**Customer Purchase Behavior Modeling with Poisson and NBD Regression**

**Sriram Srinivasan**

Table of Contents

[Part I: Replicating Models from Class 3](#_Toc182160916)

[#1. The Poisson Model 3](#_Toc182160917)

[#2. The NBD Model 5](#_Toc182160918)

[#3. The Poisson Regression 9](#_Toc182160919)

[#4. The NBD Regression 12](#_Toc182160920)

[#5. Managerial Takeaways 16](#_Toc182160921)

[Part II: Analysis of New Data 17](#_Toc182160922)

[#1. Print first and last records of both new datasets 17](#_Toc182160923)

[#2. The Poisson Model books01.csv 18](#_Toc182160924)

[#3. The Poisson Model books02.csv 18](#_Toc182160925)

[#4. The NBD Model books01.csv 22](#_Toc182160926)

[#5. The NBD Model books02.csv 22](#_Toc182160927)

[#6. Calculate the values of: 24](#_Toc182160928)

[#7. Dealing with Missing Values 25](#_Toc182160929)

[#8. Incorporate all the available customer characteristics and estimate all relevant parameters for Poisson regression using MLE. 26](#_Toc182160930)

[#9. Estimate all relevant parameters for NBD regression using MLE. 29](#_Toc182160931)

[#10. Evaluate all the models developed using the log-likelihood ratio, AIC, and BIC 32](#_Toc182160932)

[Model Recommendation: NBD Regression 32](#_Toc182160933)

[# 11. Group Discussion Board 35](#_Toc182160934)

[**Customer Purchase Behavior Insights** 36](#_Toc182160935)

[**Important Predictors: ‘Race,’ ‘Country,’ and ‘Region’** 37](#_Toc182160936)

[**Additional Variables to Consider: ‘Age’ and ‘Income’** 37](#_Toc182160937)

[**Recommendations for Future Analysis** 37](#_Toc182160938)

# Part I: Replicating Models from Class

Please Note: Instead of using `for` loops for calculations, vectorized calculations were used for computational efficiency. Instead of storing intermediate log-likelihood values in an array, used direct sum for the total negative log-likelihood was used.

## #1. The Poisson Model

1. Report your code, the estimated parameters and the maximum value of the log-likelihood:

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimated Lambda (λ) | ≈ 4.4560 |
| Maximum Log-Likelihood | ≈ -929.0439 |

1. Predict the number of people with 0, ..., 23 exposures based on the Poisson model.

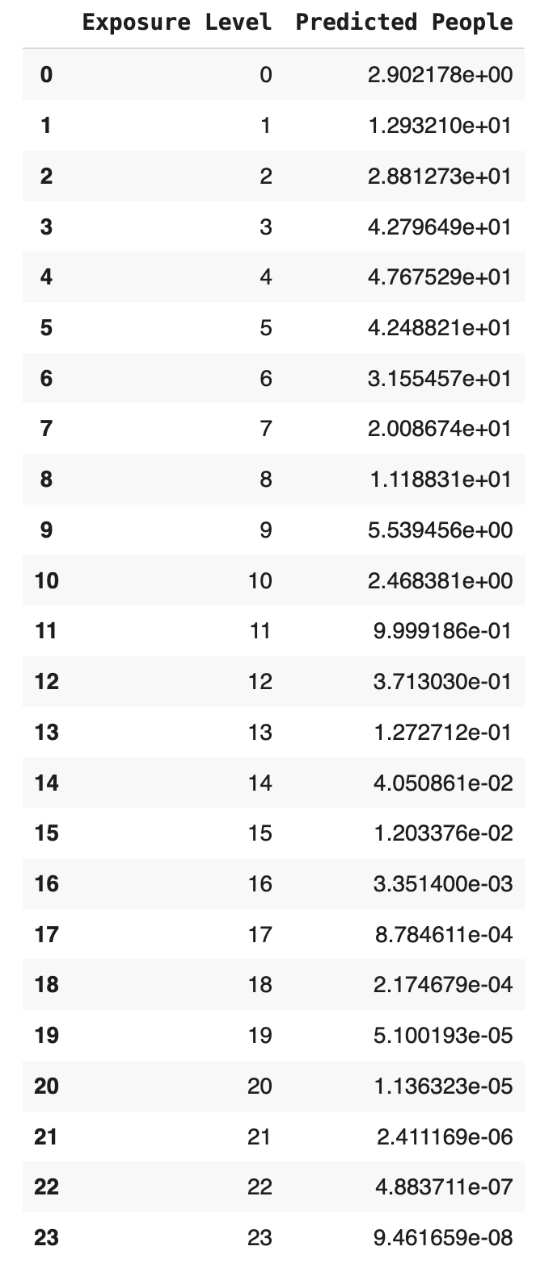


Figure : Predictions of people with 0, ..., 23 exposures based on the Poisson model

1. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations).

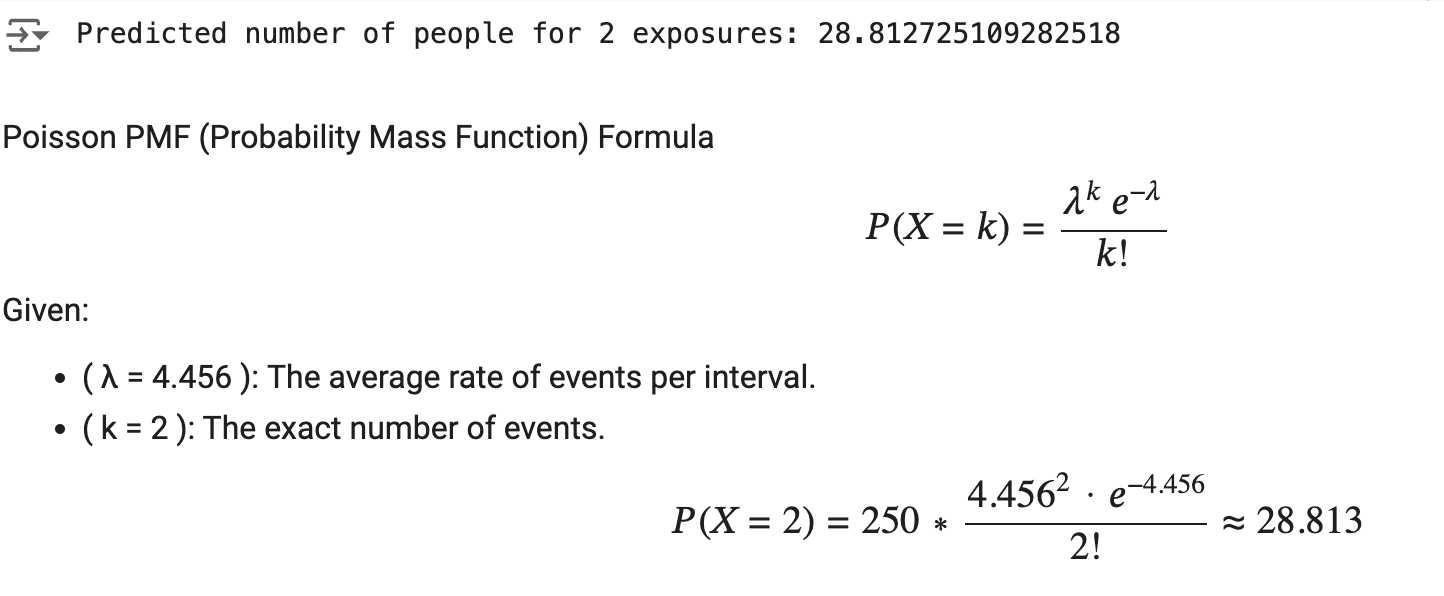


Figure : Calculations of predicted value with a case of 2 exposures using the Poisson model

1. Graph the original and predicted numbers of exposures

A graph of different colored bars

Description automatically generated

Figure : Bar chart depicting the original and predicted number of exposures using Poisson distribution

## #2. The NBD Model

1. Report your code, the estimated parameters and the maximum value of the log-likelihood

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimate of n | ≈ 0.9693 |
| Estimate of α | ≈ 0.2175 |
| Maximum Log-Likelihood | ≈ -649.6888 |

1. Evaluate the NBD model vis-à-vis the Poisson model; explain which is better and why.

* The NBD model is the better choice for this data because the predicted values align more closely with the actual observed data, leading to higher maximum log-likelihood.
* The Poisson model assumes the same rate parameter λ for everyone, while the NBD model allows λ to have a gamma distribution, which adds flexibility in handling overdispersion and capturing the declining trend in the tail.
* This plot highlights the limitations of the Poisson model in cases where the data has higher variability than the Poisson assumption can accommodate.
* The NBD model’s additional shape parameter enables it to adjust for this variability, resulting in a more accurate fit across all exposure levels.

