**ABI Project Group 6**

**Team Members**

**Sanket Panaskar**

**Srinivas**

**Mohit Sharma**

**Shouvik Pal**

**Virendra Vikram Singh**

/\*Part 1\*/

/\*question 1\*/

/\*NBD Model for Billboard Exposures data\*/

Printing first 10 rows of the data set

**proc** **print** data=mis6334.billboard(obs=**10**);**run**;



/\*NBD Model SAS Code\*/

**PROC** **NLMIXED** DATA=mis6334.billboard;

retain factor **0**; /\*retaining factor after each record\*/

parms shapeR=**0.5** alpha=**0.5**; /\*lambda is gamma distrbuted with parameters shape r and scale alpha;\*/

IF exposures = **0** THEN

DO;

factor= (alpha/(alpha + **1**))\*\*shapeR; /\*for exposures = 0\*/

ll = peoplecount\*log(factor);

END;

ELSE

DO;

ll = peoplecount \* log(factor \* ((shapeR + exposures - **1**)/(exposures\*(alpha + **1**))));

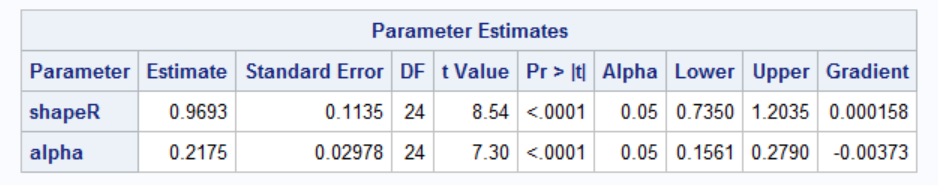
factor = (factor \* ((shapeR + exposures - **1**)/(exposures\*(alpha + **1**))));

END;

MODEL exposures ~ general(ll);

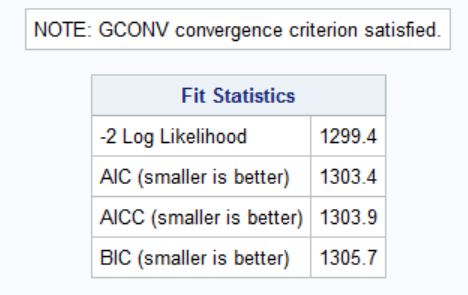
**RUN**;

/\*Reporting on the results\*/



The estimates of shape and alpha are as expected and shown in the lecture.

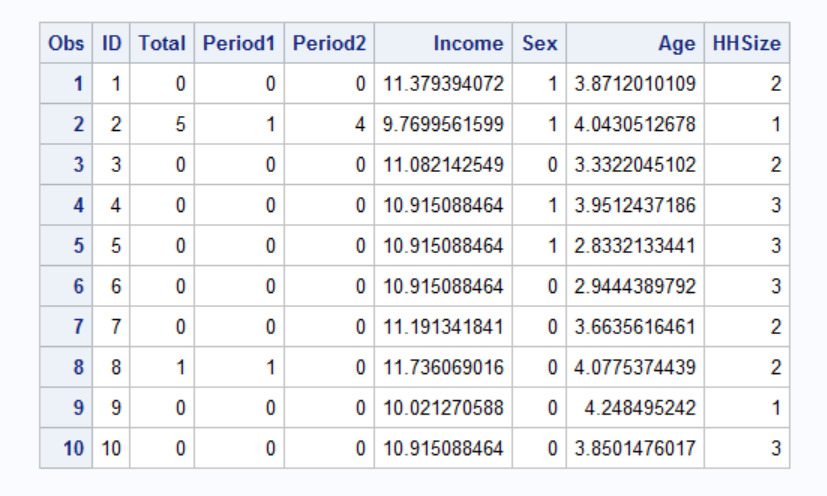
/\*Fit Statistics for the model\*/



LogLikelihood Value : - 649.688

/\*question 2\*/

First 10 records for the khakhichinos example.



/\*SAS Code for the Poisson Regression Model\*/

**proc** **nlmixed** data=mis6334.kc;

parms lambda0=**1** beta1=**0** beta2=**0** beta3=**0** beta4=**0**;

lambda=lambda0\*exp(beta1\*income+beta2\*sex+beta3\*age+beta4\*HHSize);

logprob = - lambda + total\*log(lambda) - log(fact(total)); /\*log likelihood \*/

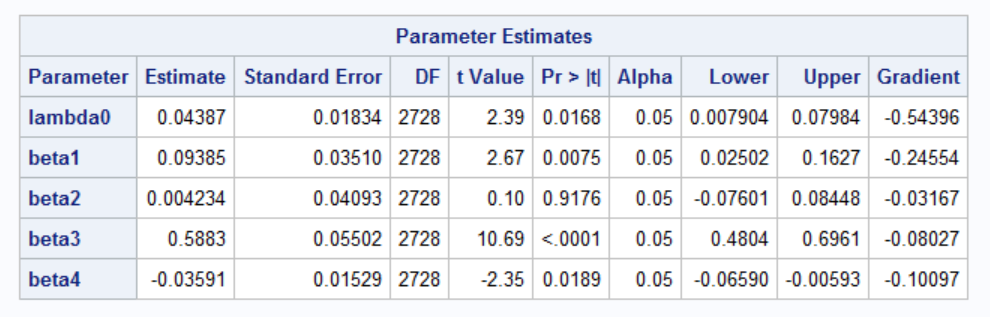
/\*One record for each customer purchase and their demographic characteristics\*/

ll = logprob;

model total ~ general(ll);

