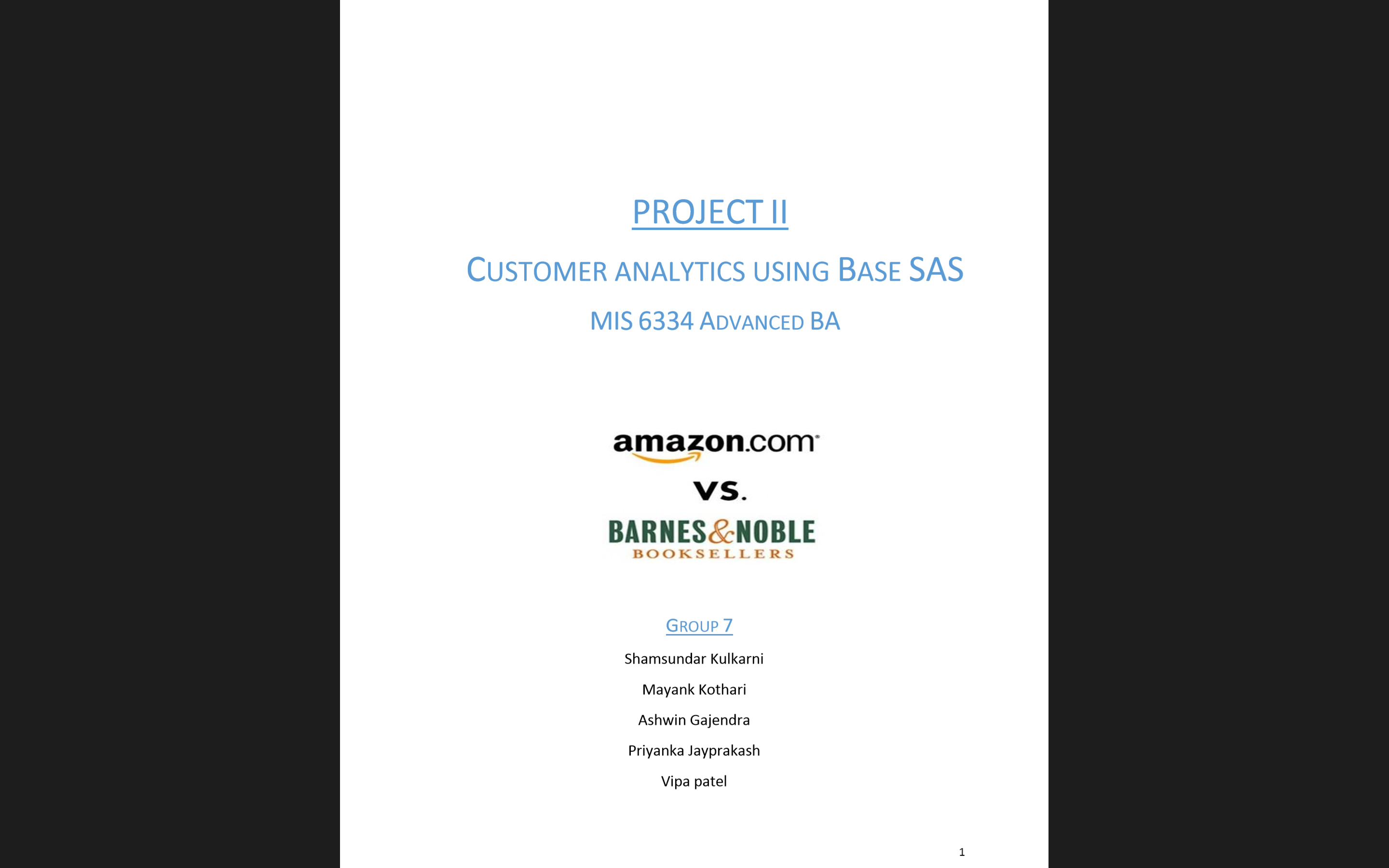
**MIS 6334.002 - Advanced Business Analytics with SAS - F17**

**Instructor: Xianjun Geng**

**Project 2**



**Group Members:** Anusha Sonnad

Sakshi Gaur

Sirisha Bommisetti

Vineet Nayak

**Contents**

**Exploring the dataset............................................................................................................................. 3**

**Part I. Modeling count data................................................................................................................... 6**

**1.1. Creation of Count Dataset ................................................................................................. 6**

**2. NBD Model ............................................................................................................................. 7**

**3. Reach, Average Frequency and GRPs ............................................................................... 10**

**4. Poisson Regression Model ....................................................................................................11**

**5. LL Formula - NBD Regression Model ............................................................................... 13**

**6. NBD Regression Model using customer characteristics....................................................14**

**7. Managerial takeaways between Poisson Regression and NBD Regression.....................16**

**8. Comparing the models - LR Test ....................................................................................... 17**

**Part II. Improving the Model ............................................................................................................. 17**

**9. Variable Selection ................................................................................................................ 17**

**10. Constructing New Variables ............................................................................................ 20**

**11. Interaction Effects ............................................................................................................ 23**

**Part III. Why Certain Customers Prefer Amazon Over BN? ......................................................... 26**

**12. Logistic Regression ............................................................................................................ 26**

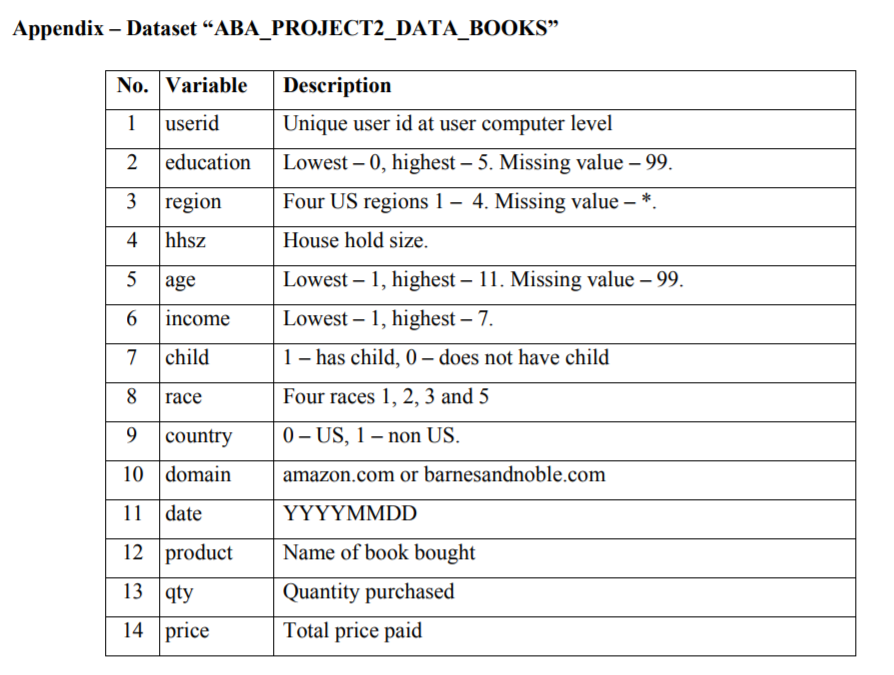
**Part IV. Summary ............................................................................................................................... 27**

**13. Insights and Learnings ...................................................................................................... 27**

**Appendix .............................................................................................................................................. 28**

**Exploring the Dataset**

**Ans.** The project is about the dataset which records customer purchases at two competing book sellers - Amazon vs BARNES&NOBLE for the year 2007. It also records variables and demographics such as education, income, and more. Below is a detailed description of all fields in this dataset.



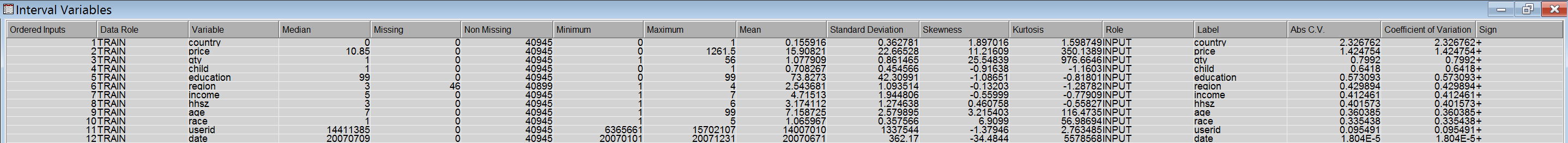
*Figure1: Dataset Appendix*

***NOTE***: *Many data fields have been transformed to protect consumer privacy (done by the company providing this data), which preserves the order of numbers (e.g., value 8 for age implies a larger true age than value 7 for age).*

We used SAS Base, SAS Enterprise Miner and MS Excel for our modelling and analysis of the dataset. Through Excel, we found that the following data are missing from the 40945 records:

1. Education has 30238 missing values: 
2. Region has 46 missing values: 
3. Age has 3 missing values: 

Since SAS-EM could identify only missing values and not values like “99”, it reported only missing values for region as shown below. Thus, we will use SAS Base later to accredit them manually through SAS code.