A graph of different colored bars

Description automatically generated

Figure :Bar chart depicting the original and comparing the predicted number of exposures using the Poisson Model vs. NBD Model

1. Predict the number of people with 0, ..., 23 exposures based on the NBD model

A table of numbers with numbers

Description automatically generated

Figure : Predicted number of people with 0...,23 exposures based on NBD model

1. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations)

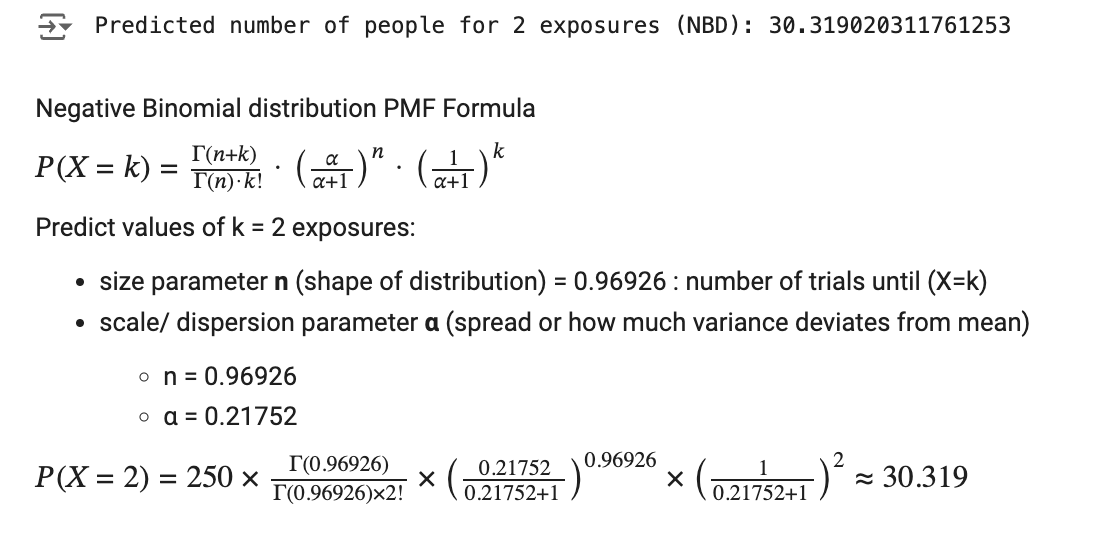


Figure :Calculations of predicted value with a case of 2 exposures using the NBD model

1. Graph the original and predicted numbers of exposures

A graph of different colored bars

Description automatically generated

Figure : Bar chart depicting the original and predicted number of exposures using NBD distribution

## #3. The Poisson Regression

1. Report your code, the estimated parameters and the maximum value of the log-likelihood.

|  |  |
| --- | --- |
| MLE Output | Value |
| λ0 = exp(*β0)* | ≈ 0.0437 |
| *β1* | ≈ 0.0941 |
| *β2* | ≈ 0.0044 |
| *β3* | ≈ 0.5887 |
| *β4* | ≈ -0.0359 |
| Maximum Log-Likelihood | ≈ -6291.4968 |

A screenshot of a computer

Description automatically generated

Figure : Summary Table of Poisson Regression Model

1. Predict the number of people with 0, ..., 23 exposures based on the Poisson regression.

A table of numbers and symbols

Description automatically generated

Figure : Predicted number of people with 0...,23 exposures based on Poisson regression

1. Explain how the predicted values are obtained using the case of 2 exposures (show your calculations).

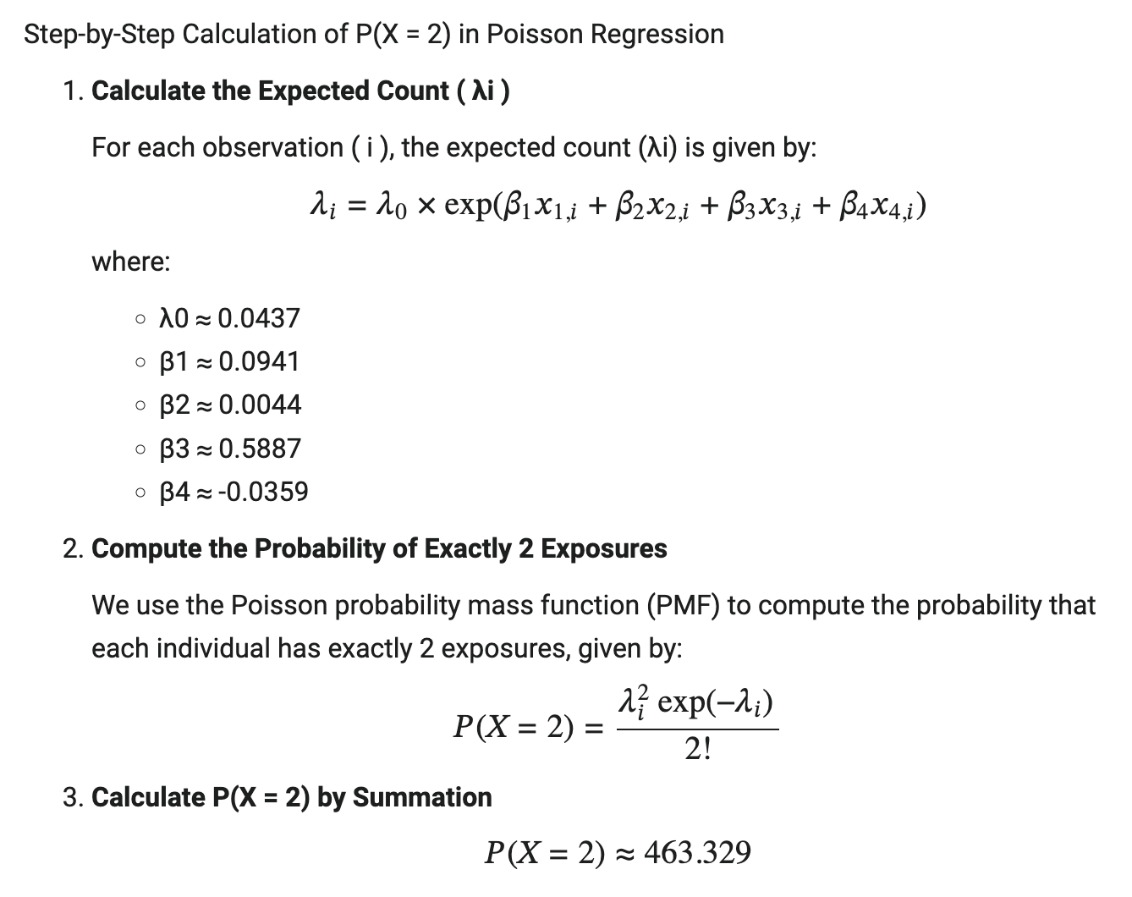


Figure :Calculations of predicted value with a case of 2 exposures using the Poisson regression

