PROJECT 2 – CUTOMER ANALYTICS



# Executive Summary:

In this project we have built customer analytics model by using base SAS on a dataset having customer purchase records of two competing book sellers: Amazon and Barnes and Noble (BN) along with the customers’ demographics. The objective of our analysis was to build customized BI model to fit existing data in order to make business predictions and to identify the consumer characteristics that drive purchase behavior for Barnes and Nobel over Amazon products.

The dataset was first cleaned of missing values by replacing them with appropriate values. Since we were working towards finding customer purchasing behavior at Barnes and Noble (BN), we considered only the purchase made from BN and considered all Amazon purchases as 0. We then created unique records for each customer and added a variable named Quantity (Total\_count) representing the total purchase made by a customer. This variable: Total\_count was used as target variable for all our models. For each customer, we considered the earliest purchase date as the Date variable.

We implemented different models on the processed dataset and analyzed the results in detail to determine what managerial implications they have. We compared the models and deliberated on the probable reason of their difference. We also implemented some model enhancement techniques and Logistic Regression to determine the odds ratio of a customer buying books from BN.

We have concluded the report with a section on all the managerial insights that we gained from the project and how such information can be used to make business strategies. We have also included a section on our learnings and the over-all experience of working on the project, at the end of the report.

Contents

[Executive Summary: 1](#_Toc436808256)

[Data Pre-processing 3](#_Toc436808257)

[Part I: Modeling Count Data 5](#_Toc436808258)

[Part II: Improving the Model 18](#_Toc436808259)

[Part III: Why certain customers prefer Amazon over BN 36](#_Toc436808260)

[Part IV: Summary 38](#_Toc436808261)

# Data Pre-processing

1. Missing Value Handling

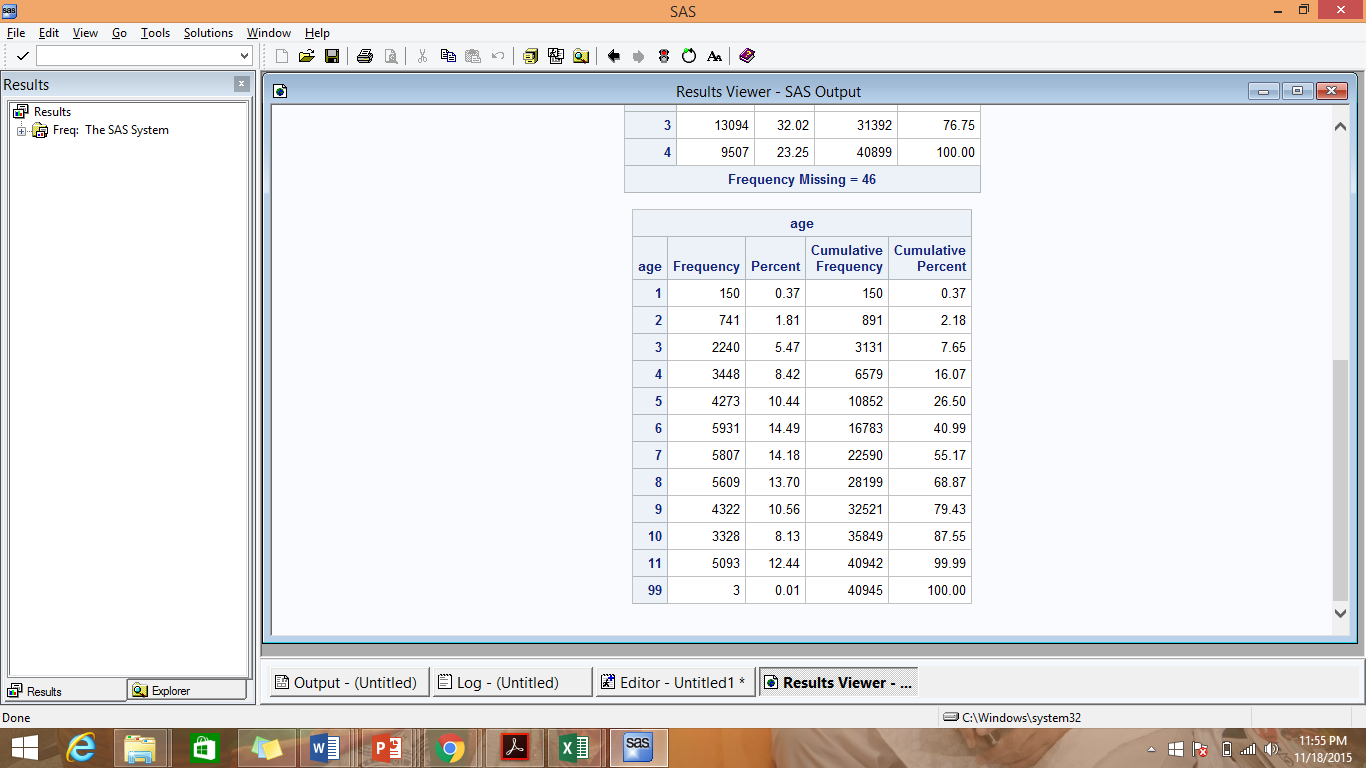
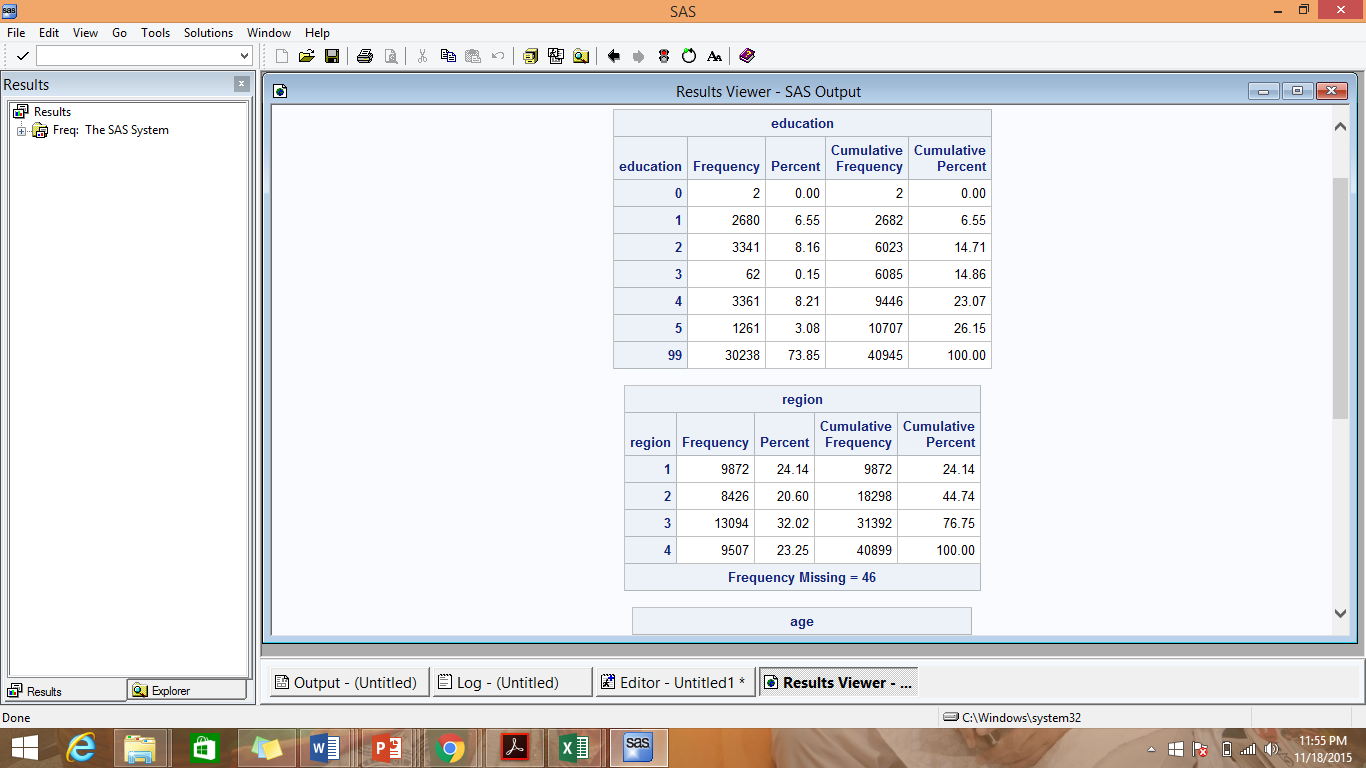
We observed that out of the all the variables, education, region and age have missing values which are represented by 99, space and 99 respectively.

