|  |
| --- |
| Close-up image showing the leaf-sides of two oversized books side-by-side on a bookshelf, with additional books in soft focus background |
| Customer analysis with Base SAS   * *Chiranjeeva, Salihdeen, Surendar, Vigneshwaran* |
| |  |  |  | | --- | --- | --- | | Group 3 |  | MIS 6334.002 Advanced Business Analytics | |

Table of Contents

[Executive Summary: 2](#_Toc473403703)

[Data pre-processing: 2](#_Toc473403704)

[Part – I: Modelling the count data 3](#_Toc473403705)

[Step 1: 3](#_Toc473403706)

[Step 2: 4](#_Toc473403707)

[Step 3: 5](#_Toc473403708)

[Step 4: 6](#_Toc473403709)

[Step 5: 7](#_Toc473403710)

[Step 6: 7](#_Toc473403711)

[Step 7: 9](#_Toc473403712)

[Step 8: 9](#_Toc473403713)

[Part – II: Improving the model: 10](#_Toc473403714)

[Step 9: 10](#_Toc473403715)

[Step 10: 11](#_Toc473403716)

[Step 11: 13](#_Toc473403717)

[Part III: Why certain customers prefer Amazon over BN? 17](#_Toc473403718)

[Step 12: 17](#_Toc473403719)

[Part IV: Summary: 20](#_Toc473403720)

[Step 13: 20](#_Toc473403721)

# Executive Summary:

The primary objective of this project is to build customer analytics model to determine consumer purchasing preference using Base SAS. To achieve this, we analysed 40,000 purchase records, from 2007, of both Barnes & Noble (BN) and Amazon (Nation’s leading booksellers) and build customer analytics model for BN. In this report we have provided detailed explanation on how we transformed raw dataset to provide meaningful customer insights.

# Data pre-processing:

Dataset which we obtained had 40,000 purchase records from both Amazon and Barnes & Noble. To determine the individual customer purchasing pattern we aggregated these records by distinct user-ids, this resulted in 9451 unique users.

Next we imputed missing values from education, age and region using SAS and replaced them with 4, 6 and 3 respectively (calculated using “mode” of the respective variables). We also observed education had missing values of 73%.

**data** mis6334.abawithoutmissingvalues;

set mis6334.Aba\_project2\_data\_books;

if education = **99** then education = **.**;

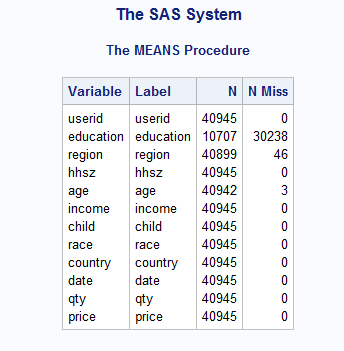
if region>**4** then region = **.**;

if age>**11** then age = **.**;

**run**;

**proc** **means** data = mis6334.abawithoutmissingvalues n nmiss;

**run**;



**data** mis6334.abaimputed;

set mis6334.abawithoutmissingvalues;

if education = **.** then education = **4**;

if region = **.** then region = **3**;

if age = **.** then age = **6**;

**run**;

Since we are only interested in BN, we created a new binary flag variable which indicates whether the given user bought from BN or not. By doing so, we discovered that majority of the users (almost 80.8% of the total users) bought only from Amazon. Only 5.6% of the user base used both the websites to make a purchase.

We have also created 2 new variable which sums the total number of book bought by users from Amazon and BN respectively. (I.e. we split quantity variable in the dataset into Amazon\_Quantity and BN\_Quantity)

**DATA** mis6334.aba;

SET mis6334.Aba\_project2\_data\_books;

IF domain = 'barnesandnoble.com' THEN barnesandnoble = qty;

ELSE barnesandnoble = **0**;

IF domain = 'amazon.com' THEN amazon = qty;

ELSE amazon = **0**;

**RUN**;

# Part – I: Modelling the count data

## Step 1:

Process the raw data using SAS to generate a count dataset in a format similar to the raw data in the "khakichinos.com" example. In other words, for each customer, count the number of books she **purchased from BN** in 2007, and keep the demographic variables. Report your code and print the first 10 records of this dataset.

**Answer:**

LIBNAME MIS6334 'C:\Users\ced150430\desktop\MIS6334';

**proc** **sort** data=mis6334.abaimputed1;

by userid;

**run**;

**PROC** **sql**;

create table mis6334.abaimputed2 as

select userid, education, region, hhsz, age, income, child, race, country, domain, barnesandnoble, amazon, min(date) AS DATE, sum(barnesandnoble) as total\_count\_barnes, sum(amazon) as total\_count\_amazon from mis6334.abaimputed1

group by userid;

**QUIT**;

**run**;

**PROC** **sql**;

create table mis6334.abaimputed3 as

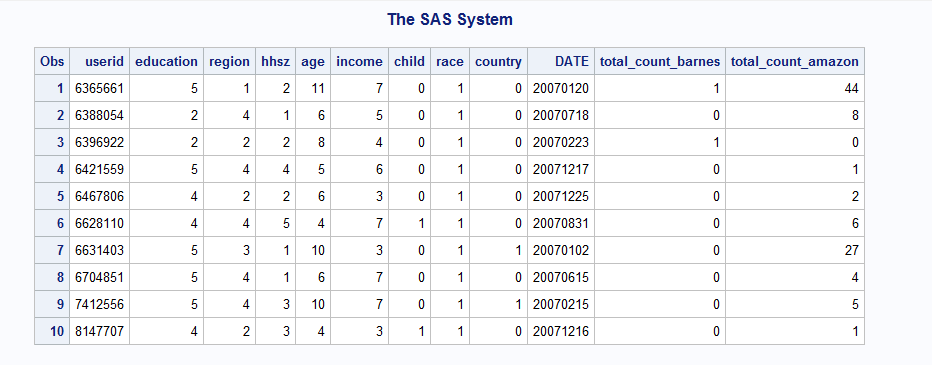
select distinct userid, education, region, hhsz, age, income, child, race, country, DATE, total\_count\_barnes, total\_count\_amazon from mis6334.abaimputed2;

**QUIT**;

**Run**

proc print data= mis6334.abaimputed3(obs= **10**);

**run**;



**Explanation:** We created two new variable namely total\_count\_barnes and total\_count\_amazon. Next we iterated through the entire dataset and assigned value of quantity to newly created variables based on domain name. To identify total number of purchase made by a single user, we grouped the dataset by userid and created a sum of newly created variables. Finally we applied distinct method to weed out duplicate user records.