1. Graph the original and predicted numbers of exposures.

A graph with numbers and text

Description automatically generated with medium confidence

Figure : Bar chart depicting the original and predicted number of exposures using Poison Regression distribution

## #4. The NBD Regression

1. Report your code, the estimated parameters and the maximum value of the log-likelihood.

|  |  |
| --- | --- |
| MLE Output | Value |
| n | ≈ 0.1387 |
| α | ≈ 8.2100 |
| *β1* | ≈ 0.0733 |
| *β2* | ≈ -0.0089 |
| *β3* | ≈ 0.9027 |
| *β4* | ≈ -0.0243 |
| Maximum Log-Likelihood | ≈ -2888.9661 |

1. Evaluate the NBD regression vis-à-vis the Poisson regression; explain which is better and why.

A screen shot of a graph

Description automatically generated

Figure : Bar chart depicting the original and comparing the predicted number of exposures using the Poisson Regression Model vs. NBD Regression Model

* The NBD regression is the better choice for this data because the predicted values align more closely with the actual observed data, leading to higher maximum log-likelihood.
* The Poisson regression assumes the same rate parameter λ0 for everyone, while the NBD model allows λ0 to vary across population through a gamma distribution, which adds flexibility in handling overdispersion and capturing the declining trend in the tail.
* This plot highlights the limitations of the Poisson model in cases where the data has higher variability than the Poisson assumption can accommodate.
* The NBD model’s additional shape parameter enables it to adjust for this variability, resulting in a more accurate fit across all exposure levels.

1. Predict the number of people with 0, ..., 23 visits based on the NBD region model.

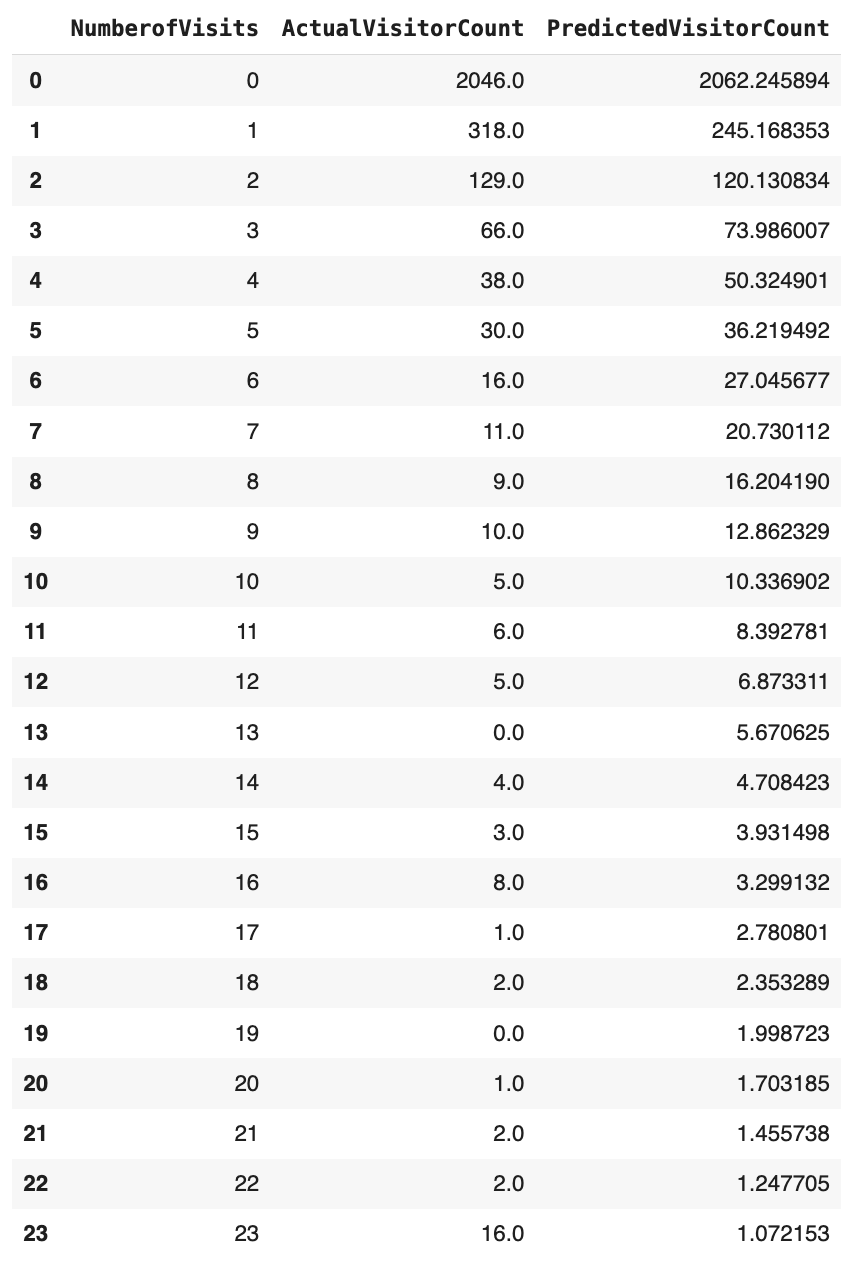


Figure :Predicted number of people with 0...,23 exposures based on NBD regression

1. Explain how the predicted values are obtained using the case of 2 visits (show your calculations).

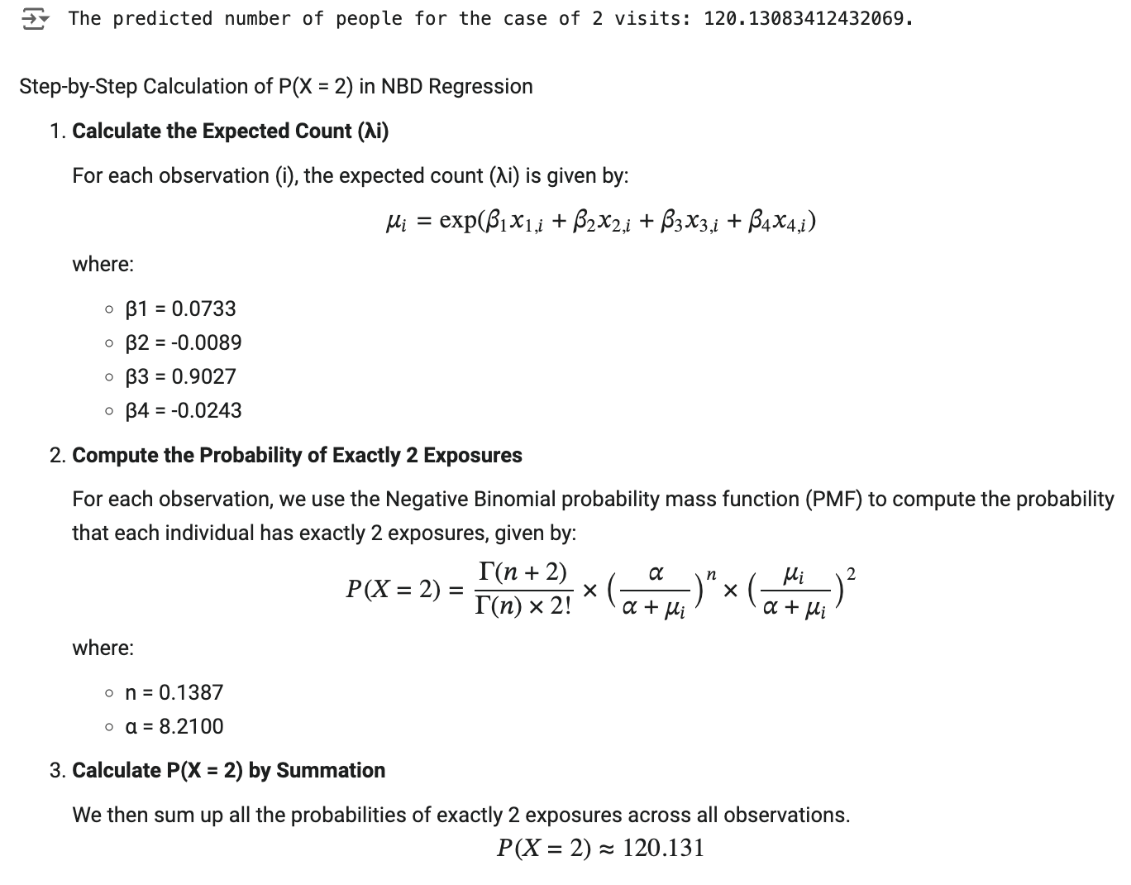


Figure : Calculations of predicted value with a case of 2 exposures using the NBD regression

