MIS 6334

ADVANCED BUSINESS ANALYTICS WITH SAS

PROJECT

GROUP 09

Bhardwaj, Tanya

Danasekaran, Vignesh

Kumar, Akshara

Pulichintala, Siri

Rajasekar, Sadhave Subathra

PART I - Examples Integrating SAS and Advanced Modelling

1. **The NBD Model:**

Consider the billboard exposures example from class. Write SAS code and conduct maximum likelihood estimation (MLE) for the NBD Model; estimate r and 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

Solution:

The Billboard dataset is loaded into SAS. The code used for Maximum Likelihood estimation using NBD model is given below.

```
LIBNAME PR 'C:\EDUCATION\COURSES\ABI WITH SAS\HOMEWORK\PROJECT';
PROC NLMIXED DATA=PR.BILLBOARD;
PARMS R=0.2 A=0.2;  *INITIALISATION OF PARAMETERS;
PART1 = LOG(GAMMA(R+EXPOSURES))-LOG(GAMMA(R)) - LOG(FACT(EXPOSURES));
*SPLITTING THE LL FORMULA TO THREE PARTS FOR EASE OF CALCUALTION;
PART2 = R*(LOG(A)-LOG(A+1));
PART3 = EXPOSURES*(LOG(1)-LOG(A+1));
LL = PEOPLECOUNT*(PART1+PART2+PART3); *LIKELIHOOD ESTIMATION;
MODEL PEOPLECOUNT ~ GENERAL(LL);
RUN;
```

Initial Parameters

r	a	Negative Log Likelihood
0.2	0.2	823.762116

Iteration History

Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	10	693.2372	130.525	278.622	-12734.7
2	13	691.8062	1.431005	409.086	-40.4918
3	17	656.6982	35.10796	52.3901	-100.062
4	21	651.2667	5.431531	73.6163	-5.30303
5	25	649.7112	1.555485	13.5674	-1.77809
6	28	649.6897	0.021502	1.12963	-0.03270
7	31	649.6889	0.000822	0.40478	-0.00120
8	34	649.6888	0.000025	0.005887	-0.00004
9	37	649.6888	9.242E-9	0.000015	-1.83E-8

NOTE: GCONV convergence criterion satisfied.

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	1.716E-6
a	0.2175	0.02978	24	7.30	<.0001	0.1561	0.2790	-0.00002

The above output is obtained. The optimized negative LL value is 649.6888. The estimated value of the parameter are alpha = 0.2175 and the corresponding p value is < 0.0001 and r = 0.9693 and its corresponding p value is < 0.0001.

2. The Poisson Regression Model:

Consider the khakichinos.com example from class. Write SAS code to estimate parameters (lambda0 and the vector beta) using MLE for the Poisson Regression Model. Report your code and the estimated values. What are some managerial takeaways?

Solution:

The Khakichinos dataset is loaded into SAS. The code used using Poisson Regression model is given below.

```
PROC NLMIXED DATA=PR.KC;
  PARMS M0=1 B1=0 B2=0 B3=0 B4=0; *INTIALISATION;
  M=M0*EXP(B1*INCOME+B2*SEX+B3*AGE+B4*HHSIZE); *LAMBDA CALCULATION;
  LL = TOTAL*LOG(M)-M-LOG(FACT(TOTAL)); *LIKELIHOOD CALCULATION;
  MODEL TOTAL ~ GENERAL(LL);
RUN;
```

The output for the code is shown below:

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
m0	0.04387	0.01834	2728	2.39	0.0168	0.007904	0.07984	-0.54396
b1	0.09385	0.03510	2728	2.67	0.0075	0.02502	0.1627	-0.24554
b2	0.004234	0.04093	2728	0.10	0.9176	-0.07601	0.08448	-0.03167
b3	0.5883	0.05502	2728	10.69	<.0001	0.4804	0.6961	-0.08027
b4	-0.03591	0.01529	2728	-2.35	0.0189	-0.06590	-0.00593	-0.10097

The maximum value of negative LL is 6291.4967. The value of Lambda0 = 0.04387. The coefficient values or the beta values are specified below:
Coefficient of income = 0.09385

Coefficient of sex = 0.004234
Coefficient of age = 0.5883
Coefficient of HHsize = -0.03591

Managerial Takeaways:
One of the key point to note in the above output is, at a 5% significance level, that the coefficient of sex is not significant. This can be due to reasons like correlation among the independent variables. Hence, this variable can be removed and the model can be run again to get better estimates.