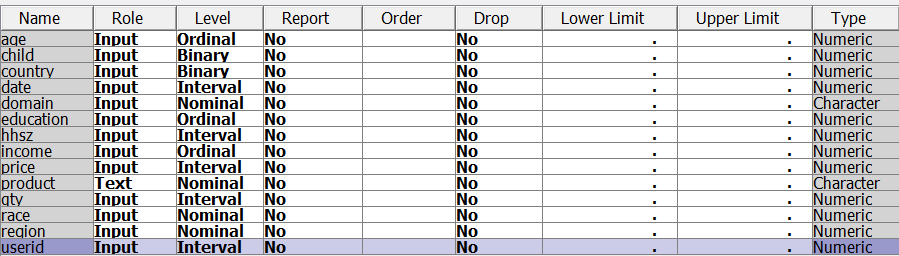
*Figure2: Stat Explore*

We classified the 14 variables as below:

**Demographic variables:** Education, Region, Household size, Age, Income, Child, Race and Country

**Count variables:** Quantity, Price, Product and Domain

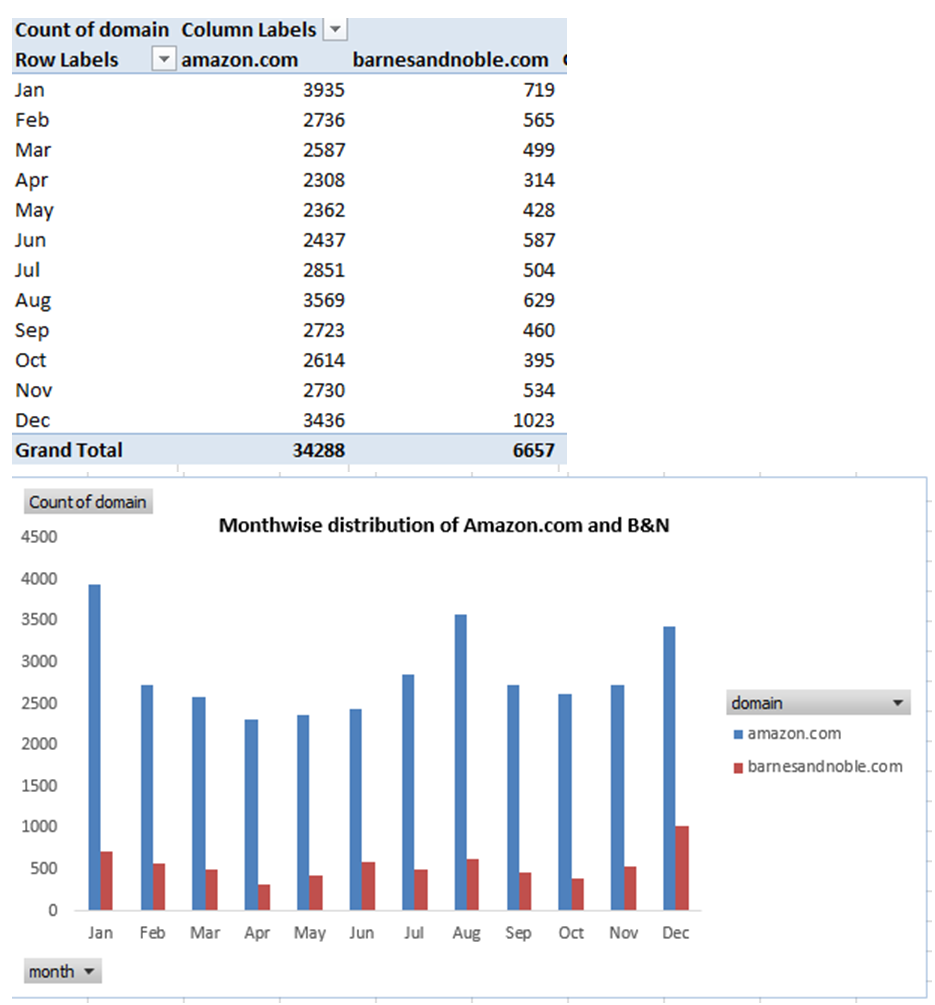
**ID variable:** UserIDand finally a Date variable



*Figure3: Variables*

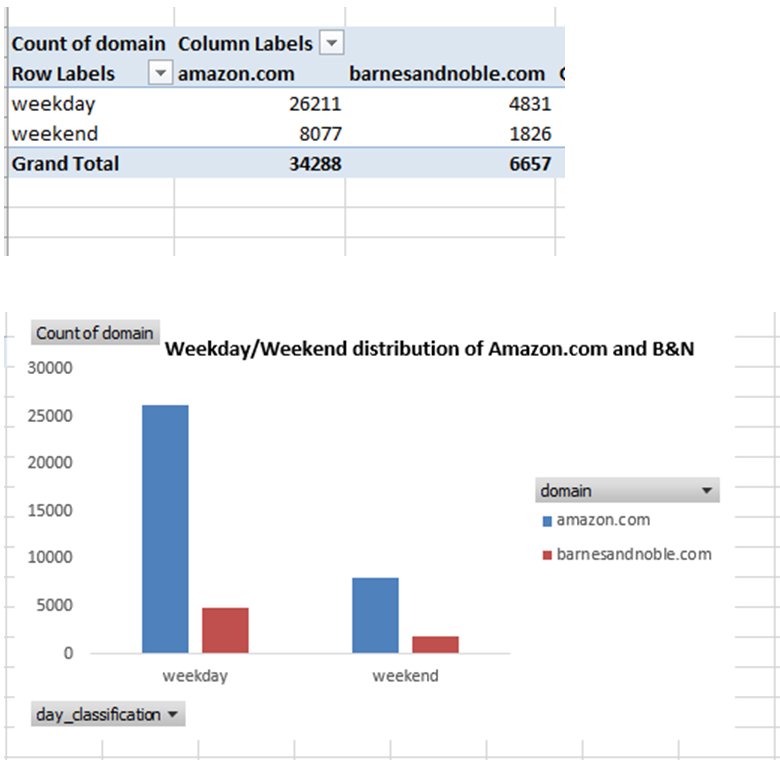
The **Excel analysis** is as shown below:

1. Monthly sales comparison between Amazon and B&N



*Figure4: Monthly Sales*

1. Weekday vs Weekend distribution between Amazon and BN



*Figure5: Weekday vs Weekend distribution*

Inferences from the above 2 graphs:

1. Monthly sales for Amazon and Barnes are high mostly in Jan and Dec, possibly because these months are near holidays, so people tend to buy gifts for their loved ones. But, Barnes has fewer sales compared to Amazon here.
2. Weekend sales are high for Amazon

**Part I. Modeling Count Data**

* 1. **Creation of Count Dataset**

**Q1.** 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.

**Ans:** We created 2 new variables; **totalbooks** which is the sum of quantity of books purchased and **totalprice** which is the sum of the price of books purchased and then we made the count of number of books purchased from B&N:

\*\*\*\*To create table with only b&n data\*\*\*\*

**PROC** **SQL**;

CREATE TABLE aba.bncount as

SELECT userid, education, region, hhsz, age, income, child, race, country, sum(qty) as totalbooks, sum(price) as totalprice

FROM aba.aba\_project2\_data\_books

WHERE domain = "barnesandnoble.com"

GROUP BY userid, education, region, hhsz, age, income, child, race, country;

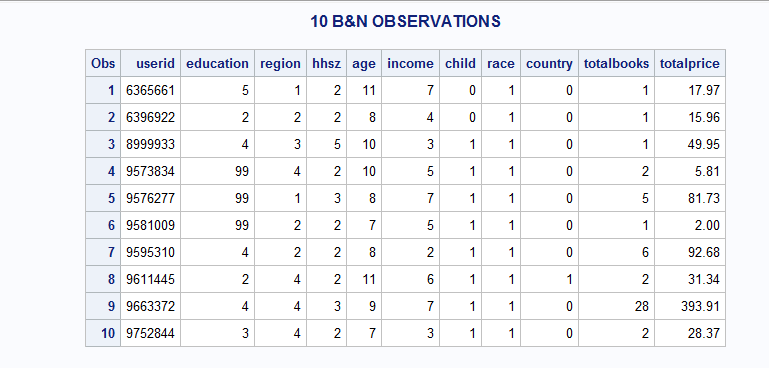
**QUIT**;

\*\*\*\*To fetch only 10 observation of b&n\*\*\*\*

**PROC** **PRINT** DATA=aba.bncount(obs=**10**);

TITLE '10 B&N OBSERVATIONS';

**RUN**;



*Figure6: Ten observations*

* 1. **NBD Model**

**Q2.** 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). (Hint: you will need to create a new dataset similar to the one on slide 5 in the count model lecture.)