## Step 2:

For now ignore the demographic information, and run the NBD Model. Report your code and the MLE results (including the optimized LL value, all the estimated parameter values, and the according p-values – same requirement for all MLE estimations in this project).

**Answer:**

**PROC** **SQL**;

create table mis6334.abaimputed4 as

select total\_count\_barnes, COUNT(userid) as NumberofCustomers from mis6334.abaimputed3

group by total\_count\_barnes;

**QUIT**;

**run**;

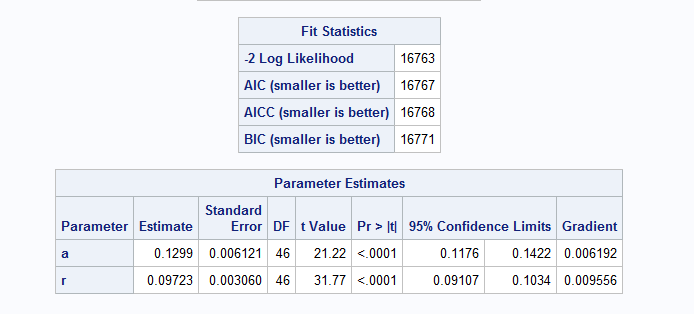
**PROC** **NLMIXED** DATA = mis6334.abaimputed4;

PARMS a=**1** r=**1**;

loglinear=NumberofCustomers\*log((gamma(r+total\_count\_barnes)/(gamma(r)\*fact(total\_count\_barnes)))\*((a/(a+**1**))\*\*r)\*((**1**/(a+**1**))\*\*total\_count\_barnes));

MODEL NumberofCustomers ~ general(loglinear);

**RUN**;



**Explanation:**

MLE: -2 Log Likelihood/2 = 16763/2 = 8381.5

A = 0.1299; R= 0.09723; and both p-values are <0.001

## Step 3:

Based on the NBD Model results, report Reach, Average Frequency and GRPs. Show your calculation.

**Answer:**

Reach = the proportion of people who purchased at least once from BN

= (Total users – Users who ordered only from Amazon) / (Total users)

= (9451 – 7639)/ (9451)

= 0.1917

Average Frequency = sum of BN book purchases / number of BN customers

= 7074/ 1812

= 3.90

GRP = 100 \* Reach \* Average frequency

= 100 \* 0.1917 \* 3.90

= 74.763

**Note:** We are calculating Reach based on total users (Both BN and Amazon) because we’re interested in finding market penetration of BN in people who interested in ordering books online. By doing so, we discovered 19% of population have ordered from BN at least once during 2007.

## Step 4:

Hereafter we will consider consumer demographic information. Run the Poisson Regression Model using the provided customer characteristics. Report your code and the MLE results. Which customer characteristics matter, i.e., what is your managerial takeaway?

**Answer:**

We have included Amazon purchase for Poisson regression model.

**proc** **nlmixed** data=mis6334.abaimputed3;

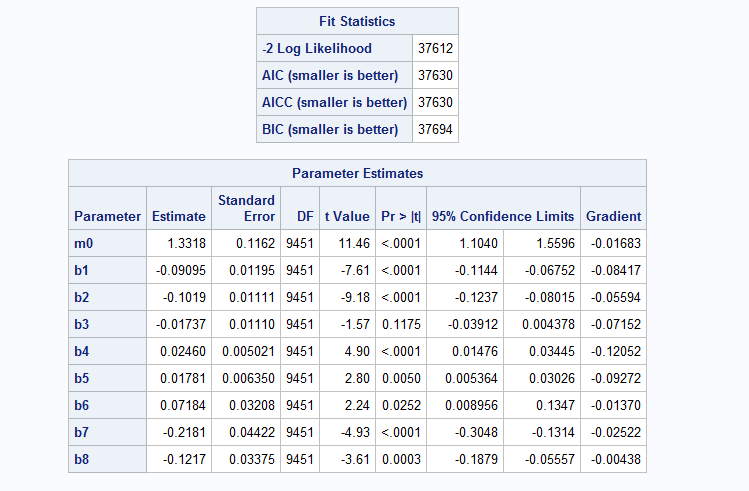
parms m0=**1** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0**;

m=m0\*exp(b1\*education+b2\*region+b3\*hhsz+b4\*age+b5\*income+b6\*child+b7\*race+b8\*country);

ll = total\_count\_barnes\*log(m)-m-log(fact(total\_count\_barnes));

model total\_count\_barnes ~ general(ll);

**run**;



**Explanation:**

MLE : -2 Log Likelihood/2 = 37612/2 = 18806

λ (m0) : 1.3318 with p-value < 0.0001

b1 (Education) : -0.09095 with p-value < 0.0001

b2 (Region) : -0.1019 with p-value < 0.0001

b3 (Hhsz) : -0.01737 with p-value 0.1175

* Since the p-value is greater than 0.05, we can ignore household size as it’s not significant.

b4 (Age) : 0.02460 with p-value < 0.0001

b5 (Income) : 0.01781 with p-value 0.0050

b6 (Child) : 0.07184 with p-value 0.0252

b7 (Race) : -0.2181 with p-value < 0.0001

b8 (Country) : -0.1217 with p-value 0.0003

Based on the above table we can infer that all demographic information other than household size plays a significant role in understanding each individual customer behaviour.