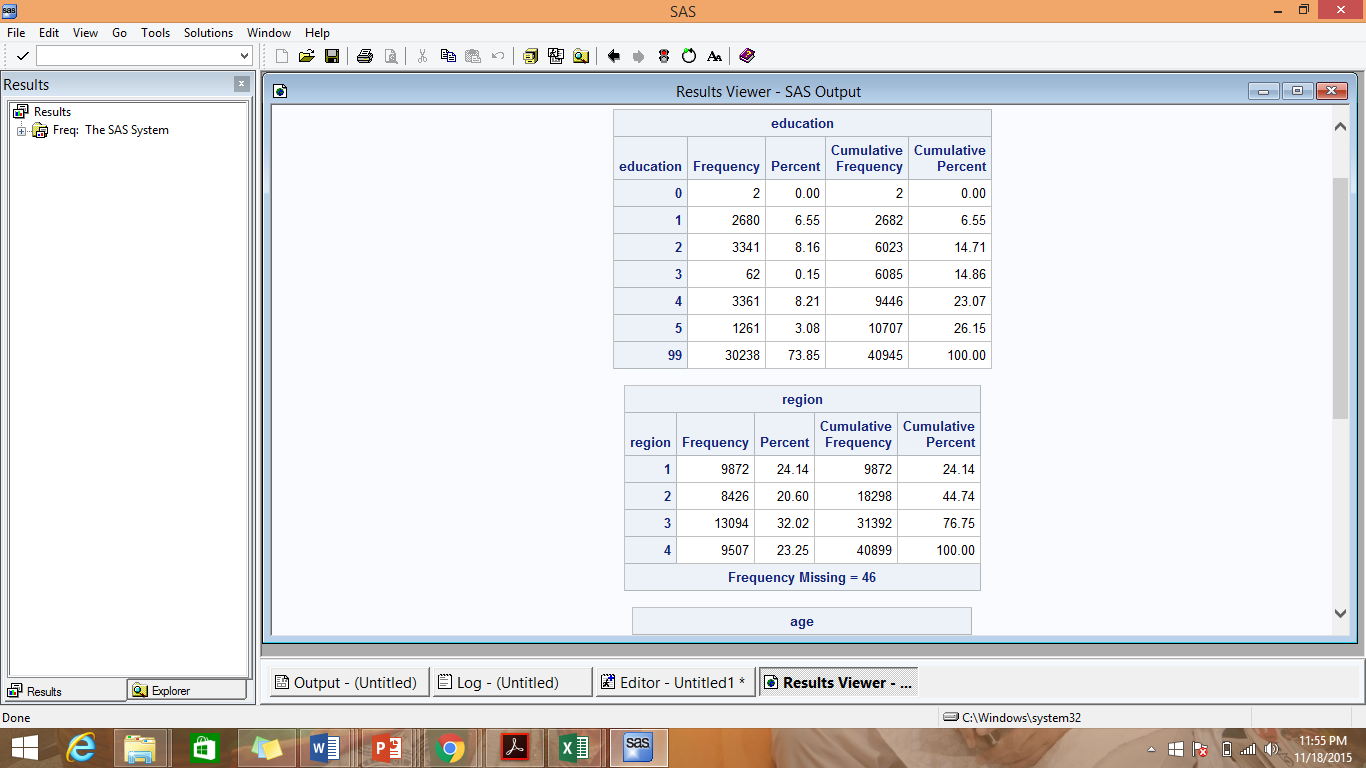
We saw that the number of missing values for education, age and region are 30238, 3 and 9507 records respectively.

As all the mentioned variables are categorical, we decided to replace the missing value data with the most frequently occurring value. We wrote the following code to determine the most frequently occurring value for each of these 3 variables.

***proc freq data = mis6334.abi\_project2\_data\_books;***

***tables education region age;***





We observed that the most common value for education, age and race are 4, 6 and 3 respective and we thus replaced the missing values with these value.

***data abi.imputedata;***

***set mis6334.abi\_project2\_data\_books;***

***if education = 99 then education = 4;***

***if age = 99 then age = 6;***

***if region = '' then region = 3;***

***run;***

1. Handling Amazon data:-

As the objective of the project is to understand the factors that affect customer purchasing behavior at Barnes and Noble(BN), we labelled the ‘qty’ for all Amazon.com records as 0. Thus all customers who bought books only from Amazon.com, their total count of books purchased would be 0 and would be equivalent to customers with 0 BN books in the count model. For customers who bought books from both BN and Amazon, we have considered only the count for BN and ignored the count for Amazon.

***data abi.bnCheck;***

***set abi.imputedata;***

***if domain = 'amazon.com' then qty = 0;***

***run;***

# Part I: Modeling Count Data

1. We used sql to create a dataset counting the number of books a customer bought from BN in 2007. All customers who have bought books only from Amazon would have count as 0.

The date field in the data is the date on which a customer did each purchase. As we created a database containing unique records for each customer displaying the total number of books bought by them, we selected the first date for each customer, i.e. we selected the date on which the customer bought a book for the first time.

However, for customers who bought books from both BN and Amazon, we selected the date from a BN purchase, i.e the date on which the customer bought a book from BN for the first time. This was done because, for customers having bought books from both Amazon and BN, we have not considered the amazon purchases.

The following code creates a database with unique records of customers along with their demographic information, total count of books purchased from BN. However customers who have bought books from both Amazon and BN have 2 records in the following dataset, each record displaying the first date of purchase from each of the 2 domains

***PROC sql;***

***create table abi.tempdata1 as***

***select userid, education, region, hhsz, age, income, child, race, country, domain, min(date) AS DATE, sum(qty) as total\_count from abi.bncheck***

***group by userid, education, region, hhsz, age, income, child, race, country, domain;***

***QUIT;***

***run;***

We now split the dataset into 2; 1 for each of the 2 domains; Amazon and BN.

The following code creates a dataset for BN:

***PROC sql;***

***create table abi.tempdatabn as***

***select userid, education, region, hhsz, age, income, child, race, country, domain, DATE, TOTAL\_COUNT***

***from abi.tempdata1***

***where domain = 'barnesandnoble.com';***

***quit;***

***run;***

The following code creates a dataset for Amazon:

***PROC sql;***

***create table abi.tempdataamazon as***

***select userid, education, region, hhsz, age, income, child, race, country, domain, DATE, TOTAL\_COUNT***

***from abi.tempdata1***

***where domain = 'amazon.com';***

***quit;***

***run;***

The following code merges the 2 datasets, with all records present in BN dataset and only those from the Amazon dataset who are not present in the BN dataset. We thus now have a dataset with ‘Date’ for customers who have bought from both Amazon and BN as the date ***of their first purchase from BN.***

***PROC sql;***

***create table abi.tempdata2 as***

***select userid, education, region, hhsz, age, income, child, race, country, DATE, TOTAL\_cOUNT***

***from abi.tempdatabn***

***union***

***select userid, education, region, hhsz, age, income, child, race, country, DATE, TOTAL\_COUNT***

***from abi.tempdataamazon***

***where userid not in (select userid from abi.tempdatabn);***

***QUIT;***

***run;***

We also observed that the date field in the dataset is in the form YYYYMMDD. We used the following code to convert date to the SAS date format. The attribute Date\_val contains the SAS dates.

***Data abi.Permdata;***

***Set abi.tempdata2;***

***Date\_val = input (put(DATE, 8.), yymmdd8.);***

***Put Date***

***@10 date\_val date9.***

***@20 date\_val 6.;***

***Run;***

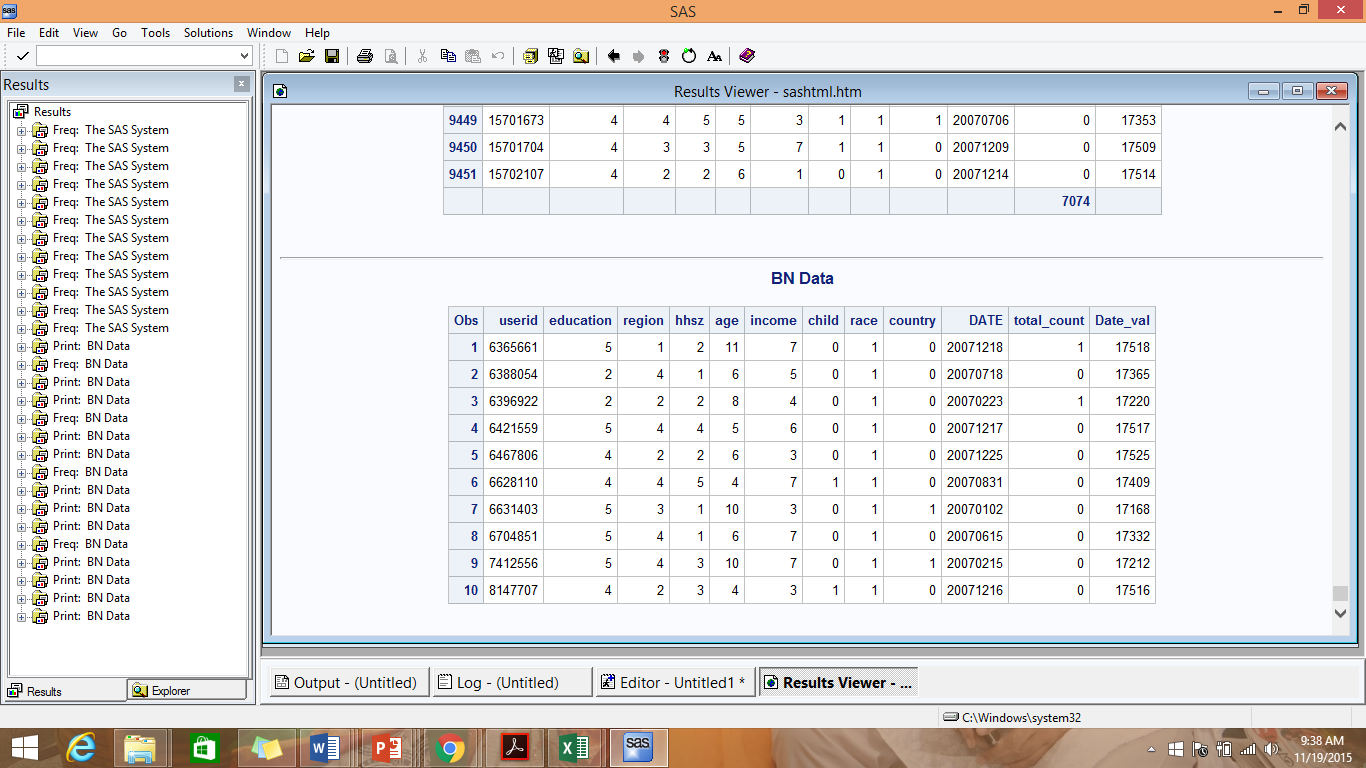
The following code prints the first 10 observations of the created dataset:-

***proc print data = abi.Permdata (obs=10);***

***title "BN Data";***

***run;***

*Note*:- Since the dataset contains data for only 2007, there was no need to add any filter for that.



