MIS 6334.001

Advanced Business Analytics with SAS

PROJECT REPORT

Submitted by:

<u>GROUP 3</u>

Bharath Kumar Natarajan

Pooja Verma

Debao Sun

Part I: Examples Integrating SAS and Advanced Modelling

1.1 The NBD Model

Problem Statement: Consider the billboard exposures example from class. Write SAS code and conduct maximum likelihood estimation (MLE) for the NBD Model; estimate r and alpha. Report your code and the estimated values. When reporting MLE results, please provide the optimized LL value, all the estimated parameter values, and the corresponding p-values. Other statistics are optional | you need report them only if you want to comment on them in some way. In addition, please add comments to your SAS code to make your code easy to understand.

SAS Code:

```
* Creating a permanent library for creating and storing the dataset;
libname project 'C:\Users\bxn180005\Desktop\Project';

* Using PROC NLMIXED to estimate the best values of parameters r and alpha
to maximize the log likelihood of the NBD model as mentioned in the problem
statement;

proc nlmixed data = project.billboard;
	parm r = 1 alpha = 1;
	bounds r > 0.000001, alpha > 0.000001;
	prob = (gamma(r + exposures)/(gamma(r)*fact(exposures)) *
	((alpha/(alpha+1))**r) * ((1/(1+alpha))**exposures);
	ll = peoplecount * log(prob);
	model peoplecount ~ general(ll);
run;
```

OUTPUT:

Fit Statistics	
-2 Log Likelihood	1299.4
AIC (smaller is better)	1303.4
AICC (smaller is better)	1303.9
BIC (smaller is better)	1305.7

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.9693	0.1135	24	8.54	<.0001	0.7350	1.2035	-0.00007
alpha	0.2175	0.02978	24	7.30	<.0001	0.1561	0.2790	0.000071

In the “Fit Statistics” section of the above outputs, we see four types of statistics: -2 Log Likelihood, AIC, AICC and BIC. -2 Log Likelihood is nothing but the double of log likelihood of the model which we are trying to maximize using the shape and the scale parameters. The Log Likelihood values can be calculated as 1299.4/2 = 649.7

AIC stands for “Akaike Information Criterion” is the loss of information when the data is subject to the given model. AICC or “Akaike Information Criterion with Correlation” is AIC subject to the penalty of correlation between the parameters used to estimate the model when the number of parameters is very small. BIC stands for “Bayesian Information Criterion” is the loss of information when the data is subject to the given model with a penalty which increases as the number of parameters increases, as it is easier to fit a model by increasing the number of parameters. These three needs to be smaller for a model to be good. Also, we need to note that these statistics are relative i.e. they come into play when we need to choose among different models. Here, we tried to maximize the log likelihood and for the model which maximizes the log likelihood the AIC, AICC, and BIC are given in the table.

Also looking at the information in the parameters estimates section of the output for all the three models, we can see that the p-value of both the parameters for all the models are less than 0.01% which implies that the values are significant. We can also see that 95% confidence interval from the table for each dataset’s shape and scale parameters. The gradient for both the parameters is very small for all the datasets which means that the values which we obtained are the best values.

1.2 The POISSON REGRESSION Model

Problem Statement: Consider the khakichinos.com example from class. Write SAS code to estimate parameters (lamba0 and the vector beta) using MLE for the Poisson Regression Model. Report your code and the estimated values. What are some managerial takeaways?

Compared to the previous model, in Poisson regression model, we are counting into account the effect of the co-variates on the customer purchases. In Poisson Regression model, we are considering lamba0 to be same for all the customers. With this assumption we are implementing the Poisson Regression model using PROC NLMIXED:

SAS Code:

```
* Using PROC NLMIXED to estimate the best values of parameters lamba0 and
the vector beta to maximize the log likelihood of the Poisson Regression
model as mentioned in the problem statement;

proc nlmixed data=project.kc;
  parms lambda_0=1 b1=0 b2=0 b3=0 b4=0;
  lambda=lambda_0*exp(b1*income+b2*sex+b3*age+b4*HHSize);
  ll = total*log(lambda)-lambda-log(fact(total));
  model total ~ general(ll);
run;
```

OUTPUT:

Fit Statistics	
-2 Log Likelihood	12583
AIC (smaller is better)	12593
AICC (smaller is better)	12593
BIC (smaller is better)	12623

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
lambda_0	0.04387	0.01832	2728	2.39	0.0167	0.007942	0.07980	-0.82204
b1	0.09385	0.03507	2728	2.68	0.0075	0.02507	0.1626	-0.37966
b2	0.004229	0.04092	2728	0.10	0.9177	-0.07602	0.08448	-0.04265
b3	0.5883	0.05501	2728	10.69	<.0001	0.4804	0.6961	-0.12534
b4	-0.03592	0.01529	2728	-2.35	0.0189	-0.06590	-0.00593	-0.14349

In the "Fit Statistics" section of the above outputs, we see four types of statistics: -2 Log Likelihood, AIC, AICC and BIC. -2 Log Likelihood is nothing but the double of log likelihood of the model which we are trying to maximize using the shape and the scale parameters. The Log Likelihood values can be calculated as 12583/2 = 6291.5

AIC stands for "Akaike Information Criterion" is the loss of information when the data is subject to the given model. AICC or "Akaike Information Criterion with Correlation" is AIC subject to the penalty of correlation between the parameters used to estimate the model when the number of parameters is very small. BIC stands for "Bayesian Information Criterion" is the loss of information when the data is subject to the given model with a penalty which increases as the number of parameters increases, as it is easier to fit a model by increasing the number of parameters. These three needs to be smaller for a model to be good. Also, we need to note that these statistics are relative i.e. they come into play when we need to choose among different models. Here, we tried to maximize the log likelihood and for the model which maximizes the log likelihood the AIC, AICC, and BIC are given in the table.