1. Graph the original and predicted numbers of exposures.

A screen shot of a graph

Description automatically generated

Figure : Bar chart depicting the original and predicted number of exposures using NBD Regression distribution

## #5. Managerial Takeaways

1. Structures of datasets

* billboard.csv
  + Advantage: more condensed, gives a better overview of the number of people for each exposure count
  + Disadvantage: can only be used for Poisson and NBD models which do not depend on other characteristics
* Khakichinos.csv
  + Advantage: can accommodate all types of models – especially regression on other independent variables, easy to obtain – data tend to be collected by transaction/customer with other characteristics
  + Disadvantage: may take time to organize if each ID may contain many records, some variables may be redundant and can confuse the models

1. Poisson Model

* Intuitive to use as a count model
* Does not do well with overdispersion because Poisson assumes mean, and variance is equal

1. NBD Model

* Introduces flexibility in the rate by allowing it to follow a gamma distribution
* More representative of an actual heterogeneous population

1. Poisson Regression

* Expansion of the Poisson model to capture characteristics of each individual within the population

1. NBD Regression

* Expansion of the NBD model to capture characteristics of each individual within the population
* Better than Poisson regression because of the ability to accommodate unobserved heterogeneity within the population

# Part II: Analysis of New Data

A black and white text with arrows

Description automatically generated

## #1. Print first and last records of both new datasets

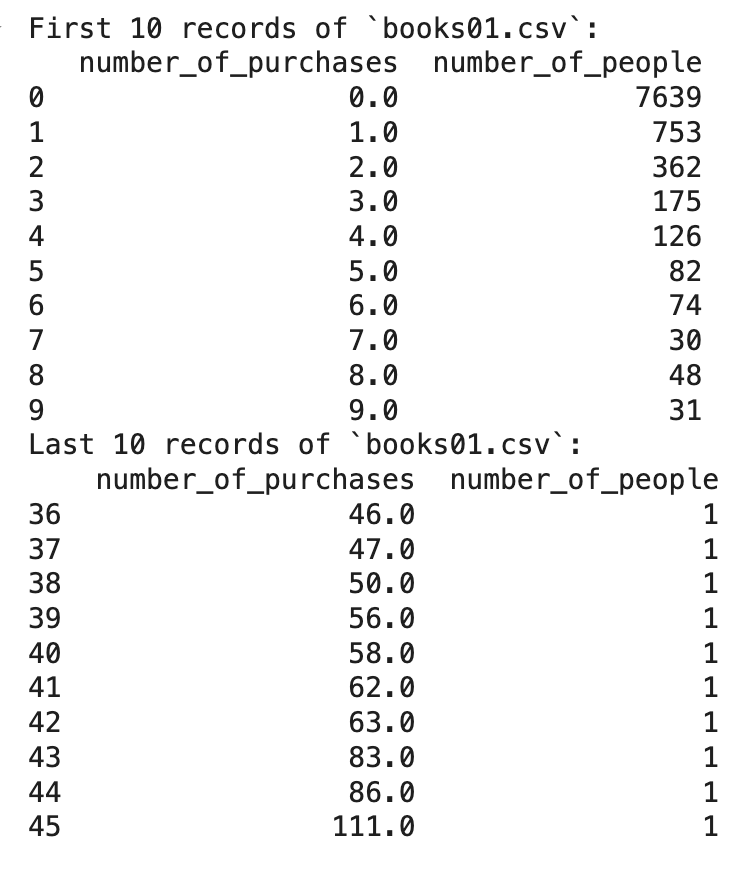


Figure : First and last 10 records of `books01.csv`

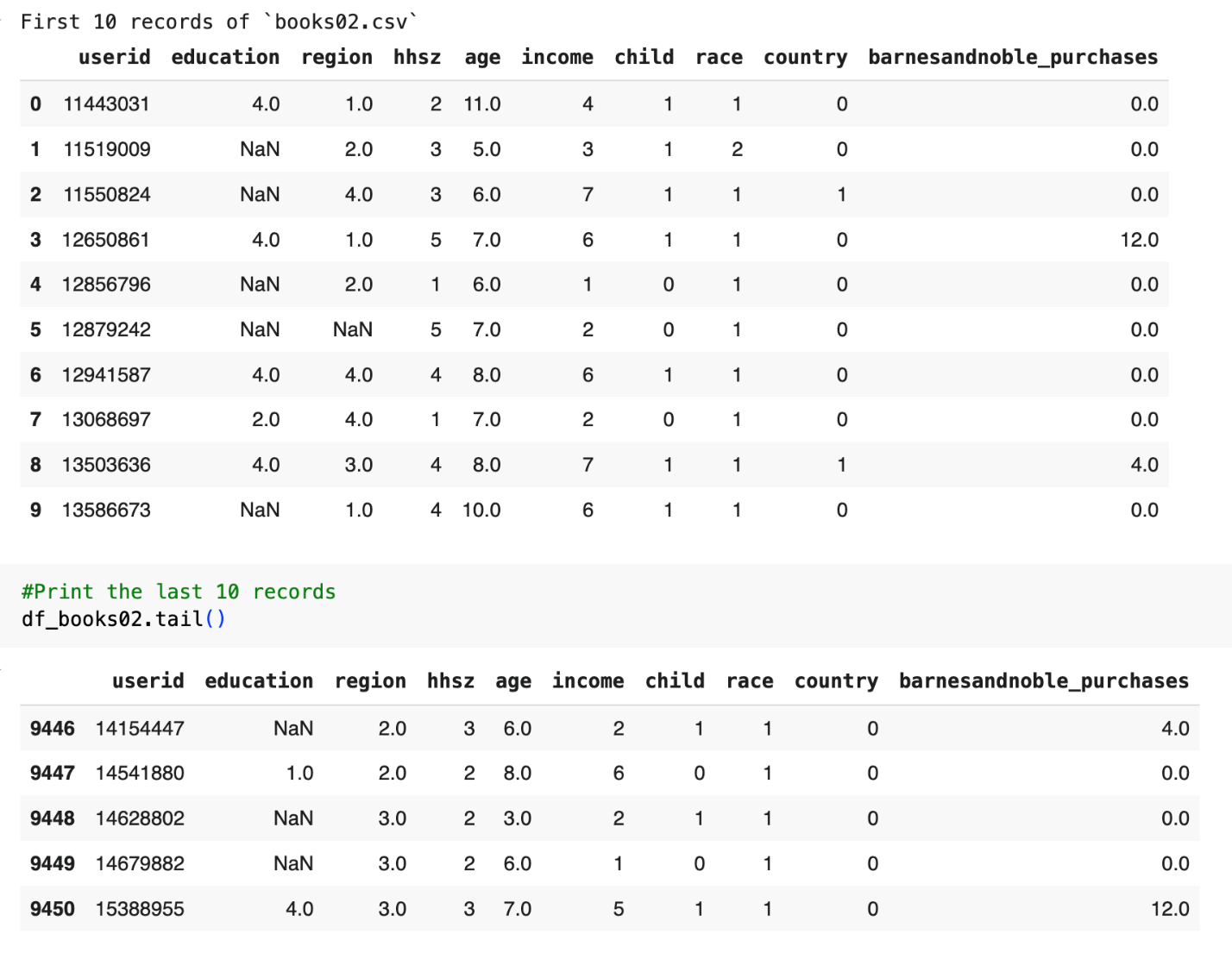


Figure : First and last 10 records of `books02.csv

## #2. The Poisson Model books01.csv

1. Report your code, the estimated parameters and the maximum value of the log-likelihood.

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimated Lambda (λ) | ≈ 0.7485 |
| Maximum Log-Likelihood | ≈ -18921.9184 |

1. Any other information you believe is relevant

## #3. The Poisson Model books02.csv

1. Report your code and confirm that the estimated parameters and the maximum value of the log-likelihood are identical to those obtained with the Poisson model developed using books01.csv.