1. NBD Model:-

***proc freq data = ABI. Permdata;***

***/\*Creates a table with total\_count as the number of books bought from BN and count as the number of customers corresponding to each number of books bought\*/***

***table total\_count / out = q2table;***

***run;***

***proc nlmixed data=q2table;***

***/\* the 2 parameters of the NBD Model (Poisson-Gamma) - r for shape and a for scale\*/***

***parms r=1 alpha=1;***

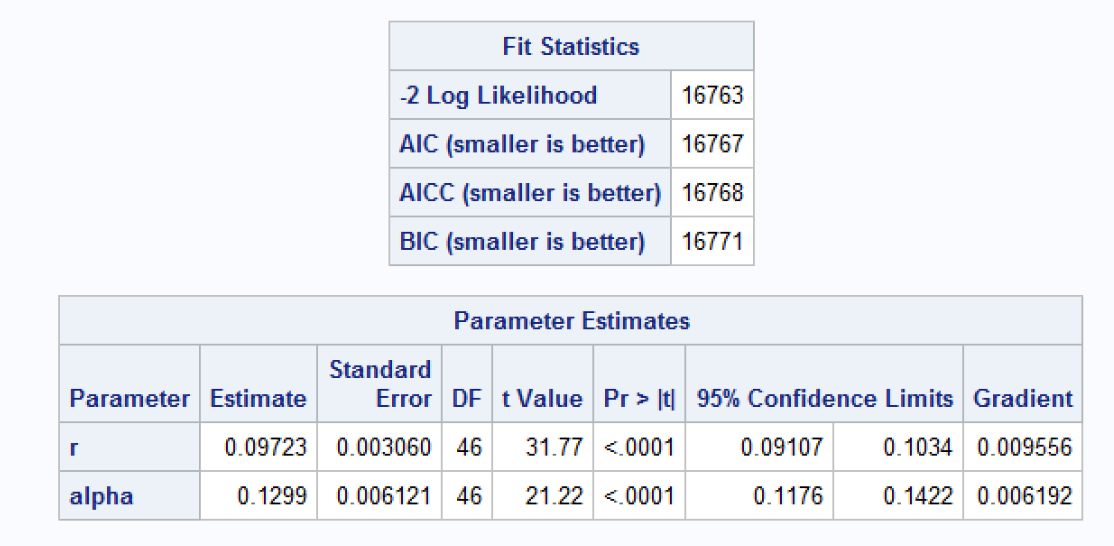
***/\*Calculation of log-likelihood\*/***

***ll = count\*log((gamma(r+total\_count)/(gamma(r)\*fact(total\_count)))\****

***((alpha/(alpha+1))\*\*r)\*((1/(alpha+1))\*\*total\_count));***

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

***run;***



1. The probability of the count being equal to 0 i.e. probability of a customer buying 0 books from BN:-

p(X=0) = (alpha/(alpha+1))^r = (0.1299/(0.1299+1))^0.09723 = 0.8103

Mean E(X) = r/alpha = 0.09723/0.1299 = 0.7485

Reach = 1- p(X=0) = 1-0.8103 = 0.1897 = 18.97%

Average Frequency = *E*[*X*] / (1 - *P*(*X*=0)) =3.946

GRPs = 100\* *E*[*X*] =74.85

1. Poisson Regression Model:-

***proc nlmixed data=ABI.Permdata;***

***/\* m stands for lamdha and b1, b2, b3, b4, b5, b6, b7 and b8 are the beta values for education, region, hhsz, age, income, child, race and country respectively\*/***

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

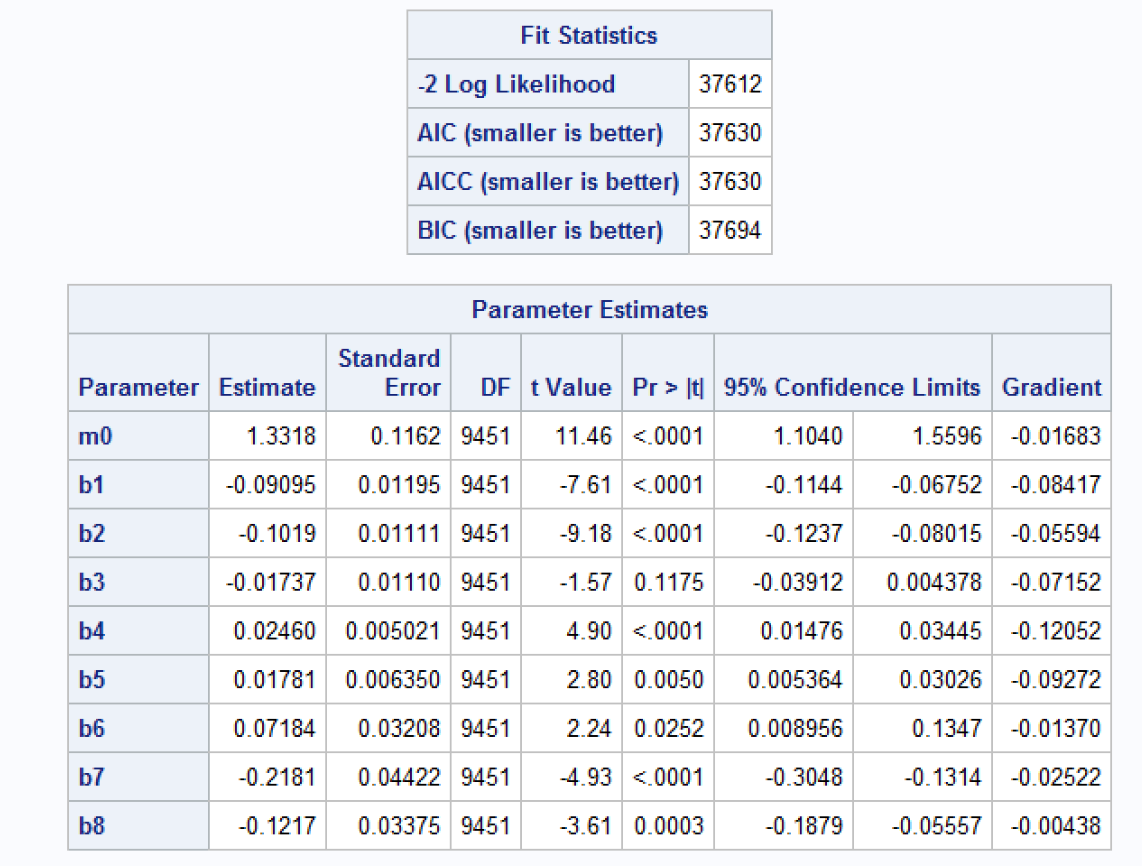
***m=m0\*exp(b1\*education+b2\*region+b3\*hhsz+b4\*age+b5\*income+b6\*child+b7\*race+b8\*country);***

***/\*Calculation of log-likelihood\*/***

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

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

***run;***



Thus the estimated values are as follows:

a) Beta value for education (b\_education) = -0.09095

b) Beta value for region (b\_region) = -0.1019

c) Beta value for Household size (b\_hhsz) = -0.01737

d) Beta value for age (b\_age) = 0.0246

e) Beta value for income (b\_income) = 0.01781

f) Beta value for child (b\_child) = 0.07184

g) Beta value for race (b\_race) = -0.2181

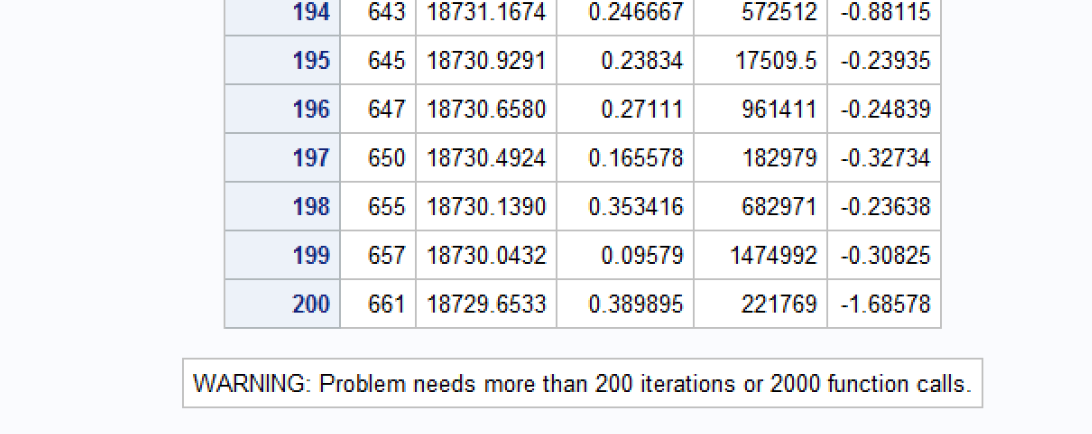
h) Beta value for country (b\_country) = -0.1217

**Managerial takeaways:**

1. If the education of a consumer is increased by 1 level (from 0 to 5), the total no. of his/her purchase of books in barnesandnoble.com site decrease by a factor of exp(0.09095)= 1.0952. Hence, an increase in the customer’s education by 1 level causes a [(x-x/1.0952)/x\*100] = 8.69% decrease in the customer’s total purchase of books in barnesandnoble.com site.
2. If the region of a consumer is changed from 1 to 4 by 1 level, the total no. of his/her purchase of books in barnesandnoble.com site decreases by a factor of exp(0.1019)=1.1072. A unit change in the customer’s region (from 1 to 4) causes a [(x-x/1.1072)/x\*100] = 9.68% decrease in the customer’s total purchase of books in barnesandnoble.com site.
3. If the household size of a consumer is increased by 1, the total no. of his/her purchase of books in barnesandnoble.com site decreases by a factor of exp(0.01737)=1.0175. Hence, a unit increase in the customers household size causes a [(x-x/1.0175)/x\*100] = 1.71% decrease in the customer’s total purchase of books in barnesandnoble.com site.
4. If the age of a consumer is increased by 1 level (from 1 to 11), the total no. of his/her purchase of books in barnesandnoble.com site increases by a factor of exp(0.0246)=1.0249. An increase in the customer’s age by 1 level causes a 2.49% increase in the customer’s total purchase of books in barnesandnoble.com site.
5. If the income of a consumer is increased by 1 level (from 1 to 7), the total no. of his/her purchase of books in barnesandnoble.com site increases by a factor of exp(0.01781)=1.0179. An increase in the customer’s income by 1 level causes a 1.79% increase in the customer’s total purchase of books in barnesandnoble.com site.
6. If a consumer has children, the total no. of his/her purchase of books in barnesandnoble.com site is exp(0.07184)=1.0744 times the total no. of purchase of books of a consumer who has no children. Thus, a customer having children buys 7.44% books in barnesandnoble.com site more than a customer having no children.
7. If the race of a consumer is changed from 1 to 5 by 1 level,, the total no. of his/her purchase of books in barnesandnoble.com site decreases by a factor of exp(0.2181)=1.2437. Hence a unit change in the customer’s region (from 1 to 5) causes a [(x-x/1.2437)/x\*100] = 19.59% decrease in the customer’s total purchase of books in barnesandnoble.com site.
8. If the country of a consumer is not US, the total no. of his/her purchase of books in barnesandnoble.com site is exp(0.1217)=1.1294 times less than the total no. of purchase of books of a consumer whose country is U.S. Thus, a customer whose country is not U.S buys [(x-x/1.1294)/x\*100] = 11.45% less books in barnesandnoble.com site than a customer whose country is U.S.
9. Also from the p-values, we observed that Household size and Child are not significant in determining the total number of books that a customer buys from buys from barnesandnoble.com site.