**Ans:** We need to come up with a logic that shows the number of people and the purchases they have made from B&N. For instance, 0 number of purchases has been made by x number of people, 1 purchase is made by y number of people and so on.

***Step 1***: First, we need to figure out how many people bought no books from B&N (totalbooks=0). For this we will need the total count of unique user ids who either bought from Amazon, B&N or both. This can be done by using the PROC SQL procedure:

\*\*\*\*To count total records (both amazon and b&n users)\*\*\*\*

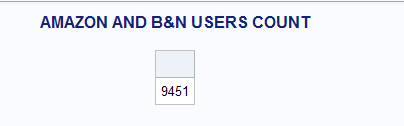
**PROC** **SQL**;

SELECT count(UNIQUE(userid))

FROM aba.aba\_project2\_data\_books;

TITLE 'AMAZON AND B&N USERS COUNT';

**QUIT**;



Next, we find out how many users did not buy from B&N. For this we need a count of who did buy after which we subtract this value from the total:

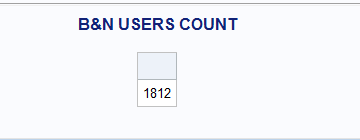
\*\*\*\*To count b&n users;\*\*\*\*

**PROC** **SQL**;

SELECT count(unique(userid)) FROM aba.bncount;

TITLE 'B&N USERS COUNT';

**QUIT**;



Thus, the unique UserIds who did not buy a book from Barnes&Noble are 9451-1812 = **7639**

***Step 2***: Now we create a dataset (based on hint)

\*\*\*\*To create user purchasing frequency of b&n users;\*\*\*\*

**PROC** **SQL**;

CREATE TABLE aba.bnfreq AS

SELECT totalbooks, count(userid) as totalusers

FROM aba.bncount

GROUP BY totalbooks;

INSERT INTO aba.bnfreq values(**0**,**7639**);

**QUIT**;

\*\*\*\*To sort and print the user purchasing frequency of b&n users;\*\*\*\*

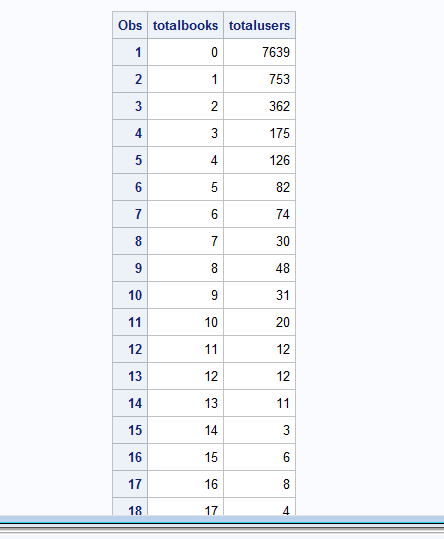
**PROC** **SORT** DATA=aba.bnfreq;

BY totalbooks;

**RUN**;

**PROC** **PRINT** DATA=aba.bnfreq;

**RUN**;



*Figure7: TotalBooks and TotalUsers count*

\*\*\*\*Implementing NBD model;\*\*\*\*

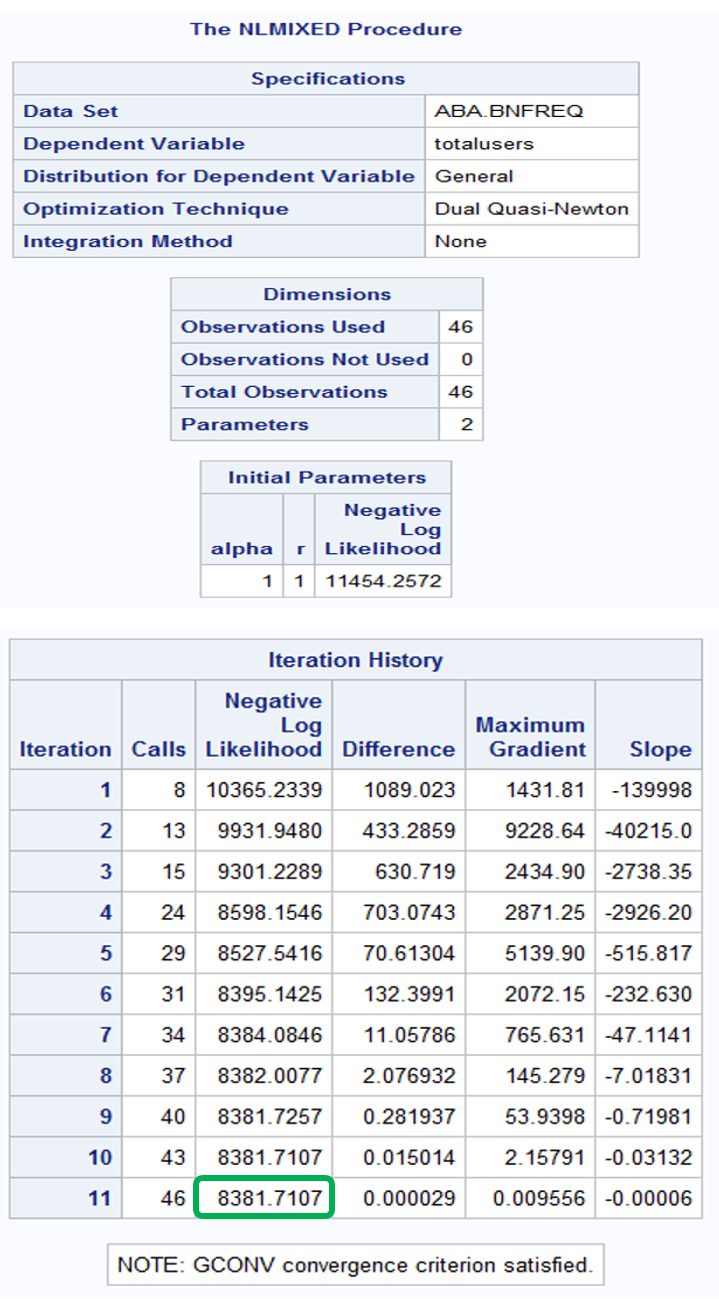
**PROC** **NLMIXED** data = aba.bnfreq;

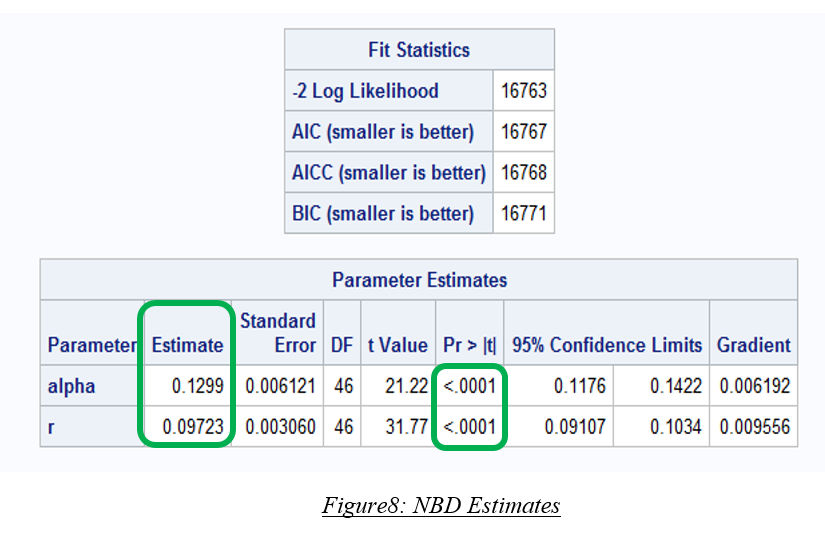
parms alpha=**1** r=**1**; /\*Parameter values used for alpha and r\*/

ll = totalusers\*log((GAMMA(r+totalbooks))/(GAMMA(r)\*fact(totalbooks))\*((alpha/(alpha+**1**))\*\*r)\*((**1**/(alpha+**1**))\*\*totalbooks));

Model totalusers~general(ll);

**run**;





*Figure8: NBD Estimates*

As per the above MLE observations:

1. Maximum LL = -8381.7107
2. α = 0.1299 and r = 0.09723
3. Both p-values are significant
   1. **Reach, Average Frequency and GRPs**