**Note:** The positive or negative value of estimate explains how much effect each variable has over the frequency of customer purchase (Assuming other variables remains constant). For e.g. b2 (Region) has an estimate of -0.09095, which means independent variable region can decrease the chance of customer purchasing book from Barnes & Noble by 10%. (Assuming all other independent variables remains constant)

**Note 2:** We didn’t include date in Poisson regression model, as Poisson distribution is itself based on fixed time intervals. By adding date to Poisson would make it a time series which we believe will add to further complexity to the problem. Hence we didn’t include date in Poisson calculation.

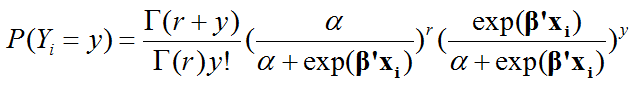
**Managerial takeaway:** Based on Poisson test results, it’s clear that the household size doesn’t play significant role in customer purchase preferences. To improve this we suggest BN run promotions targeted towards families (such as for every purchase of cook book, BN could offer some discount on children books to target Mothers) or we could offer free shipping to any address which orders more than certain number of books (say 3) within a month (which encourage users to order from BN instead of rivals to enjoy free shipping).

If you look at the estimate for race, it’s having a negative correlation. That can actually mean that Barnes and Noble has a thin customer base when it comes to race 2, 3 and 4. Also for country the estimate is again negative. That can be improved by selling more international books that will attract more customers from outside US and also Barnes and Noble have to develop a good strategy towards shipping internationally which could further increase their chances of achieving this goal.

## Step 5:

For the NBD Regression Model, what is the formula for LL? Write it down in your report. Getting this math formula clearly written will help your follow-up coding.

**Answer:**



## Step 6:

Run the NBD Regression Model using the provided customer characteristics. Report your code and the MLE results. Which customer characteristics matter, i.e., what is your managerial takeaway?

**Answer:**

**proc** **nlmixed** data = mis6334.abaimputed3;

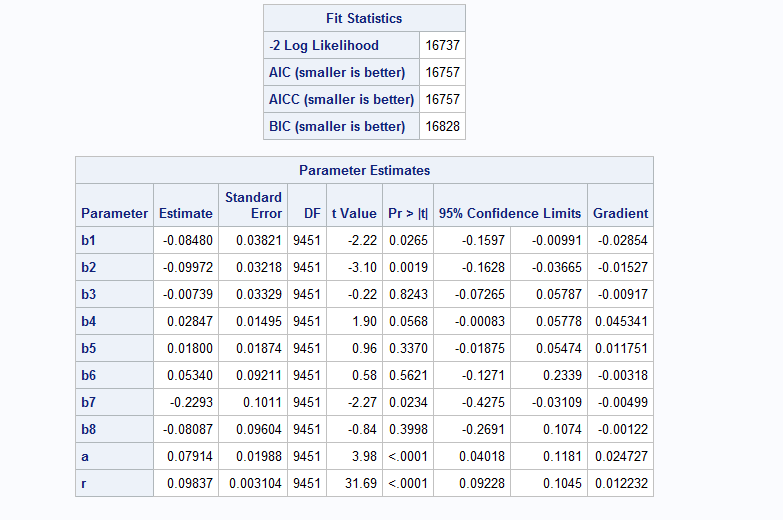
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0** a=**1** r=**1**;

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

ll = log(gamma(r+total\_count\_barnes))-log(gamma(r)\*fact(total\_count\_barnes))+(r\*(log(a)-log(a+m)))+(total\_count\_barnes\*(log(m)-log(a+m)));

model total\_count\_barnes ~ general(ll);

**RUN**;



**Explanation:**

MLE : -2 log Likelihood/2 = 16737/2 = 8368.5

b1 (Education) : -0.08480 with p-value 0.0265

b2 (Region) : -0.09972 with p-value 0.0019

b3 (Hhsz) : -0.00739 with p-value 0.8243

b4 (Age) : 0.02847 with p-value 0.0568

b5 (Income) : 0.01800 with p-value 0.3370

b6 (Child) : 0.05340 with p-value 0.5261

b7 (Race) : -0.2293 with p-value 0.0234

b8 (Country) : -0.08087 with p-value 0.3998

a : 0.07914 with p-value < 0.0001

r : 0.09837 with p-value < 0.0001

Based on NBD regression model results, we can infer that only three independent variable have p-value less than 0.05. This means only these independent variables (Education, region and age) play a significant role in determining the customer preference. The other independent variable have p-value greater than 0.05, hence they’re insignificant in determining customer preference.

**Note:** Estimate for each independent variable is a measure of how much each variable affects customer preference assuming all other independent variable are constant. For e.g. b2 (Region) has an estimate of -0.09 which means region can decrease the frequency of customer purchase assuming all other independent variable remain constant.

**Managerial takeaway:** To know more about the effect of race on customer preference we decided to count frequency of race in BN customer list. Due to which we discovered 96% of BN customer were from race category 1. This may be due to lack of diverse titles which will entice customer from other races to make a purchase. Our suggestion to BN would be to include more book from major languages of other races (and their culture) to their library.

## Step 7:

Any noticeable difference regarding the managerial takeaways between Poisson Regression and NBD Regression? If yes, what exactly is the difference? (Optional) Any thought on why the difference?

**Answer:**

* There is a noticeable difference between Poisson regression and NBD regression. Comparing the p-values of the parameter estimates from both the models we observe that, Poisson has more significant (P-value <0.05) independent variable, six, compared to three from NBD.
  + This implies Poisson model is dependent on more independent variable for predict customer preference compared to NBD
  + Comparing both the models, Poisson will help a manager who is trying to improve Barnes and Noble’s business because it returned more significant variables for the manager to infer from
* The difference exist because of over dispersion factor as denoted by alpha value (A =0.07). Both the models (NBD and Poisson) would be equal if alpha is equal to 0.
  + **Over dispersion:** In Poisson model where mean is equal to variance, there exists a 1:1 relationship between mean and variance. However there are certain instances in which the variability in data may exceed these expectations. For example in the case NBD regression model, the variance increase greater than the expected 1:1 relationship under Poisson. This is called over dispersion.
* The log likelihood of the Poisson model is much greater than the log likelihood of NBD model.