We recommend not using date in the regression due to the following reasons:-

1. The number of distinct values of the variable Date is too many, for it to be considered as a categorical variable
2. If the Date variable is considered as an Interval variable, it is still not suitable for analysis because the data set does not have data that is spread over a long period of time to use date as an efficient predictor of the dependent variable.
3. When the date is converted into SAS date and used for regression, SAS gives a warning that “Problem needs more than 200 iterations or 2000 function calls.”



1. NBD Regression Model:-

Formula used for LL:

Where

r = shape

alpha = scale

Y = total\_count

Beta = Beta Values for all the input Parameters (Xi) - education, region, household size, age, Income, child, race and country; in our project

1. NBD Regression Model code:-

***proc nlmixed data= abi.Permdata;***

***/\* r stands for shape, alpha for scale and b\_education, b\_region, b\_hhsz, b\_age, b\_income, b\_child, b\_race and b\_country are the beta values for education, region, hhsz, age, income, child, race and country respectively\*/***

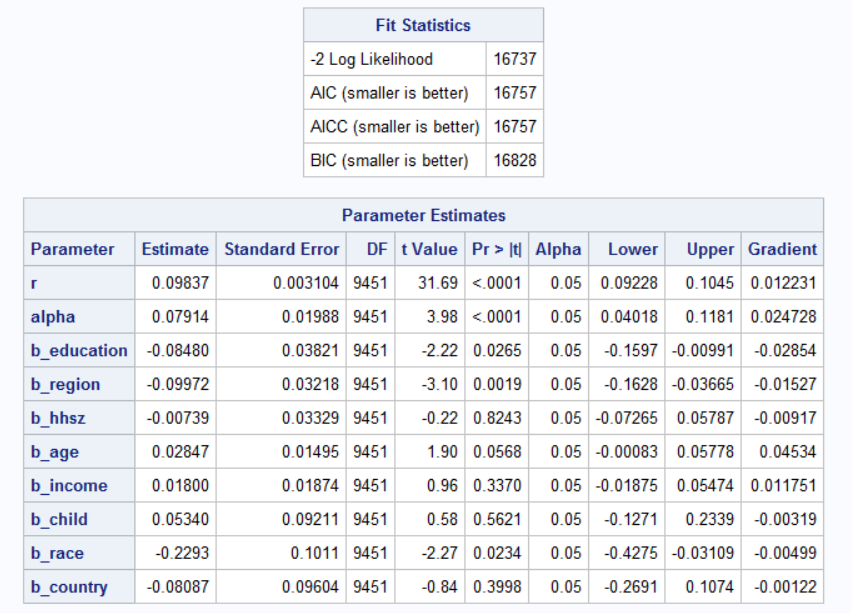
***parms r=1, alpha=1, b\_education=0, b\_region=0, b\_hhsz=0, b\_age=0, b\_income=0, b\_child=0, b\_race=0, b\_country=0;***

***expon=exp(b\_education\*education+b\_region\*region+b\_hhsz\*hhsz+b\_age\*age+b\_income\*income+b\_child\*child+b\_race\*race+b\_country\*country);***

***ll = log((gamma(r+total\_count)/(gamma(r)\*fact(total\_count)))\*((alpha/(alpha+expon))\*\*r)\*((expon/(alpha+expon))\*\*total\_count));***

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

***run;***



Thus the estimated value are as follows:-

1. Beta value for education (b\_education) = -0.08480
2. Beta value for region (b\_region) = -0.09972
3. Beta value for Household size (b\_hhsz) = -0.00739
4. Beta value for age (b\_age) = 0.02847
5. Beta value for income (b\_income) = 0.01800
6. Beta value for child (b\_child) = 0.05340
7. Beta value for race (b\_race) = -0.2293
8. Beta value for country (b\_country) = -0.08087

Lamda = exp (b\_education\*education + b\_region\*region + b\_hhsz\*hhsz + b\_age\*age + b\_income\*income + b\_child\*child + b\_race\*race + b\_country\*country );

**Managerial takeaways:**

1. If the education of a consumer is increased by 1 level (from 0 to 5), the total no. of his/her purchase of books in barnesandnoble.com site decrease by a factor of exp(0.0848)= 1.088467. Hence, an increase in the customer’s education by 1 level causes a [(x-x/1.0885)/x\*100] = 8.13% decrease in the customer’s total purchase of books in barnesandnoble.com site.
2. If the region of a consumer is changed from 1 to 4 by 1 level, the total no. of his/her purchase of books in barnesandnoble.com site decreases by a factor of exp(0.09972)=1.1048. A unit change in the customer’s region (from 1 to 4) causes a [(x-x/1.1048)/x\*100] = 9.49% decrease in the customer’s total purchase of books in barnesandnoble.com site.
3. If the household size of a consumer is increased by 1, the total no. of his/her purchase of books in barnesandnoble.com site decreases by a factor of exp(0.00739)=1.0074. Hence, a unit increase in the customers household size causes a [(x-x/1.0074)/x\*100] = 0.73% decrease in the customer’s total purchase of books in barnesandnoble.com site.
4. If the age of a consumer is increased by 1 level (from 1 to 11), the total no. of his/her purchase of books in barnesandnoble.com site increases by a factor of exp(0.02847)=1.0289. An increase in the customer’s age by 1 level causes a 2.88% increase in the customer’s total purchase of books in barnesandnoble.com site.
5. If the income of a consumer is increased by 1 level (from 1 to 7), the total no. of his/her purchase of books in barnesandnoble.com site increases by a factor of exp(0.018)=1.0181. An increase in the customer’s income by 1 level causes a 1.81% increase in the customer’s total purchase of books in barnesandnoble.com site.
6. If a consumer has children, the total no. of his/her purchase of books in barnesandnoble.com site is exp(0.0534)=1.0548 times the total no. of purchase of books of a consumer who has no children. Thus, a customer having children buys 5.48% books in barnesandnoble.com site more than a customer having no children.
7. If the race of a consumer is changed from 1 to 5 by 1 level,, the total no. of his/her purchase of books in barnesandnoble.com site decreases by a factor of exp(0.2293)=1.2577. Hence a unit change in the customer’s region (from 1 to 5) causes a [(x-x/1.2577)/x\*100] = 20.49% decrease in the customer’s total purchase of books in barnesandnoble.com site.
8. If the country of a consumer is not US, the total no. of his/her purchase of books in barnesandnoble.com site is exp(0.08087)=1.0842 times less than the total no. of purchase of books of a consumer whose country is U.S. Thus, a customer whose country is not U.S buys [(x-x/1.0842)/x\*100] = 7.76% less books in barnesandnoble.com site than a customer whose country is U.S.
9. Also from the p-value, we observed that Household size, child, country and income are not significant in determining the total number of books that a customer buys from barnesandnoble.com site.
10. We also plotted a gamma distribution with the estimated r = 0.09837 and alpha = 0.07914 value.

***DATA random1;***

***Do i=1 to 1000;***

***x=0.07914\*rand('GAMMA', 0.09837);***

***output;***

***end;***

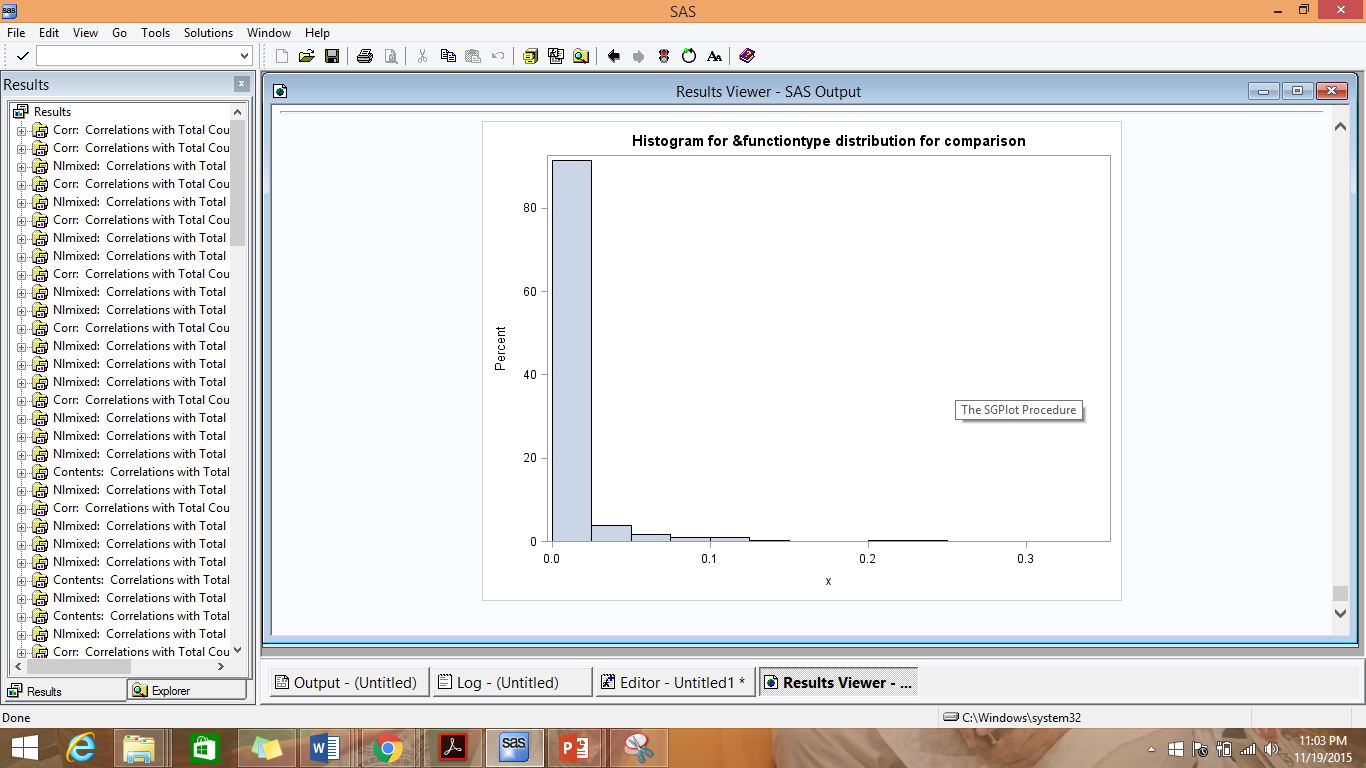
***RUN;***

***proc sgplot data=&functiontype;***

***histogram x;***

***title "Histogram ";***

***run;***



We thus observed that the lambda value is high for less number of customers i.e. few customers buy books from BN frequently.