In the parameter estimates part of the output for the Poisson regression model, we can see that the most important parameter b3 which is the age of the customer. This is evident from the p-value and the confidence limits. The estimate for b1 is 0.09385. This is the coefficient for ln(actual income), stored in the dataset as income. This means for a 10% increase in actual income there will be approx. 0.09385*ln(1.10) = 0.00895 increase in the visit to the website. The estimate for lambda-0 is 0.04387 which will be same for each customer. It is significant at 5% significant level as interpreted from its low p-value. The estimate of b2 is very small and doesn't have any significant impact which can be said from it p-value and significance levels.

Managerial Takeaways:

As per the Poisson regression model that has been implemented, following are the managerial insights on the given dataset:

- It is clear that when age increases, chances of visiting the website increases.
- Another factor can be the income. Most of the people who visit the websites are people with some good income
- The number of people visiting the websites decreases with increase in the number of households.
- There is no significant impact of gender visiting the websites.

1.3 The NDB REGRESSION Model

Problem Statement: Consider the khakichinos.com example again. Write SAS code to estimate parameters (r, alpha and the vector beta) using MLE for NBD Regression Model. Report your code and the estimated values. What are some managerial takeaways? Explain the difference in results between the NBD and the Poisson Regression Model.

In Poisson regression model, we considered that lambda0 to be same for everyone. In NBD regression model, let lambda0 vary across population according to a gamma distribution with parameters shape (r) and scale (alpha). The NBD Regression model has been implemented on the given dataset using PROC NLMIXED using the below code:-

SAS Code:

```
* Using PROC NLMIXED to estimate the best values of parameters r, alpha and
the vector beta to maximize the log likelihood of the NBD Regression model
as mentioned in the problem statement;

proc nlmixed data=project.kc;
  parms r=1 alpha=1 b1=0 b2=0 b3=0 b4=0;
  expBX=exp(b1*income+b2*sex+b3*age+b4*HHSize);
  ll = log(gamma(r+total))-log(gamma(r))-log(fact(total))
       +r*log(alpha/(alpha+expBX))+total*log(expBX/(alpha+expBX));
  model total ~ general(ll);
run;
```

OUTPUT:

Fit Statistics	
-2 Log Likelihood	5777.9
AIC (smaller is better)	5789.9
AICC (smaller is better)	5790.0
BIC (smaller is better)	5825.4

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.1388	0.007269	2728	19.09	<.0001	0.1245	0.1530	0.065001
alpha	8.1958	9.4524	2728	0.87	0.3860	-10.3389	26.7304	-0.00117
b1	0.07339	0.09723	2728	0.75	0.4505	-0.1173	0.2640	0.10240
b2	-0.00928	0.1212	2728	-0.08	0.9390	-0.2469	0.2284	0.005835
b3	0.9022	0.1676	2728	5.38	<.0001	0.5736	1.2307	0.032846
b4	-0.02432	0.04272	2728	-0.57	0.5692	-0.1081	0.05945	0.029598

In the "Fit Statistics" section of the above outputs, we see four types of statistics: -2 Log Likelihood, AIC, AICC and BIC. -2 Log Likelihood is nothing but the double of log likelihood of the model which we are trying to maximize using the shape and the scale parameters. The Log Likelihood values can be calculated as 5777.9/2 = 2888.95

AIC stands for "Akaike Information Criterion" is the loss of information when the data is subject to the given model. AICC or "Akaike Information Criterion with Correlation" is AIC subject to the penalty of correlation between the parameters used to estimate the model when the number of parameters is very small. BIC stands for "Bayesian Information Criterion" is the loss of information when the data is subject to the given model with a penalty which increases as the number of parameters increases, as it is easier to fit a model by increasing the number of parameters. These three needs to be smaller for a model to be good. Also, we need to note that these statistics are relative i.e. they come into play when we need to choose among different models. Here, we tried to maximize the log likelihood and for the model which maximizes the log likelihood the AIC, AICC, and BIC are given in the table.

On analysing the parameter estimates of NBD Regression model, the parameter b3 has a major impact on the customer visiting the website. Parameters b1 and b4 also has some impact on the customer Parameters b1 and b3 has positive impact while parameter b4 has negative impact.

Managerial Takeaways:

As per the Poisson regression model that has been implemented, following are the managerial insights on the given dataset:

- Factors like gender and number of people in the household has no significant impact on increasing the number of visits
- Factors like age play an important role in increasing the number of visits to the website. To increase the number of visits to the website, some campaigns can be run for the people towards the higher age.

- Other factors can be income. With increase in income of the customers, there are high chances to visit the website.

Explain the difference in results between the NBD and the Poisson Regression Model.

As per the implemented Poisson regression model and the NBD regression model, we saw that age plays an important role in increasing the number of visits to the website. We have got almost same characteristics that shows impact on the number of visits when comparing the results from both implemented models.

However, in the Poisson regression model, we have tried to find the parameters of a Poisson regression model which will help us to determine the total number of transactions carried out by an individual. For this we assumed that the mean frequency of transaction is different for everyone. Even though we are trying to model individual level heterogeneity using the attributes given in the dataset, we are assuming that lambda0 is the same for all individuals. There is still some unobserved heterogeneity which we are not able to capture as we are assuming that the value of lambda-0 is same for all individuals.