## Step 8:

Does NBD Regression fit the data better than Poisson Regression? (Hint: use the LR test – i.e., likelihood ratio test – on slide 29 in the count model lecture.)

**Answer:**

LR = -2 (LLB-LLA) where LLB is log likelihood of Poisson and LLA is log likelihood NBD

LR = 37612 – 16737

LR = 20875

Now for **hypothesis**

**H0:** Model A is not different from Model B

**H1:** Model A is different from Model B

Now according to LR test if **LR > χ2 (.05,k)** reject the null. (K denotes the degree of freedom which in this case is 8). Substituting k in the equation we get:

20875 > 15.5

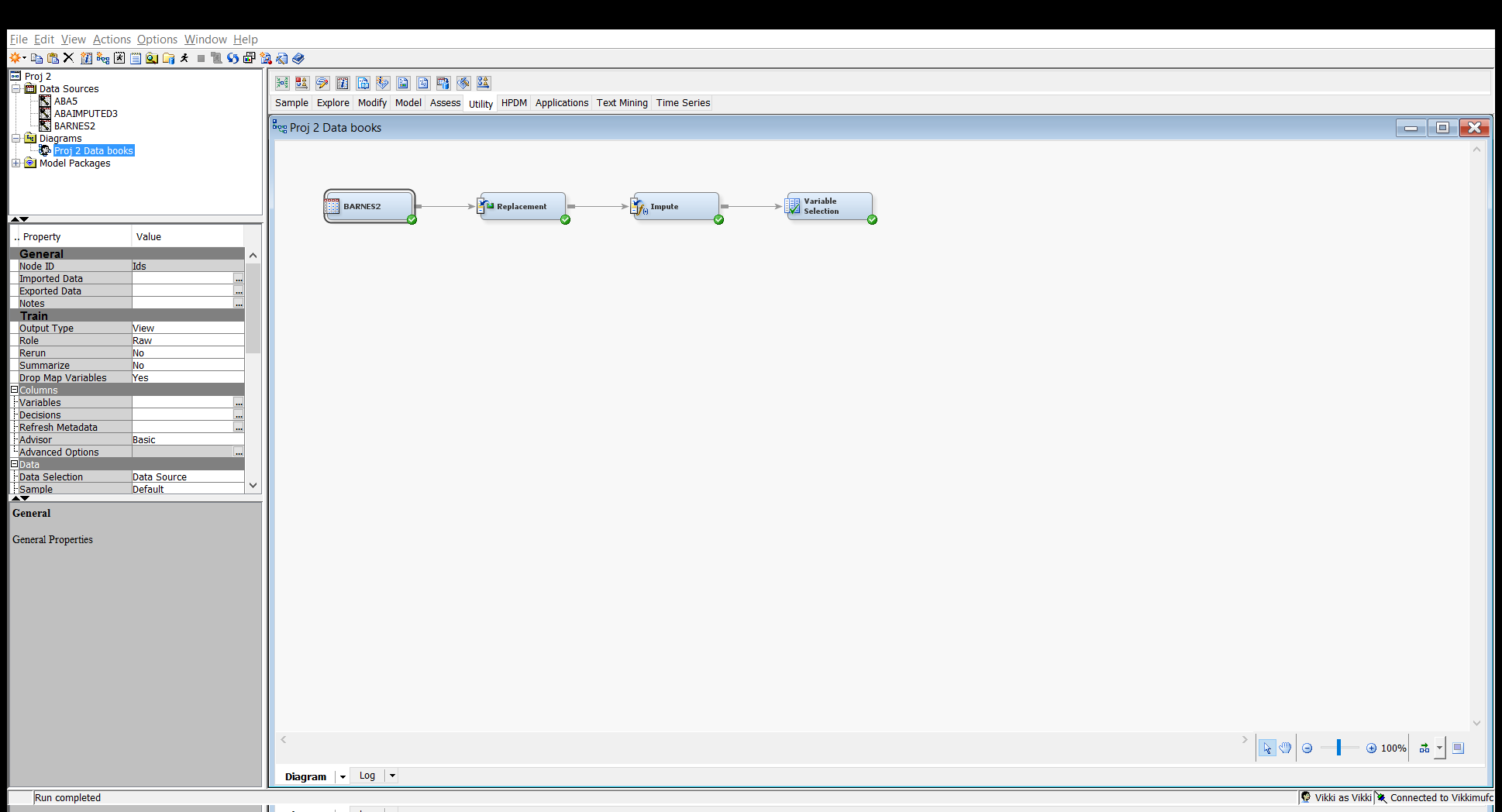
Thus by using LR test we can conclude that NBD is better than LR model.