1. Difference regarding the managerial takeaways between Poisson Regression and NBD Regression:-

|  |  |  |
| --- | --- | --- |
|  | Poisson Regression | NBD Regression |
| Increase in customer’s education by 1 level | 8.69% decrease | 8.13% decrease |
| The region of consumer changed from 1 to 4 by 1 level | 9.68% decrease | 9.49% decrease |
| The household size of a consumer is increased by 1 | **1.71% decrease** | 0.73% decrease |
| Increase in the customer’s age by 1 level | 2.49% increase | 2.88% increase |
| Increase in the customer’s income by 1 level | 1.79% increase | 1.81% increase |
| Customer having children | 7.44% more | 5.48% more |
| A unit change in the customer’s race (from 1 to 5) | 19.59% decrease | 20.49% decrease |
| A customer whose country is not U.S | **11.45% less** | 7.76% less |

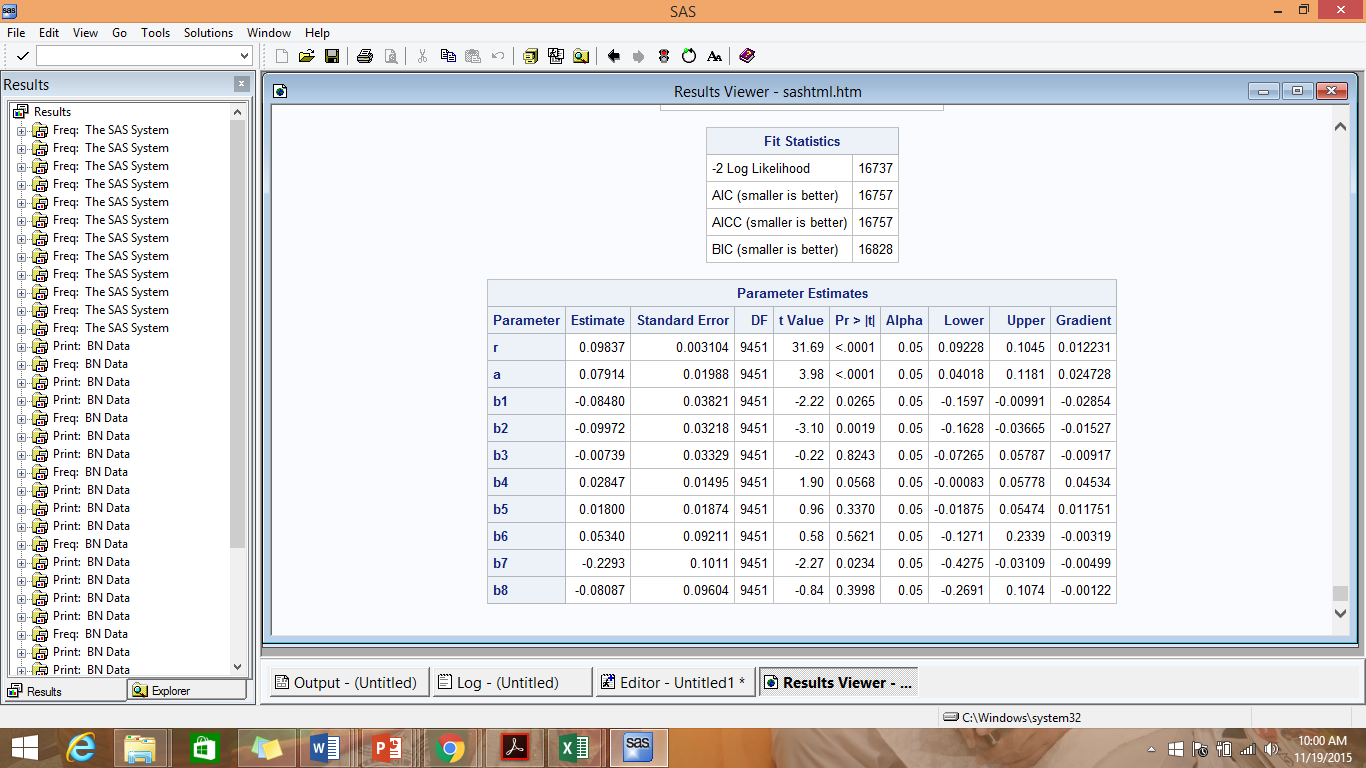
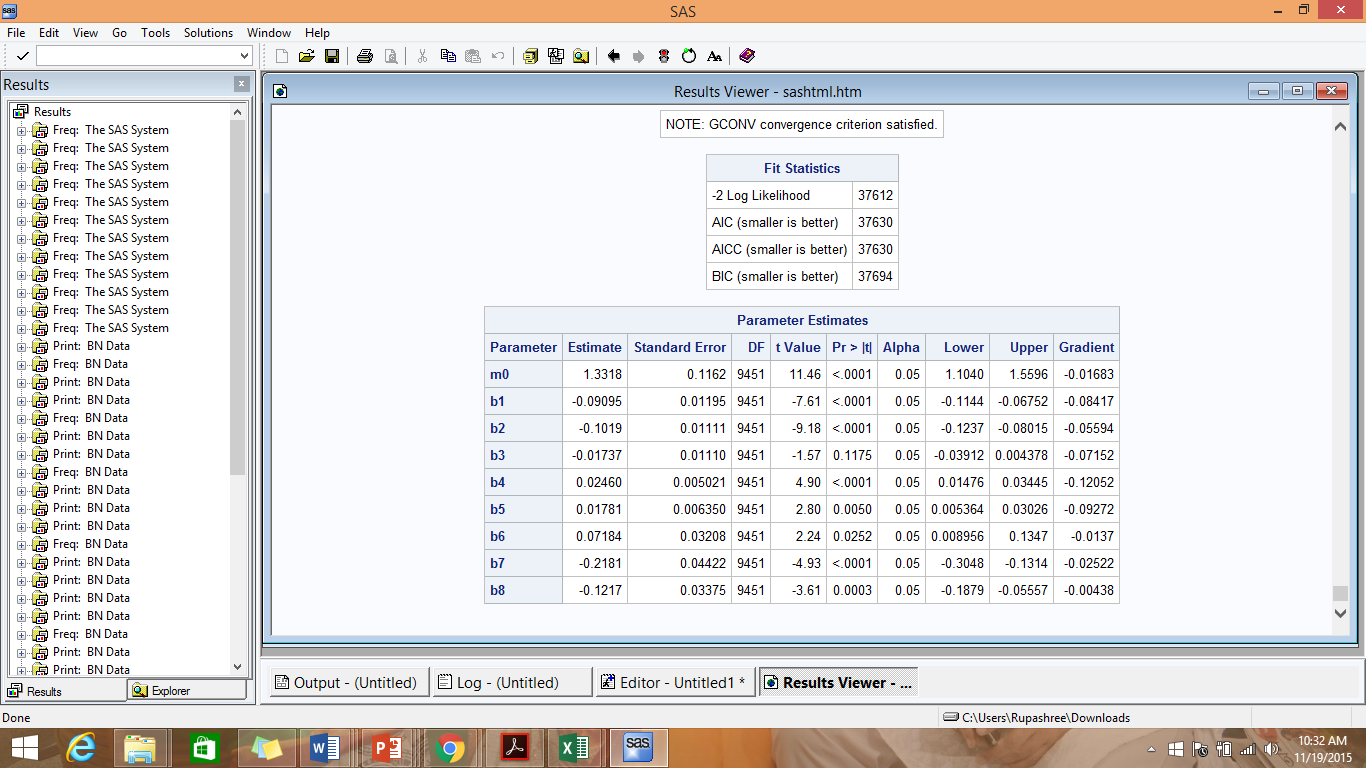
We observed that the effect of each input variable is similar for both the models.

However, the effect of household size is more pronounced in the Poisson Regression than the NBD Regression. A unit increase in the household size causes a 1.71% decrease in lambda in the Poisson Regression and only 0.73% (less than half) in the NBD Regression. Also, the country of the customer has more of an effect in determining lambda in the Poisson Model. A non-US customer buys 7.76% lesser books than an US customer as per the NBD model whereas as per Poisson Regression, he/she buys 11.45% lesser than an US customer.

In the NBD Regression Model, that Household size, child, country and income are not significant in determining the total number of books that a customer buys from barnesandnoble.com site. Whereas in the Poisson Regression Model, only the Household size and child are not significant.

1. Comparison of whether NBD Regression fit the data better than Poisson Regression:-

NBD Regression Model Poisson Regression Model

ll for Poisson Regression Model = -18,806

ll for NBD Regression Model = -8368.5

LR = −2(LLB − LLA) =-2(-18,806vv-- -- 8368.5)

= -2(-18,806+8368.5)

= -2(-10,437.5)

=20,875

χ2 (.05,k) for df=1 (because the NBD Regression has 1 extra parameter a and r instead mo of Poisson Regression) 🡪 3.841

Hence, we observed that LR > χ2; signifying that the complicated model; NBD Regression in this case fits the data better.

**Conclusion:**

The NBD Regression Model fits the data better than Poisson Regression i.e the performance of the more complicated model (10 attributes) is better than the performance of the simpler model (9 attributes).