**run**;

/\*Reporting on the results\*/

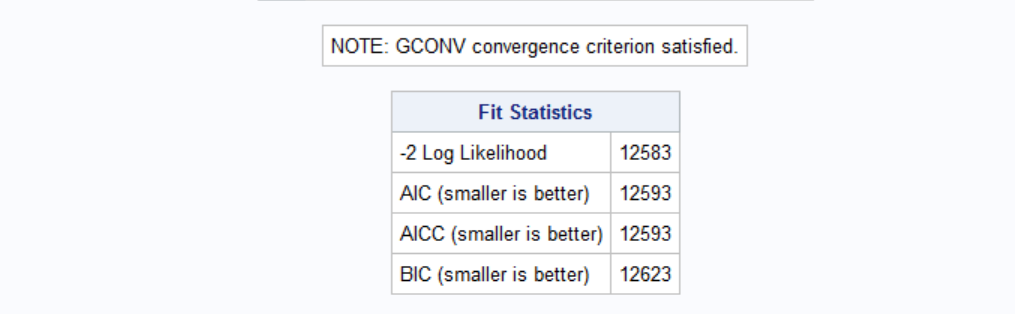


Apart from beta2 , all other betas are significant at 95%.

Beta4 = -0.03591, When HHSize increases by 1, the expression exp(beta1\*income+beta2\*sex+beta3\*age+beta4\*HHSize) value decreases by 0.03591

keeping all other parameters constant and hence making it less likely the customer to visit the site.

Model Fit Statistics



Log Likelihood : -6291.49675

Part 1

Answer 3

/\*NBD Regression Model\*/

/\*SAS Code \*/

**proc** **nlmixed** data=mis6334.kc;

parms shapeR=**1** alpha=**1** beta1=**0** beta2=**0** beta3=**0** beta4=**0**;

expBX=exp(beta1\*income+beta2\*sex+beta3\*age+beta4\*HHSize);

component1 = log(gamma(shapeR+total))- log(gamma(shapeR)) - log(fact(total));

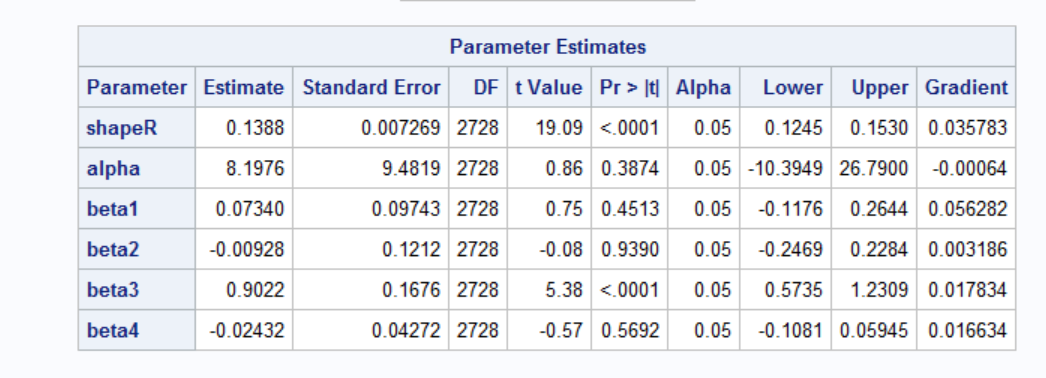
component2 = shapeR \* log(alpha/(alpha+expBX));

component3 = total \* log(expBX/(alpha+expBX));

ll = component1 + component2 + component3;

model total ~ general(ll);

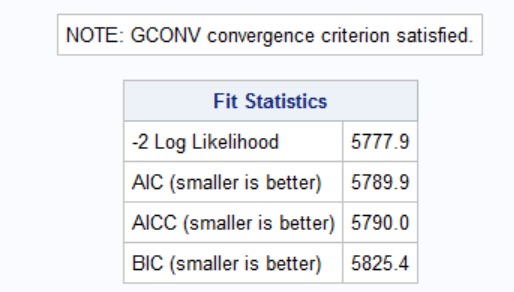
**run**;



Estimates are as expected in the lecture.

In this model, only beta3 is significant at 95% confidence.

/\*Fit Statistics\*



LL = -2888.96611

/\*Part 2\*/

**Part2**

Answer1

1. **Writing data set in sas:**

libname mis "C:\Users\virendsi\Desktop\ABI project"; **run**;

**proc** **import** datafile="C:\Users\virendsi\Google Drive\Study\3rd sem\ABI\HW\project\books.txt" out=mis.booksnew dbms=dlm replace;

delimiter='09'x;

getnames=yes;

**run**;

/\*barnesandnoble.com truncated to barnesandn but that works since we have two values only\*/

**data** mis.booksnew;

set mis.booksnew(drop= var15);**run**;

/\*Since the data had one extra tab after last variable, after reading it, it was taking one more variable of character length 1 and value as blank\*/

/\*Hence removed\*/

/\* extra column dropped\*/

Data cleaning:

1 row with amazon.com domain has quantity as 0 so removing the variable.

2. 46 rows with region as \* value convert into missing.