From the above output we can conclude that, the variables income and age has a positive effect on the output variable. As the income and age of the customers are more, they have more chances of visiting the Khakichinos website compared to young people and people with low income. On the other hand, if the household size of a customer is more, then they have a less chance of visiting the website compared to the customers whose household size is less.

3. The NBD Regression Model:

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.

Solution:

The Khakichinos dataset is loaded into SAS. The code used for NBD regression model is specified below:

```
PROC NLMIXED DATA=PR.KC;
  PARMS R=1 A=1 B1=0 B2=0 B3=0 B4=0;    *INITIALISATION OF PARAMETERS;
  EXPBX=EXP(B1*INCOME+B2*SEX+B3*AGE+B4*HHSIZE); *EXPONENTIAL BETA VALUES;
  LL = LOG(GAMMA(R+TOTAL))-LOG(GAMMA(R))-
LOG(FACT(TOTAL))+R*LOG(A/(A+EXPBX))+TOTAL*LOG(EXPBX/(A+EXPBX)); *LIKELIHOOD
CALCULATION;
  MODEL TOTAL ~ GENERAL(LL);
RUN;
```

The output values are as follows:
Maximum value of negative Log likelihood = 2888.9661
Alpha = 8.1976
R = 0.1388
Coefficient of income = 0.07340

Coefficient of sex = -0.00928
Coefficient of age = 0.9022
Coefficient of HHsize = -0.02432

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	0.035783
a	8.1976	9.4819	2728	0.86	0.3874	-10.3949	26.7900	-0.00064
b1	0.07340	0.09743	2728	0.75	0.4513	-0.1176	0.2644	0.056282
b2	-0.00928	0.1212	2728	-0.08	0.9390	-0.2469	0.2284	0.003186
b3	0.9022	0.1676	2728	5.38	<.0001	0.5735	1.2309	0.017834
b4	-0.02432	0.04272	2728	-0.57	0.5692	-0.1081	0.05945	0.016634

Managerial Takeaway:

From the above model we can see that at 5% significance level, the coefficients of income, sex and HHsize are not significant. Correlation between the variables can be one of the reasons for such high p-values. The model has to be fine tuned to get better estimates.

The positive coefficient value of Age denotes that as the age is more, people are more likely to visit the Khakichinos website compared to others who are young.

Difference in Results between NBD and Regression and Poisson Regression Model:

- The results of the NDB model fits the data better than Poisson Regression model.
- In Poisson regression, almost all the beta values except sex are significant at 5% significance level. Whereas in NBD regression, only the coefficient of age is significant.
- The negative log likelihood value for NBD regression (2888.9661) is higher compared to Poisson Regression (6291.4967).
- The AIC value (5789.9) is smaller in the case of NBD model. This proves that this is a better model compared to Poisson Regression Model (AIC = 12593)

Part II: Analysis of New Real Data

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 rest 10 records of this dataset.

Solution:

We first tried using the proc import statement but there were many ASCII values which were not imported properly and that threw errors while running the codes. Hence, we decided to use the data statement and modify the values that are taken as inputs.