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimated Lambda (λ) | ≈ 0.7485 |
| Maximum Log-Likelihood | ≈ -18921.9184 |

1. Predict the number of people with 0, ..., 20, 20+ purchases based on the Poisson model.

A table of numbers and symbols

Description automatically generated

Figure : Predicted purchase levels from 0 - 20, 20+ for `books02.csv` using Poisson Distribution

1. Explain how the predicted values are obtained using 2 purchases (show your calculations)

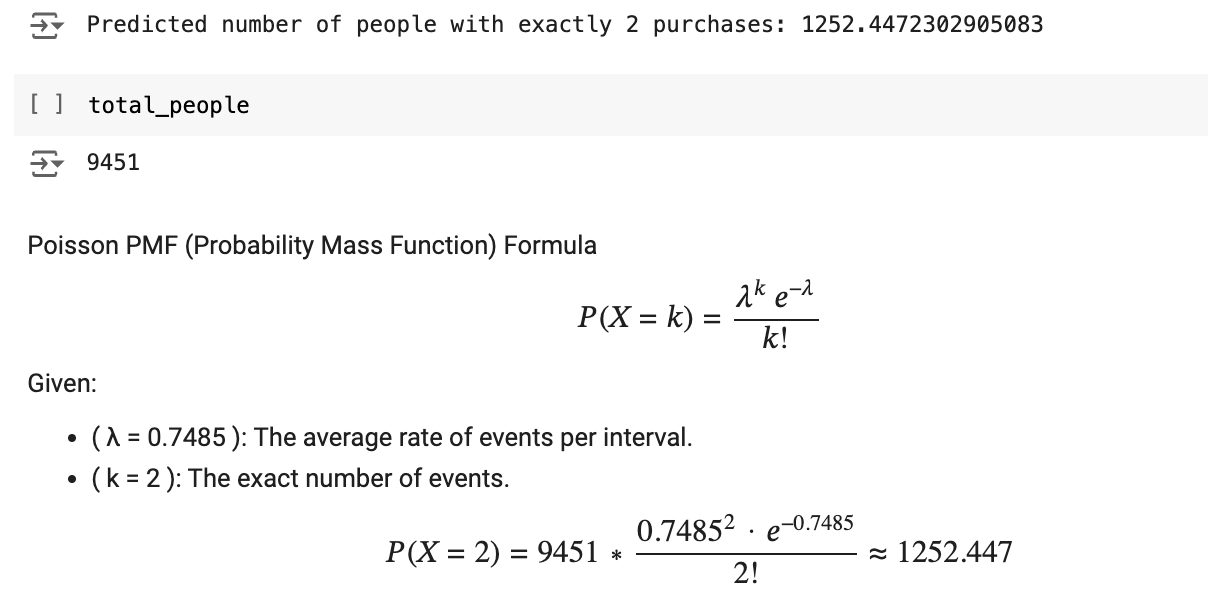


Figure :Calculations of predicted value with a case of 2 purchases using the Poisson model

1. Graph the original and predicted numbers of purchases.

A graph with numbers and text

Description automatically generated

Figure : Bar chart depicting the original and predicted number of exposures using Poison distribution for `books02.csv`

## #4. The NBD Model books01.csv

1. Report your code, the estimated parameters and the maximum value of the log-likelihood

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimate of n | ≈ 0.0813 |
| Estimate of α | ≈ 0.2192 |
| Maximum Log-Likelihood | ≈ -8603.4979 |

1. (and any other information you believe is relevant).

## #5. The NBD Model books02.csv

1. Report your code and confirm that the estimated parameters and the maximum value of the log-likelihood are identical to those obtained with the NBD model developed using books01.csv.

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimate of n | ≈ 0.0813 |
| Estimate of α | ≈ 0.2192 |
| Maximum Log-Likelihood | ≈ -8603.4982 |

1. Predict the number of people with 0, ..., 20, 20+ purchases based on the NBD model.

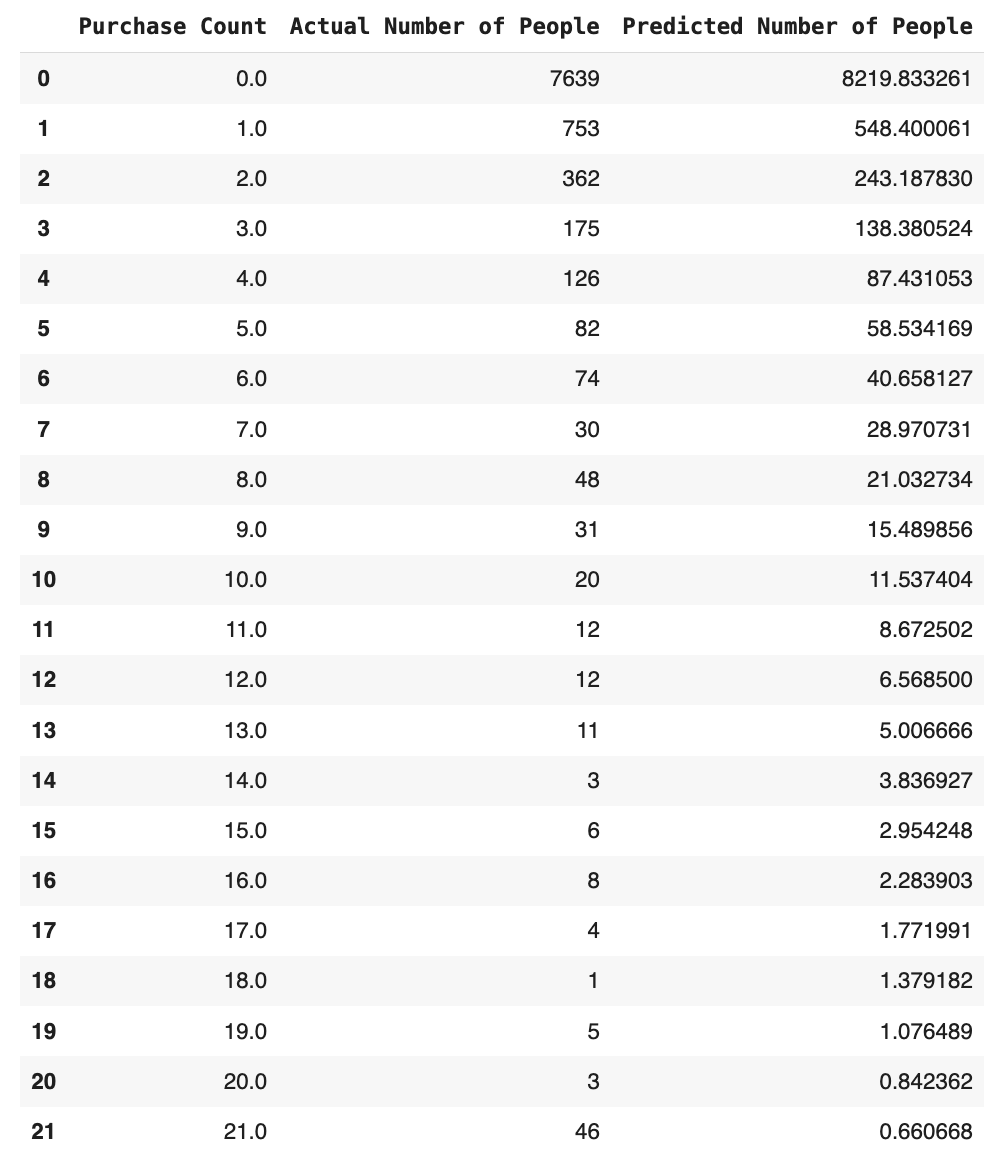


Figure : Predicted Purchase levels from 0 - 20, 20+ for `books02.csv` using NBD model