**data** miss;

set mis.booksnew;

if region = '\*' then region= **.**;

if qty= **.** then delete;

**run**;

/\*question 1\*./

Sorting by user id

**proc** **sort** data =miss; by userid;**run**;

**summation:**

**data** books\_1(drop = domain date product qty price);

set miss;

by userid;

if first.userid then t\_qty=**0**; /\*total quantity assigned to zero for the first record of the group\*/

if domain = "barnesandn" then t\_qty + qty; /\*total quantity summed up successively for "barnesandn"\*/

if last.userid; /\*Taking only last value of the group which is based on userid\*/

**run**;

Printing first ten observations.

| **Obs** | **userid** | **education** | **region** | **hhsz** | **age** | **income** | **child** | **race** | **country** | **t\_qty** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 6365661 | 5 | 1 | 2 | 11 | 7 | 0 | 1 | 0 | 0 |
| **2** | 6388054 | 2 | 4 | 1 | 6 | 5 | 0 | 1 | 0 | 0 |
| **3** | 6628110 | 4 | 4 | 5 | 4 | 7 | 1 | 1 | 0 | 0 |
| **4** | 6631403 | 5 | 3 | 1 | 10 | 3 | 0 | 1 | 1 | 0 |
| **5** | 6704851 | 5 | 4 | 1 | 6 | 7 | 0 | 1 | 0 | 0 |
| **6** | 7412556 | 5 | 4 | 3 | 10 | 7 | 0 | 1 | 1 | 0 |
| **7** | 8985187 | 2 | 4 | 3 | 11 | 5 | 1 | 1 | 0 | 0 |
| **8** | 9350810 | 4 | 3 | 2 | 6 | 7 | 0 | 1 | 0 | 0 |
| **9** | 9363405 | 2 | 3 | 3 | 2 | 5 | 0 | 2 | 0 | 0 |
| **10** | 9552099 | 5 | 2 | 2 | 5 | 7 | 0 | 1 | 0 | 0 |

/\*question 2\*/

For the NBD model, we ignored the demographic variables.

As the t\_qty was not sequential in & probability is missing for many t\_qty>20 we are taking data where t\_qty<=20;

The probability calculation is based on previous iteration probability.

**proc** **sort** data=books\_1; by t\_qty;**run**; /\*sorting by total quantity\*/

**data** books\_2(drop = userid education region hhsz age income child race country);

set books\_1;

by t\_qty;

if first.t\_qty then peoplecount=**0**; /\*first person in the group of t\_qty = 0 ,1 ,2, 3.. etc assigned peoplecount = 0\*/

peoplecount + **1**; /\*successively added for t\_qty\*/

if last.t\_qty; /\*taken the last record in the group t\_qty with the total number of people for that t\_qty\*/

**run**;

/\* keeping only T\_QTY <20 as data is not sequential after that\*/

**data** mis.books\_2;

set books\_2;

if t\_qty <=**20**;

/\*NBD Model\*/

NBD model:

**PROC** **NLMIXED** DATA=mis.books\_2;

retain factor **0**;

parms shapeR=**0.5** alpha=**0.5**; /\*lambda is gamma distrbuted with parameters shape r and scale alpha;\*/

IF t\_qty = **0** THEN /\*for quantity=0\*/

DO;

factor=((alpha/(alpha + **1**)) \*\* shapeR);

ll=peoplecount\*log(factor);

END;

ELSE

DO; /\*For Quantity > 0 \*/

ll = peoplecount \* log(factor \* ((shapeR + t\_qty - **1**)/(t\_qty\*(alpha + **1**))));

factor = (factor \* ((shapeR + t\_qty - **1**)/(t\_qty\*(alpha + **1**))));

END;

MODEL t\_qty ~ general(ll);

**RUN**;

/\*Reporting on the results\*/

| **Fit Statistics** | |
| --- | --- |
| **-2 Log Likelihood** | 12694 |
| **AIC (smaller is better)** | 12698 |
| **AICC (smaller is better)** | 12699 |
| **BIC (smaller is better)** | 12700 |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | | **Gradient** |
| **shaper** | 0.1119 | 0.004276 | 21 | 26.17 | <.0001 | 0.1030 | 0.1208 | -0.00009 |
| **Alpha** | 0.2197 | 0.01164 | 21 | 18.88 | <.0001 | 0.1955 | 0.2439 | 0.003054 |

-------------------------------------------------------------------------------------------------------------------------------------------

/\*Question 3\*/

Effectiveness Measures

We have taken yearly data . hence t=1

Probablity (X=0/ shaper, alpha) = ((alpha/alpha + 1) \*\* shaper) = 0.8254

E(X=0) = (shaper\*t/alpha) = 0.5093

Reach = 100 \*( 1- Probablity(X=0/ shaper, alpha)) = 17.56;

Average Freq = (0.5093/17.56)\* = 2.90

The values are low , since the data is dominated by purchasers from amazon.

-----------------------------------------------------------------------------------------------------------------------------

/\*Question 4\*./

Since we identified the variables as categorical, we went for dummy variable poisson regression.

Here the catch is for a variables having 4 levels, we have created 3 categories regressing with the reference which is the fourth level.

Similarly we have created dummy variables depending on the levels of the categorical variables.

/\*Code for dummy variables\*/

**data** books\_3;

set books\_1;

if region =**.** then delete;

/\* creating dummy variable for region\*/

if region = **2** then reg2 =**1**;else reg2 =**0**;

if region = **3** then reg3 =**1**;else reg3 =**0**;

if region = **4** then reg4 =**1**;else reg4 =**0**;

/\*Created dummy variables for education categorical variable\*/

if education = **0** then e0 = **1**; else e0 = **0**;

if education = **1** then e1 = **1**; else e1 = **0**;

if education = **2** then e2 = **1**; else e2 = **0**;

if education = **3** then e3 = **1**; else e3 = **0**;

if education = **4** then e4 = **1**; else e4 = **0**;

if education = **5** then e5 = **1**; else e5 = **0**;

/\*Created dummy variables for income categorical variable\*/

/\*if income = 1 then i1 = 1;

else i1 = 0;\*/

if income = **2** then i2 = **1**; else i2 = **0**;

if income = **3** then i3 = **1**; else i3 = **0**;

if income = **4** then i4 = **1**; else i4 = **0**;

if income = **5** then i5 = **1**; else i5 = **0**;

if income = **6** then i6 = **1**; else i6 = **0**;

