Part – I

1. **NBD Model**

Answer: The Code is as follows:-

libname project "E:\Study\Adv BI with SAS\Project";

**PROC** **NLMIXED** DATA=project.billboard;

RETAIN l;

PARMS alpha=**0.5**, r = **0.5**; /\*Initiating Parameters to be optimized\*/

IF exposures=**0** THEN

DO;

l = (alpha/(alpha+**1**))\*\*r; /\*probability of zero exposures\*/

END;

else do;

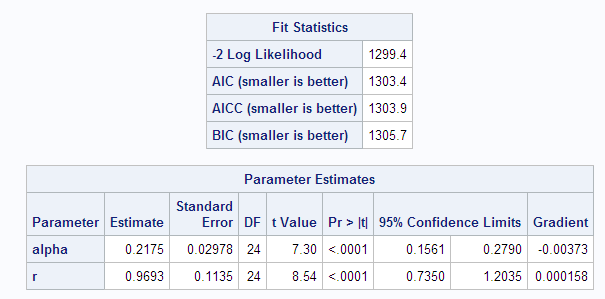
l = l\*((r+exposures-**1**)/(exposures\*(alpha+**1**))); /\*probability of exposures >= 1\*/

end;

ll = log(l)\*peoplecount; /\*calculating values for log likelihood fucntion\*/

MODEL peoplecount ~ general(ll); /\*optimizing values of alpha and r to get minimum log liklihood or maximum -2log likelihood\*/

**RUN**;



1. **The Poisson Regression Model**

Answer: SAS code –

**proc** **nlmixed** data=project.kc;

/\* m stands for lambda \*/

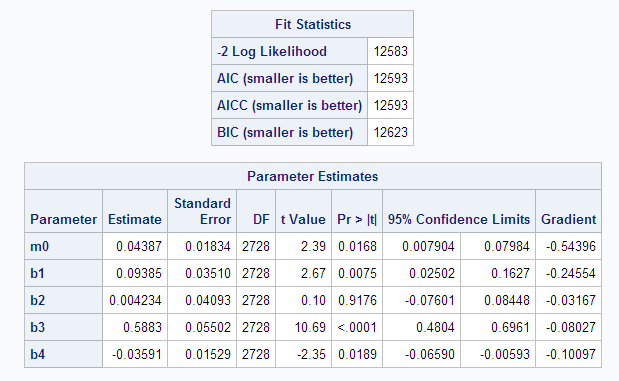
parms m0=**1** b1=**0** b2=**0** b3=**0** b4=**0**;/\*Initiating Parameters to be optimized\*/

m=m0\*exp(b1\*income+b2\*sex+b3\*age+b4\*HHSize); /\*inidividual's mean(Poisson disttribution) related to observable characteristics\*/

ll = total\*log(m)-m-log(fact(total)); /\*calculating values for log likelihood fucntion\*/

model total ~ general(ll); /\*optimizing PARAMS to get minimum log liklihood or maximum -2log likelihood\*/

**run**;



Managerial takeaways at significance level of 0.05 are:-

1. Individuals with higher income are more likely to visit
2. Sex of individual is not significant enough to explain frequency of visits
3. Older individuals are more likely visit frequently
4. As number of people in a household increases they are less likely to visit the online store
5. **The NBD regression model**

Answer: SAS code –

**proc** **nlmixed** data=project.kc;

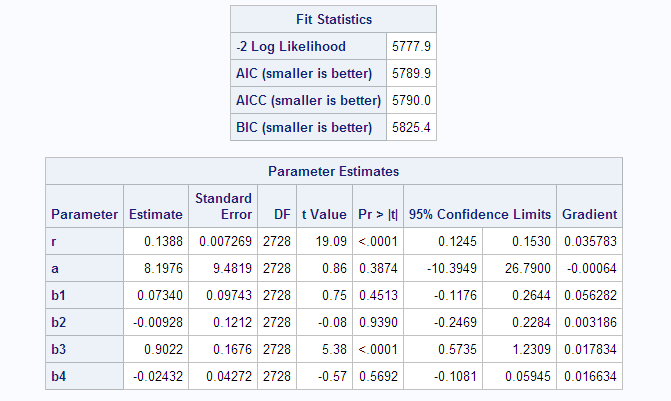
parms r=**1** a=**1** b1=**0** b2=**0** b3=**0** b4=**0**; /\*Initiating Parameters to be optimized\*/