Code:

```
libname pr 'C:\Education\Courses\ABI with SAS\Homework\project';

DATA pr.books (drop=dummy);
*skipped the header using firstobs;
*read the '1A' ASCII character using IGNOREDOSEOF keyword;
*record length specified using lrecl;
infile 'C:\Education\Courses\ABI with SAS\Homework\project\books.txt'
delimiter='09'x MISSOVER DSD lrecl=50000 firstobs=2 IGNOREDOSEOF;
*instructing SAS to read the record in the given format;
informat userid best32. ;
informat education best32. ;
informat region best32. ;
informat hhsz best32. ;
informat age best32. ;
informat income best32. ;
informat child best32. ;
informat race best32. ;
informat country best32. ;
informat domain $20. ;
informat date best32. ;
informat product $132. ;
informat qty best32. ;
informat price best32. ;
informat dummy $1. ;
*instructing SAS to display the record in the given format;
format userid best12. ;
format education best12. ;
format region best12. ;
format hhsz best12. ;
format age best12. ;
format income best12. ;
format child best12. ;
format race best12. ;
format country best12. ;
format domain $20. ;
format date best12. ;
format product $132. ;
format qty best12. ;
format price best12. ;
format dummy $1. ;
*input field names;
```

```
Input userid  education region hhsz age income child race country domain $
date  product $
qty price VAR15 $;
RUN;


*method 1;
proc sql;
create table pr.books_agg as
select userid, education, region, hhsz, age, income, child, race, country,
sum(qty) as qty
from pr.books
where domain = 'barnesandnoble.com'
group by userid, education, region, hhsz, age, income, child, race, country;
quit;


*method 2;

data bnb;
set pr.books;
if domain = "barnesandnoble.com";
run;

proc means data=bnb NOPRINT;
class userid;
id education region hhsz income child race country age;
output out=bnb_agg2
sum(qty) = tot;
run;


Data bnb_agg2;
set bnb_agg2
 (drop = _TYPE_ _FREQ_);
if userid = . then delete;
run;

PROC PRINT data=bnb_agg2 (obs=10);
run;


proc sql;
select count(*) from pr.books_agg ;
select count(*) from bnb_agg2;
quit;
```

Output:
There are two methods that can be used to arrive at the same output as given below. There are 1812 customers in total who purchased 7074 books from Barnes and Nobles.com

Obs	userid	education	region	hhsz	income	child	race	country	age	tot
1	6365661	5	1	2	7	0	1	0	11	1
2	6396922	2	2	2	4	0	1	0	8	1
3	8999933	4	3	5	3	1	1	0	10	1
4	9573834	99	4	2	5	1	1	0	10	2
5	9576277	99	1	3	7	1	1	0	8	5
6	9581009	99	2	2	5	1	1	0	7	1
7	9595310	4	2	2	2	1	1	0	8	6
8	9611445	2	4	2	6	1	1	1	11	2
9	9663372	4	4	3	7	1	1	0	9	28
10	9752844	3	4	2	3	1	1	0	7	2

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.)

The above codes have data only for B&N. Similar set of codes can be written for Amazon.com and the data can be merged. The code for that is shown below. Again, we have tried two different methods here to arrive at the same data. We have used both Proc SQL and a round about technique using proc means to arrive at the same data set.

Codes:

```
*method 1;
proc sql;
create table pr.books_all as
select userid, education, region, hhsz, age, income, child, race, country,
sum(qty) as qty
from pr.books
group by userid, education, region, hhsz, age, income, child, race, country;
quit;
```

```
*method 2;

data amazon;
set pr.books;
if domain = "amazon.com";
run;

proc means data=amazon NOPRINT;
class userid;
id education region hhsz income child race country age;
output out=amazon_agg
sum(qty) = tot_ama;
run;


Data amazon_agg;
set amazon_agg (drop = _TYPE_ _FREQ_);
if userid = . then delete;
run;



data overall;
merge amazon_agg bnb_agg2;
by userid;
if tot = . then tot = 0;
run;
```

The above data gives overall data for the books dataset. For simplicity, we are planning to model the probability distribution for the Barnes and Nobles total purchase of books. Hence we are only considering the total books purchased for B&N per customer and preparing the dataset needed for NBD model.

```
proc means data=overall NOPRINT;
class tot;
output out=data
n(userid) = peoplecount;
run;

data data;
set data (drop= _TYPE_ _FREQ_);
if tot = . then delete;
run;

*model;

proc nlmixed data=data;
  parms alpha=.5 r=.5; *Our decision variables;
  part1 = (gamma(r+tot)/(gamma(r)*fact(tot)));
  part2 = ((alpha/(alpha+1))**r);
  part3 = (1/(alpha+1))**tot;
  calc = part1*part2*part3;
  ll = peoplecount*log(calc); *sum of ll is what we are trying to maximize;
  model peoplecount ~ general(ll);
run;
```