This heterogeneity is removed in the NBD regression model by assuming that the value of lambda0 is different for individuals, as the characteristics that determine the value of lambda0 are varied accordingly. The lambda-0 is determined by gamma distribution with shape(r) and scale(alpha) parameters.

Part II: Analysis of New Real Data

In this part of the project we are developing the the count models and applying them to the given dataset "books.txt".

Dataset Description: - The dataset records customer purchases at two competitors, Amazon.com and BARNES & NOBLE (B&N) in 2007. Some customer demographic variables such as education, household size (hhsz), income, and race are also included in the dataset. The dataset contains the details of the purchase transactions made by each of the customers. Each customer can have multiple transactions with B&N and amazon as per the dataset. The description of each of features in the given dataset are as follows: -

Sl No.	Feature Name	Description	Column Type
1	userid	ID of customer making the transaction	Numeric
2	education	Education background of the customer	Numeric
3	region	Region of the customer	Numeric
4	hhsz	household size	Numeric
5	age	age of the cutomer	Numeric
6	income	income of the customer	Numeric
7	child	whether the customer is an adult or not	Numeric
8	race	Race of the customer	Numeric
9	country	Customer belongs to home country or not	Numeric
10	domain	Whether the book is purchased from amazon or B&N	Text
11	date	Date of the transaction	Date
12	product	Product name of the product purchased by customer	Text
13	qty	Quantity of the product purchased by the customer	Numeric
14	price	Amount the customer needs to pay for the product	Numeric

Not all the features in the given dataset will be used to build the models. Certain columns are not required for implementing the analysis. For example:- the column "product" in the given dataset contains string values with special characters and also it is not relevant to the models that we develop. Also, when we tried to import the dataset including this column, all the rows in the dataset were not included in the final dataset. There was a loss of data around 6000 rows. To avoid this, we have removed this column from the dataset and then imported the file into SAS. We were able to import all the data successfully without any data loss.

Also, since we are going to have some managerial insights regarding the purchases made by the customers at Amazon.com and BARNES & NOBLE (B&N), we are going to use three datasets throughout this assignment. One dataset will be the given original dataset while the other two datasets will be for amazon.com specific purchases and B&N specific purchases. These two datasets were created by splitting the given original dataset as per the domain column in the dataset.

1. **Write a SAS program that reads the data in books.txt and generates a count dataset (similar to that used in the khaki chinos example). That is, for each customer count the number of books purchased from B&N in 2007, while keeping the demographic variables. Print the first 10 records of this dataset.**

SAS Code:

```
*Importing the given dataset books.txt using PROC IMPORT;

proc import datafile= 'C:\Users\bxn180005\Desktop\Project\books.txt'
      out= project.books
      dbms = dlm
      replace;
      delimiter= '09'x;
run;


* Creating two separate datasets for amazon.com and barnesandnoble.com
domains using domain variable. Also creating a new numeric column
domain_numeric based on the domain column;
data project.books;
      set project.books;
      if domain = 'amazon.com' then domain_numeric = 1; *Changing the
domain to numeric variable;
      else domain_numeric = 0;
run;

* Create two datasets one for amazon.com and one for barnesandnoble.com;

data project.amazonbooks project.bandnbooks;
      set project.books;
      if domain_numeric = 1 then output project.amazonbooks;
      else output project.bandnbooks;
run;


* sorting the data by userid for amazon, bandn and books datasets;
proc sort data = project.amazon;
      by userid;
run;
proc sort data = project.bandn;
      by userid;
run;
proc sort data = project.books;
      by userid;
run;


* creating count dataset for amazon purchasers;

data project.amazondata (drop= qty price);
      set project.amazon;
      by userid;
      retain transactioncount 0;
      retain total_qty 0;
```

```
      retain total_price 0;
      if first.userid then do;
            transactioncount = 1;
            total_qty = qty;
            total_price = price;
      end;
      else do;
            transactioncount = transactioncount + 1;
            total_qty = total_qty + qty;
            total_price = total_price + price;
      end;
            if last.userid then output project.amazondata;
run;

* creating count dataset for bandn purchasers;

data project.bandndata (drop= qty price);
      set project.bandn;
      by userid;
      retain transactioncount 0;
      retain total_qty 0;
      retain total_price 0;
      if first.userid then do;
            transactioncount = 1;
            total_qty = qty;
            total_price = price;
      end;
      else do;
            transactioncount = transactioncount + 1;
            total_qty = total_qty + qty;
            total_price = total_price + price;
      end;
            if last.userid then output project.bandndata;
run;

* creating count dataset for the original full books dataset;

data project.booksdata (drop= qty price);
      set project.books;
      by userid;
      retain transactioncount 0;
      retain total_qty 0;
      retain total_price 0;
      if first.userid then do;
            transactioncount = 1;
            total_qty = qty;
            total_price = price;
      end;
      else do;
            transactioncount = transactioncount + 1;
            total_qty = total_qty + qty;
            total_price = total_price + price;
      end;
            if last.userid then output project.booksdata;
run;

* Print the first 10 observations from bandndata dataset;

proc print data = project.bandndata(obs=10);
run;
```

OUTPUT:

The following output is the purchases of the customers of B&N:-

The SAS System

Obs	userid	education	region	hhsz	age	income	child	race	country	domain	date	transactioncount	total_qty	total_price
1	6365661	5	1	2	11	7	0	1	0	barnesandn	20071218	1	1	17.97
2	6396922	2	2	2	8	4	0	1	0	barnesandn	20070223	1	1	15.96
3	8999933	4	3	5	10	3	1	1	0	barnesandn	20070608	1	1	49.95
4	9573834	99	4	2	10	5	1	1	0	barnesandn	20071217	2	2	5.81
5	9576277	99	1	3	8	7	1	1	0	barnesandn	20070228	5	5	81.73
6	9581009	99	2	2	7	5	1	1	0	barnesandn	20070106	1	1	2.00
7	9595310	4	2	2	8	2	1	1	0	barnesandn	20071217	4	6	92.68
8	9611445	2	4	2	11	6	1	1	1	barnesandn	20070506	2	2	31.34
9	9663372	4	4	3	9	7	1	1	0	barnesandn	20070927	9	28	393.91
10	9752844	3	4	2	7	3	1	1	0	barnesandn	20071118	2	2	28.37

2. **Build an NBD model, ignoring the demographic variables. Report your results. (Hint: you will need to create a data set similar to that used in the billboard exposures example.)**

As given in the question, we will be developing an NBD model with the given dataset. Before the model implementation, we need to create a count dataset similar to the billboard exposures example which we saw in the class. For this we are using PROC FREQ to generate the count dataset. As mentioned earlier, We will be implementing the model for three the datasets: amazon.com, B&N and full original dataset. The model has been implemented using PROC NLMIXED. Below is the code that has been used to create the count dataset and the model implementation: -

SAS Code:

```
*Using PROC FREQ to create a dataset that has number of books bought and
its frequency to implement the count model;

*For bandn dataset;

proc freq data = project.bandndata;
      tables total_qty /
      out = project.bandncount (drop = percent rename=(Count =
peoplecount));
run;

* For amazon dataset;

proc freq data = project.amazondata;
      tables total_qty /
      out = project.amazoncount (drop = percent rename=(COUNT =
peoplecount));
run;

* For original complete books dataset;

proc freq data = project.booksdata;
```

```
	tables total_qty /
	out = project.bookscount (drop = percent rename=(COUNT =
peoplecount));
run;

* Removing two rows from the generated count dataset as PROC NLMIXED was
not able to run successfully with these values and these values are the
purchase details of just two customers. So deleting two customer details
when compared to the entire dataset does not have any significant impact on
analysis;

data project.amazoncount;
set project.amazoncount;
if total_qty=197 then delete;
if total_qty=317 then delete;
run;
data project.bookscount;
set project.bookscount;
if total_qty=197 then delete;
if total_qty=317 then delete;
run;

*Using PROC NLMIXED to implement the NBD model as mentioned in the
question;

* For bandn dataset;

proc nlmixed data = project.bandncount;
	parm r = 1 alpha = 1;
	bounds r > 0.000001, alpha > 0.000001;
	prob = (gamma(r + total_qty)/(gamma(r)*fact(total_qty)))*
	((alpha/(alpha+1))**r) * ((1/(1+alpha))**total_qty);
	ll = peoplecount * log(prob);
	model peoplecount ~ general(ll);
run;

* For amazon dataset;

proc nlmixed data = project.amazoncount;
	parm r = 1 alpha = 1;
	bounds r > 0.000001, alpha > 0.000001;
	prob = (gamma(r + total_qty)/(gamma(r)*fact(total_qty)))*
	((alpha/(alpha+1))**r) * ((1/(1+alpha))**total_qty);
	ll = peoplecount * log(prob);
	model peoplecount ~ general(ll);
run;

*For original complete books dataset;

proc nlmixed data = project.bookscount;
	parm r = 1 alpha = 1;
	bounds r > 0.000001, alpha > 0.000001;
	prob = (gamma(r + total_qty)/(gamma(r)*fact(total_qty)))*
	((alpha/(alpha+1))**r) * ((1/(1+alpha))**total_qty);
	ll = peoplecount * log(prob);
	model peoplecount ~ general(ll);
run;
```

OUTPUT:

The following screenshots demonstrates the Fit Statistics and Parameter Estimates of the output of NBD model using PROC NLMIXED for Amazon.com, B&N and given full dataset.

B&N dataset:

Fit Statistics	
-2 Log Likelihood	8966.3
AIC (smaller is better)	8970.3
AICC (smaller is better)	8970.6
BIC (smaller is better)	8974.0

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.2024	0.04687	45	25.65	<.0001	1.1080	1.2968	0.000071
alpha	0.3080	0.01418	45	21.72	<.0001	0.2794	0.3366	-0.00024

Amazon Dataset:

Fit Statistics	
-2 Log Likelihood	42413
AIC (smaller is better)	42417
AICC (smaller is better)	42417
BIC (smaller is better)	42422

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.2117	0.02184	77	55.48	<.0001	1.1682	1.2552	-0.00608
alpha	0.2709	0.005768	77	46.97	<.0001	0.2594	0.2824	0.004405

Original full Books Dataset:

Fit Statistics	
-2 Log Likelihood	49614
AIC (smaller is better)	49618
AICC (smaller is better)	49618
BIC (smaller is better)	49623

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.1847	0.01964	84	60.31	<.0001	1.1456	1.2237	0.021856
alpha	0.2566	0.005050	84	50.82	<.0001	0.2466	0.2667	-0.03814

In the "Fit Statistics" section of the above outputs, we see four types of statistics: -2 Log Likelihood, AIC, AICC and BIC. -2 Log Likelihood is nothing but the double of log likelihood of the model which we are trying to maximize using the shape and the scale parameters. The Log Likelihood values can be calculated as follows: -

For B&N Dataset: 8966.3/2 = 4483.15

For Amazon.com Dataset: 42413/2 = 21206.5

For original full Dataset: 49614/2 = 24807

AIC stands for "Akaike Information Criterion" is the loss of information when the data is subject to the given model. AICC or "Akaike Information Criterion with Correlation" is AIC subject to the penalty of correlation between the parameters used to estimate the model when the number of parameters is very small. BIC stands for "Bayesian Information Criterion" is the loss of information when the data is subject to the given model with a penalty which increases as the number of parameters increases, as it is easier to fit a model by increasing the number of parameters. These three needs to be smaller for a model to be good. Also, we need to note that these statistics are relative i.e. they come into play when we need to choose among different models. Here, we tried to maximize the log likelihood and for the model which maximizes the log likelihood the AIC, AICC, and BIC are given in the table.