/\*Created dummy variables for age categorical variable\*/

/\*if age = 1 then a1 = 1;

else a1 = 0;\*/

if age = **2** then a2 = **1**; else a2 = **0**;

if age = **3** then a3 = **1**; else a3 = **0**;

if age = **4** then a4 = **1**; else a4 = **0**;

if age = **5** then a5 = **1**; else a5 = **0**;

if age = **6** then a6 = **1**; else a6 = **0**;

if age = **7** then a7 = **1**; else a7 = **0**;

if age = **8** then a8 = **1**; else a8 = **0**;

if age = **9** then a9 = **1**; else a9 = **0**;

if age = **10** then a10 = **1**; else a10 = **0**;

if age = **11** then a11 = **1**; else a11 = **0**;

/\*if race = 11 then r1 = 1;

else r = 0;\*/

if race = **2** then r2 = **1**; else r2 = **0**;

if race = **3** then r3 = **1**; else r3 = **0**;

if race = **4** then r4 = **1**; else r4 = **0**;

**run**;

/\*Model for Poisson Regression\*/

proc nlmixed data=books\_3;

parms lambda0=1 b0=0 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0 b9=0 b10=0 b11=0 b12=0 b13=0 b14=0

b15=0 b16=0 b17=0 b18=0 b19=0 b20=0 b21=0 b22=0.5 b23=0 b24=0 b25=0 b26 = 0 b27 =0 b28 = 0 b29 =0 b30 =0 ;

lambda=lambda0\*exp(b0+b1\*hhsz + b2\*child + b3\*country + b4\*e0 + b5\*e1 + b6\*e2 + b7\*e3 + b8\*e4 +

b9\*e5 + b10\*a2 + b11\*a3 + b12\*a4 + b13\*a5 + b14\*a6 + b15\*a7 + b16\*a8 + b17\*a9 + b18\*a10 + b19\*a11

+ b20\*r2 + b21\*r3 + b22\*r4 + b23\*reg2 + b24\*reg3 + b25\*reg4 + b26\*i2 + b27\*i3+ b28\*i4+ b29\*i5 + b30\*i6);

logprob = - lambda + t\_qty\*log(lambda) - log(fact(t\_qty));

ll = logprob;

model t\_qty ~ general(ll);

run;

Some Results of the model:

| **Fit Statistics** | |
| --- | --- |
| **-2 Log Likelihood** | 29686 |
| **AIC (smaller is better)** | 29750 |
| **AICC (smaller is better)** | 29751 |
| **BIC (smaller is better)** | 29974 |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | | **Gradient** |
| **lambda0** | 0.6466 | . | 8134 | . | . | . | . | 0.30706 |
| **b0** | -0.3711 | . | 8134 | . | . | . | . | 0.19853 |
| **b1** | -0.00285 | 0.01261 | 8134 | -0.23 | 0.8209 | -0.02756 | 0.02186 | 0.56605 |
| **b2** | 0.09791 | 0.03694 | 8134 | 2.65 | 0.0081 | 0.02549 | 0.1703 | -0.25687 |
| **b3** | -0.06178 | 0.03809 | 8134 | -1.62 | 0.1048 | -0.1364 | 0.01288 | -0.05235 |
| **b4** | -0.3866 | 1.6542 | 8134 | -0.23 | 0.8152 | -3.6292 | 2.8559 | 0.36611 |
| **b5** | 0.2631 | 0.05026 | 8134 | 5.24 | <.0001 | 0.1646 | 0.3616 | -0.19155 |
| **b6** | 0.3373 | 0.04486 | 8134 | 7.52 | <.0001 | 0.2494 | 0.4252 | 0.22134 |
| **b7** | -0.1609 | 0.3983 | 8134 | -0.40 | 0.6862 | -0.9417 | 0.6198 | -0.66808 |
| **b8** | -0.2972 | 0.05649 | 8134 | -5.26 | <.0001 | -0.4079 | -0.1865 | -0.03284 |
| **b9** | -0.2079 | 0.08943 | 8134 | -2.32 | 0.0201 | -0.3832 | -0.03261 | -0.14873 |
| **b10** | 0.04004 | 0.3042 | 8134 | 0.13 | 0.8953 | -0.5563 | 0.6364 | -0.41773 |
| **b11** | 0.4944 | 0.2855 | 8134 | 1.73 | 0.0833 | -0.06520 | 1.0540 | 0.16789 |
| **b12** | 0.4455 | 0.2827 | 8134 | 1.58 | 0.1151 | -0.1087 | 0.9997 | 0.12064 |
| **b13** | 0.3743 | 0.2822 | 8134 | 1.33 | 0.1847 | -0.1789 | 0.9275 | -0.09073 |
| **b14** | 0.7606 | 0.2806 | 8134 | 2.71 | 0.0067 | 0.2105 | 1.3107 | 0.036910 |
| **b15** | 0.3432 | 0.2815 | 8134 | 1.22 | 0.2227 | -0.2085 | 0.8950 | -0.11035 |
| **b16** | 0.5098 | 0.2812 | 8134 | 1.81 | 0.0698 | -0.04134 | 1.0610 | 0.14248 |
| **b17** | 0.6989 | 0.2815 | 8134 | 2.48 | 0.0131 | 0.1470 | 1.2507 | 0.41377 |
| **b18** | 0.2592 | 0.2846 | 8134 | 0.91 | 0.3624 | -0.2986 | 0.8171 | 0.54794 |
| **b19** | 0.6492 | 0.2816 | 8134 | 2.31 | 0.0212 | 0.09718 | 1.2012 | -0.49429 |
| **b20** | -0.5857 | 0.1108 | 8134 | -5.29 | <.0001 | -0.8028 | -0.3686 | 0.29042 |
| **b21** | -0.2706 | 0.1269 | 8134 | -2.13 | 0.0330 | -0.5194 | -0.02178 | 0.25760 |
| **b22** | 0.5000 | 0 | 8134 | Infty | <.0001 | -Infty | Infty | 0 |
| **b23** | -0.1814 | 0.03863 | 8134 | -4.69 | <.0001 | -0.2571 | -0.1056 | -0.09742 |
| **b24** | -0.3236 | 0.03563 | 8134 | -9.08 | <.0001 | -0.3935 | -0.2538 | -0.10095 |
| **b25** | -0.3440 | 0.04009 | 8134 | -8.58 | <.0001 | -0.4225 | -0.2654 | -0.17565 |
| **b26** | -0.07363 | 0.06164 | 8134 | -1.19 | 0.2323 | -0.1945 | 0.04721 | 0.11980 |
| **b27** | -0.2260 | 0.05790 | 8134 | -3.90 | <.0001 | -0.3395 | -0.1125 | 0.019061 |
| **b28** | 0.02839 | 0.04541 | 8134 | 0.63 | 0.5320 | -0.06064 | 0.1174 | 0.013860 |
| **b29** | 0.1496 | 0.03649 | 8134 | 4.10 | <.0001 | 0.07811 | 0.2212 | 0.025943 |
| **b30** | 0.1496 | 0.03997 | 8134 | 3.74 | 0.0002 | 0.07121 | 0.2279 | 0.040580 |