The above model yields the following results:

Iteration History					
Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	10	9161.2705	810.6365	1588.86	-306461
2	13	8839.6058	321.6647	4714.58	-52120.8
3	15	8721.5997	118.0061	14409.1	-787.050
4	17	8481.4675	240.1322	4653.64	-4308.92
5	20	8463.3040	18.16348	3128.07	-94.2939
6	24	8389.8814	73.42265	957.757	-96.2154
7	27	8382.8463	7.035079	306.726	-8.78148
8	30	8381.7612	1.085078	66.6914	-1.81031
9	33	8381.7110	0.050228	5.19184	-0.10809
10	36	8381.7107	0.000291	0.080373	-0.00059
11	39	8381.7107	1.413E-7	0.049863	-2.38E-7

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	16763
AIC (smaller is better)	16767
AICC (smaller is better)	16768
BIC (smaller is better)	16771

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
alpha	0.1299	0.006121	46	21.22	<.0001	0.1176	0.1422	-0.01909
r	0.09723	0.003060	46	31.77	<.0001	0.09107	0.1034	0.049863

Solution:

The optimal parameters for the NBD model are as follows:

Optimized Negative LL value: -8381.7107

Estimated alpha value: 0.1299

Estimated r value: 0.09723

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

Based on the above output, we can calculate the reach, average frequency and GPR values as follows:

P(X(t=0) | r, alpha) = (alpha/ alpha + t) ^ r = (0.1299/0.1299+1)^.09723 = .8103
E(X(1)) = r*t / alpha = (.09723*1)/.1299 = .74849

Reach:
100 * (1-P(X(t=0)|r, alpha)) = 100 * (1-.8103) =**<u>18.97%</u>**

Average Frequency:
E(X(t)) / (1-P(X(t=0)|r,alpha)) = E(X(1))/(1-P(X(t)=0)) = .748498845/(10.810324813) = **<u>3.95</u>**

GRP:
100* E(X(t)) = 100*E(X(1)) = 100 * .748498845 = **<u>74.85</u>**

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?
Optional: You have flexibility in choosing the variables to include | if you wish to do so, you can choose to eliminate some (via feature selection, for example) or create new ones (from the variables you have available, for example, fraction of weekend purchases). This is optional for this project, but if you do anything along these lines, please provide your justification

We are using the same dataset that we used for other models before. Here instead of ignoring the demographic variables we are including them for the B&N purchases to understand their impact.

```
Data poisson;
set overall (drop=tot_ama);
run;
```

Upon analysing the variables, there are many unknown values and missing values. We decided to correct them. The following code shows that there are nearly around 6900 records with missing values. Since most of the data has missing values, this variable can be deleted.

```
Proc means data=poisson N;
class education;
var education;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : education		
education	N Obs	N
0	1	1
1	638	638
2	772	772
3	13	13
4	811	811
5	302	302
99	6914	6914

There are 11 junk values present in the region variable. To avoid errors, this variable being a categorical variable, the missing values are converted to the mode. Here majority of the data has 3 as the region. Hence the missing values are converted to 3. There are no further missing values in the dataset.