expBX=exp(b1\*income+b2\*sex+b3\*age+b4\*HHSize);

ll = log(gamma(r+total))-log(gamma(r))-log(fact(total)) +r\*log(a/(a+expBX)) +total\*log(expBX/(a+expBX)); /\*calculating values for log likelihood fucntion\*/

model total ~ general(ll);

**run**;



Managerial takeaways:-

1. Age is the only factor that is significant enough to determine if an induvial is likely to visit the online store
2. Older individuals are more likely to visit the store

Differences in Poisson Reg Model and NBD Reg Model

1. NBD Model is a better fit since it has smaller values for -2 Log Likelihood value, AIC, AICC, BIC
2. Income and HHsize are not significant indicators as per NBD Reg Model at significance of 0.05

PART – II

Step 1 - Write a SAS program that reads the data in books.txt and generates a count dataset. 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.

Answer: - SAS Code:

**proc** **import** datafile="E:\Study\Adv BI with SAS\Project\books.txt" out=project.books dbms=tab replace;

getnames=yes;

**run**;

**Data** Books;

set project.books;

Amazon = **0**;

BN = **0**;

if domain = "amazon.com" then amazon = qty;/\*creating flag for Amazon.com\*/

else BN = qty;/\*creating flag for B&N\*/

if region = '\*' then region = **5**;/\*Replacing \* in region with numeric value 5\*/

regions = region\***1**; /\*creating numeric column for region\*/

**run**;

**proc** **means** data = books; /\*exploring the dataset\*/

**run**;

/\*SQL statment to count the number of books purchased from B&N website while retaing all demographic variables for each user\*/

**PROC** **SQL**;

create table freqcount as

SELECT userid,avg(education) as education,avg(regions) as region,avg(hhsz) as hhsze,avg(age) as age,avg(income) as income,avg(child) as child,avg(race) as race,avg(country) as country,sum(BN) as BNbookspurch

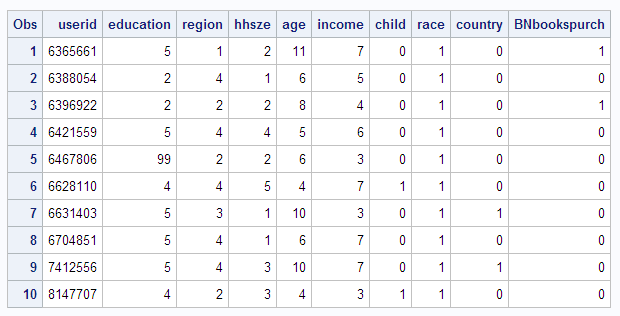
FROM books

GROUP BY userid;

**QUIT**;

**proc** **print** data = freqcount(OBS = **10**);

**run**;



Step 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.)

Answer:- SAS Code:

/\*Creating dataset similar to Billboard exposures for fitting NBD model\*/

**Proc** **SQL**;

create table purchasecount as

select BNbookspurch, count(userid)as usercount

from freqcount

group by BNbookspurch;

**quit**;

/\*BUilding an NBD model for number of books purchased at B&N\*/

**PROC** **NLMIXED** DATA=purchasecount;

RETAIN l;

PARMS alpha=**0.5**, r = **0.5**; /\*Initiating Parameters to be optimized\*/

IF BNbookspurch=**0** THEN

DO;

l = (alpha/(alpha+**1**))\*\*r; /\*probability of zero purchases\*/

END;

else do;