**Q3.** Based on the NBD Model results, report Reach, Average Frequency and GRPs. Show your calculation.

**Ans:**  P(X(t) = 0) = [α/(α+t)] ^r = [0.1299/ (0.1299+1)] ^0.09723 = 0.8103

E(X(t)) = rt/α = (0.09723\*1)/0.1299 = 0.7485

**Reach** = 1 – P(X(t)=0) = 1-0.8103 = 0.1897 =**18.97%**

**Average Frequency** = E[X(t)]/(1-P(X(t)=0)) = 0.7485/0.1897 = **3.95**

**GRPs** = 100 **\*** reach \* avg frequency = 100 \* 0.1897 \* 3.95 = **74.9315**

**[or** GRP = 100\* E(X(t)) = 100\*0.7485= **74.85]**

* 1. **Poisson Regression Model**

**Q4.** 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? (Hint: should you “date” in this regression? Why?)

**Ans:** We first create a dataset and then run the model accordingly:

\*\*\*\*To append a new variable (if domain is amazon assigns 0 else 1);\*\*\*\*

**DATA** aba.bncount;

SET aba.aba\_project2\_data\_books;

IF domain = "amazon.com" THEN DO books\_purchased = **0**;

END;

ELSE DO books\_purchased = **1**;

END;

**RUN**;

\*\*\*\*To create a table with book purchasing count of b&n users;\*\*\*\*

**PROC** **SQL**;

CREATE TABLE aba.newbncount as

SELECT DISTINCT userid, education, region, hhsz, age, income, child, race, country, date, sum(books\_purchased) as totalbooks

FROM aba.bncount

GROUP BY userid, education, region, hhsz, age, income, child, race, country, date;

**QUIT**;

\*\*\*\*Implementing Poisson Regression Model;\*\*\*\*

**PROC** **NLMIXED** DATA=aba.newbncount;

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

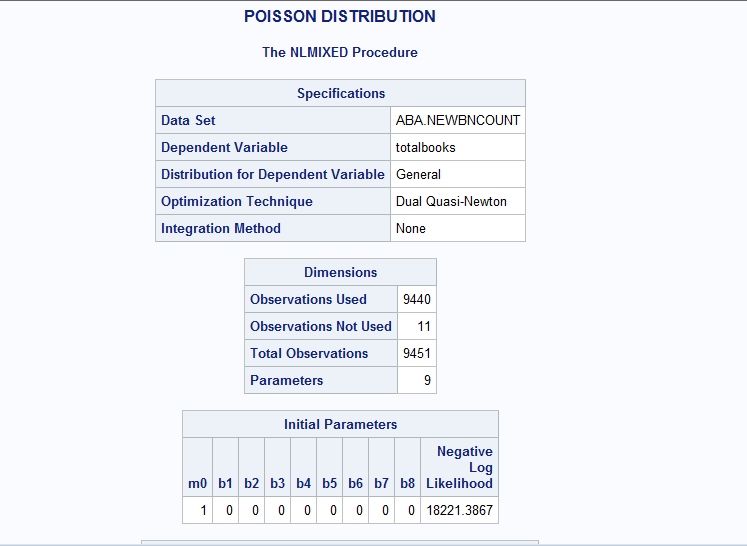
m=m0\*exp(b1\*education+b2\*region+b3\*hhsz+b4\*age+b5\*income+b6\*child+b7\*race+b8\*country);

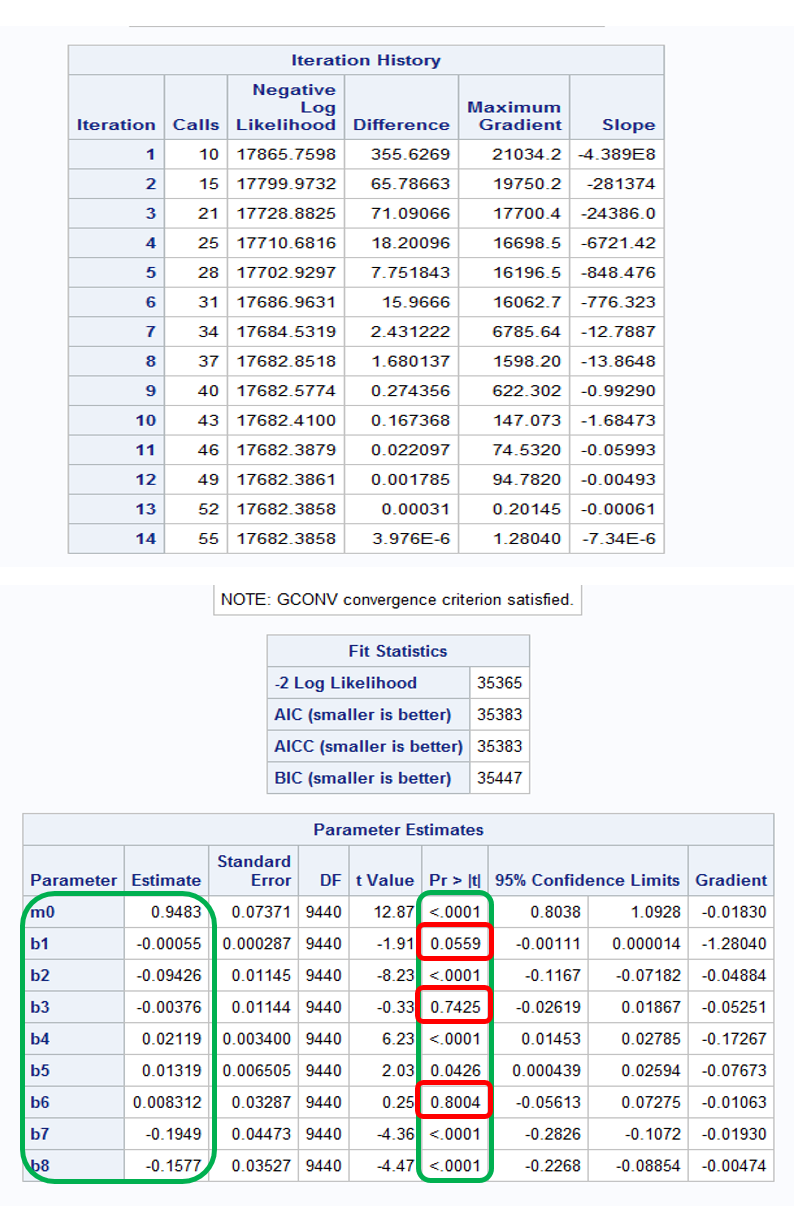
ll = totalbooks\*log(m)-m-log(fact(totalbooks));

MODEL totalbooks ~ general(ll);

TITLE 'POISSON DISTRIBUTION';

**RUN**;





*Figure9: Poisson Regression Estimates*

*MLE Observations:*

* Maximum LL = -17682.3858
* Value of “λ” (m0) = 0.9483 and is significant (with p-value < 0.05)
* b1=education; b2=region; b3=hhsz; b4=age; b5=income; b6=child; b7=race; b8=country
* We can infer that all the variable estimates are significant **except education, household size and child** on the Poisson Model as the former “p-values” are all less than 5% and latter cannot determine purchasing behavior of customers
* Date should not be used because by using date value (as shown in excel earlier), rows are getting duplicated. **Using date value results in a bad model, and can lead to wrong predictions.** Hence, we will not be making use of date in the regression as it cannot add any value to our analysis unless it is broken down into sales on special days like weekends or holidays (vs weekday)

***Managerial Takeaway***

1. Increase in region (b2), race and country (b7 and b8 ordinal values) will decrease the probability of book purchase on Barnes&Noble. So, B&N should put more efforts to attract customers and target segments who live in USA in a small number region, and belongs to a small number race.
2. People who have high incomes or are of older age are more likely to buy books from B&N. That makes sense because people with high incomes are not so sensitive to the price. People with greater ages are more likely to buy books from bookstores directly rather than online. We can put more promotions and advertisement on these customers to gain more of their business.
3. If we increase the region of a customer by 1 level, then customer purchase at Barnes and Noble will ***reduce*** by a factor of ***1.0988 (exp(0.09426)***). Similarly, for each increase in age, customer purchases at Barnes and Noble will also ***increase*** by a factor of 1.1708 (exp(0.1577)) and so on.
   1. **LL Formula - NBD Regression Model**

**Q5.** 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

**Ans:** The LL for NBD is: (“Negbin II” model)



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

**ll = log((GAMMA(r+totalbooks)/(GAMMA(r)\*fact(totalbooks)))\*((alpha/(alpha + m))\*\*r) \*((m/( alpha+m ))\*\*totalbooks));**

* 1. **NBD Regression Model using customer characteristics**

**Q6.** 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?

**Ans:** The NBD Regression is as follows:

\*\*\*\*Implementing NBD regression;\*\*\*\*

**PROC** **NLMIXED** DATA=aba.newbncount;

PARMS alpha=**1** r=**1** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0**;