Also looking at the information in the parameters estimates section of the output for all the three models, we can see that the p-value of both the parameters for all the models are less than 0.01% which implies that the values are significant. We can also see that 95% confidence interval from the table for each dataset's shape and scale parameters. The gradient for both the parameters is very small for all the datasets which means that the values which we obtained are the best values.

3. Calculate the values of (i) Reach, (ii) Average Frequency, and (iii) Gross Ratings Points (GRPs) based on the NBD Model. Show your work.

For B&N dataset:

Probability of zero transactions in time t given r and alpha and the expected value of number of transactions in given time t.

$$P(X(t) = 0|r,\, \alpha) = \left(\frac{\alpha}{\alpha + t}\right)^r$$

$$= \left(\frac{0.3080}{0.3080+1}\right)^{1.2024} = 0.1757$$

the expected value of number of transactions in given time t.

$$E(X(t)) = \frac{rt}{\alpha} = \frac{1.2024\times 1}{0.3080} = 3.9038$$

Reach:

Reach = 100 * (1 – P (X(t) = 0)) = 100 * (1 – 0.1757) = 82.43%

Average Frequency:

Average Frequency = $\frac{E(X(t))}{(1-\mathrm{P}(X(t)=0))}$ = 3.9038/ (1-0.1757) = 4.7358

GRP:

GRP = 100 * E(X(t)) = 100 * 3.9038 = 390.38

For Amazon.com dataset:

Probability of zero transactions in time t given r and alpha and the expected value of number of transactions in given time t.

$$P(X(t) = 0|r,\, \alpha) = \left(\frac{\alpha}{\alpha + t}\right)^r$$

$$= \left(\frac{0.2709}{0.2709+1}\right)^{1.2177} = 0.1522$$

the expected value of number of transactions in given time t.

$$E(X(t)) = \frac{rt}{\alpha} = \frac{1.2177\times 1}{0.2709} = 4.495$$

Reach:

Reach = 100 * (1 – P (X(t) = 0)) = 100 * (1 – 0.1522) = 84.78%

Average Frequency:

Average Frequency = $\frac{E(X(t))}{(1-\mathrm{P}(X(t)=0))}$ = 4.495/ (1-0.1522) = 5.3

GRP:

GRP = 100 * E(X(t)) = 100 * 4.495 = 449.5

For original full dataset:

Probability of zero transactions in time t given r and alpha and the expected value of number of transactions in given time t.

$$P(X(t) = 0|r,\ \alpha) = \left(\frac{\alpha}{\alpha + t}\right)^r$$

$$= \left(\frac{0.2566}{0.2566+1}\right)^{1.1847} = 0.1523$$

the expected value of number of transactions in given time t.

$$E(X(t)) = \frac{rt}{\alpha} = \frac{1.1847\times 1}{0.2566} = 4.6169$$

Reach:

Reach = 100 * (1 – P (X(t) = 0)) = 100 * (1 – 0.1523) = 84.77%

Average Frequency:

Average Frequency = $\frac{E(X(t))}{(1-\mathrm{P}(X(t)=0))}$ = 4.6169/ (1-0.1523) = 5.4464

GRP:

GRP = 100 * E(X(t)) = 100 * 4.6169 = 461.69

4 **Build a Poisson regression model using the demographic information (customer characteristics) provided. Report your results. What are the managerial takeaways | which customer characteristics seem to be important?**

As mentioned in this question we will be developing a Poisson Regression model on the given dataset. This time we will be considering some of the demographic information as well. As like before, we will be implementing this model for all the three datasets. We won't be considering the variable "Education" as it has some high frequency value of 99 that doesn't make any sense. Also, we found some missing values in the region column for few of the rows which was removed from the dataset. PROC NLMIXED was used to implement Poisson Regression model on the three datasets. The following code was used for implementing the same: -

SAS Code:

```
* Implementing Poisson Regression Model;

*Inorder to implement Poisson regression Model, the variable "Education"
needs to be removed since this variable has a high frequency value "99"
which dosent make any sense when compared to other values. Therefore we are
not considering this variable for the Poisson regression Model;

*Also we find some missing values for the variable "region" and only few
rows has this missing values. So we are deleting these records;

* For amazon dataset;
data project.amazondata;
      set project.amazondata;
      if region = '*' then delete;
run;

* For bandn dataset;
data project.bandndata;
      set project.bandndata;
      if region = '*' then delete;
run;

*For original full dataset;
data project.booksdata;
      set project.booksdata;
      if region = '*' then delete;
run;

* Model Implementation;

* Removing two rows from the generated amazondata and booksdata datasets as
PROC NLMIXED was not able to run successfully with these values and these
values are the purchase details of just two customers. So deleting two
customer details when compared to the entire dataset does not have any
significant impact on analysis;

data project.amazondata;
set project.amazondata;
if total_qty=197 then delete;
if total_qty=317 then delete;
run;
data project.booksdata;
```

```
set project.booksdata;
if total_qty=197 then delete;
if total_qty=317 then delete;
run;
* On bandn dataset;

proc nlmixed data=project.bandndata;
  parms lambda_0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;

lambda=lambda_0*exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*
country);
  ll = total_qty*log(lambda)-lambda-log(fact(total_qty));
  model total_qty ~ general(ll);
run;

* On amazon dataset;

proc nlmixed data=project.amazondata;
  parms lambda_0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;

lambda=lambda_0*exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*
country);
  ll = total_qty*log(lambda)-lambda-log(fact(total_qty));
  model total_qty ~ general(ll);
run;

*For original full dataset;

proc nlmixed data=project.booksdata;
  parms lambda_0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;

lambda=lambda_0*exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*
country+b8*domain_numeric);
  ll = total_qty*log(lambda)-lambda-log(fact(total_qty));
  model total_qty ~ general(ll);
run;
```