# Part – II: Improving the model:

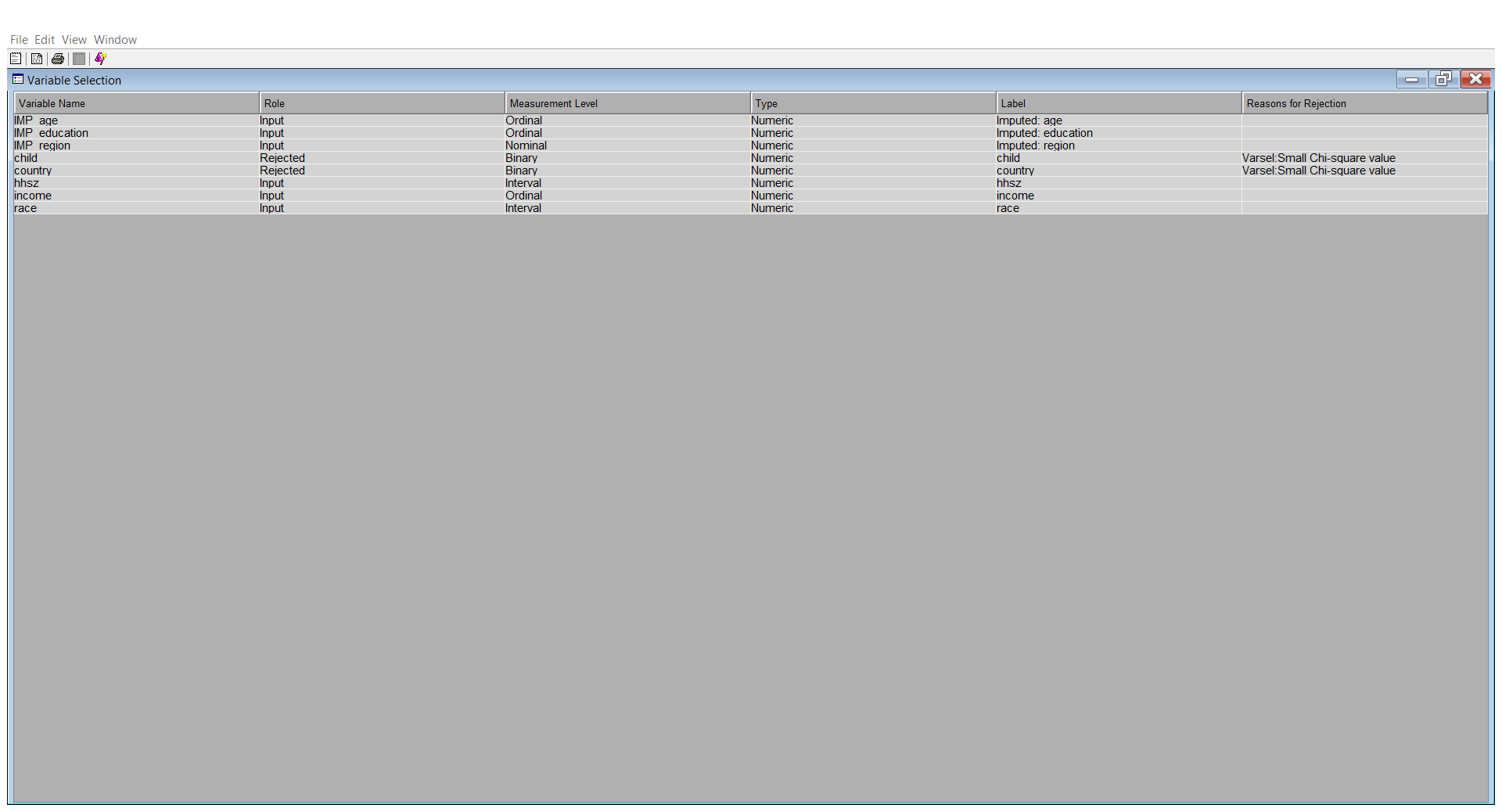
## Step 9:

Similar to what you found out in Project 1, not all variables are always useful. Please try feature election (i.e. selecting only a subset of customer characteristics), and report your findings.

**Answer:** For 9 child and country are ignored because of chi-square stats



* Using SAS EM we loaded the data using file import and imputed the missing values in the dataset using mode. Then we proceed with the variable selection to identify key variables influencing the dependent variable.



* As shown in the above result we have rejected child and county value based on chi-square stats. We picked chi-square value of 3.84 which gave us 6 input variables and rejected 2 variables.
* We also tried running variable selection with minimum R-square value of 0.005 but just gave 2 input variables, hence we moved forward with the chi-square estimation.
* We have also rejected education as it contained more than 70% data as missing values.
* We made use of the remaining 5 values for analysis

After performing variable selection, we found that child and country were not significant and got rejected. Also we excluded education as it has more than 70% missing values. For further analysis, we don’t include education, child and country columns in the input data set.

## Step 10:

You can also construct some variables on your own (e.g. convert date to weekday/weekend, or to holiday/non-holiday, or to seasons, construct percentage of weekend purchases, degree of loyalty to BN etc. -- totally your call and just try 2-3 ideas). Report your code (including the code for constructing the new variables) and the MLE results. Which newly constructed variables matter, i.e., what is your new managerial takeaway?

We created a new variable called loyalty.

Loyalty= number of books purchased from BN/(number of books purchased from BN+ number of books purchased from Amazon)\*100

**data** mis6334.loyalty;

set mis6334.abaimputed3;

loyalty=(total\_count\_barnes/(total\_count\_barnes+total\_count\_amazon))\***100**;

**run**;

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

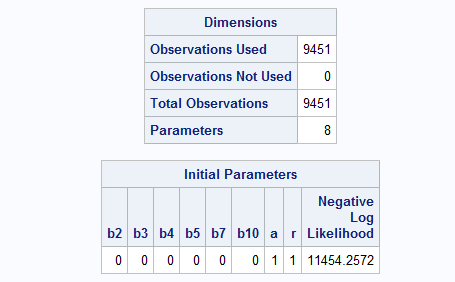
parms b2=**0** b3=**0** b4=**0** b5=**0** b7=**0** b10=**0** a=**1** r=**1**;

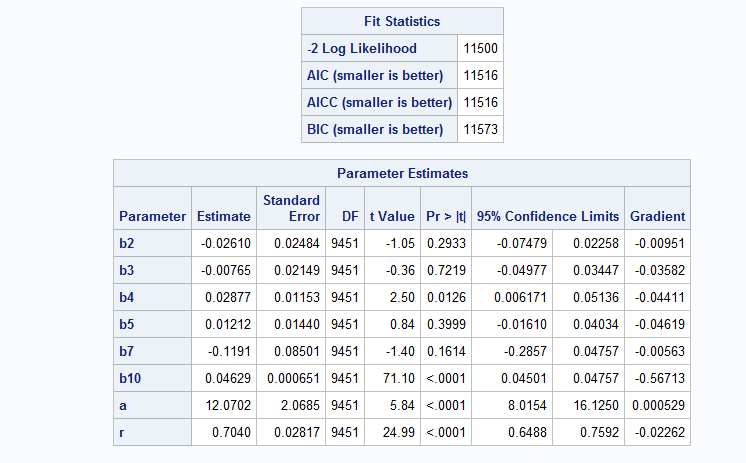
m=exp(b2\*region+b3\*hhsz+b4\*age+b5\*income+b7\*race+b10\*loyalty);

ll = log(gamma(r+total\_count\_barnes))-log(gamma(r)\*fact(total\_count\_barnes))+(r\*(log(a)-log(a+m)))+(total\_count\_barnes\*(log(m)-log(a+m)));

model total\_count\_barnes ~ general(ll);