Here we found some of the important characteristics which are significant according to the model.

Important variables are:

child when child=1 has positive effect compared to child=0,

education when education = 1 has positive effect compared to education =99 ( we have taken reference as education = 99)

education when education =2 has a positive effect compared to education =99

education when education = 4 has a negative effect compared to education =99

education when education = 5 has a negative effect compared to education =99

age when age =6,8,9,11 all have positive effects compared to age 1

race when race =2,3 ( both have negative effect on purchase),4 (positive effect) compare to race 1

region 1 has positive effect compared to region (2,3,4)

when income is = 3 has a negative effect compared to income=1

when income =5,6 has positive effect compared to income =1

**So Based on above result targeting customers with child, age( value=6,8,9,11), education(value =1,2,) race (value =4) with income(values=5,6) will result into higher probablity of purchases as compared to baseline variable shown above.**

Now some predictions and checking the model fit. How good is our model.

/\*Code for checking the model fit\*/

%let lambda0= 0.6466;%let b0 = -0.3711;%let b1 = -0.00285;

%let b2=0.09791;

%let b3 = -0.06178;

%let b4 = -0.3866;

%let b5 = 0.2631;

%let b6 = 0.3373;

%let b7 = -0.1609;

%let b8 = -0.2972;

%let b9 = -0.2079;

%let b10 = 0.04004;

%let b11 = 0.4944;

%let b12 = 0.4455;

%let b13 = 0.3743;

%let b14 = 0.7606;

%let b15 = 0.3432;

%let b16 = 0.5098;

%let b17 = 0.6989;

%let b18 = 0.2592;

%let b19 = 0.6492;

%let b20 = -0.5857;

%let b21 = -0.2706;

%let b22 = 0.5;

%let b23 = -0.1814;

%let b24 = -0.3236;

%let b25 = -0.344;

%let b26 = -0.07363;

%let b27 = -0.226;

%let b28 =0.02839;

%let b29 =0.1496;

%let b30 = 0.1496;

**data** BNProb;

set books\_3;

lambda=&lambda0\*exp(&b0+&b1\*hhsz + &b2\*child + &b3\*country + &b4\*e0 + &b5\*e1 + &b6\*e2 + &b7\*e3 + &b8\*e4 +

&b9\*e5 + &b10\*a2 + &b11\*a3 + &b12\*a4 + &b13\*a5 + &b14\*a6 + &b15\*a7 + &b16\*a8 + &b17\*a9 + &b18\*a10 + &b19\*a11

+ &b20\*r2 + &b21\*r3 + &b22\*r4+&b23\*reg2+&b24\*reg3+&b25\*reg4 +&b26\*i2+&b27\*i3+&b28\*i4+&b29\*i5+&b30\*i6);

array prob (**11**) prob0 - prob10; /\* prob(y+1)=proby \*/

prob0=poisson(lambda,**0**);

prob10=**1**-prob0; /\* prob of visited 10+ times. \*/

do y=**1** to **9**;

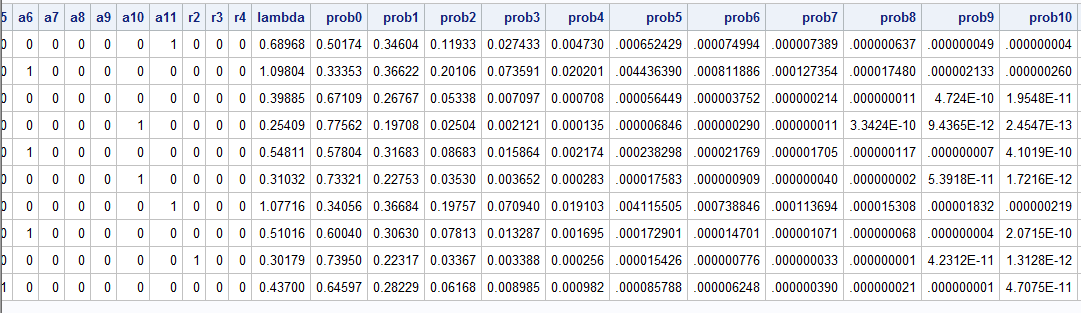
prob(y+**1**)=poisson(lambda,y)-poisson(lambda,y-**1**);

prob10=prob10-prob(y+**1**);

end;

**run**;

above code gives probability of buying 0 to more than 10 books for each customer based on customer characteristics.