```
data poisson;
set poisson;
if region=. then region=3;
```

We ran the proc means procedure to all the other variables to check for any abnormalities. All the other variables looked fine and hence we decided to keep them as such. The proc means output for some of those are pasted below:

Analysis Variable : region		
region	N Obs	N
1	2160	2160
2	2055	2055
3	3151	3151
4	2074	2074

Analysis Variable : race		
race	N Obs	N
1	9039	9039
2	250	250
3	140	140
5	22	22

Analysis Variable : hhsz		
hhsz	N Obs	N
1	533	533
2	3059	3059
3	2303	2303
4	1889	1889
5	1153	1153
6	514	514

```
proc nlmixed data=poisson;
  parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;
  m=m0*exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*country);
  ll = tot*log(m)-m-log(fact(tot));
  model tot ~ general(ll);
run;
```

The NLMIXED Procedure

Specifications	
Data Set	WORK.PBOOKS
Dependent Variable	NumBooks
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	9440
Observations Not Used	11
Total Observations	9451
Parameters	9

Initial Parameters

m0	b1	b2	b3	b4	b5	b6	b7	age	Negative Log Likelihood
1	0	0	0	0	0	0	0	1	19236.828

Iteration History

Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	7	18962.5174	274.3106	4051.94	-2364256
2	10	18886.4114	76.10599	2538.70	-18914.5
3	13	18878.3277	8.083737	2647.36	-1354.94
4	15	18861.1736	17.15414	2245.40	-434.233
5	18	18852.6848	8.488761	1882.25	-96.9432
6	20	18839.7208	12.96399	988.285	-70.8889
7	23	18834.9472	4.773614	129.786	-38.2950
8	27	18834.7271	0.220123	74.6045	-3.05547
9	29	18834.5460	0.181111	39.5440	-0.58042
10	32	18834.5187	0.027272	7.72368	-0.05919
11	35	18834.5179	0.000802	0.46338	-0.00164
12	38	18834.5179	7.672E-6	0.057536	-0.00002

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	37669
AIC (smaller is better)	37687
AICC (smaller is better)	37687
BIC (smaller is better)	37751

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
m0	1.0743	1.8464	9440	0.58	0.5607	-2.5450	4.6937	0.008342
b1	-0.1034	0.01110	9440	-9.31	<.0001	-0.1251	-0.08161	0.042813
b2	-0.01560	0.01107	9440	-1.41	0.1591	-0.03730	0.006112	0.039661
b3	0.03839	0.4064	9440	0.09	0.9247	-0.7582	0.8350	0.008728
b4	0.01856	0.006292	9440	2.95	0.0032	0.006228	0.03089	0.057536
b5	0.08093	0.03201	9440	2.53	0.0115	0.01818	0.1437	0.000781
b6	-0.2082	0.04422	9440	-4.71	<.0001	-0.2949	-0.1215	0.003590
b7	-0.1201	0.03374	9440	-3.56	0.0004	-0.1862	-0.05395	0.009762
age	0.9739	10.3094	9440	0.09	0.9247	-19.2349	21.1826	0.000344

Solution:

Optimized LL value: -18821.9064
At 5% significance level, income, country, region and race are significant variables. Region, race, and country have a negative relationship with a customer's numbers of purchases at barnesandnoble.com. Age, income, and child have a positive relationship with a customer's numbers of purchases at barnesandnoble.com

5. Next, we start the setup for developing an NBD regression model. What is the formula for the log-likelihood expression, LL?
From the variables that are currently exisiting in our final table, the log likelihood calculation can be shown as below:
LL=log((gamma(r+tot)/(gamma(r)fact(tot)))*((alpha/(alpha+e^{bx}))**r)*((e^{bx}/(alpha+e^{bx}))**tot))
Where e^^{bx} = exp((b1*region)+(b2*hhsz)+(b3*age)+(b4*income)+(b5*child)+(b6*race)+(b7*country))

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? Optional: As with the Poisson regression, you have exibility in choosing the variables to include | if you wish to do so, you can choose to eliminate some (via feature selection, for example) or create new ones (from the variables you have available | for example, fraction of weekend purchases).

We have continued with the same dataset for the NBD regression.