OUTPUT:

B&N dataset:

Fit Statistics	
-2 Log Likelihood	14398
AIC (smaller is better)	14414
AICC (smaller is better)	14414
BIC (smaller is better)	14458

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
lambda_0	3.9590	0.2825	1810	14.01	<.0001	3.4049	4.5132	0.001451
b1	-0.00562	0.01090	1810	-0.52	0.6062	-0.02699	0.01575	0.009318
b2	0.009806	0.01124	1810	0.87	0.3832	-0.01224	0.03186	0.012827
b3	0.005251	0.003260	1810	1.61	0.1074	-0.00114	0.01164	0.040013
b4	0.01794	0.006364	1810	2.82	0.0049	0.005455	0.03042	0.000939
b5	0.02471	0.03240	1810	0.76	0.4457	-0.03883	0.08825	0.003288
b6	-0.1322	0.04360	1810	-3.03	0.0025	-0.2177	-0.04667	0.008186
b7	-0.2049	0.03383	1810	-6.06	<.0001	-0.2712	-0.1385	0.002135

Amazon Dataset:

Fit Statistics	
-2 Log Likelihood	69935
AIC (smaller is better)	69951
AICC (smaller is better)	69951
BIC (smaller is better)	70007

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
lambda_0	3.3624	0.1026	8162	32.76	<.0001	3.1612	3.5636	0.000468
b1	0.008268	0.004914	8162	1.68	0.0925	-0.00136	0.01790	0.030596
b2	-0.00171	0.004884	8162	-0.35	0.7262	-0.01128	0.007864	-0.01514
b3	0.03457	0.002189	8162	15.79	<.0001	0.03028	0.03886	0.068483
b4	0.001077	0.002751	8162	0.39	0.6954	-0.00432	0.006470	0.044147
b5	-0.00748	0.01393	8162	-0.54	0.5913	-0.03478	0.01982	-0.01271
b6	0.02381	0.01441	8162	1.65	0.0986	-0.00444	0.05206	0.014168
b7	-0.00835	0.01432	8162	-0.58	0.5600	-0.03642	0.01973	-0.01516

Original full Books Dataset:

Fit Statistics	
-2 Log Likelihood	83635
AIC (smaller is better)	83653
AICC (smaller is better)	83654
BIC (smaller is better)	83718

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
lambda_0	4.1249	0.1180	9438	34.95	<.0001	3.8936	4.3562	0.083872
b1	0.006259	0.004491	9438	1.39	0.1634	-0.00254	0.01506	0.97301
b2	-0.00136	0.004465	9438	-0.31	0.7599	-0.01012	0.007387	1.21405
b3	0.02048	0.001388	9438	14.75	<.0001	0.01776	0.02320	2.50419
b4	0.006274	0.002523	9438	2.49	0.0129	0.001329	0.01122	1.50824
b5	-0.00440	0.01278	9438	-0.34	0.7303	-0.02945	0.02064	0.25741
b6	0.007851	0.01384	9438	0.57	0.5706	-0.01928	0.03499	0.35076
b7	-0.04896	0.01319	9438	-3.71	0.0002	-0.07481	-0.02311	0.076545
b8	-0.08597	0.01266	9438	-6.79	<.0001	-0.1108	-0.06116	0.27264

In the “Fit Statistics” section of the above outputs, we see four types of statistics: -2 Log Likelihood, AIC, AICC and BIC. -2 Log Likelihood is nothing but the double of log likelihood of the model which we are trying to maximize using the shape and the scale parameters. The Log Likelihood values can be calculated as follows: -

For B&N Dataset: 14398/2 = 7199

For Amazon.com Dataset: 69935/2 = 34967.5

For original full Dataset: 83635/2 = 41817.5

AIC stands for “Akaike Information Criterion” is the loss of information when the data is subject to the given model. AICC or “Akaike Information Criterion with Correlation” is AIC subject to the penalty of correlation between the parameters used to estimate the model when the number of parameters is very small. BIC stands for “Bayesian Information Criterion” is the loss of information when the data is subject to the given model with a penalty which increases as the number of parameters increases, as it is easier to fit a model by increasing the number of parameters. These three needs to be smaller for a model to be good. Also, we need to note that these statistics are relative i.e. they come into play when we need to choose among different models. Here, we tried to maximize the log likelihood and for the model which maximizes the log likelihood the AIC, AICC, and BIC are given in the table.

For **B&N dataset**, using parameter estimates in the output we can understand that the parameters lambda0, b6 and b7 has some significant impact on the number of books purchased by a customer while parameters b2, b4 and b5 has some impact on the customer purchase. Parameters b6 and b7 has negative impact on the customer purchase while b2, b4 and b5 has positive impact.

For **Amazon.com dataset**, using parameter estimates in the output we can understand that the parameters lambda0, b3, b6 and b7 has some significant impact on the number of books purchased by a customer. Parameters b7 has negative impact on the customer purchase while b3, b6 has positive impact.