/\*average the probabilities over the whole population\*/

**proc** **means** data=BNProb;

var prob0-prob10;

output out=mean\_var mean=;

**run**;

**PROC** **TRANSPOSE** DATA=mean\_var OUT=transpose NAME=Prob;

ID \_FREQ\_;

VAR prob0-prob10;

**RUN**;

**PROC** **PRINT** data=transpose;**run**;

/\*Removing the variables created in proc means and multiplying the probablities by the population\*/

**data** transpose(drop=\_8134);

set transpose;

\_8134 = \_8134\***8134**;

rename \_8134 = value;

**run**;

**proc** **print** data=transpose;**run**;

**proc** **sgplot** data=transpose;

TITLE "Bar chart of Counts";

vbar prob / response=value;

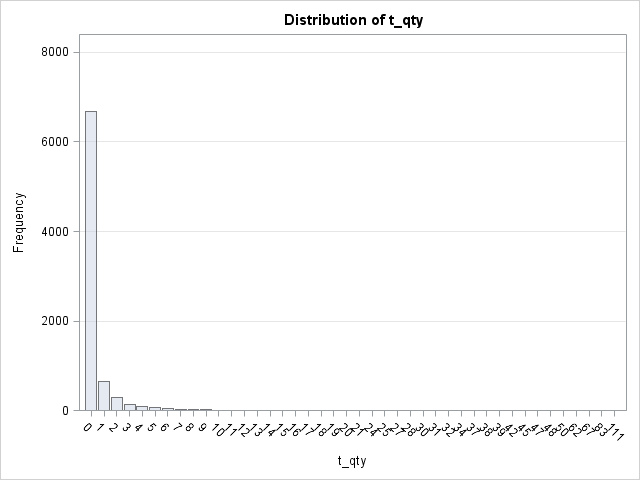
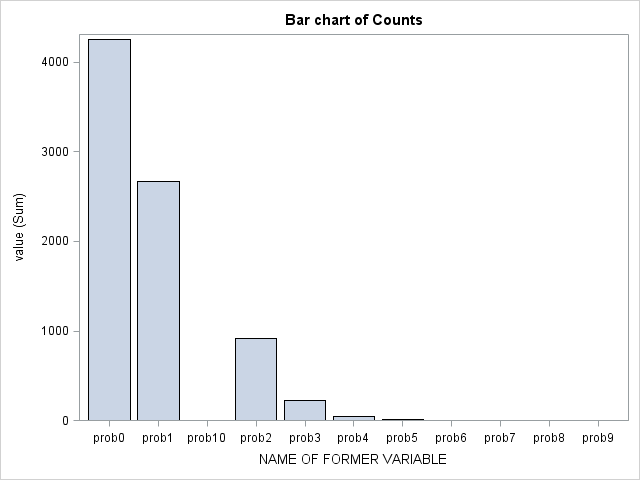
**run**;

**proc** **freq** data=books\_3;

TITLE "Frequency chart of t\_qty";

tables t\_qty / plots=freqplot;**run**;

Difference between actual vs predicted

Actual Predicted

The model fit is not good as you can see the Actual vs Predicted. There is huge difference in the predictions.

Here prob0 is the count of the customers making zero purchases and similarly other probs1 .. prob10.

Reason: We have kept baseline lambda(lambd0) same for each customer. We have incorporated observed heterogeneity through covariates but that is not enough to capture the behavior of the customers as you can see from the model fit.

To solve this problem we will implement NBD regression.

--------------------------------------------------------------------------------------------------------------------------------------

/\*Question 5\*./

Log Likelihood setup for NBD Regression.

Formula for log likelihood:

LL = nlog(gamma(nbdr+t\_qty))-log(gamma(nbdr))-log(fact(t\_qty))+nbdr\*log(alpha/(alpha+eBx))+t\_qty\*log(eBx/(alpha+eBx));

We will see the actual implementation of the formula in the sas code.

/\*Question 6\*./

NBD Regression SAS Code:

**proc** **nlmixed** data=books\_3;

parms nbdr=**1** alpha=**1** b0=**0** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0** b9=**0** b10=**0** b11=**0** b12=**0** b13=**0**

b14=**0** b15=**0** b16=**0** b17=**0** b18=**0** b19=**0** b20=**0** b21=**0** b22=**0.5** b23=**0** b24=**0** b25=**0** b26 = **0** b27 =**0** b28 = **0** b29 =**0** b30 =**0**;

eBx =exp(b0+b1\*hhsz + b2\*child + b3\*country + b4\*e0 + b5\*e1 + b6\*e2 + b7\*e3 + b8\*e4 + b9\*e5+ b10\*a2 +

b11\*a3 + b12\*a4 + b13\*a5 + b14\*a6 + b15\*a7 + b16\*a8 + b17\*a9 + b18\*a10 + b19\*a11 + b20\*r2 + b21\*r3

+ b22\*r4+b23\*reg2+b24\*reg3+b25\*reg4+b26\*i2+b27\*i3+b28\*i4+b29\*i5+b30\*i6);

logprob = log(gamma(nbdr+t\_qty))-log(gamma(nbdr))-log(fact(t\_qty))+nbdr\*log(alpha/(alpha+eBx))+t\_qty\*log(eBx/(alpha+eBx));

ll =logprob;

model t\_qty ~ general(ll);