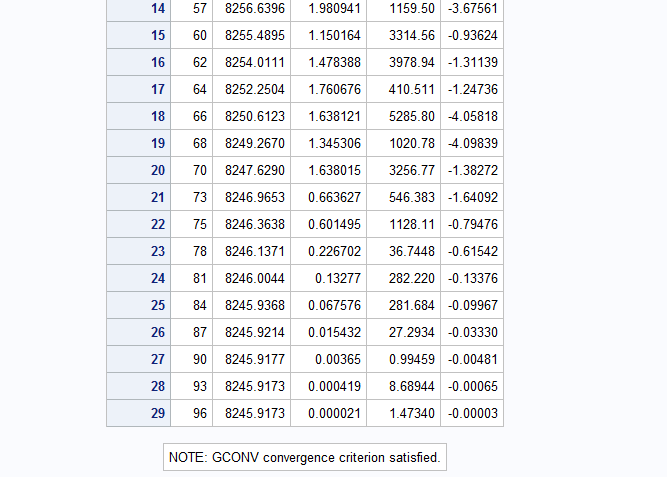
m= exp(b1\*education+ b2\*region + b3\*hhsz + b4\*age + b5\*income + b6\*child + b7\*race +b8\*country );

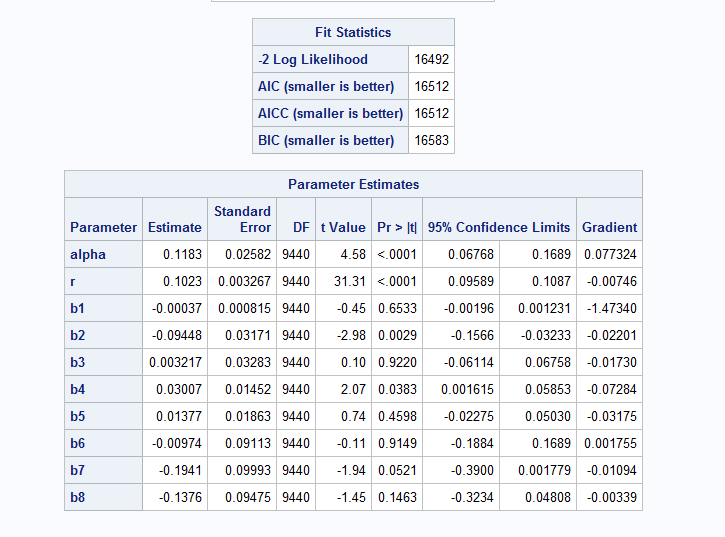
ll = log((GAMMA(r+totalbooks)/(GAMMA(r)\*fact(totalbooks)))\*((alpha/(alpha + m))\*\*r)\*((m/(alpha+m))\*\*totalbooks));

MODEL totalbooks ~ general(ll);

TITLE 'NBD REGRESSION DISTRIBUTION';

**RUN**;





*Figure10: NBD Regression Estimates*

*MLE Observations:*

* Maximum LL = -8245.9173
* Value of “α” = 0.1183 and “r” = 0.1023 and both are significant (with p-value < 0.05)
* b1=education; b2=region; b3=hhsz; b4=age; b5=income; b6=child; b7=race; b8=country
* We can infer that all the variable estimates are significant **except education, household size, income, child, race and country** as the former “p-values” are all less than 5% and latter cannot determine purchasing behavior of customers

***Managerial Takeaway***

1. In this model, the log likelihood is improved however the number of significant variables has reduced.
2. If we increase the region of a customer by 1 level, then probability of book purchase at Barnes and Noble will ***reduce***. Similarly, for each increase in age, customer purchases at Barnes and Noble will also ***increase***, and so on.
   1. **Managerial takeaways between Poisson Regression and NBD Regression**

**Q7.** 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?)

**Ans:** Noticeable difference are:

|  |  |  |
| --- | --- | --- |
|  | **Poisson Regression** | **NBD Regression** |
| **LL** | -17682.3858 | -8245.9173 |
| **Significant variables** | b2 (region), b4 (age), b5 (income), b7 (race), b8 (country) | b2 (region), b4 (age) |
| **Not-significant** | b1 (education), b3 (hhsz), b6 (child) | b1 (education), b3 (hhsz), b5 (income), b6 (child), b7 (race), b8 (country) |

We observe above that there is a significant difference between the two results. Negative log likelihood value of NBD regression model is significantly lower than the Poisson regression model. From this we can infer that NBD Regression is a good model as it has better maximum log likelihood value.

***Possible explanation***: In Poisson regression model, we considered only how people differ on a set of available explanatory variables (We keep the lambda constant), while in NBD regression model, we try to capture the unobserved variables which might influence the target variables along with the available explanatory variables. (i.e. we varied the λ0 across the customers according to a gamma distribution with parameters r and α).

* The change in output of NBD regression might be a phenomenon of omitted variables’ bias (unobserved variables). These omitted variables might be correlated with the included independent variables and the dependent variable. In other words, in the regression the error term might be correlated with the independent variables and hence the difference
* This means that probability of predicting the actual data, given the parameters, is higher for NBD model than the Poisson regression. Although the log likelihood has improved in this NBD model, the number of significant variables has drastically reduced compared to that of Poisson regression model

***Why the difference***: Since this data is over-dispersed, variance is higher than the mean. Poisson regression assumes that mean and the variance are similar, but here it is not the case. On the other hand, NBD model has one additional parameter alpha which can be used to adjust the variance independent of the mean. NBD model takes less number of significant variable and fits to the data in a better manner and this along with dispersion parameter makes NBD model work better for such discrete data.

* We can arrive at model B by placing k constraints on the parameters of model A (we can say model B is nested within the model A). In model A: parameters α,γ capture large amount of differences among observations. This makes parameters b1,b3,b5,b6,b7,b8 capture less difference in model A. So, they become not significant in model A (NBDR) while they are significant in model B (PR).
* The probability distribution of two models:
  + Model B

* + Model A



* 1. **Comparing the models - LR Test**

**Q8.** 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.)

**Ans:** The null hypothesis is that model A is not different from model B

Computing test statistic: Likelihood Ratio = −2[LL(Poisson) –LL(NBD)]

**LR = −2(LLB − LLA)**

~= -2(-17682-(-8245))

**LR = 18874**

χ2 (.05, k) = χ2 (.05,1)

**χ2 =3.8415**

Since, LR > χ2 (.05, k), we ***reject null hypothesis***. Hence, we state that Model A is better than model B. Therefore, **NBD model fits and performs better than Poisson regression model**:

χ2 value for p=0.05 and df=1 (Since the NBD Regression has 1 extra parameter as compared to Poisson Regression i.e. alpha and r instead of m0) and considering λ0 constant as a constraint (k=1) for Poisson regression

**Part II. Improving the Model**

* 1. **Creation of Count Dataset**

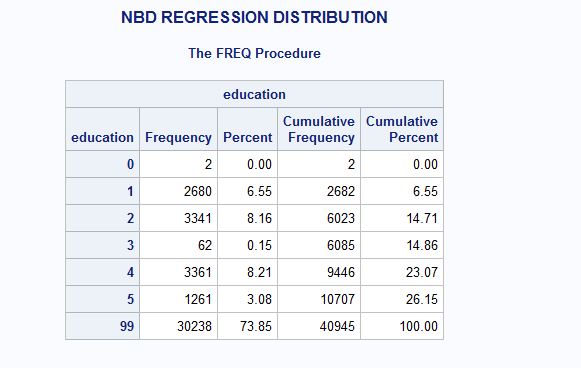
**Q9.** Similar to what you found out in Project 1, not all variables are always useful. Please try feature selection (i.e. selecting only a subset of customer characteristics), and report your findings. (Hint: You can use Enterprise Miner to get some ideas on which variables to keep/remove, or, you can use the built-in variable selection mechanisms in SAS statistical procedures.)

**Ans**: From above question calculatios, we came to know education as a variable was not very significant every time. Hence, we tried finding variable worth using SAS EM.