**RUN**;





We created 3 new variables called weekday (for purchases made Monday through Friday), weekend (for purchases made Saturday and Sunday), and season (1 being Spring, 2 being summer, 3 being fall, and 4 being winter)

**data** mis6334.dateimputed;

set mis6334.abaimputed3;

year=substrn(date,**1**,**4**);

month=substrn(date,**5**,**2**);

day=substrn(date,**7**,**2**);

date1=mdy(month,day,year);

weekday1=weekday(date1);

if weekday1 = **1** or weekday1= **7** then weekend = **1**;

else weekend = **0**;

if weekday1 = **2** or weekday1= **3** or weekday1 = **4** or weekday1= **5** or weekday1 = **6** then weekday = **1**;

else weekday = **0**;

if month = **03** or month = **04** or month = **05** then season = **1**;

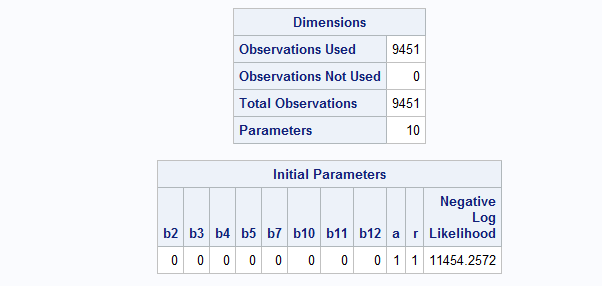
else if month = **06** or month = **07** or month = **08** then season = **2**;

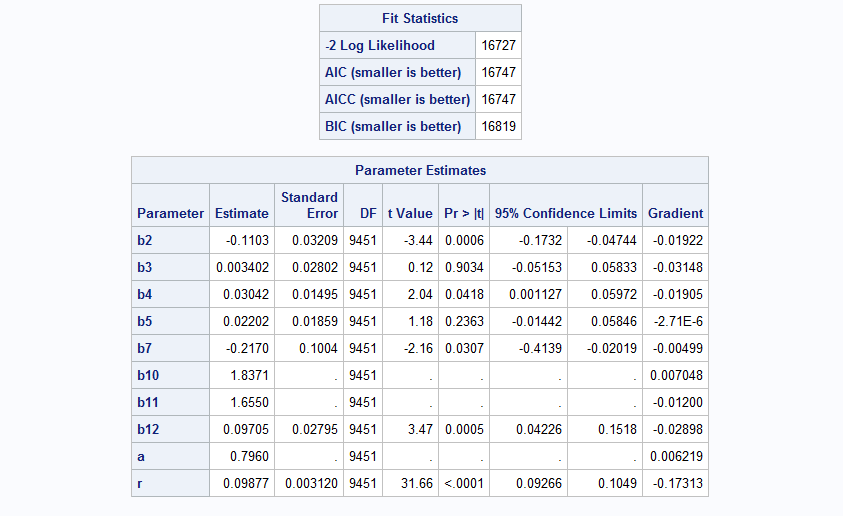
else if month = **09** or month = **10** or month = **11** then season = **3**;

else if month = **12** or month = **01** or month = **02** then season = **4**;

drop year day month weekday1 date1;

**run**;





We tried creating new fields like weekend/weekday, loyalty % and season of purchase in our dataset. After running NBD regression on this new dataset, we found that based on loyalty % our model improved considerably but with the weekend/weekday flag and season of puchase, the model performance didn’t change much.

## Step 11:

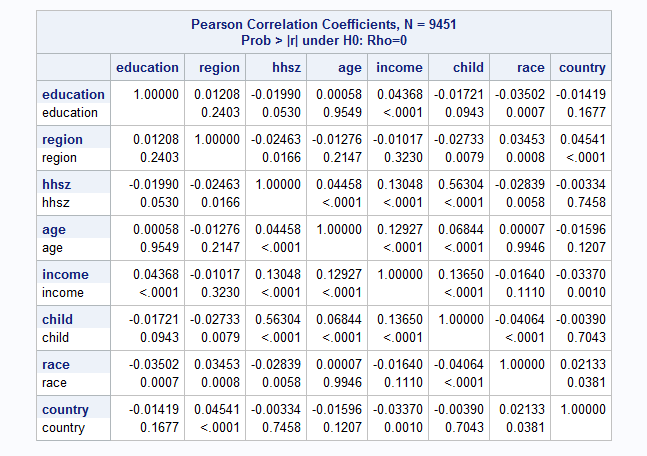
Researchers often try to improve a model by considering interaction effects (e.g., age\*income) in the regression. Try 2-3 interaction effects you think are likely. Report your findings.