For **original complete dataset**, using parameter estimates in the output we can understand that the parameters lambda0, b3, b7 and b8 has some significant impact on the number of books purchased by a customer. Parameters b7 and b8 has negative impact on the customer purchase while b3 has positive impact.

Managerial Takeaways:

As per the Poisson regression model that has been implemented, following are the managerial insights on the given dataset:

- It is clear that the domain of the transaction plays an important role in the customer purchase. Customer tends to buy more from amazon.com
- Another important factor is whether the customer belongs to the home country or not. This also plays an important role in the number of books purchased by a customer
- Age of the customer also plays an important role in the customer purchase.

Therefore as per the implemented Poisson Regression model, the important characteristics of customer purchase are age of the customer, country of the customer and the domain of the transaction.

5. Next, we start the setup for developing an NBD regression model. What is the formula for the log-likelihood expression, LL?

Let λ_0 vary across population according to a Gamma distribution with parameters r and α,

$$P(Y_i = y) = \frac{\Gamma(r+y)}{\Gamma(r)y!}\left(\frac{\alpha}{\alpha + e^{\beta x}}\right)^r \left(\frac{e^{\beta x}}{\alpha + e^{\beta x}}\right)^y$$

The log-likelihood expression (LL) for NBD regression model for the given dataset is given as below: -

```
expBX=exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*cou
ntry+b8*domain_numeric);
ll = log(gamma(r+total_qty))-log(gamma(r))-log(fact(total_qty))
     +r*log(alpha/(alpha+expBX))+total_qty*log(expBX/(alpha+expBX));
```

Where,
B1-b8 = coefficients of respective variables.
ll = log likelihood
r = shape parameter of the gamma distribution
alpha = scale parameter of the gamma distribution
transactioncount = frequency of number of transactions.

6. Build a NBD regression model using the demographic information provided. Report your results. What are the managerial takeaways | which customer characteristics seem to be important?

As mentioned in the question, we are implementing NBD regression model on the given dataset. As usual, we are implementation NBD regression on all the three datasets. Let λ_0 vary across population according to a Gamma distribution with parameters r and α. The code for the implementation of the NBD regression model using PROC NLMIXED is give below:-

<u>SAS Code:</u>

```
*Implementing NBD Regression Model;

* On bandn dataset;

proc nlmixed data=project.bandndata;
  parms r=1 alpha=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;

expBX=exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*country);
  ll = log(gamma(r+total_qty))-log(gamma(r))-log(fact(total_qty))
       +r*log(alpha/(alpha+expBX))+total_qty*log(expBX/(alpha+expBX));
  model total_qty ~ general(ll);
```

```
run;

* On amazon dataset;

proc nlmixed data=project.amazondata;
  parms r=1 alpha=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;

expBX=exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*country);
  ll = log(gamma(r+total_qty))-log(gamma(r))-log(fact(total_qty))
       +r*log(alpha/(alpha+expBX))+total_qty*log(expBX/(alpha+expBX));
  model total_qty ~ general(ll);
run;

* On original full dataset;

proc nlmixed data=project.booksdata;
  parms r=1 alpha=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;

expBX=exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*country+b8
*domain_numeric);
  ll = log(gamma(r+total_qty))-log(gamma(r))-log(fact(total_qty))
       +r*log(alpha/(alpha+expBX))+total_qty*log(expBX/(alpha+expBX));
  model total_qty ~ general(ll);
run;
```

OUTPUT:

B&N dataset:

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	8941.7
AIC (smaller is better)	8959.7
AICC (smaller is better)	8959.8
BIC (smaller is better)	9009.2

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.2135	0.04746	1810	25.57	<.0001	1.1204	1.3066	0.070233
alpha	0.3130	0.04677	1810	6.69	<.0001	0.2212	0.4047	-0.20946
b1	-0.00833	0.02236	1810	-0.37	0.7096	-0.05218	0.03552	0.15594
b2	0.01086	0.02344	1810	0.46	0.6431	-0.03510	0.05682	0.20747
b3	0.007514	0.008743	1810	0.86	0.3902	-0.00963	0.02466	0.26687
b4	0.01749	0.01309	1810	1.34	0.1815	-0.00817	0.04316	0.33922
b5	0.02211	0.06502	1810	0.34	0.7338	-0.1054	0.1496	0.047048
b6	-0.1219	0.07794	1810	-1.56	0.1180	-0.2748	0.03096	0.074318
b7	-0.1982	0.06601	1810	-3.00	0.0027	-0.3277	-0.06878	0.006756

<u>Amazon Dataset:</u>

Fit Statistics	
-2 Log Likelihood	42310
AIC (smaller is better)	42328
AICC (smaller is better)	42328
BIC (smaller is better)	42391

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.2213	0.02208	8162	55.32	<.0001	1.1780	1.2646	-0.00056
alpha	0.3607	0.02453	8162	14.70	<.0001	0.3126	0.4088	-0.00328
b1	0.007971	0.01047	8162	0.76	0.4466	-0.01256	0.02850	0.002077
b2	0.000309	0.01068	8162	0.03	0.9769	-0.02062	0.02123	0.003386
b3	0.03420	0.004714	8162	7.25	<.0001	0.02496	0.04344	0.004512
b4	0.000473	0.005918	8162	0.08	0.9362	-0.01113	0.01207	0.006708
b5	-0.01102	0.03008	8162	-0.37	0.7141	-0.07000	0.04795	0.001112
b6	0.01962	0.03163	8162	0.62	0.5351	-0.04238	0.08161	0.000678
b7	-0.00888	0.03089	8162	-0.29	0.7739	-0.06943	0.05168	0.001027