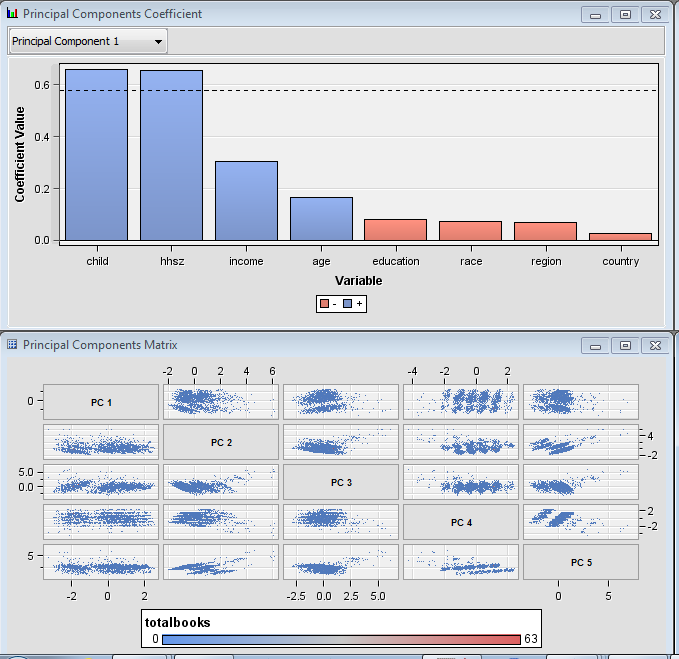
*Figure11: Variable worth screenshot*

It shows age, region, education, income, hhsz, race and country are all worthy variable. But if we closely observe the frequency distribution of education it shows that it has >50% missing values. Thus, education cannot be considered as worthy variable. Even if we try to impute missing values of education with mean (or median) it will induce skewness in the dataset.



*Figure12: Education variable evaluation*

Also, just to verify the same, we decided to run PCA on the variables. And we got the similar worthy variables (here, we rejected country as well it is the subset of variable region).



*Figure13: PCA analysis*

Now, we run NBD model distribution on just 4 worthy variables: age, region, income, hhsize (in NBD model, the p-value for child is >0.05, thereby this variable is measured insignificant)

\*\*\*\*Implementing NBD on only 4 variables;

**PROC** **NLMIXED** DATA=aba.newbncount;

PARMS alpha=**1** r=**1** b1=**0** b2=**0** b3=**0** b4=**0**;

m= exp(b1\*age+ b2\*region + b3\*income + b4\*hhsz);

ll = log((GAMMA(r+totalbooks)/(GAMMA(r)\*fact(totalbooks)))\*((alpha/(alpha + m))\*\*r)\*((m/(alpha+m))\*\*totalbooks));

MODEL totalbooks ~ general(ll);

TITLE 'NBD REGRESSION FOR MOST FREQUENT VARIABLES';

**RUN**;



*Figure14: Improvised model*

There is not much improvement after variable reduction:

Old negative LL ~= **8245**

New negative LL (modified by dropping few variables) = **8248.8107**

Thus, this transformation is of no use and is being rejected.

* 1. **Constructing New Variables**

**Q10.** 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?

**Ans**: Considering the “Date” parameter in our analysis, date can be further broken down into Weekdays and Weekends to give more meaning or understandability. We can draw insights if the purchases were divided into weekdays and weekends.

We created 4 new variables: (a) weekend =1 or 0 (means it is a weekend or not)

(b) quarter = 1-4 (indicating the quarter of the year)

(c) month = 1-12 (indicating the month number)

(d) holidaytime = 0 or 1 (indicating whether a person bought in Nov /Dec)

\*\*\*\*Performing date conversion;\*\*\*\*

**DATA** aba.newbncount;

SET aba.newbncount;

newdate = input(put(date, Z8.), yymmdd8.);

put newdate downame.;

**RUN**;

\*\*\*\*assigning 1 for weekend and 0 for weekday;\*\*\*\*

**DATA** aba.newbncount;

SET aba.newbncount;

IF weekday(newdate)=**7** or weekday(newdate)=**1** THEN weekend = **1**;

ELSE weekend = **0**;

**RUN**;

\*\*\*\*assigning nov and dec as holidaytime;\*\*\*\*

**DATA** aba.newbncount1;

SET aba.newbncount;

quarter = qtr(newdate);

month = month(newdate);

IF month = **11** or month = **12** THEN holidaytime = **1**;

ELSE holidaytime = **0**;

**RUN**;

\*\*\*\*implementing NBD with new date variables;\*\*\*\*

**PROC** **NLMIXED** DATA=aba.newbncount1;

PARMS alpha=**1** r=**1** b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** b6=**0** b7=**0** b8=**0** b9=**0** b10=**0** b11=**0**;

m= exp(b1\*region+ b2\*hhsz + b3\*age + b4\*income +b5\*child + b6\*race + b7\*country + b8\*weekend + b9\*quarter + b10\*month + b11\*holidaytime);

ll = log((GAMMA(r+totalbooks)/(GAMMA(r)\*fact(totalbooks)))\*((alpha/(alpha + m))\*\*r)\*((m/(alpha+m))\*\*totalbooks));

model totalbooks ~ general(ll);

**RUN**;

We can see that before adding new variables, negative Log LL for NBD model was ~= -8245, but after adding new variables, **LL for NBD is -12026.0542** (as shown below in Figure15)

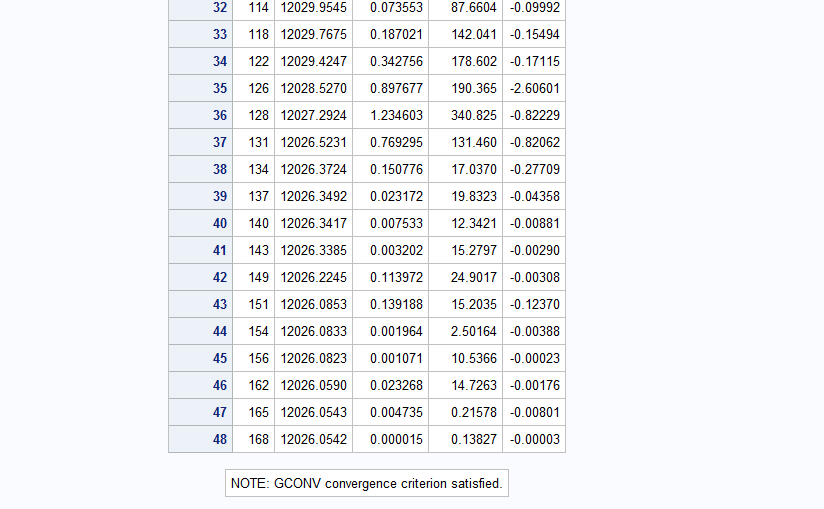
**Likelihood Ratio = −2[LL(m1) – LL(m2)]**

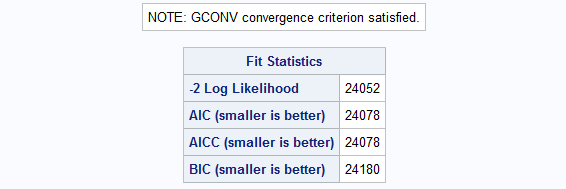
= -2((-12026.0542)-( -8245.9173) )

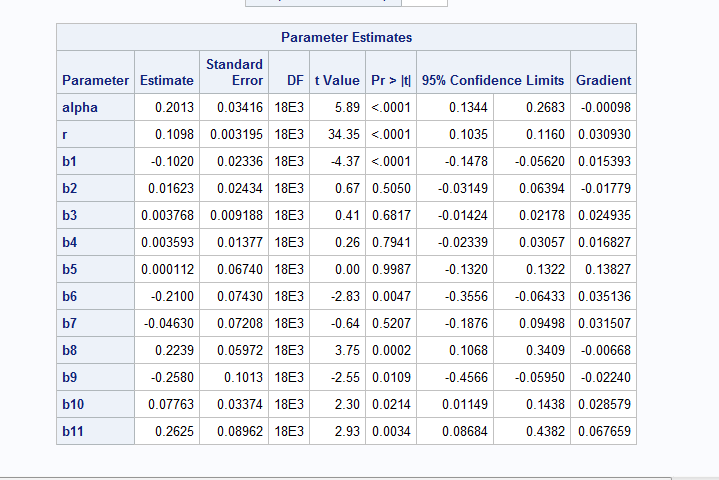
= -2(-3780.1369)

= **7560.2738**

* χ2 value for p=0.05 and df=4 (Since we added *4 new variables* in the analysis) is **9.488**







*Figure15: Improvised model with new variables*

***Managerial Takeaway***

* Since, LR > χ2 (.05, k), we ***reject null hypothesis***. Hence, we state that the NBD Model did not perform better after adding new date variables as compared to the model with lesser variables
* The new variables are all significant, but few older variables remain insignificant
* This unexpected outcome occurred may be because of shopaholic users (who bought books on weekend as well as weekdays), which results in duplicate user count
  1. **Interaction Effects**