1. Explain how the predicted values are obtained using the case of 2 purchases (show your calculations).

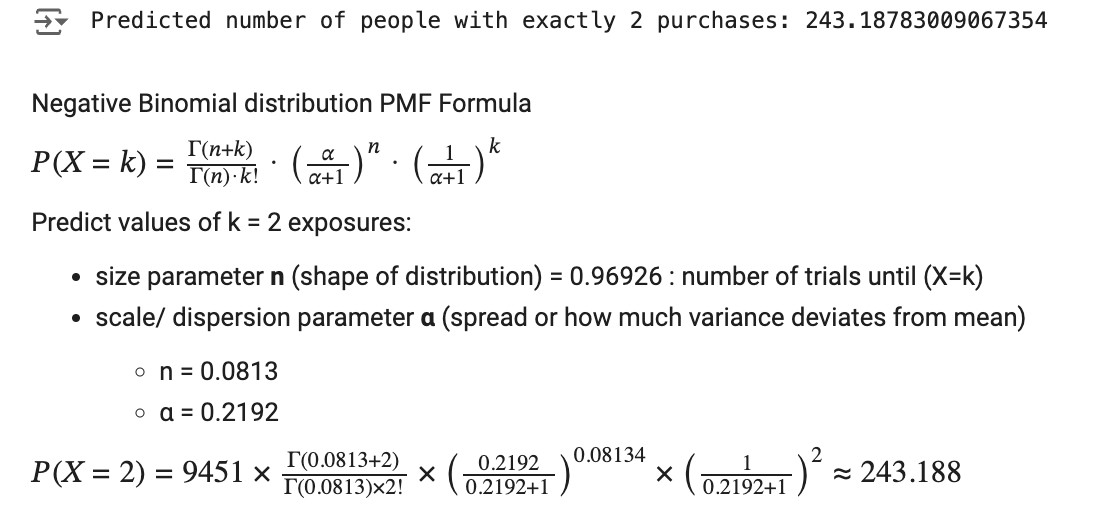


Figure : Calculations of predicted value with a case of 2 purchases using the NBD model

1. Graph the original and predicted number of purchases.

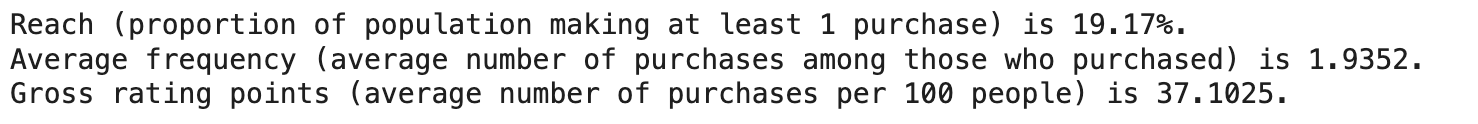
A graph with numbers and a number of people

Description automatically generated

Figure : Bar chart depicting the original and predicted number of exposures using NBD distribution for `books02.csv

## #6. Calculate the values of:

1. Reach – extract the actual number of people making at least 1 purchase
2. Average frequency – divide the sum product of the number of purchases and the number of people by reach
3. Gross ratings points (GRPs) - multiply reach in % by average frequency, where reach in % is reach divided by the total number of people and multiplied by 100.



## #7. Dealing with Missing Values

1. Identify all independent variables with missing values. How many values are missing in each?

A screenshot of a computer code

Description automatically generated

1. Drop any variable with many missing values (specify how you are defining “many”). If the number of missing values is small (again, specify how you are defining “small”), delete the rows involved.
2. For the remaining variables (if any), replace the missing values with the means of the corresponding variables. Explain the steps taken; report your code.

A screenshot of a computer

Description automatically generated

* Dropped ‘education’ column because the majority (73.16%) of the values are missing.
* Removed rows with missing values in ‘region’ and ‘age’ columns
  + ‘region’ has 11 records (0.12%) of NA values and is categorical data (values 1 to 4) and replacing using mean would be difficult to estimate and could affect accuracy.
  + There is only 1 record with the ‘age’ variable with NA value so just removed.

## #8. Incorporate all the available customer characteristics and estimate all relevant parameters for Poisson regression using MLE.

1. Report your code, the estimated parameters and the maximum value of the log-likelihood (and any other information you believe is relevant).

|  |  |
| --- | --- |
| MLE Output | Value |
| Estimated Lambda (λ0) | ≈ 0.9502 |
| *β1* | ≈ -0.1022 |
| *β2* | ≈ -0.0154 |
| *β3* | ≈ 0.0250 |
| *β4* | ≈ 0.0150 |
| *β5* | ≈ 0.0744 |
| *β6* | ≈ -0.2072 |
| *β7* | ≈ -0.1196 |
| Maximum Log-Likelihood | ≈ -18,819.0504 |

1. What are the managerial takeaways – which customer characteristics seem to be important?

* 'race’, ‘country’, and ‘region’ seem to be the most important characteristics based on the magnitude of the corresponding betas.