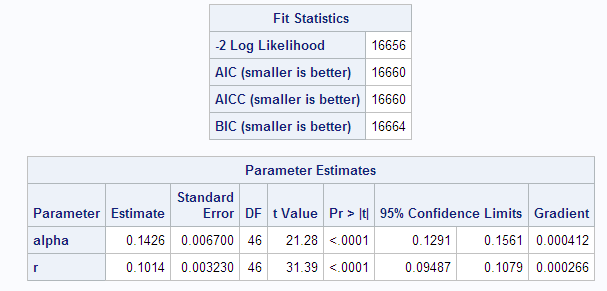
l = l\*((r+BNbookspurch-**1**)/(BNbookspurch\*(alpha+**1**))); /\*probability of every purchase >= 1\*/

end;

ll = log(l)\*usercount; /\*calculating values for log likelihood fucntion\*/

MODEL usercount ~ general(ll); /\*optimizing values of alpha and r to get minimum log liklihood or maximum -2log likelihood\*/

**RUN**;



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

Answer: - We can assume t=1 as the time frame of data provided is 1 year and we are not required to predict the measure for beyond 2007.

|  |  |  |
| --- | --- | --- |
| Measures | Formula | Value |
| Alpha | NBD Model result | 0.142588 |
| R | NBD Model result | 0.10137 |
| P(X(t) = 0|r,alpha) | (alpha/(alpha+t))^r | 0.809762 |
| E(X(t)) | rt/alpha | 0.71108 |
| Reach: | 100\*(1-P(X(t) = 0|r,a)) | 19.0238% |
| Average Frequency | E(X(1))/(1-P(X(t) = 0|r,a)) | 3.737839 |
| Gross Rating Points | 100\*E(X(t)) | 71.10799 |

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

Answer: - SAS Code:

\* The Poisson Regression Model;

\* for Books;

**proc** **nlmixed** data=freqcount;

/\* m stands for lambda \*/

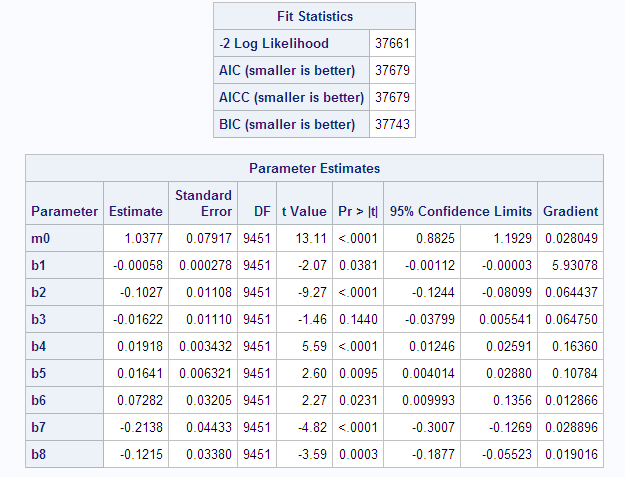
parms m0=**0.5** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0**;/\*Initiating Parameters to be optimized\*/

m=m0\*exp(b1\*education + b2\*region + b3\*hhsze + b4\*age + b5\*income + b6\*child + b7\*race + b8\*country); /\*inidividual's mean(Poisson disttribution) related to observable characteristics\*/

ll = bnbookspurch\*log(m)-m-log(fact(bnbookspurch)); /\*calculating values for log likelihood function\*/

model bnbookspurch ~ general(ll); /\*optimizing PARAMS to get minimum log liklihood or maximum -2log likelihood\*/

**run**;



Managerial Takeaways (significance level 0.05):-

1. The significant variable are education, region, age, income, child, race and country.
2. Education, region, age, income and child have positive impact on probability of an individual purchasing from B&N
3. Race and country have a negative impact on probability of purchase from B&N
4. Since the given variables are either categorical variables or flags we will need to look at data dictionary to derive more accurate interpretation of the estimates from the regression

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

Answer: - LL = P(Y=y1 |α,r) + P(Y=y2 |α,r) + P(Y=y3 |α,r) + P(Y=y4 |α,r) ……………. P(Y=yn |α,r) where yn = books purchased from B&N by nth customer. P(Y=yi |α,r) = (Γ(r+ yi )/ Γ(r) yi!)(α/ α+eβx )(eβx/ α+eβx)

Step 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?