**Q11.** 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.

**Ans**: We tried few interactions as shown below:

1. **Age and Income**

**Proc** **NLMIXED** DATA=project2.bn;

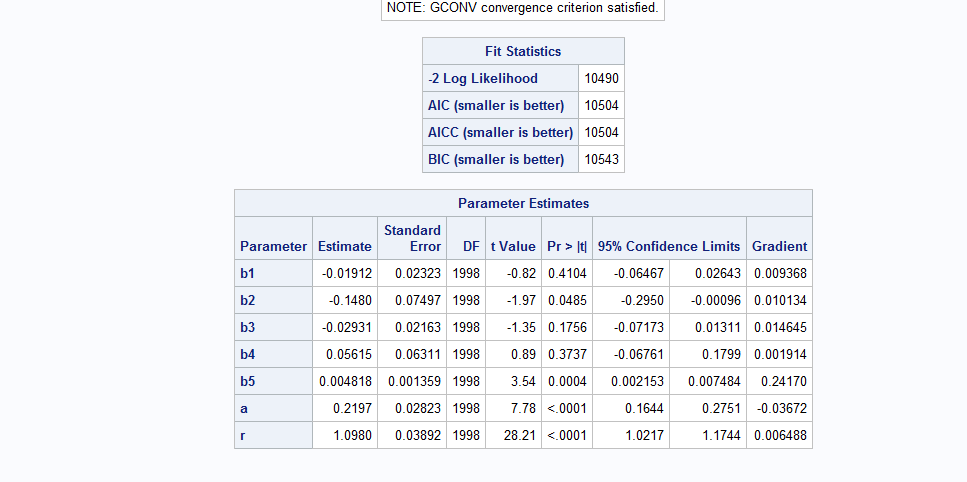
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*REGION+b4\*CHILD+b5\*AgeIncome);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



*Figure16: Interaction between Age and Income (b5)*

From the above results, we can say that the interaction of AGE and INCOME gives us significant value because the p-value of b5 (age\*income) is 0.0004 (< 0.05)

**(b) Age and Child**

**Proc** **NLMIXED** DATA= aba.newbncount;

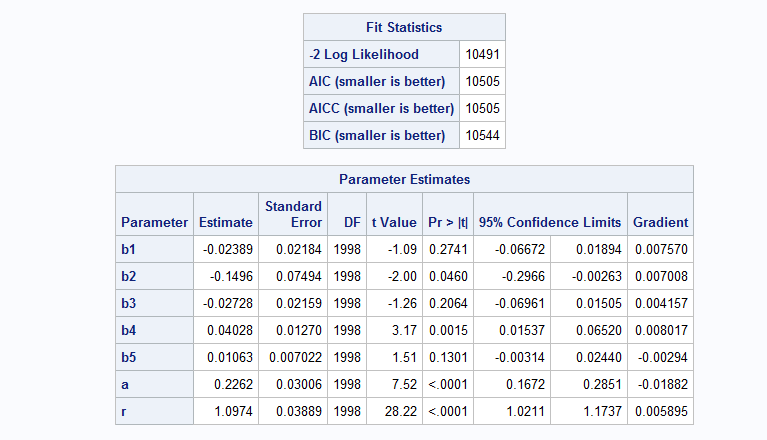
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*REGION+b4\*Income+b5\*AgeChild);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



*Figure17: Interaction between Age and Child (b5)*

From the above results, we can say that the interaction of AGE and CHILD does not give us significant value because the p-value of b5 (age\*child) is 0.1301 (> 0.05)

**(c) Child and Region**

**Proc** **NLMIXED** DATA= aba.newbncount;

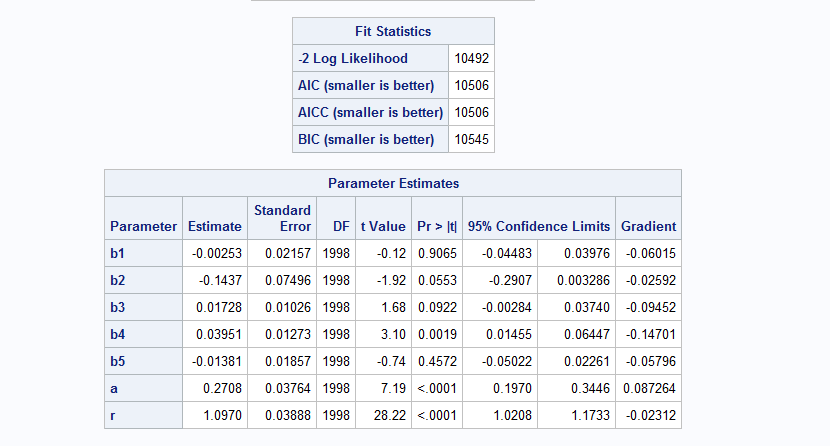
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*AGE+b4\*Income+b5\*ChildRegion);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



*Figure18: Interaction between Child and Region*

From the above results, we can say that the interaction of CHILD and REGION does not give us significant value because the p-value of b5 is 0.4572 (> 0.05)

**(d) Region and Income**

**Proc** **NLMIXED** DATA= aba.newbncount;

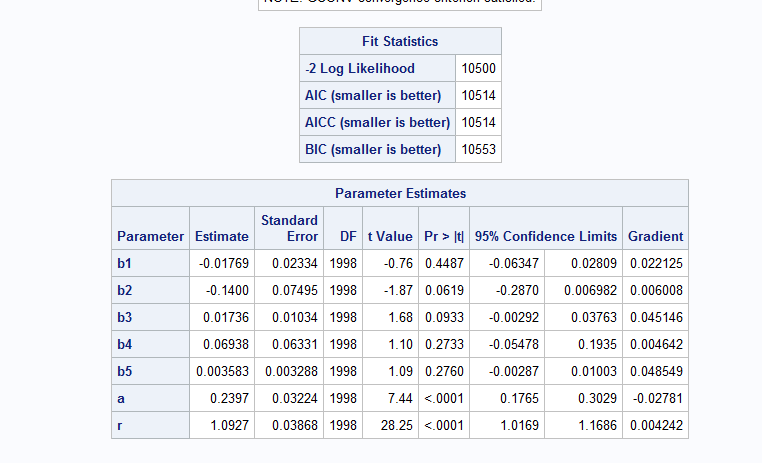
parms b1=**0** b2=**0** b3=**0** b4=**0** b5=**0** a=**0.1** r=**0.1**;

m=exp(b1\*HHSZ+b2\*RACE+b3\*AGE+b4\*Child+b5\*RegionIncome);

ll=log(gamma(r+COUNT\_BN)/(gamma(r)\*fact(COUNT\_BN))\*(a/(a+m))\*\*r\*(m/(a+m))\*\*COUNT\_BN);

MODEL COUNT\_BN~general(ll);

**run**;



*Figure19: Interaction between Region and Income*

From the above results, we can say that the interaction of REGION and INCOME is not significant because the p-value of b5 is 0.2760 (> 0.05).

**Part III. Why Certain Customers Prefer Amazon Over BN?**

* 1. **Logistic Regression**

**Q12.** 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 3 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?).

**Ans**: **Data** aba.logistic;

set aba.Bncount ;

if BarnesNoble > **0** then flag\_bn = **1**;

else flag\_bn = **0**;

if AMAZON > **0** then flag\_am = **1**;

else flag\_am = **0**;

**run**;

**proc** **logistic** data = aba.logistic;

model flag\_bn = REGION HHSZ AGE INCOME CHILD RACE COUNTRY/expb;

