Question 1:

Consider the hard candy example from class. The associated data is in the file candy.csv. Develop the following, models discussed in class using maximum likelihood estimation (MLE):

a. the Poisson model,

b. the NBD model,

c. the Zero-Inflated NBD model, and

d. Finite Mixture models for 2, 3, and 4 segments.

Report your code and all relevant details, including the estimated values of the parameters for each model and the corresponding log-likelihood values. Please add comments to your code to make it easy to understand.

Answer:

1. The Poisson model

* Data Type:

Packs int64

People int64

dtype: object

* Shape of dataframe: (21, 2)

# First five records in the dataset

Packs People

0 0 102

1 1 54

2 2 49

3 3 62

4 4 44

* Poisson Model (Candy Dataset) Estimated Values:

Estimated Poisson parameter (lambda): 3.9912278692038283

Maximum Log-Likelihood (Poisson Model): -1544.996390448969

* Predict the number of people for k exposures
* Plot for Actual and Predicted values using Poisson Model (Billboard Data)

A graph of different colored bars

Description automatically generated

1. The NBD model:

* Initial guess for NBD parameters (alpha, shape)

(1,0.5)

* Perform Maximum Likelihood Estimation for NBD

NBD Model (Candy Dataset) Estimated Values:

Estimated NBD parameters (alpha, shape):

0.2499636938580261 0.9976641804701156

Maximum Log-Likelihood (NBD): -1140.0237461879096

* Graph for original and predicted number of exposures

A graph of a graph of value

Description automatically generated with medium confidence

1. the Zero-Inflated NBD model:

* Initial guess for ZI NBD parameters (pi, alpha, shape)

(0.02, 1, 0.5)

* Perform Maximum Likelihood Estimation for NBD

Zero Inflated NBD Model (Candy Dataset) Estimated Values:

Estimated Zero Inflated NBD parameters (pi, alpha, shape): 0.11310370076049828 0.3341876580996791 1.5039185087655875

Maximum Log-Likelihood (ZI NBD): -1136.1656408319284

* Predict the number of people for k exposures based on the ZI NBD model
* Graph for original and predicted number of exposures

A graph of different colored bars

Description automatically generated

1. Finite Mixture models for 2, 3, and 4 segments.

2 Segment Model

* Initial guess for Finite Mixture 2 Segment Model parameters (lambda1, lambda2, pi1) - 0.01, 0.02, 0.01
* Perform Maximum Likelihood Estimation for Finite Mixture 2 Segment Model

Finite Mixture 2 Segment Model (Candy Dataset) Estimated Values:

Estimated Finite Mixture 2 Segment Model parameters (lambda1, lambda2, pi1): 1.8021535211489903 9.120686057554176 0.7008860871584063

Maximum Log-Likelihood (ZI NBD): -1188.8328271724242

* Predict the number of people for k exposures based on the Finite Mixture 2 Segment model
* Graph for original and predicted number of exposures

A graph of different colored bars

Description automatically generated

**3 Segments Model**

* Initial guess for Finite Mixture 3 Segment Model parameters (lambda1, lambda2, lambda3, theta1, theta2) - 0.05, 0.05, 0.02, 0.01, 0.02
* Perform Maximum Likelihood Estimation for Finite Mixture 3 Segment Model

Finite Mixture 3 Segment Model (Candy Dataset) Estimated Values:

Estimated Finite Mixture 3 Segment Model parameters (lambda1, lambda2, lambda3, theta1, theta2): 3.48337070933276 11.215912174077538 0.2905151612455488 0.6744646396142554 -0.4304445259727132

Maximum Log-Likelihood (Finite Mixture 3 Segment Model): -1132.0429844099829

* Predict the number of people for k exposures based on the Finite Mixture 3 Segment model
* Graph for original and predicted number of exposures

A graph of different colored bars

Description automatically generated

**4 Segments fmm**

* Initial guess for Finite Mixture 4 Segment Model parameters (lambda1, lambda2, lambda3, lambda4, theta1, theta2, theta3) - 0.01, 0.01, 0.01, 0.01, 0.01, 0.001, 0.02
* Perform Maximum Likelihood Estimation for Finite Mixture 4 Segment Model

Finite Mixture 4 Segment Model (Candy Dataset) Estimated Values:

Estimated Finite Mixture 4 Segment Model parameters (lambda1, lambda2, lambda3, theta1, theta2, theta3): 7.418270719589816 3.0019202302244943 12.872583216092373 0.20470930000653137 -0.478018639706003 0.7221211606514268 -0.8759818166183413

Maximum Log-Likelihood (Finite Mixture 4 Segment Model): -1130.0705911210619

* Predict the number of people for k exposures based on the Finite Mixture 2 Segment model

A graph of different colored bars

Description automatically generated

Question 2:

Evaluate the models developed; explain which of them is best, and why. Are there any significant differences among the results from these models? If so, what exactly are these differences? Discuss what you believe could be causing the differences.