**run**;

/\*Reporting on the results\*/

| **Fit Statistics** | |
| --- | --- |
| **-2 Log Likelihood** | 13414 |
| **AIC (smaller is better)** | 13480 |
| **AICC (smaller is better)** | 13480 |
| **BIC (smaller is better)** | 13711 |

| **Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | | **Gradient** |
| **Nbdr** | 0.09484 | 0.003373 | 8134 | 28.12 | <.0001 | 0.08823 | 0.1014 | -4.80802 |
| **Alpha** | 0.3686 | . | 8134 | . | . | . | . | 0.32912 |
| **b0** | 0.7185 | . | 8134 | . | . | . | . | -0.12132 |
| **b1** | 0.003782 | 0.03770 | 8134 | 0.10 | 0.9201 | -0.07012 | 0.07768 | -0.39887 |
| **b2** | 0.09451 | 0.1038 | 8134 | 0.91 | 0.3625 | -0.1089 | 0.2979 | -0.24891 |
| **b3** | -0.02508 | 0.1090 | 8134 | -0.23 | 0.8180 | -0.2387 | 0.1886 | -0.05411 |
| **b4** | -0.2191 | 8.4608 | 8134 | -0.03 | 0.9793 | -16.8044 | 16.3662 | 0.077791 |
| **b5** | 0.1799 | 0.1603 | 8134 | 1.12 | 0.2616 | -0.1343 | 0.4942 | 0.12889 |
| **b6** | 0.2858 | 0.1466 | 8134 | 1.95 | 0.0513 | -0.00160 | 0.5732 | -0.33457 |
| **b7** | -0.5376 | 0.9331 | 8134 | -0.58 | 0.5645 | -2.3668 | 1.2915 | -0.22295 |
| **b8** | -0.3686 | 0.1467 | 8134 | -2.51 | 0.0120 | -0.6562 | -0.08094 | 0.003520 |
| **b9** | -0.1937 | 0.2300 | 8134 | -0.84 | 0.3996 | -0.6446 | 0.2571 | 0.17411 |
| **b10** | -0.1435 | 0.7173 | 8134 | -0.20 | 0.8414 | -1.5496 | 1.2626 | -0.21909 |
| **b11** | 0.3535 | 0.6809 | 8134 | 0.52 | 0.6036 | -0.9812 | 1.6883 | -0.14258 |
| **b12** | 0.3267 | 0.6712 | 8134 | 0.49 | 0.6265 | -0.9891 | 1.6424 | -0.03223 |
| **b13** | 0.2040 | 0.6685 | 8134 | 0.31 | 0.7603 | -1.1064 | 1.5144 | -0.03473 |
| **b14** | 0.5837 | 0.6657 | 8134 | 0.88 | 0.3806 | -0.7213 | 1.8886 | 0.004862 |
| **b15** | 0.2349 | 0.6661 | 8134 | 0.35 | 0.7244 | -1.0709 | 1.5406 | 0.073974 |
| **b16** | 0.3604 | 0.6666 | 8134 | 0.54 | 0.5888 | -0.9463 | 1.6671 | -0.01338 |
| **b17** | 0.6027 | 0.6698 | 8134 | 0.90 | 0.3682 | -0.7102 | 1.9156 | -0.08909 |
| **b18** | 0.1134 | 0.6744 | 8134 | 0.17 | 0.8665 | -1.2086 | 1.4353 | -0.00217 |
| **b19** | 0.5262 | 0.6693 | 8134 | 0.79 | 0.4318 | -0.7858 | 1.8381 | -0.13302 |
| **b20** | -0.6416 | 0.2521 | 8134 | -2.54 | 0.0110 | -1.1357 | -0.1474 | 0.048553 |
| **b21** | -0.4138 | 0.3292 | 8134 | -1.26 | 0.2087 | -1.0591 | 0.2314 | 0.053136 |
| **b22** | 0.5000 | 0 | 8134 | Infty | <.0001 | -Infty | Infty | 0 |
| **b23** | -0.2540 | 0.1180 | 8134 | -2.15 | 0.0314 | -0.4852 | -0.02273 | -0.09482 |
| **b24** | -0.3979 | 0.1071 | 8134 | -3.72 | 0.0002 | -0.6078 | -0.1880 | -0.07321 |
| **b25** | -0.3527 | 0.1178 | 8134 | -2.99 | 0.0028 | -0.5837 | -0.1217 | -0.07910 |
| **b26** | -0.04520 | 0.1715 | 8134 | -0.26 | 0.7922 | -0.3814 | 0.2910 | -0.13426 |
| **b27** | -0.1954 | 0.1513 | 8134 | -1.29 | 0.1965 | -0.4920 | 0.1011 | -0.17564 |
| **b28** | 0.05010 | 0.1298 | 8134 | 0.39 | 0.6995 | -0.2043 | 0.3045 | -0.01683 |
| **b29** | 0.1395 | 0.1083 | 8134 | 1.29 | 0.1979 | -0.07286 | 0.3518 | 0.029337 |
| **b30** | 0.1580 | 0.1206 | 8134 | 1.31 | 0.1904 | -0.07851 | 0.3945 | 0.023666 |

/\*Some of the important characteristics which can make it more likely that the customer will purchase a book from b&n.

There has been a decrease in the number of significant variables as compared to poisson regression.

Education when education = 2 has a positive effect compared to education =99

Education = 4 has a negative effect compared to education =99

Race = 2 has a negative effect on purchase compared to race =1

Race 4 has a positive effect of .5 compared to race 1

Region 1 has a positive effect on purchase compared to region 2,3,4 with a significant estimate.