**Probable Reason:**

It is often assumed that a more complex model i.e one with more number of attributes perform better than a simpler model as it would have more predictor attributes resulting better prediction performance,.

The [Poisson distribution](http://en.wikipedia.org/wiki/Poisson_distribution) assumes that the mean and variance are the same. However, sometimes data show extra variation that is greater than the mean. This situation is called [over-dispersion](http://en.wikipedia.org/wiki/Overdispersion) and negative binomial regression is more flexible in that regard than Poisson regression. The [negative binomial distribution](http://en.wikipedia.org/wiki/Negative_binomial_distribution#Overdispersed_Poisson) has one parameter more than the Poisson regression that adjusts the variance independently from the mean.

Thus, only the nature of the data and the questions of interest can determine which of these regressions are best for the situation and one cannot conclude on which model is universally better.

Our data fits the NBD Regression better because it may be a case of over-dispersion where the variance is more than the mean.

# Part II: Improving the Model

1. We performed 3 tests to determine the variables that are not useful and can be dropped from analysis

A) Correlation

We ran correlation of each of the variable with Total\_count:

***DATA abi.corrdata;***

***set abi.Permdata12;***

***RUN;***

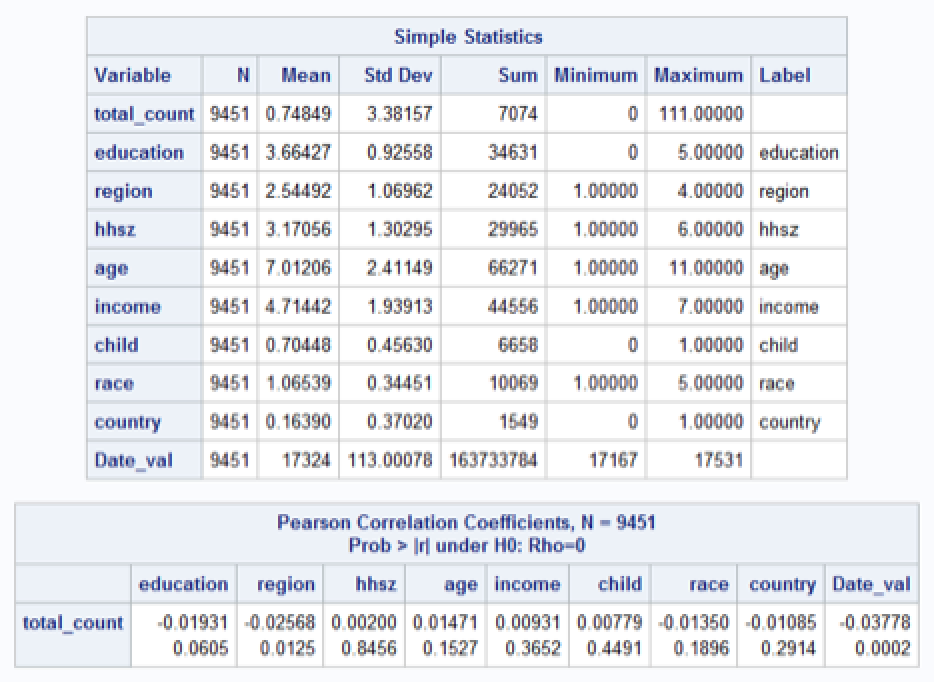
***PROC CORR DATA = abi.corrdata;***

***VAR education region hhsz age income child race country date\_val;***

***WITH total\_count;***

***TITLE 'Correlations with Total Count';***

***RUN;***

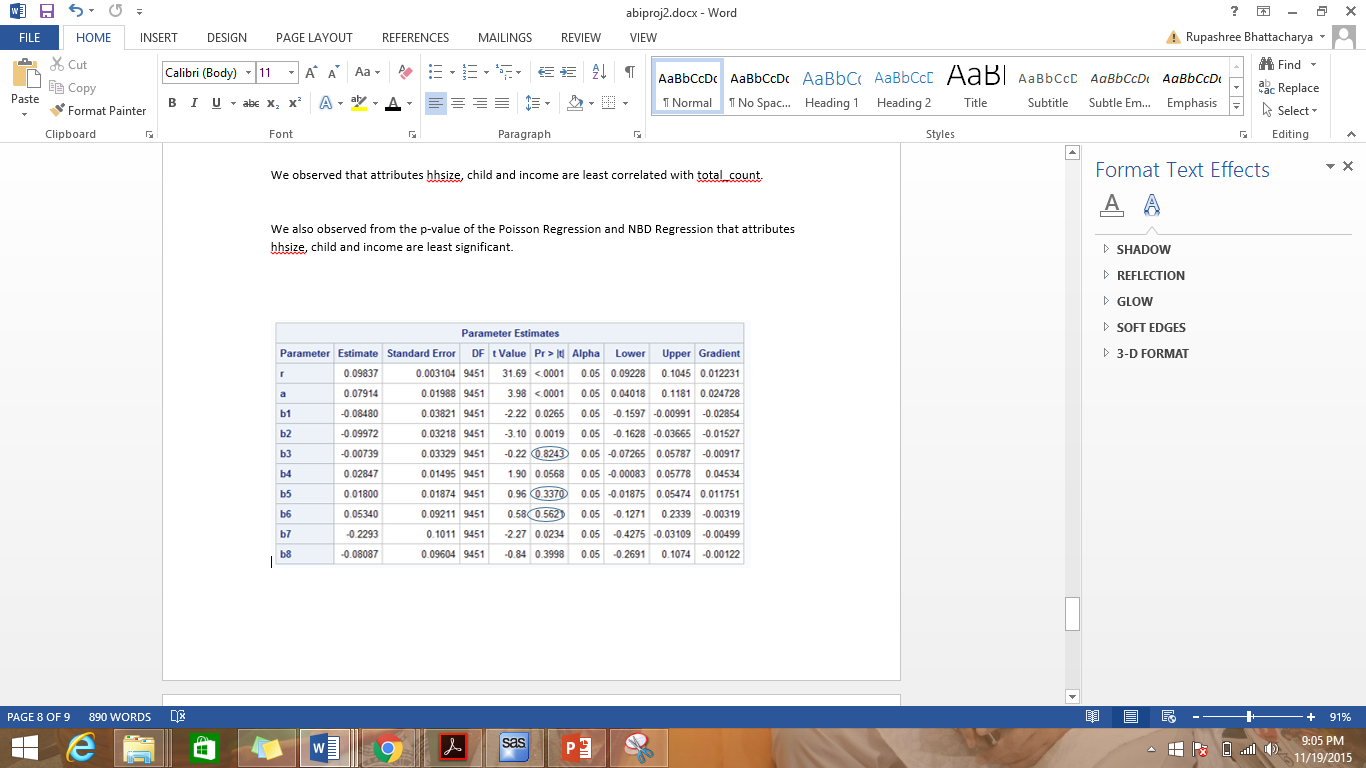


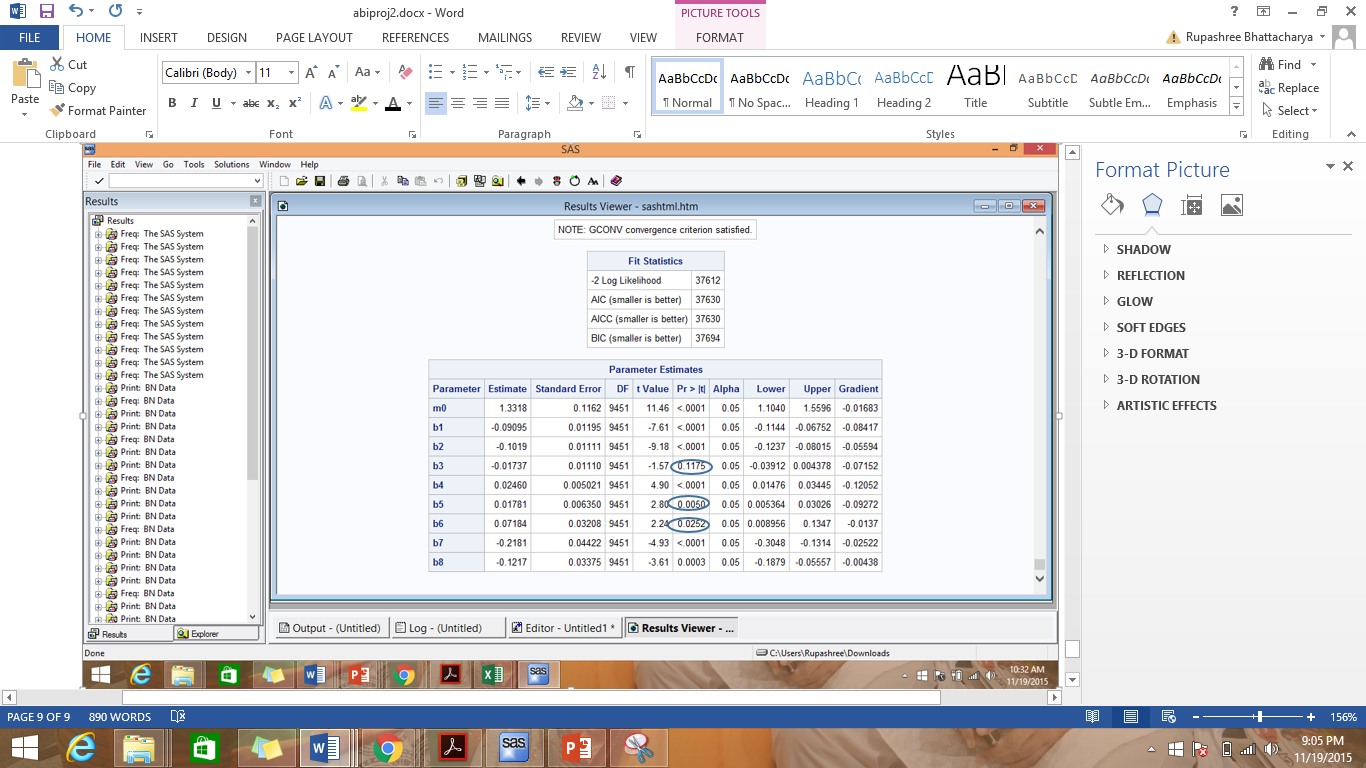
**Conclusion:**

We observed that attributes hhsize, child and income are least correlated with total\_count.