Answer:

The evaluation of models depends on several factors, and choosing the "best" model involves considering both statistical metrics and practical relevance to the specific problem. Let's analyze the models and their respective results:

A black and white screen with numbers and a black background

Description automatically generated  
**Comparison and Evaluation:**

The models with lower AIC and BIC values are generally preferred, as these criteria penalize models with more parameters. Lower AIC and BIC values indicate a better trade-off between model fit and complexity.

The 3-segment Finite Mixture Model (3 SEG FM) has the lowest AIC and BIC among the finite mixture models, suggesting that it might be the preferred model among the considered mixture models.

**Differences among Models:**

The Poisson, Negative Binomial (NBD), and Zero-Inflated Negative Binomial (ZI-NBD) models have fewer parameters but may not capture the complexity of the data as well as the finite mixture models.

The finite mixture models (2 SEG FM, 3 SEG FM, 4 SEG FM) have more parameters and might offer a better fit to the data, especially the 3-Segment model.

**Possible Causes of Differences:**

The data may exhibit different patterns or behaviors that are better captured by models with different complexities.

The finite mixture models allow for more flexibility in capturing heterogeneous subpopulations within the data, potentially leading to better fit.

In summary, based on the model evaluation, the 3-segment Finite Mixture Model appears to be the preferred model due to lower AIC and BIC values and a higher log-likelihood.

Question 3:

Based on the 2, 3, and 4-segment finite mixture models, how many packs are the following customers likely to purchase over the next 8 weeks?

1. customer who purchased 5 packs in the past week, and

Answer:

2 Seg FM Model

* Predicting Posterior Probability - 2 Seg FM Model

A graph of a number of packs purchased

Description automatically generated

* customer who purchased 5 packs in the past weeks: 49.71471996469466
* customer who purchased 9 packs in the past weeks: 72.8924509398868

3 Seg FM Model

* Predicting Posterior Probability - 3 Seg FM Model

A graph of a number of packs purchased

Description automatically generated

customer who purchased 5 packs in the past weeks: 30.825246830490574

customer who purchased 9 packs in the past weeks: 80.06282859753176

4 Seg FM Model

* Predicting Posterior Probability - 4 Seg FM Model

A diagram of a number of packs purchased

Description automatically generated

customer who purchased 5 packs in the past weeks: 33.67147248894754

customer who purchased 9 packs in the past weeks: 69.45336986379394

**Analysis of New Data**

articles.csv contains the number of publications by 915 doctoral candidates (articles), along with five predictors:

* + - 1. female: 1 if candidate was female, 0 otherwise
      2. married: 1 if candidate was married, 0 otherwise
      3. kids: number of children aged ≤ 5 4.
      4. prestige: prestige of the candidate's department (higher is better)
      5. mentorpubs: number of publications by the candidate's mentor over the past 3 years

Your task is to build a model to predict the number of articles as a function of the five independent variables. Specifically, you need to do the following:

1. Estimate all relevant parameters for Poisson regression using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood. What are the managerial takeaways – which customer characteristics seem to be important? Predict the number of people purchasing 0, …, 20, 20+ packs based on the Poisson regression. Explain how the predicted values are obtained using the case of 5 packs (show your calculations). Graph the original and predicted number of packs.

Predict the number of people purchasing 0, …, 20, 20+ packs based on the Poisson regression. Explain how the predicted values are obtained using the case of 5 packs (show your calculations). Graph the original and predicted number of packs.

Answer:

* Bringing in Articles Dataset
  + Data Type:

articles int64

female int64

married int64

kids int64

prestige float64

menpubs int64

dtype: object

Shape of dataframe: (915, 6)

articles female married kids prestige menpubs

0 0 0 1 0 2.52 7

1 0 1 0 0 2.05 6

2 0 1 0 0 3.75 6

3 0 0 1 1 1.18 3

4 0 1 0 0 3.75 26

* Estimated Poisson Regression Parameters: 1.3561737432427203 -0.22462603961190555 0.15525498258537043 -0.18489610215243077 0.012806750630080894 0.025543378552394587
* Maximum Log-Likelihood (Poisson Regression): -1651.0563166022598
* Predicting Number of Visitors

A graph with purple and grey bars

Description automatically generated

* **Calculation:**

To obtain the predicted number of people **for Case 5** of Exposures,

We have, NumberOfVisits = 5, NumberOfPeople = 27, total\_people = 915

* First, I calculated the mean values of X’s corresponding to each visit.

mean(x1\_5) = 0.4074

mean(x2\_5) = 0.7777

mean(x3\_5) = 0.4074

mean(x4\_5) = 3.3825

mean(x5\_5) = 13.407

Then, I calculated lambdai corresponding to each article.