**Answer:**

**proc** **corr** data = mis6334.abaimputed3;

var education region hhsz age income child race country;

**run**;



**We created 3 new variables:**

**data** mis6334.abaimputed4;

set mis6334.abaimputed3;

ageandincome=age\*income;

ageandhhsz=age\*hhsz;

hhszandincome=hhsz\*income;

**run**;

**Age\*Income:**

**proc** **nlmixed** data = mis6334.abaimputed4;

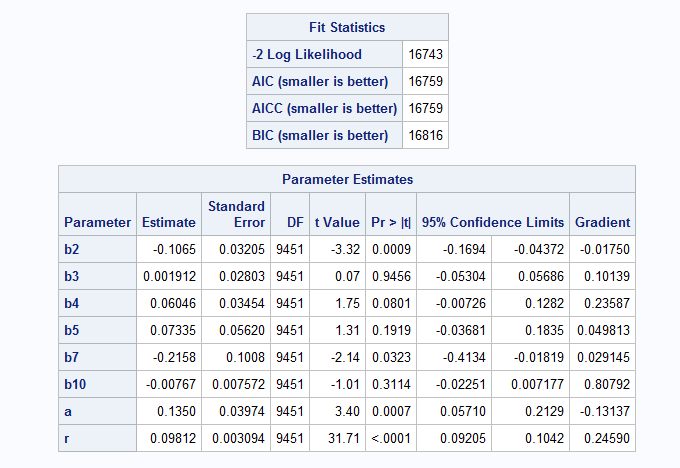
parms b2=**0** b3=**0** b4=**0** b5=**0** b7=**0** b10=**0** a=**1** r=**1**;

m=exp(b2\*region+b3\*hhsz+b4\*age+b5\*income+b7\*race+b10\*ageandincome);

ll = log(gamma(r+total\_count\_barnes))-log(gamma(r)\*fact(total\_count\_barnes))+(r\*(log(a)-log(a+m)))+(total\_count\_barnes\*(log(m)-log(a+m)));

model total\_count\_barnes ~ general(ll);

**RUN**;



**Age\*Hhsz:**

**proc** **nlmixed** data = mis6334.abaimputed4;

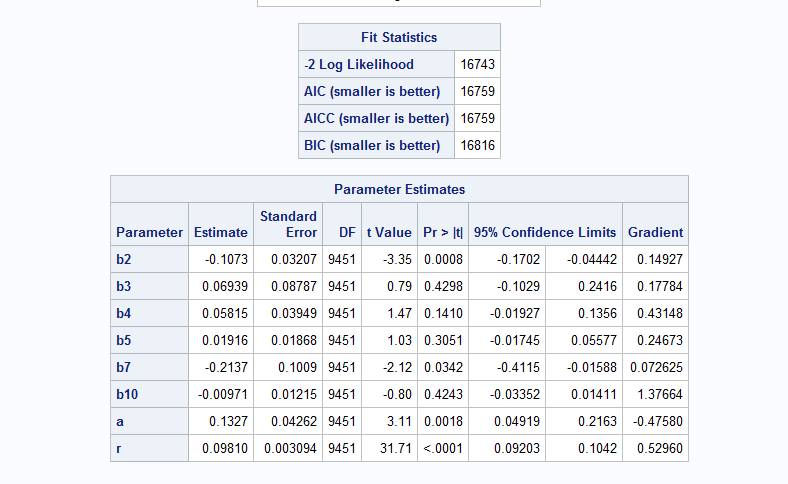
parms b2=**0** b3=**0** b4=**0** b5=**0** b7=**0** b10=**0** a=**1** r=**1**;

m=exp(b2\*region+b3\*hhsz+b4\*age+b5\*income+b7\*race+b10\*ageandhhsz);

ll = log(gamma(r+total\_count\_barnes))-log(gamma(r)\*fact(total\_count\_barnes))+(r\*(log(a)-log(a+m)))+(total\_count\_barnes\*(log(m)-log(a+m)));

model total\_count\_barnes ~ general(ll);

**RUN**;



**Hhsz\*Income:**

**proc** **nlmixed** data = mis6334.abaimputed4;

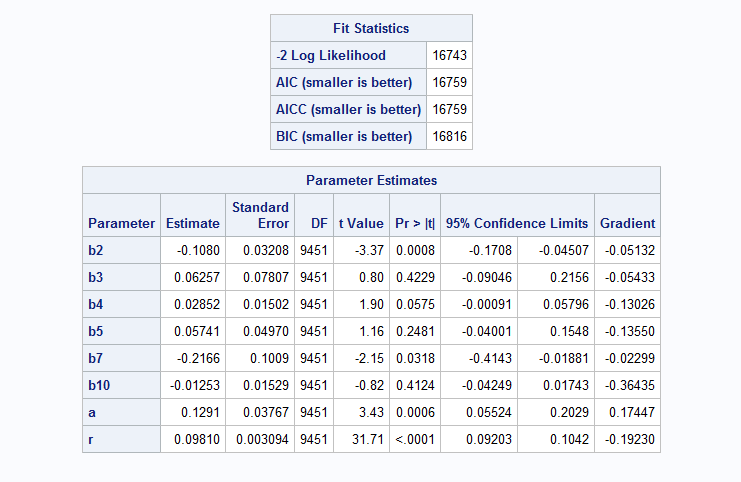
parms b2=**0** b3=**0** b4=**0** b5=**0** b7=**0** b10=**0** a=**1** r=**1**;

m=exp(b2\*region+b3\*hhsz+b4\*age+b5\*income+b7\*race+b10\*hhszandincome);

ll = log(gamma(r+total\_count\_barnes))-log(gamma(r)\*fact(total\_count\_barnes))+(r\*(log(a)-log(a+m)))+(total\_count\_barnes\*(log(m)-log(a+m)));

model total\_count\_barnes ~ general(ll);

**RUN**;



**Explanation:**

To create an interaction effect, we first created a correlation matrix (as shown in the first figure in step 11) and compared their respective p-value with the 0.05. By doing so we identified 3 combination which less p-values than 0.05 i.e. these combinations are significant in determining the customer preference. Those combinations are:

* Age\*Income: MLE Results: 16743/2 = 8371.5
* Age\*Hhsz: MLE Results: 16743/2 = 8371.5
* Hhsz\*Income: MLE Results: 16743/2 = 8371.5

However we didn’t observe any noticeable difference in the model performance due to these interaction effects. Hence we decided not to proceed further with this.

# Part III: Why certain customers prefer Amazon over BN?

## Step 12:

Now let’s study why certain customers prefer Amazon over BN and vice versa. We will apply the concepts of a choice model – logistic regression. For each customer, you need to generate a binary dependent variable indicating whether a user has made a purchase at BN (denote yes as 1 and 0 otherwise). Then use Proc Logistic to run a logistic regression model, report the results and your takeaways. (Optional: Using the data to answer this question: should you do variable selection?)