Answer:- SAS code:

\* The NBD Regression Model

\* for Books;

**proc** **nlmixed** data=freqcount;

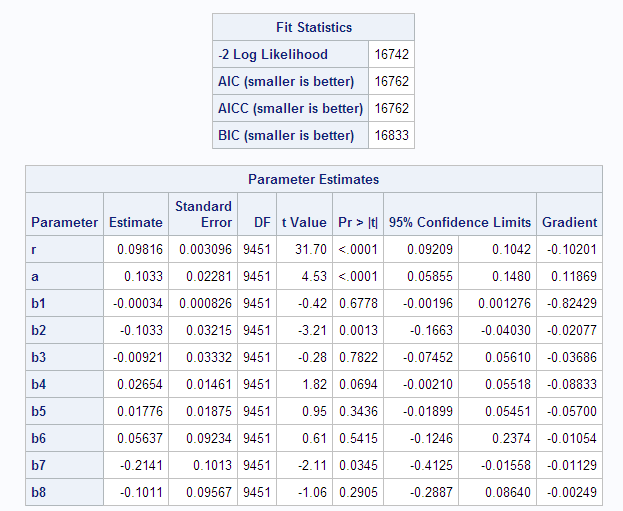
parms r=**1** a=**0** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0**; /\*Initiating Parameters to be optimized\*/

expBX=exp(b1\*education + b2\*region + b3\*hhsze + b4\*age + b5\*income + b6\*child + b7\*race + b8\*country);

ll = log(gamma(r+bnbookspurch))-log(gamma(r))-log(fact(bnbookspurch))+r\*log(a/(a+expBX))+bnbookspurch\*log(expBX/(a+expBX)); /\*calculating values for log likelihood fucntion\*/

model bnbookspurch ~ general(ll);

**run**;



Managerial Takeaways (significance level of 0.05):

1. The significant variables are Region and Race.
2. Both variables have a negative impact on probability of an individual buying from B&N
3. If we relax the significance level to 0.1 then age also becomes a significant indicator of customer behavior and has appositive impact purchases from B&N

Step 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.

Answer: - Yes, there are significant differences between the Poisson and NBD regression model.

1. NBD regression model has much smaller values for Fit statistics indicating it is more accurate than Poisson regression model.
2. NBD model indicates there are only two significant variables, Region and Race. Poisson regression model indicates there are seven significant variables.

Reason for difference in results:-

1. Poisson regression depends on independent variables to generate heterogeneity for each individual i.e.  **λi** (lambda-i) but they do not do a good job in explaining the complete heterogeneity between each individual i.e. observed and unobserved heterogeneity.
2. NBD regression model accounts both observed and unobserved heterogeneity as **λ0** varies for each individual and depend on r and α. Hence, it is able to capture more heterogeneity in the data thus indicating from the results that unobserved heterogeneity has bigger impact on likelihood of an individual buying from B&N than observed heterogeneity i.e. the independent variables.

Step 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 modeling tool (SAS), the modeling process, insights from the modeling, any managerial takeaways that were insightful to you, and so on.

Answer:-

1. Basics of using SAS for data manipulation, regression, clustering, simulation, macros and IML
2. Enhanced understanding of Bayesian Classifier, model enhancements and probability distributions
3. Building models with respect to time, count and choice. Modification of models to account for both observed and unobserved heterogeneity