lambdai (5) = 1.90455

P(X=k) = 0.031092

Predicted number of people = 0.03109 \* 915 = 28.44

**Actual number of people = 27**

**Predicted number of people = 28.44**

The below screenshot shows the calculation for case 5 of exposure,

**A grid of numbers and symbols

Description automatically generated**

* The managerial takeaways:
  + **Female (coefficient: -0.225):** A negative coefficient suggests that the expected number of articles published by female candidates is less compared to the articles published by male candidates.
  + **Married (coefficient: 0.155)**: A positive coefficient suggests that married candidates publish more articles compared to PHD candidates who are not married.
  + **Kids (coefficient: -0.185)**: A negative coefficient suggests that having kids is associated with a decrease in the expected number of articles published by the candidate.
  + **Prestige (coefficient: 0.013)**: A positive coefficient suggests that higher prestige is associated with an increase in the expected number of articles published by the candidate.
  + **Menpubs (coefficient: 0.026)**: A positive coefficient suggests that more publications by the candidate’s mentor are associated with an increase in the expected number of articles published by the candidate.
  + **Managerial Takeaway:**
  + **Customer Segmentation:** Consider segmenting customers based on their characteristics to tailor marketing strategies. For example, targeting married customers or those with higher prestige might result in more engagement.
  + **Demographic Insights:** The negative coefficient for *female* and *kids* indicates that these groups might be less active in terms of consuming articles. Understanding their preferences and adapting content might be beneficial.
  + **Content Strategy:** The positive coefficient for *menpubs* suggests that more publications by the candidate’s mentors are associated with an increase in the expected number of articles. Focusing on content preferences can be a key strategy.

Question 2:

Estimate all relevant parameters for NBD regression using MLE. Report your code, the estimated parameters, and the maximum value of the log-likelihood. What are the managerial takeaways? which customer characteristics seem to be important?

Predict the number of people purchasing 0, …, 20, 20+ packs based on the NBD regression. Explain how the predicted values are obtained using the case of 5 packs (show your calculations). Graph the original and predicted number of packs.

Answer:

NBD Regression

* Estimated NBD Regression Parameters: 2.264588170585852 1.7528169085375391 -0.2164925264234801 0.15048373390087377 -0.17638293524692333 0.015270092575949945 0.02908168379167232
* Maximum Log-Likelihood (NBD Regression): -1560.9583398020238
* **Calculation:**

To obtain the predicted number of people **for Case 5** of Exposures,

We have, NumberOfVisits = 5, NumberOfPeople = 27, total\_people = 915

* First, I calculated the mean values of X’s corresponding to each visit.

mean(x1\_5) = 0.4074

mean(x2\_5) = 0.7777

mean(x3\_5) = 0.4074

mean(x4\_5) = 3.3825

mean(x5\_5) = 13.407

* Then, I calculated exp\_beta and other parts of the nbd regression probability formula corresponding to each article.

Exp\_beta (5) = 1.4884, part1 = 8.66, part2 = 0.2485, part3 = 0.0204

* P(X=k) = 0.0439
* Predicted number of people = 0.0439 \* 915 = 40.23
* **Actual number of people = 27**
* **Predicted number of people = 40.23**

The below screenshot shows the calculation for case 5 of exposure,

**A screen shot of a computer

Description automatically generated**

* Graph for original and predicted number of exposures

A graph with purple and grey bars

Description automatically generated

* **Managerial Takeaway:**
  + ***Female*:** A negative coefficient (-0.216) for *females* suggests that being female is associated with a decrease in the expected number of articles published.
  + ***Married*:** A positive coefficient (0.150) for *married* suggests that being married is associated with an increase in the expected number of articles published. Marketing strategies could be tailored to this demographic.
  + ***Kids*:** A negative coefficient (-0.176) for *kids* suggests that having kids is associated with a decrease in the expected number of articles published. Understanding and addressing the preferences of customers with children may be important.
  + ***Prestige*:** A positive coefficient (0.015) for *prestige* suggests that candidates in the higher prestigious department are more likely to publish an increased number of articles.
  + ***Menpubs*:** A positive coefficient (0.029) for *menpubs* suggests that more publications by the candidate’s mentor are associated with an increase in the expected number of articles published by the candidate.
* **General Insights:**

**Demographic Differences:** The coefficients for demographic variables (*female*, *married*, *kids*) indicate that customer characteristics play a role in purchasing behavior.

**Content Preferences:** Preferences for certain content types, such as *menpubs*, influence customer engagement. Adjusting content based on preferences can impact purchasing decisions.

**Variability in Purchasing Behavior:** The estimated dispersion parameter (α) suggests that there is variability in purchasing behavior not accounted for by a Poisson distribution. This may reflect heterogeneity among customers.

