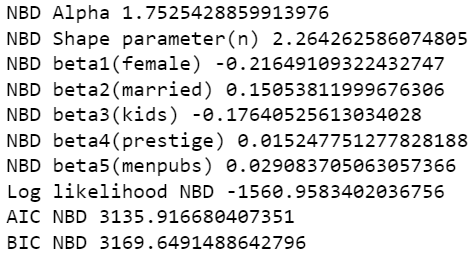
Married independent variable seems to be more important as it can increase the count variable in the Poisson Regression Model



Q2. The optimized log likelihood value for NBD Regression is -1560.9583402036756

The other associated parameters for NBD Regression are alpha = 1.7525428859913976, Shape Parameter value n = 2.264262586074805, β1= -0.21649109322432747, β2 = 0.15053811999676306, β3 = -0.17640525613034028, β4= 0.015247751277828188 and β5= 0.029083705063057366, AIC NBD 3135.916680407351  
BIC NBD 3169.6491488642796

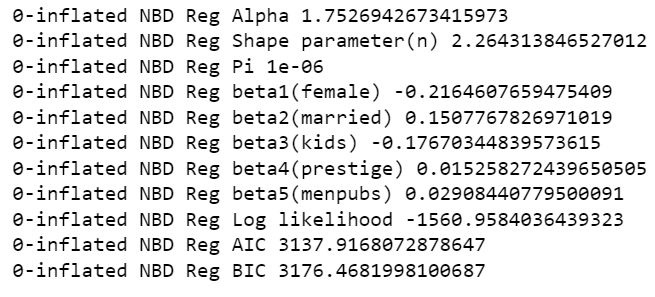
Married independent variable seems to be more important as it can increase the count variable in the NBD Regression Model



Q3. The optimized log likelihood value for Zero Inflated NBD Regression is -1560.9584036439323

The other associated parameters for Zero Inflated NBD Regression are alpha = 1.7526942673415973, Shape Parameter value n = 2.264313846527012, π= 1e-06, β1= -0.2164607659475409, β2 = 0.1507767826971019, β3 = -0.17670344839573615, β4= 0.015258272439650505 and β5= 0.02908440779500091

Married independent variable seems to be more important as it can increase the count variable in the Zero inflated NBD Regression



logic used in developing the model :

The 0-inflated NBD regression is a combination of 4 different assumptions. First, we assume that publishing occurs with a poisson distribution - that is, there is an odds (lambda) that is the average publish rate, and that different people are likely to produce k amount of articles distributed across that rate.

Second, we assume that, rather than the population having their own lambda, each person has a lambda unique to them. This distribution of individual lambdas falls along a gamma distribution, described by a shape and scale variable. This is the NBD model. It incorporates heterogeneity in the model, and is generally fairly effective.

Third, we add in additional variables. We assume that we can further explain each person's individual lambda via characteristics, each with their own beta variable. This is the NBD regression. It allows us to incorporate known quantities within the dataset into our analysis.

Finally, we inflate the 0’s. We assume that the model can be split into two groups. One group will (in this case) never publish an article, and the other will. We find an optimal probability - pi - that someone belongs to the group that will never publish. This allows us to account for scenarios where a large count at 0 throws off the whole model.

We found that the most important characteristics are if the candidate is female, married, or has kids. Female candidates are less likely to publish, as are candidates with children. Married candidates are more likely to publish. Further analysis could be done on the covariance between married candidates and candidates with children. This tells us that the quality of the candidate's school or scholarship of their mentor is less impactful than their social situation (candidates with kids may not have the time to publish alongside other work).

Q4. We found that the AIC and BIC for the 0-inflated model are lower than the previous models. The reason for this is the incredibly low value of pi after optimization - the model assumes that almost all doctoral candidates will publish at some point in their lifetime. This is a reasonable assumption given the rigor of a doctoral program and the desire to get work published. However, this means the model performs similarly to a NBD regression, but the additional variable lowers the AIC.

We can see very similar values for all of the betas across all of the models. This means that we are confident in the findings above – being female or having kids lowers the number of papers produced, while being married raises the number. We recommend the Poisson Regression then. The log-likelihood ratio is roughly 100 between it and the other models. However, the fact that it can create similar conclusions with less variables means it has the lowest AIC and BIC and is therefore our best model. If one was concerned only about accuracy and not about computational concerns, the NBD regression would still be perfered to the 0-inflated NBD, as the latter adds little to no information.