**Answer:**

**First we looked at customers who purchased Barnes and Nobel’s books only:**

**Customer who prefer BN over Amazon:**

**Data** mis6334.abaimputedlogsitic;

set mis6334.abaimputed3;

if total\_count\_barnes = **0** then logistic=**0**;

else logistic=**1**;

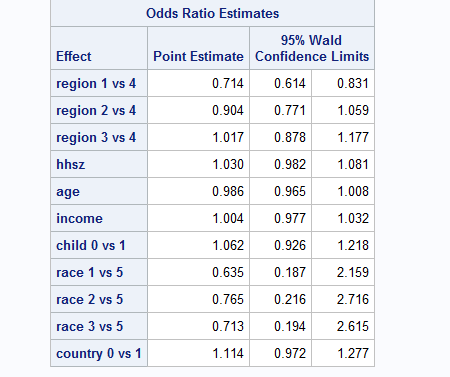
**run**;

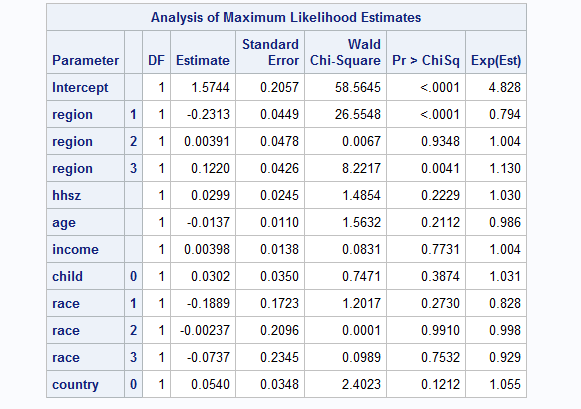
**proc** **logistic** data=mis6334.abaimputedlogsitic;

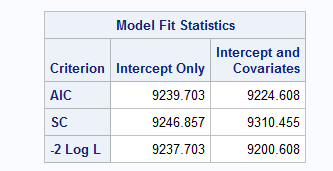
class region child race country;

model logistic= region hhsz age income child race country / expb;

**run**;







**Then we looked at customers who purchased Amazon over Barnes and Nobel’s:**

**Customer who prefer Amazon over BN:**

**data** mis6334.abaimputedlogsitic;

set mis6334.abaimputed3;

if total\_count\_amazon = **0** then logistic=**0**;

else logistic=**1**;

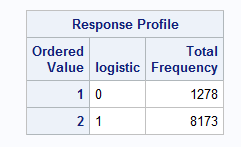
**run**;

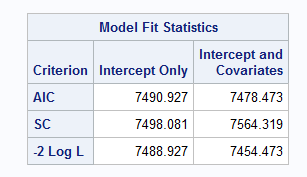
**proc** **logistic** data=mis6334.abaimputedlogsitic;

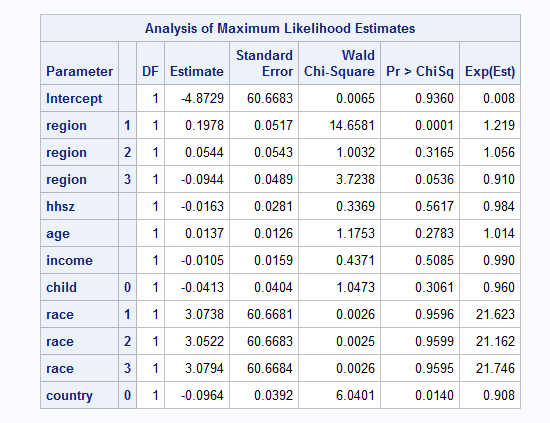
class region child race country;

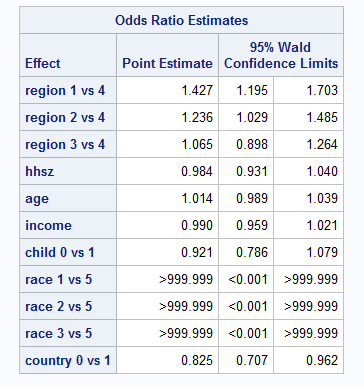
model logistic= region hhsz age income child race country / expb;

**run**;









**Explanation:**

In both variations of the logsitic regression the only significant variables we find are region 1 (compared to 4) and country (a binary variable). All other variables have a p value of greater than 0.05 and therefore are not significant. When looking at customers who choose BN over Amazon it looks like being from region 3 verses being from region 4 increases the log odds of choosing BN over Amazon by 0.122. When looking at the logistic regression for customers who choose Amazon over BN it looks like being from region 1 verses being from region 4 increases the log odds of choosing Amazon over BN by 0.1978. Furthermore, being from outside the United States decreases the log odds of choosing Amazon over BN by 0.09.

# Part IV: Summary:

## Step 13:

Summarize what you learned from this project -- it can be key managerial insights you got, BA techniques or SAS skills you learned from this project, new perspective of BA you got by doing hands-on, or anything you feel worthwhile to summarize. Be concise.

**Answer:**

1. Barnes and Noble has a long way to go before becoming a household name like Amazon when it comes to selling books. It's evident from the number of purchases on both these websites and the portion of users who bought more books on Barnes and Noble than Amazon (less than 2%).
2. There is a lot of avenues for improvement from Barnes and Noble's point of view. Better advertising, better listings, supporting new authors, striking deals with small scale publishers are some things the marketing team of Barnes and Noble can do to develop a unique user base that can be loyal to them.
3. Majortity of BN customers are from one race category, which implies BN didn’t have appeal among other races. So BN can work towards appealing to those races by means of adding other major languages in the country, and books signifying their heritage.
4. Through this project we learned that builing a customer analytics model to understand customer perference (using various demographic information) is much difficult than making a business prediction.
5. An important BA perspective we take away from this course if not the project is that more effort has to be put into preprocessing the data rather than building and running a prediction model because the data is what drives the model we build and if that's clear and clean then there won't be any problem while building and testing the model.