1. Predict the number of people with 0, ..., 20, 20+ purchases based on the Poisson regression.

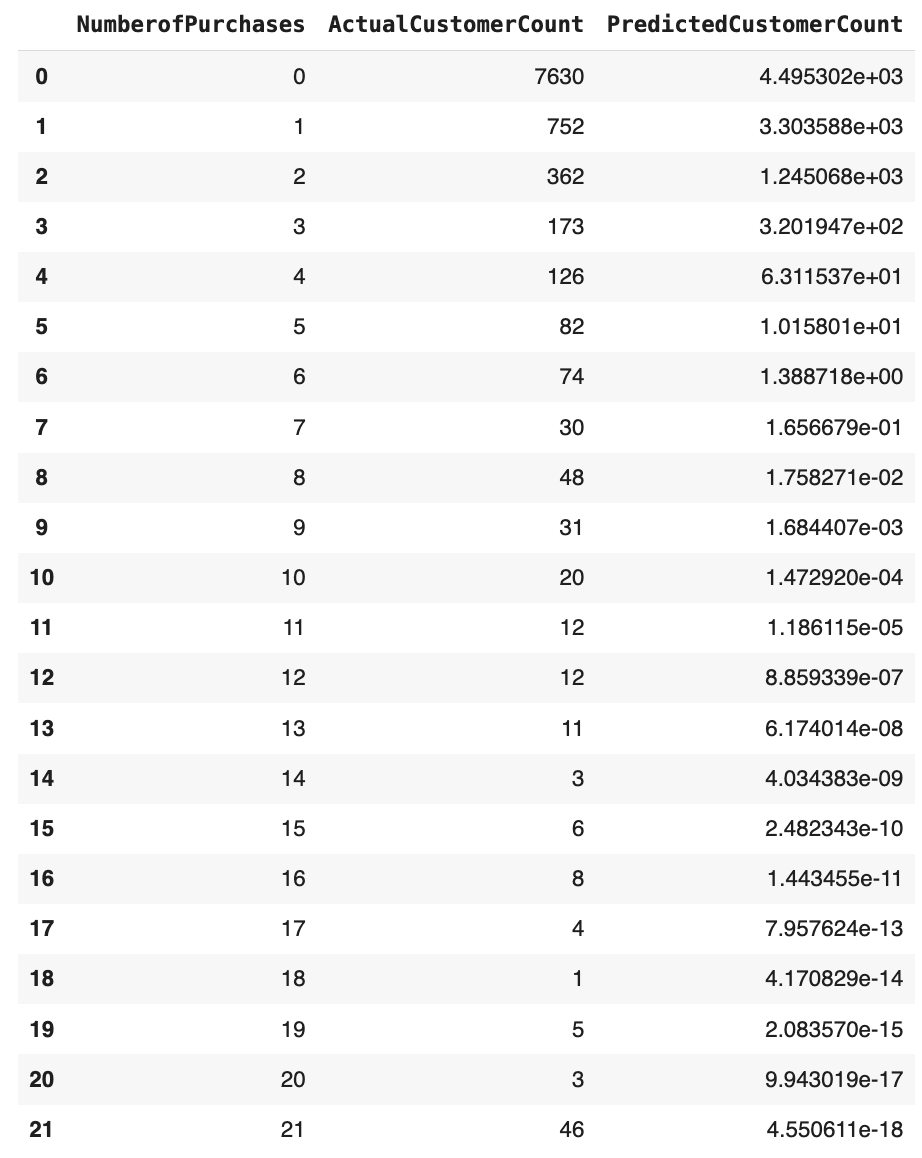


Figure : Predicted Purchase levels from 0 - 20, 20+ for `books02.csv` using Poisson Regression model

1. Explain how the predicted values are obtained using the case of 2 purchases (show your calculations).

A screenshot of a math problem

Description automatically generated

Figure : Calculations of predicted value with a case of 2 purchases using the Poisson Regression Model

1. Graph the original and predicted number of purchases.

A graph with numbers and text

Description automatically generated

Figure :Bar chart depicting the original and predicted number of exposures using Poisson Regression distribution for `books02.csv

## #9. Estimate all relevant parameters for NBD regression using MLE.

1. Report your code, the estimated parameters and the maximum value of the log-likelihood (and any other information you believe is relevant).

|  |  |
| --- | --- |
| MLE Output | Value |
| n | ≈ 0.0980 |
| α | ≈ 0.1080 |
| *β1* | ≈ -0.1033 |
| *β2* | ≈ -0.0083 |
| *β3* | ≈ 0.0284 |
| *β4* | ≈ 0.0176 |
| *β5* | ≈ 0.0580 |
| *β6* | ≈ -0.2088 |
| *β7* | ≈ -0.1026 |
| Maximum Log-Likelihood | ≈ -8358.7724 |

1. What are the managerial takeaways – which customer characteristics seem to be important?

* ‘race’, ‘country’, and ‘region’ seem to be the most important characteristics based on the magnitude of the corresponding betas – consistent with our conclusion using Poisson regression.

1. Predict the number of people with 0, ..., 20, 20+ purchases based on the NBD regression.

A table of numbers with numbers on it

Description automatically generated

Figure :Predicted Purchase levels from 0 - 20, 20+ for `books02.csv` using NBD Regression model

1. Explain how the predicted values are obtained using the case of 2 purchases (show your calculations).

A screenshot of a math test

Description automatically generated

Figure :Calculations of predicted value with a case of 2 purchases using the NBD Regression Model

1. Graph the original and predicted number of purchases.

A graph with numbers and text

Description automatically generated

Figure :Bar chart depicting the original and predicted number of exposures using NBD Regression distribution for `books02.csv

## #10. Evaluate all the models developed using the log-likelihood ratio, AIC, and BIC

1. What are your recommendations on which model to use based on each of these criteria? Are the recommendations consistent?

### Model Recommendation: NBD Regression

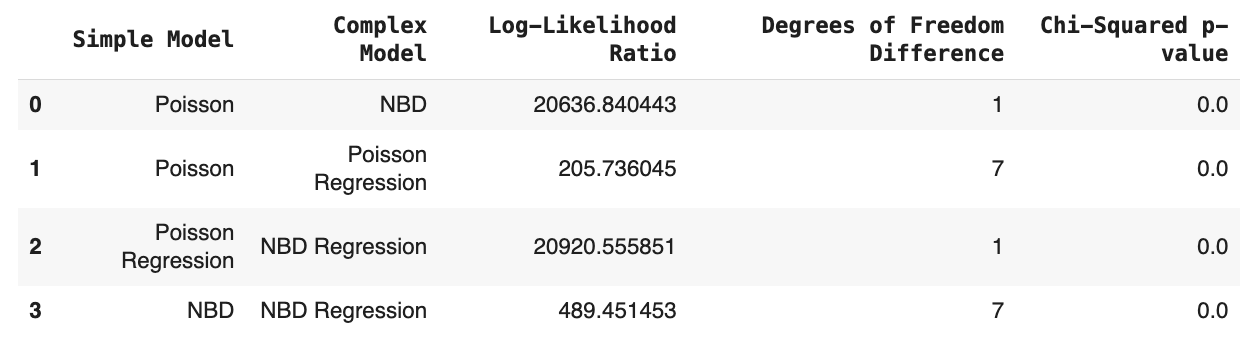


Figure : Comparative Model Fit Analysis using Log-Likelihood Ratios and Statistical Significance for Poisson, NBD and Regression Models

1. **Log-Likelihood Ratio**: The NBD model demonstrates a higher log-likelihood compared to Poisson, indicating a better fit for the data. Since the dataset is large, the log-likelihood ratio alone can lead to very low p-values due to the model’s complexity, potentially causing overfitting. This is why additional criteria like AIC and BIC are essential to validate model selection.
2. **AIC and BIC**: NBD Regression scores lowest on both AIC and BIC, suggesting it is the most appropriate model. AIC and BIC offer more reliable comparisons when log-likelihood values are large, as they add a penalty for model complexity and prevent overfitting. The similarity in AIC and BIC values in this case reinforces the robustness of the NBD Regression.
3. **Handling Overdispersion**: The NBD model is well-suited for data exhibiting overdispersion (variance greater than the mean), which is common in count data for customer purchases. Poisson models often assume equidispersion (mean equals variance), which may not hold for this data, leading to less accurate estimates. The NBD model’s dispersion parameter (α) allows it to flexibly accommodate overdispersion, making it a more reliable choice.

* **Log-Likelihood Ratio**:
  + The **NBD Regression** model has the highest log-likelihood (-8358.77), indicating a better fit compared to the other models. This suggests that NBD Regression can more accurately capture the underlying data patterns.
  + Poisson and Poisson Regression models show significantly lower log-likelihood values, indicating they may not capture the variability in purchase behavior as well as the NBD models.
  + **Recommendation**: Based on the log-likelihood criterion, NBD Regression is preferred.
* **AIC (Akaike Information Criterion)**:
  + The NBD Regression model has the lowest AIC (16735.54), followed closely by the basic NBD model (17211.00). This low AIC value implies that NBD Regression balances model fit and complexity better than the alternatives.
  + AIC penalizes model complexity, and despite having more parameters, NBD Regression’s low AIC suggests it is the most efficient model.
  + **Recommendation**: Based on AIC, NBD Regression is recommended for optimal performance.
* **BIC (Bayesian Information Criterion)**:
  + NBD Regression also has the lowest BIC (16799.93), reinforcing its selection as the best model. BIC, with a stronger penalty for complexity than AIC, supports NBD Regression over simpler models like Poisson, where BIC is considerably higher.
  + **Recommendation**: BIC results are consistent with AIC, favoring the NBD Regression model.

b) are the recommendations consistent?

* The recommendations across log-likelihood, AIC, and BIC are consistent, with all criteria favoring the **NBD Regression model**. This consistency strengthens the recommendation, as the NBD Regression model demonstrates the best balance of fit and complexity across multiple model selection metrics.

c) Explain why you are recommending the model you have selected.