```
proc nlmixed data=poisson;
  parms r=1 a=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+tot))-log(gamma(r))-
log(fact(tot))+r*log(a/(a+expBX))+tot*log(expBX/(a+expBX));
  model tot ~ general(ll);
run;
```

8	44	8425.9264	34.81796	1983.46	-420.314
9	46	8395.4631	30.4633	582.551	-217.059
10	48	8369.9971	25.46602	296.769	-41.1563
11	51	8364.9080	5.089024	132.332	-10.1543
12	54	8364.5687	0.339353	45.1954	-1.19629
13	57	8364.4351	0.133611	47.9777	-0.22095
14	60	8364.4063	0.028741	16.3711	-0.02266
15	63	8364.3914	0.01494	19.0022	-0.01864
16	66	8364.3879	0.003541	1.10853	-0.00575
17	69	8364.3878	0.000031	0.11670	-0.00006

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	16729
AIC (smaller is better)	16749
AICC (smaller is better)	16749
BIC (smaller is better)	16820

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.09796	0.003090	9440	31.70	<.0001	0.09190	0.1040	-0.11670
a	0.1321	.	9440	.	.	.	.	0.079130
b1	-0.1019	0.03214	9440	-3.17	0.0015	-0.1649	-0.03890	-0.03537
b2	-0.01021	0.03335	9440	-0.31	0.7595	-0.07559	0.05517	-0.04720
b3	0.6747	9.7502	9440	0.07	0.9448	-18.4379	19.7873	-0.00600
b4	0.02043	0.01868	9440	1.09	0.2742	-0.01619	0.05704	-0.04569
b5	0.06814	0.09222	9440	0.74	0.4600	-0.1126	0.2489	-0.00754
b6	-0.2097	0.1007	9440	-2.08	0.0374	-0.4071	-0.01227	-0.00817
b7	-0.1053	0.09575	9440	-1.10	0.2713	-0.2930	0.08235	-0.00254
age	0.5736	8.2888	9440	0.07	0.9448	-15.6742	16.8214	-0.00705

Optimal LL value: -8362.5473
From above, we can see that at 5% significance level, the coefficients of region and race are significant. They both are negatively related to the number of books purchased.

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.

Below are some of the differences in the results of the Poisson and NBD Regression:

- In Poisson Regression model, we have observed that almost all the variables are significant but in the NBD Regression model only the race and the region variables are significant. Also, the race and the region variables negatively impact the number of books purchased in both the Poisson and NBD Regression models.
- The negative LL value for the NBD regression model (-8362.5473) has a maximum value compared to the poisson regression (-18821.9064). This proves that NBD model performs better than Poisson model.
- The AIC value of the NBD regression model is lower compared to the AIC value of the Poisson regression model. It is always better to have a low AIC value.
- Though the demographic variable captured individual differences, we could not arrive at proper prediction values using Poisson Regression. This was due to the presence of unobserved heterogeneity, i.e. there are still several other factors that influence the purchase of the books for any individual other than the demographic factors mentioned in the dataset. This is captured by the NBD regression model.
- Few demographics that were significant in the Poisson regression model were no more significant in the NBD model. We capture the unobserved component of differences among individuals in NBD model which explain the model well compared to the original demographics given.

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.

- This project helped us implement some of the best regression techniques in use and helped us understand them. This project also taught us that data pre-processing is a very important step in modelling. We implemented some of those techniques learnt in class to process the data and convert it to the required form.

- For the Poisson regression model, we did not get exact results as expected. This is due to the presence of the unobserved heterogeneity present in the data set. This was later taken into consideration in our NBD model. Hence from that approach we can assume that we can significantly improve our prediction model.

- There are a few variables that are not significant when we built the model. This might be due to a variety of reasons like the presence of multicollinearity. This can be eliminated by removed the variables that are highly correlated to give better estimates of the coefficients.

- If we had more time, we would have tried on creating dummy variables for all the categorical variables. Here in the above models, the categorical variables are simply used as numbers and there is not the right approach. We would have tried different models by creating dummy variables for each categorical variable and then run the model.

- This Project also helped us to play around with a lot of SAS Procedures. There are a variety of techniques to arrive at the answers. We have used some proc sql steps also to arrive at the same numbers as the other procedures. It was really interesting to learn new aspects of the prediction modelling and to understand the applications and usage of various distribution models. This project helped us understand how to go about modelling data from scratch. We would really love to explore further on this project in the future.