B) p-value from NBD Regression and Poisson Regression

We observed the p-value of the Poisson Regression and NBD Regression to determine the variables that are least significant.

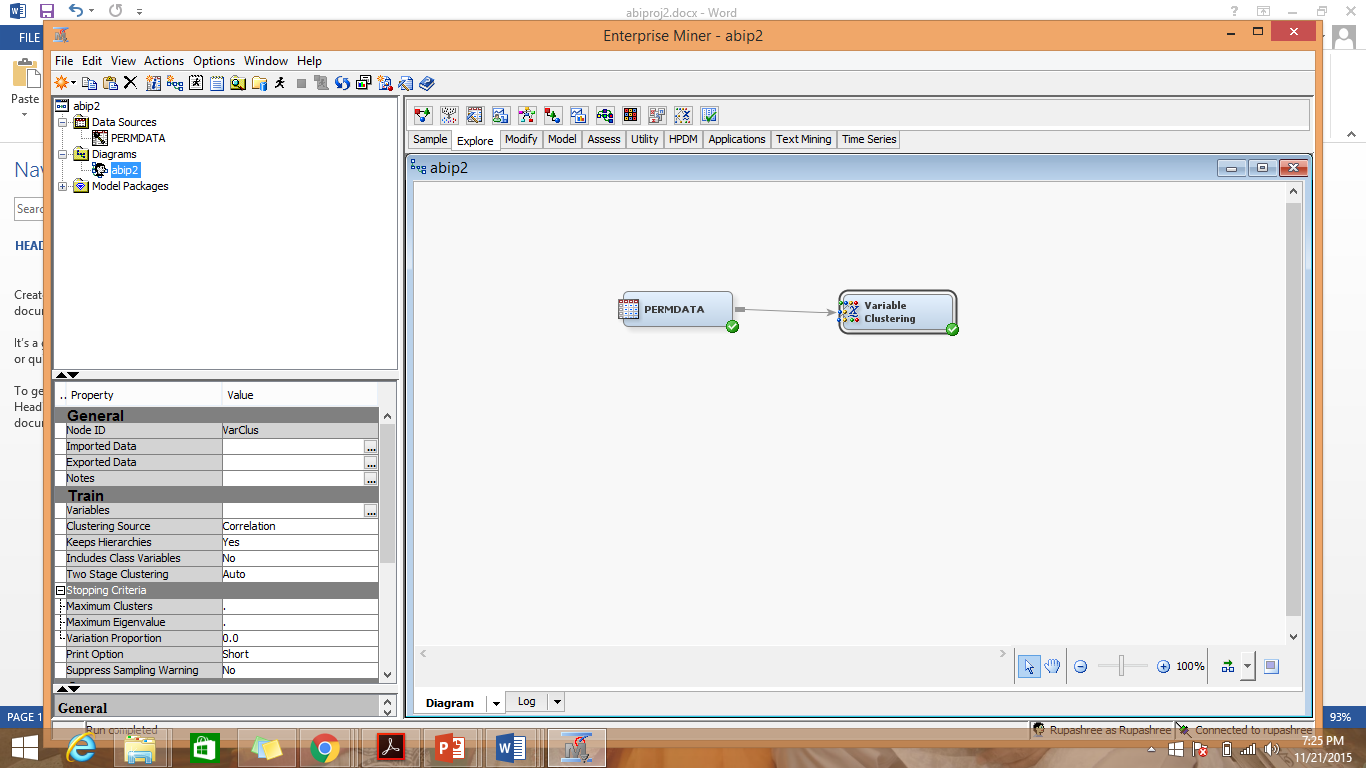


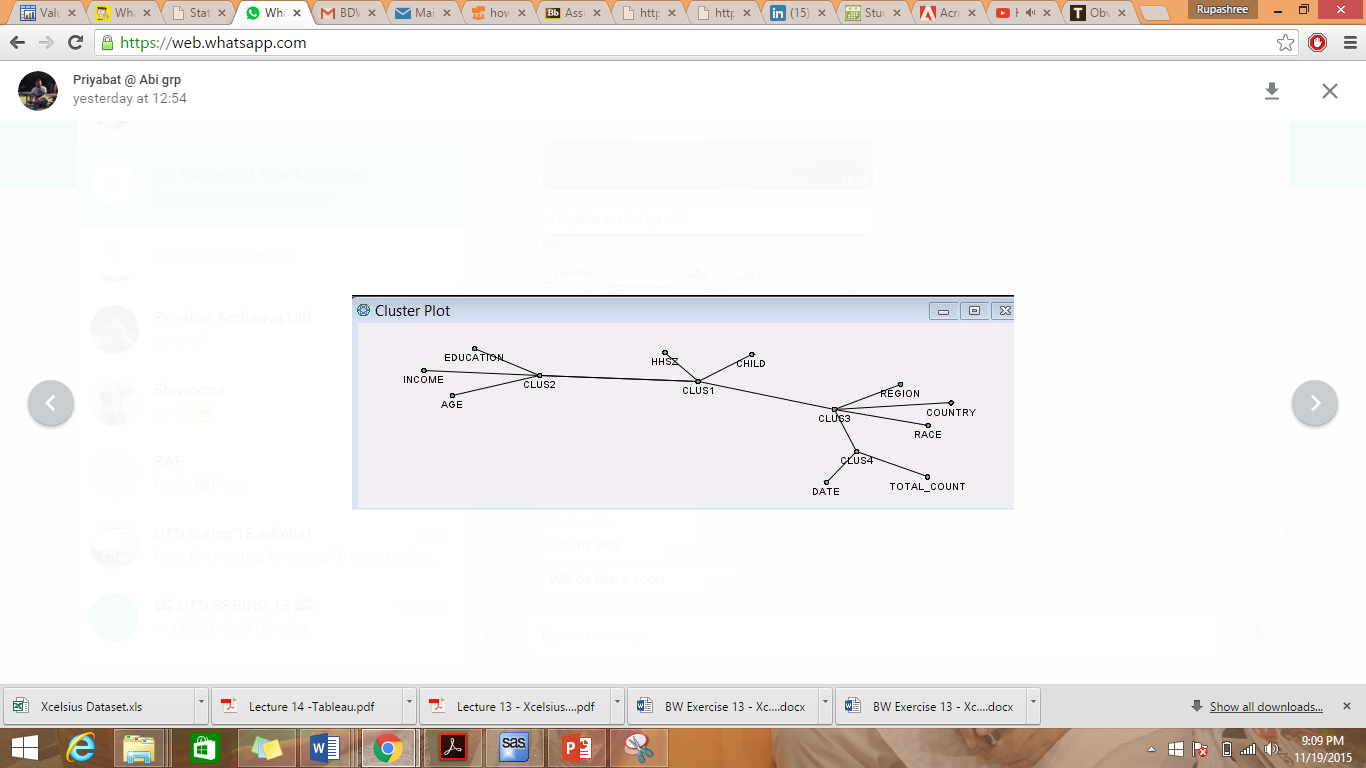


**Conclusion:** We observed that the attributes; hhsize, child and income are least significant.

C) Variable Clustering on Enterprise Miner

We executed ‘Variable Clustering’ on the dataset in SAS EM:-





**Conclusion:**

We observed that the variables Education, Income, HHSZ, Child and Age are far from Total\_Count.

**Conclusion from the 3 tests:**

All the 3 tests showed that the variables; Education, Income and Age are not significant in determining Toatal\_Count and hence, we decided to drop them from the analysis.

**Poisson Regression code with reduced number of variables:**

***proc nlmixed data=abi.Permdata;***

***/\* m stands for lamdha \*/***

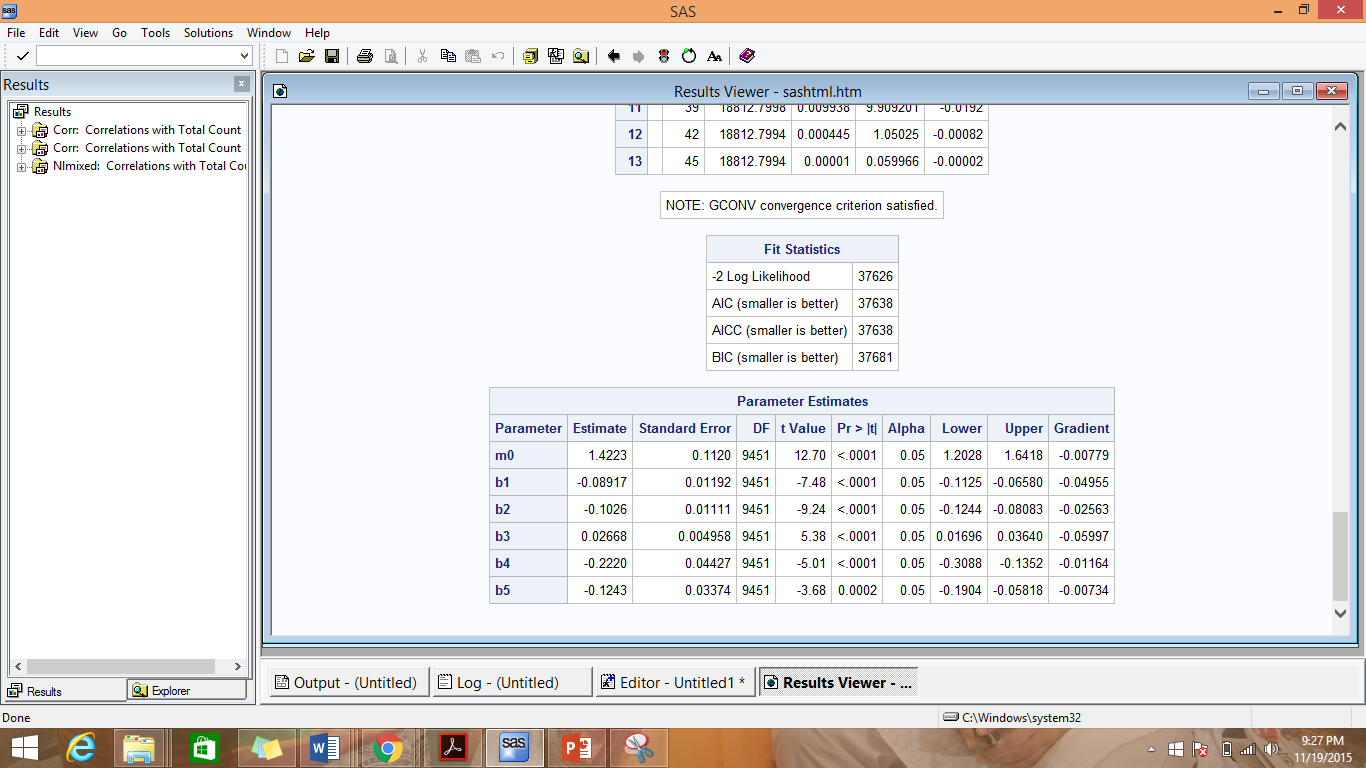
***parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0;***