<u>Original full Books Dataset:</u>

Fit Statistics	
-2 Log Likelihood	49496
AIC (smaller is better)	49516
AICC (smaller is better)	49516
BIC (smaller is better)	49587

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.1933	0.01984	9438	60.15	<.0001	1.1544	1.2322	0.13965
alpha	0.3034	0.02041	9438	14.87	<.0001	0.2634	0.3435	-0.02798
b1	0.005956	0.009782	9438	0.61	0.5426	-0.01322	0.02513	0.17937
b2	-0.00020	0.009996	9438	-0.02	0.9836	-0.01980	0.01939	0.35199
b3	0.02949	0.004403	9438	6.70	<.0001	0.02086	0.03812	-0.17781
b4	0.004375	0.005553	9438	0.79	0.4308	-0.00651	0.01526	-0.26272
b5	-0.00743	0.02814	9438	-0.26	0.7917	-0.06258	0.04772	0.060692
b6	0.004112	0.03007	9438	0.14	0.8912	-0.05483	0.06306	-0.08785
b7	-0.04773	0.02875	9438	-1.66	0.0969	-0.1041	0.008619	-0.03829
b8	-0.09142	0.02858	9438	-3.20	0.0014	-0.1474	-0.03540	0.077680

In the “Fit Statistics” section of the above outputs, we see four types of statistics: -2 Log Likelihood, AIC, AICC and BIC. -2 Log Likelihood is nothing but the double of log likelihood of the model which we are trying to maximize using the shape and the scale parameters. The Log Likelihood values can be calculated as follows: -

For B&N Dataset: 8941.7/2 = 4470.85

For Amazon.com Dataset: 42310/2 = 21155

For original full Dataset: 49496/2 = 24748

AIC stands for “Akaike Information Criterion” is the loss of information when the data is subject to the given model. AICC or “Akaike Information Criterion with Correlation” is AIC subject to the penalty of correlation between the parameters used to estimate the model when the number of parameters is very small. BIC stands for “Bayesian Information Criterion” is the loss of information when the data is subject to the given model with a penalty which increases as the number of parameters increases, as it is easier to fit a model by increasing the number of parameters. These three needs to be smaller for a model to be good. Also, we need to note that these statistics are relative i.e. they come into play when we need to choose among different models. Here, we tried to maximize the log likelihood and for the model which maximizes the log likelihood the AIC, AICC, and BIC are given in the table.

For **B&N dataset**, using parameter estimates in the output we can understand that the parameters r, alpha, b2, b4, b5, b6 and b7 has some significant impact on the number of books purchased by a customer. Parameters b6 and b7 has negative impact while b2, b4 and b5 has positive impact.

For **Amazon.com dataset**, using parameter estimates in the output we can understand that the parameters r, alpha, b3, b5 and b6 has some significant impact on the number of books purchased by a customer. Parameters b5 has negative impact on the customer purchase while b3, b6 has positive impact.

For **original complete dataset**, using parameter estimates in the output we can understand that the parameters r, alpha, b3, b7 and b8 has some significant impact on the number of books purchased by a customer. Parameters b7 and b8 has negative impact on the customer purchase while b3 has positive impact.

Managerial Takeaways:

As per the NBD regression model that has been implemented, following are the managerial insights on the given dataset:

- It is clear that the domain of the transaction plays an important role in the customer purchase. Customer tends to buy more from amazon.com
- Another important factor is whether the customer belongs to the home country or not. This also plays an important role in the number of books purchased by a customer
- Age of the customer also plays an important role in the customer purchase.

Therefore, as per the implemented NBD Regression model, the important characteristics of customer purchase are age of the customer, country of the customer and the domain of the transaction.

7. Are there any significant differences between the results from the Poisson and NBD regressions? If so, what exactly is the difference? Discuss what you believe about the cause(s) of the difference.

On comparing the results from the Poisson and NBD regression models, we can see that the customer characteristics that have a significant impact on the customer purchase are almost the same. However, in the Poisson regression model, we have tried to find the parameters of a Poisson regression model which will help us to determine the total number of transactions carried out by an individual. For this we assumed that the mean frequency of transaction is different for each individual. Even though we are trying to model individual level heterogeneity using the attributes given in the dataset, we are assuming that lambda0 is the same for all individuals. There is still some unobserved heterogeneity which we are not able to capture as we are assuming that the value of lambda-0 is same for all individuals.

This heterogeneity is removed in the NBD regression model by assuming that the value of lambda0 is different for individuals, as the characteristics that determine the value of lambda0 are varied accordingly. The lambda-0 is determined by gamma distribution with shape(r) and scale(alpha) parameters.

8. Briefly summarize what you learned from this project. This is an open-ended question, so please include anything you found worthwhile | relating to the modelling tool (SAS), the modelling process, insights from the modelling, any managerial takeaways that were insightful to you, and so on.

This project has helped us to learn a lot about SAS tool. The key takeaways are as follows: -

- Hands on experience using SAS to implement various models on a real time dataset.
- How to approach the problem statement through model implementation using SAS.
- Various functions and PROC's available in SAS.
- To build Count Data in SAS from a real time dataset in order to implement the count models.
- Built NBD model, Poisson regression mode and NBD regression model on a real time data to determine the purchasing behaviour of the customers.
- To predict the customer behaviour with respect to various factors which are understood through the count models that was developed.
- Found the significant variables that affect the purchasing behaviour of customers in B&N and amazoncom using the NBD, Poisson and NBD regression models
- To use PROC NLMIXED to maximize the log likelihood of a particular model using various parameters of the distribution.