Question 03:

In this question, you will apply the ideas learned in this course to build a model that you have not seen before – the Zero-Inflated NBD regression.

First, recall that zero-inflated models view 0s as coming from 2 sources –

(i) from a fraction π who is 0 “by type” (in the context of this problem, these are candidates who will never publish), and

(ii) from the remaining fraction (1 – π) who are likely to eventually become nonzero (these are candidates who will publish at some point, but have not done so yet). You can assume that the candidates in the latter group are distributed as a negative binomial (making the NBD regression appropriate for them).

Explain the logic used in developing the model in detail. (hint: you do not need anything beyond what you have learned in the class to do this.)

Report your code, the estimated parameters and the maximum value of the log-likelihood. What are the managerial takeaways – which customer characteristics seem to be important?

Predict the number of people purchasing 0, …, 20, 20+ packs based on the Poisson regression. Explain how the predicted values are obtained using the case of 5 packs (show your calculations). Graph the original and predicted number of packs.

Answer:

* Estimated ZI NBD Regression Parameters: 2.2646131265702336 1.7533444917662122 -0.21648238236708722 0.15057135236052457 -0.17635103890555914 0.015333190003517187 0.02908264503119874 1e-06
* Maximum Log-Likelihood (ZI NBD Regression): -1560.9583926419843
* **Calculation:**

To obtain the predicted number of people **for Case 5** of Exposures,

We have, NumberOfVisits = 5, NumberOfPeople = 27, total\_people = 915

First, I calculated the mean values of X’s corresponding to each visit.

mean(x1\_5) = 0.4074

mean(x2\_5) = 0.7777

mean(x3\_5) = 0.4074

mean(x4\_5) = 3.3825

mean(x5\_5) = 13.407

Then, I calculated exp\_beta and other parts of the nbd regression probability formula corresponding to each article.

Exp\_beta (5) = 1.4884, part1 = 8.66, part2 = 0.2485, part3 = 0.0204

P(X=k) = 0.04

Predicted number of people = 0.04 \* 915 = 40.22

**Actual number of people = 27**

**Predicted number of people = 40.22**

The below screenshot shows the calculation for case 5 of exposure,

**A screenshot of a computer

Description automatically generated**

* Graph for original and predicted number of exposures

A graph with purple and grey bars

Description automatically generated

* **Logic used to develop the model:**

We introduce zero-inflated models, when there are more zeros in the model.

Here, we assume zeros come from two sources,

* 1. candidates who will never publish (π)
  2. candidates who will publish in the future but haven’t done so yet (π-1).

Similar to NBD regression code, except that I have introduced π and π-1, without making any other changes.

Parameters in the model, **params = n, alpha, betas, pi**

Code changes, **prob.append((pi if v[i] == 0 else 0) + (1-pi) \* part1[i]\*part2[i]\*part3[i])**

Zero Inflated NBD Regression results are very similar to NBD regression results.

* **Managerial Takeaway:**

***Female*:** A negative coefficient (-0.216) for *females* suggests that being female is associated with a decrease in the expected number of articles. This could be an important consideration in targeted marketing efforts.

***Married*:** A positive coefficient (0.150) for *married* suggests that being married is associated with an increase in the expected number of articles. Marketing strategies could be tailored to this demographic.

***Kids*:** A negative coefficient (-0.176) for *kids* suggests that having kids is associated with a decrease in the expected number of articles. Understanding and addressing the preferences of customers with children may be important.

***Prestige*:** A positive coefficient (0.015) for *prestige* suggests that higher prestige is associated with an increase in the expected number of articles. This could be a target group for premium content or offers.

***Menpubs*:** A positive coefficient (0.029) for *menpubs* suggests that more publications by the candidate’s mentor are associated with an increase in the expected number of articles. Tailoring content to specific interests may be beneficial.

4. Evaluate the models developed; explain which of them is best, and why. Are there any significant differences among the results from these models? If so, what exactly are these differences? Discuss what you believe could be causing the differences.

Answer:  
**A screenshot of a computer

Description automatically generated**

* We find that the NBD Regression model is the best since this model contains lower AIC and BIC values compared to other models.
* We could find a significant difference between the Poisson regression and the NBD regression models; this difference is because of the heterogeneity incorporated by the NBD regression model.
* Whereas we don’t find much difference between NBD regression and ZI-NBD regression models, even the estimated parameters of both models look very similar.

This makes us choose the NBD regression model that includes heterogeneity and uses fewer parameters compared to ZI-NBD regression, and still performs best.

**Summary:**

* Part 1 – 3 Seg FM model is the best model with lower AIC and BIC values.
* Part 2 – The NBD Regression model is the best model with lower AIC and BIC values.
* Part 2 – The demographics of the candidate play a significant role with females and kids having a negative correlation with articles published, whereas married people have a positive correlation with articles published.