* The **NBD Regression** model is preferred because:
* It provides **higher accuracy** in parameter estimates and better model fit, especially important for data exhibiting overdispersion (where variance exceeds the mean).
* The model's **dispersion parameter** allows it to handle variability more flexibly than Poisson models, which assume constant variance. This flexibility makes NBD Regression more suitable for real-world purchase data that may not follow a strict Poisson distribution.

d) Are there any significant differences? If so what exactly are the differences?

* **AIC and BIC Similarity**: The AIC and BIC values for each model are close to each other, indicating minimal differences in model selection using these criteria. This similarity likely stems from the relatively small logarithmic adjustment factor, ln(9451), which is much smaller than the log-likelihood values. Consequently, the complexity penalty has a minimal effect, and both criteria favor the same model.
* **Model Fit and Overdispersion**: The primary difference arises from the NBD models’ ability to accommodate overdispersion, resulting in better fit metrics (log-likelihood, AIC, BIC). Poisson models, which assume equidispersion, perform poorly in comparison, as they cannot capture the full variability in customer purchase behavior.

e) Discuss what you believe is causing the differences.

* **Overdispersion in Data**: The NBD models outperform the Poisson models due to their ability to handle overdispersion. The variability in purchase behavior is better modeled by the NBD’s dispersion parameter, which allows for more accurate predictions.
* **Complexity of Customer Behavior**: Customer purchase patterns are complex and influenced by multiple factors, making simpler models like Poisson inadequate. NBD Regression’s additional parameters allow it to capture this complexity, leading to a better fit.

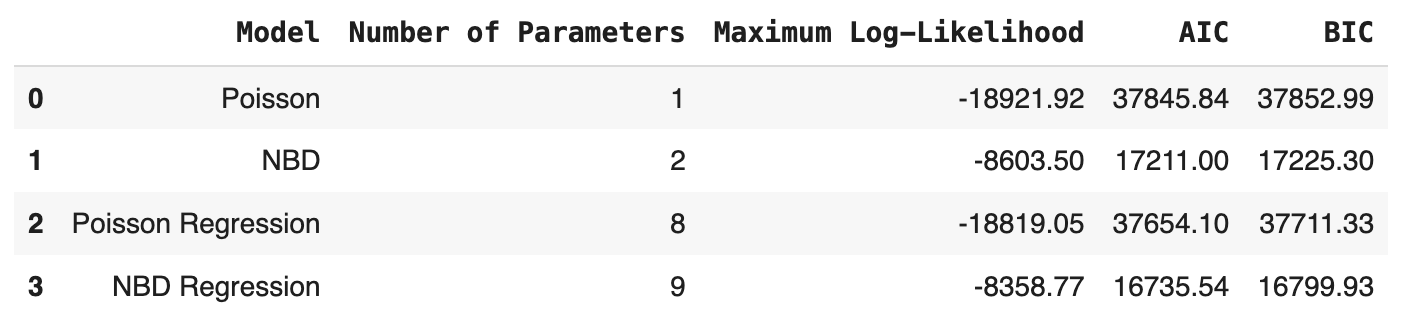


Figure : Model Selection Criteria: Comparison of Poisson, Negative Binomial (NBD), and Regression Models Based on Log-Likelihood, AIC, and BIC Values

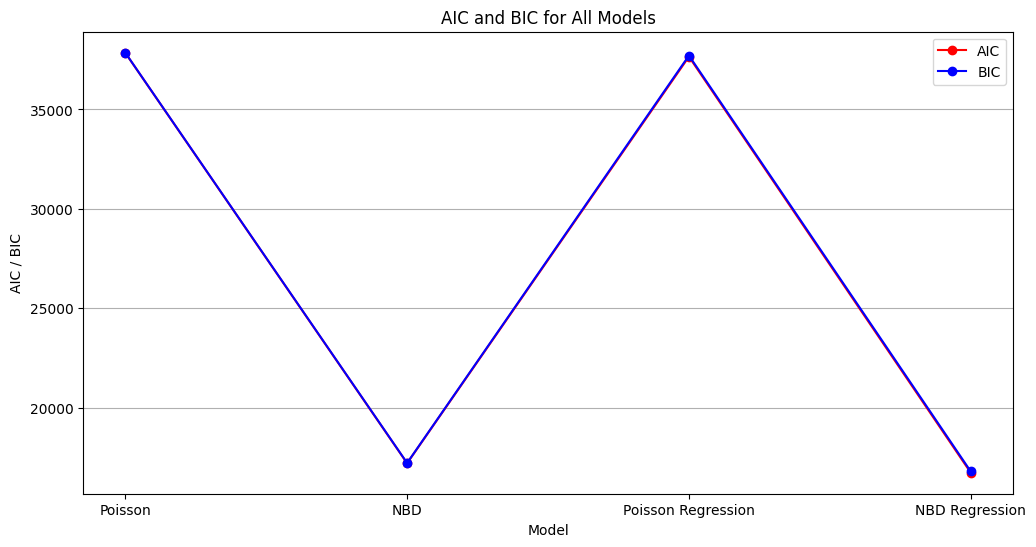


Figure : Line graph depicting the AIC and BIC of Poisson, NBD and Regression Models

Note: AIC & BIC are very similar in value which is why only BIC is visible in visualization

## # 11. Discussion Board

Briefly summarize what you learned from this project. This is an open-ended question, so please include anything you found worthwhile – relating to the modeling process, insights from the process and models, any managerial takeaways that were insightful to you, and so on.

A graph of a person with a red and blue line

Description automatically generated

Figure : Combination graph using bar chart and line graph to compare the predictions Poisson, NBD and Regression Models

* NBD models, including regression, generally outperform their Poisson counterparts due to their flexibility in handling overdispersion in the data.
* The dispersion parameter, α, provides added adaptability, allowing NBD models to better accommodate the variability in purchase behavior that Poisson models struggle with.
* Additionally, regression models capture a range of customer characteristics, which further improves predictive accuracy by identifying the factors that influence purchase frequency.

### **Customer Purchase Behavior Insights**

* Only **19.17%** of customers who purchased at least one book did so through barnesandnoble.com, indicating **low online engagement**.
* Online customers averaged **1.9 books** purchased, showing **low repeat frequency**.
* **GRPs** (reach x frequency) indicates **37 purchases per 100 people**. Increasing visibility and customer loyalty is recommended.

### **Important Predictors: ‘Race,’ ‘Country,’ and ‘Region’**

* In Poisson and NBD models, **‘race,’ ‘country,’ and ‘region’** are significant predictors.
* Negative coefficients in Poisson suggest **lower purchase frequency** for higher variable values; target regions 1-2, races 1-2, and country = 0.
* NBD model’s dispersion parameter offers **nuanced predictions**, maintaining the same trends.

### **Additional Variables to Consider: ‘Age’ and ‘Income’**

* **‘Age’ and ‘income’** are not significant now but could be useful in future models.
* Amazon leverages age diversity and income sensitivity (e.g., Kindle, affordable options). Barnes and Noble could similarly use these variables to refine strategies.

### **Recommendations for Future Analysis**

To streamline the modeling process, we could focus on the core predictors (‘race,’ ‘country,’ and ‘region’), given their importance in determining purchase behavior. However, it would also be beneficial to explore the potential effects of ‘age’ and ‘income’ in future iterations, especially as Barnes and Noble consider strategic pivots to expand its reach and competitiveness against platforms like Amazon.

In summary, Barnes and Noble can optimize customer engagement by:

1. **Improving visibility and reach** among potential customers who have yet to engage with barnesandnoble.com.
2. **Encouraging repeat purchases** from existing customers through loyalty programs or targeted promotions.
3. **Focusing on specific demographic segments** (regions 1-2, races 1-2, and country = 0) for tailored marketing efforts.
4. **Considering age and income** factors for a nuanced competitive strategy against Amazon.