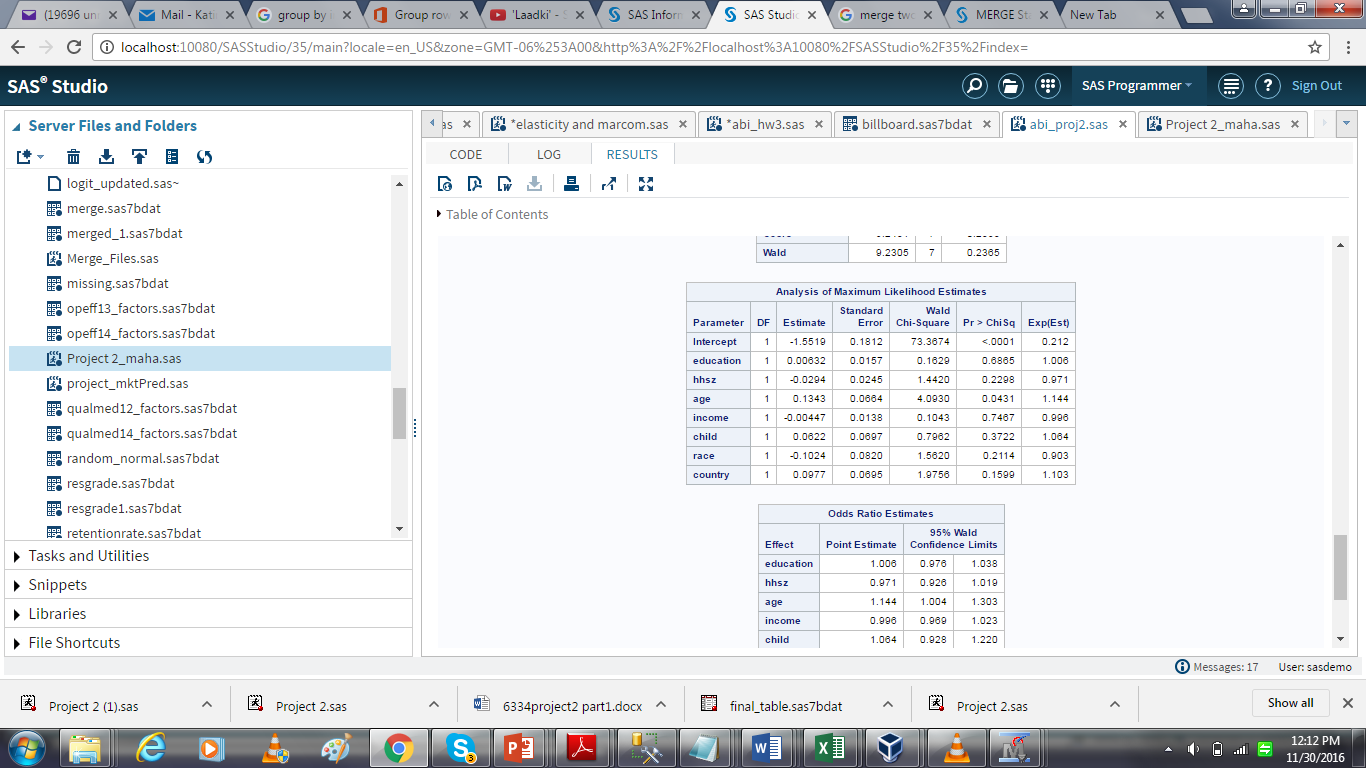
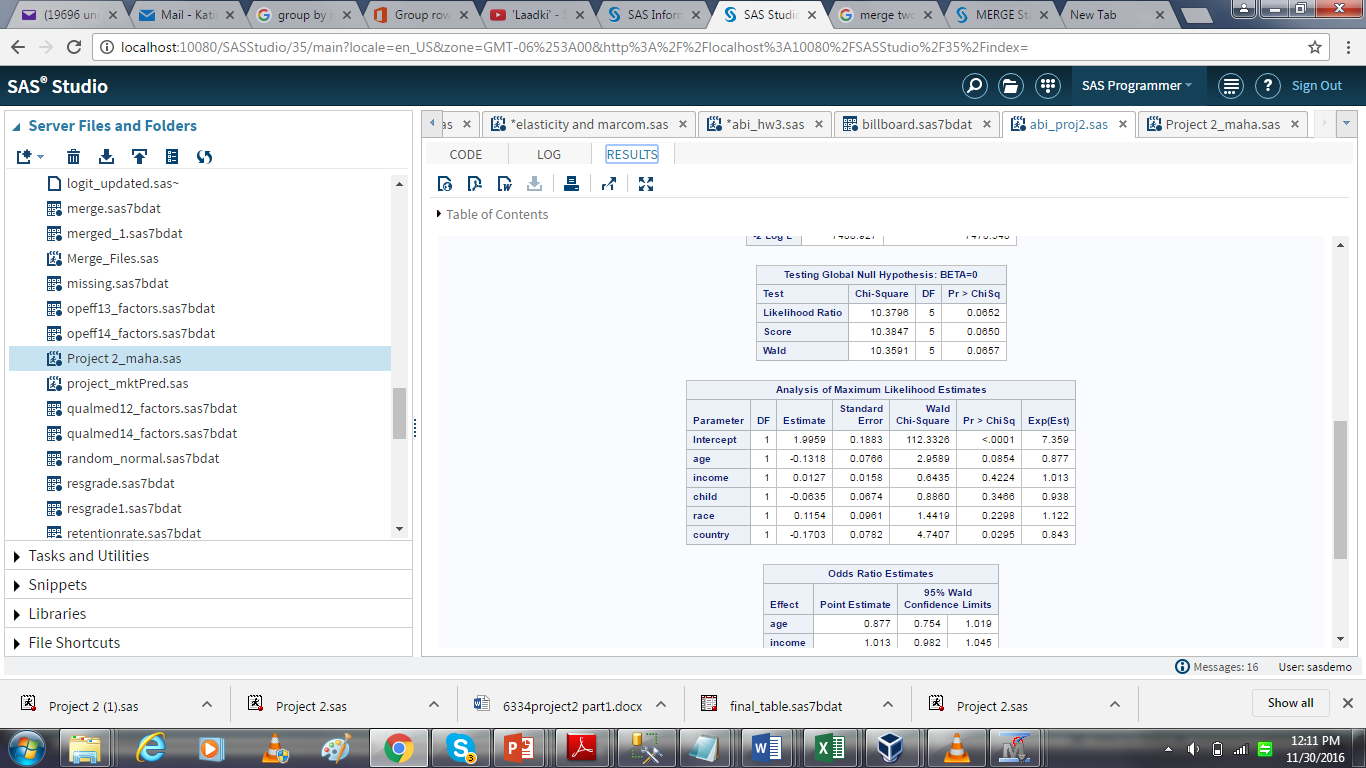
**run**;

**proc** **logistic** data = aba.logistic;

model flag\_am = REGION HHSZ AGE INCOME CHILD RACE COUNTRY/expb;

**run**;

*Figure20:* Barnes And Noble VS Amazon

***Managerial Takeaway***

* From the above results, we can easily state that that for Barnes & Nobel “Age” is the only variable that is significant whereas for amazon both the variables “Age” and “Country” are significant
* This result is justified because Amazon is a bigger firm than Barnes and Nobel which not only sells book but other articles as well and it I famous across other countries as well whereas a Barnes is an in-store branch which majorly operates in the US
* Younger people prefer Amazon over BN  People with high income as compared to others tend to buy from Amazon  People from U.S. prefer Amazon over BarnesandNobles  People in Region 4 tend to buy more from Amazon as compared to people from Region 1-3. This might be because BarnesandNobles has not many operations in Region 4 and Amazon dominates BarnesandNobles in that Area

**Part IV. Summary**

* 1. **Insights**

**Q13.** 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.

**Ans**:

In an era where data is the new gold, companies are thereby focussed more on analytics as they can make use of the abundant data collected and build complex analytics model to target the right customer and predict growth of the company based on different parameters. Customer analytics is the branch of analytics which help companies to understand their customer better so that they can target customers aggressively by designing targeted marketing campaigns or targeted offers for different types of customer.

* In this project, we got better understanding of how different demographics factor impact the purchasing behaviour of the customer. Building customer analytics model based on the raw data and improving it further incorporating models such as NBD and Poisson-Regression with new custom variables explained us how to interpret data and make changes to our model and raw data based on new findings and better understanding of data.
* Extensive use of PROC SQL served better understanding of difficulties faced in real world projects to create custom datasets as per the business problems and how to model the dataset based on requirement and intuition for better performance of the model. We also used different functions such as ‘weekday’ to have better insights for the customer purchasing behaviour.
* Learning curve during this was exponential as we not only learned how to implement different customer analytics model such as NBD, poison regression and choice model i.e. logistic regression but we also developed a better understanding of all the nitty gritty involved in these models and improving models by changing different parameters.
* With the new generation opting for online retail options, Barnes should invest and market this to get the young crowd to purchase from Barnes. This is evident from the model that older people buy buys from Barnes while the younger people prefer Amazon.
* Barnes may want to consider marketing strategies and more of sale offers to customers. One of our analysis also showed that no new customers bought from Barnes during the second half of the semester and only the recurring customers purchased from Barnes. So they may want to introduce strategies to attract new customers.
* Learned that even though we think that we can explain the dependent variable using all possible independent variables, we are wrong; we always miss on some aspects or variables that can explain the dependent variable
* Understood the difference between Poisson Regression and NBD Regression Model and how to compare the models using LR test. Learnt about the various count models and choice models. Dataset is biased towards amazon which leads to very small data part to be worked upon for findings on BN

To conclude, working on this project helped us to consider the business insights. It helped everyone individually to attain the Business Intelligence skills which we are sure to help us a lot in the longer run and open our doors to pursue a career in Business Analytics field.

**Appendix**

Additional things that were tried which were not of any value:

1. Using Regression in SAS enterprise miner to determine variable significance (stepwise)
2. The Contour plot to find correlation amongst variables:

proc glm data=WeekQtyData;

model totOrders= totPrice|weQty / solution;

store glm\_totOrders;

run;