Recommendation:

Customer with education(value =2),race(value =4),Region (value =1) results in more probability making a purchase as compared to reference values.

/\*Code for making predictions and getting the below graphs are on similar lines as above\*/

**data** NBIIProb;

set books\_3;

eBx =exp(&b0+&b1\*hhsz + &b2\*child + &b3\*country + &b4\*e0 + &b5\*e1 + &b6\*e2 + &b7\*e3 + &b8\*e4 + &b9\*e5+ &b10\*a2 +

&b11\*a3 + &b12\*a4 + &b13\*a5 + &b14\*a6 + &b15\*a7 + &b16\*a8 + &b17\*a9 + &b18\*a10 + &b19\*a11 + &b20\*r2 + &b21\*r3

+ &b22\*r4+&b23\*reg2+&b24\*reg3+&b25\*reg4+&b26\*i2+&b27\*i3+&b28\*i4+&b29\*i5+&b30\*i6);

array prob(**11**) prob0 - prob10; /\* prob(y+1)=proby \*/

prob0=((&alpha/(&alpha+eBx))\*\* &shapeR);

prob10=**1**-prob0; /\* prob of visited 10+ times. \*/

do y=**1** to **9**;

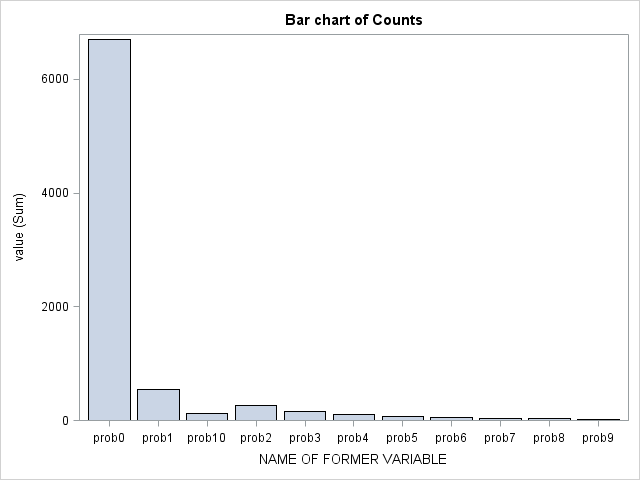
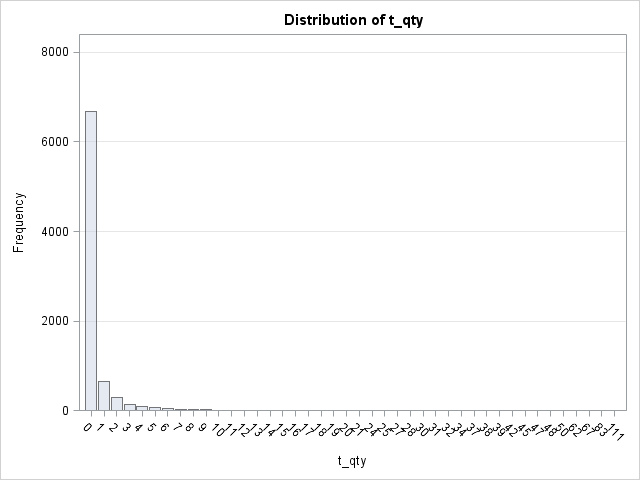
prob(y+**1**)= (gamma(&shapeR+y)/(gamma(&shapeR)\*fact(y))) \* ((&alpha/(&alpha+eBx))\*\* &shapeR) \* ((eBx/(&alpha+eBx)) \*\* y);

prob10=prob10-prob(y+**1**);

end;

**run**;

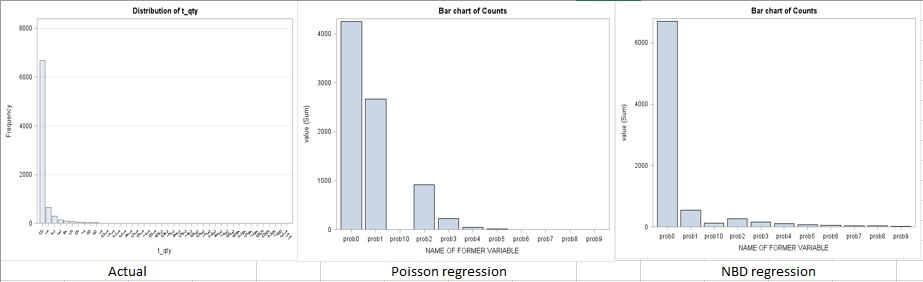
Result comparison



Actual value Predicted value

The model fit has improved once we had incorporated unobserved heterogeneity apart from observed heterogeneity.

/\*Question 7\*/



There is a significant difference between Poisson regression and NBD regression as the NBD regression shows more accurate results compared to Poisson regression.

Significance of the parameters also have changed when we go from Poisson Regression to NBD regression.

The cause of the difference is that we are only taking observed heterogeneity for poisson regression as compared to both observed and unobserved heterogeneity for NBD regression .

The NBD regression is more accurate in predicting the behavior of the customer purchase for B&N.

/\*Question 8 Things Learnt\*/

/\*Things taken from \*Learnt to model mixture models in SAS with the help of The NLMIXED procedure fits nonlinear mixed models, models in which both fixed and random effects are permitted to have a nonlinear relationship to the response variable.

/\*MLE Estimation and Model Convergence\*/

/Chance to incorporate dummy variable regression\*/

/\*Different SAS procedures and their implementation\*/