***m=m0\*exp(b1\*education+b2\*region+b3\*age+b4\*race+b5\*country);***

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

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

***run;***



**Comparison of Poisson Regression on all the variables with the Poisson Regression on reduced number of variables:-**

ll of the Poisson Regression on all the variables = -18,806

ll of the Poisson Regression on the reduced number of variables = -18,813

LR = - 2(-18,813 - - 18,806) = -2(-7) =14

χ2 (.05,k) for df=3 (because we dropped 3 variables from the analysis) 🡪 7.815

Hence LR> χ2; signifying that the more complicated model I.e. the Poisson Model with all the variables is indeed better than the model with reduced number of variables.

**NBD Regression code with reduced number of variables:-**

***proc nlmixed data= abi.Permdata;***

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

***m=exp(b1\*education+b2\*region+b3\*age+b4\*race+b5\*country);***

***num1 = gamma(r+total\_count);***

***denom1 = gamma(r) \* fact(total\_count);***

***term1 = num1 / denom1;***

***denom2 = a+m;***

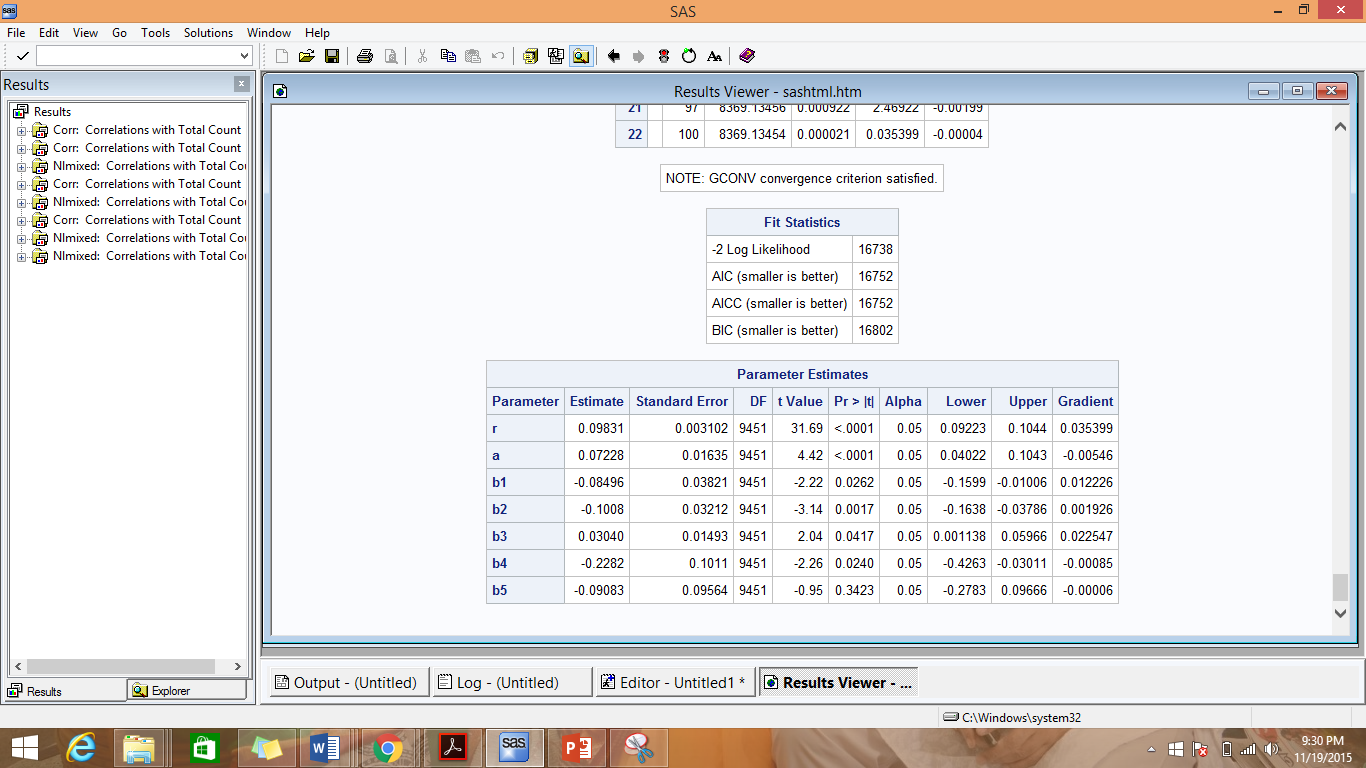
***term2 = (a / denom2)\*\*r;***

***term3 = (m/denom2)\*\*total\_count;***

***ll = log(term1\*term2\*term3);***

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

***run;***



**Comparison of NBD Regression on all the variables with the NBD Regression on reduced number of variables:-**

ll of the NBD Regression on all the variables = -8,368.5

ll of the NBD Regression on the reduced number of variables = -8,369

LR = - 2(-8,369- -8,368.5) = -2(-0.5) =3

χ2 (.05,k) for df=3 (because we dropped 3 variables from the analysis) 🡪 7.815

Hence LR< χ2; signifying that the more complicated model I.e. the NBD Model with all the variables is not different from the model with reduced number of variables. Hence the simpler model can be used.

**Conclusion of the Variable Reduction Analysis:**

The Poisson Regression fits the original dataset i.e with the one with 8 input variables better than the dataset with only 5 variables. However for the NBD Regression model, there is no difference in the fit for the 2 datasets and hence a simpler dataset i.e. the one with 5 variables can be used. We can thus conclude that there is no overfitting for Poisson Model whereas there was a case of overfitting for NBD model and thus one can use a simpler dataset for it.

1. Construction of 3 new input variables in the analysis:-

a) Loyalty of Consumer – The variable *Loyalcust* is set to 1 if the customer has purchased the books only from Barnes and nobel. It is set to 0 if the customer has bought from both Barnes and nobel and from amazon. It is set to 0 if the customer has bought from amazon alone.

Code to add the variable Loyalcust:-

***PROC sql;***

***create table abi.permdata\_q10 as***

***select userid, education, region, hhsz, age, income, child, race, country, domain, date, total\_count ,'1' as LoyalCust***

***from abi.permdata***

***QUIT;***

***Run;***

In the above code, we are simply creating another database with one additional column ‘LoyalCust’ whose value is set to 1.

***PROC sql;***

***update abi.permdata\_q10***

***set LoyalCust = '0'***

***where***

***userid in (select userid from abi.tempdataamazon);***

***run;***

In the above code, we have set the variable ‘LoyalCust’ to 0 for those customers who have bought books from Amazon.com. For q1, we had split the original database to create abi.tempdataamazon which contains customers who have bought books from Amazon (both who have bought only from Amazon and those who have bought from both BN and Amazon).

Thus by setting the ‘LoyalCust’ of customers who are present in the database abi.tempdataamazon to 0, we have maintained the ‘LoyalCust’ as 1 only for those customers who have bought exclusively from BN.

b) Seasons – If the month of purchase is December, Jan or Feb, the season is considered ‘Winter’ and the variable *Season* is set to 1. If the month of purchase is March, April or May, the season is considered ‘Spring’ and the variable *Season* is set to 2. If the month of purchase is June, July or August, the season is considered ‘Summer’ and the variable *Season* is set to 3. If the month of purchase is September, October or November, the season is considered ‘Fall’ and the variable *Season* is set to 4.

c) Weekends – If the day of purchase is Saturday or Sunday, it is considered weekend and the variable *Weekend* is set to 1 , otherwise it is set to 0.

Code to add Season and Weekend

***Data abi.Permdata\_q10\_date;***

***Set abi.Permdata\_q10;***

***Weekend = 0;***

***If weekday(date\_val) = 1 or weekday(date\_val) = 7 then do;***

***weekend = 1;***

***end;***

***If month(date\_val) = 1 or month(date\_val) = 2 or month(date\_val) = 12 then do;***

***Season = 1;***

***end;***

***else if month(date\_val) = 3 or month(date\_val) = 4 or month(date\_val) = 5 then do;***

***Season = 2;***

***end;***

***else if month(date\_val) = 6 or month(date\_val) = 7 or month(date\_val) = 8 then do;***

***Season = 3;***

***end;***

***else do;***

***Season = 4;***

***end;***

***Run;***

In the above code, we set the variable weekend = 1 for all those records whose weekday of the date of purchase is 1 or 7 i.e. Sunday or Saturday.

We have also set the variable Season to Winter or ‘1’ for months December, Jan and Feb; to Spring or ‘2’ for months March, April and May; to Summer or ‘3’ for months June, July and August; and to Fall or ‘4’ for September, October or November.

**We now executed the Poisson Regression model on the dataset with the additional 3 variables:**

***proc nlmixed data= abi.Permdata\_q10\_date;***

***/\* m stands for lamdha \*/***

***parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0, b8=0, b9=0, b10=0, b11=0;***

***m=m0\*exp(b1\*education+b2\*region+b3\*hhsz+b4\*age+b5\*income+b6\*child+b7\*race+b8\*country+b9\*LoyalCust+b10\*Weekend+b11\*Season);***

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

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

***run;***



**Comparison between the Poisson Regression Model on the original dataset and the Poisson Regression on the dataset with the 3 additional attributes:**

ll for the original dataset = -18,806

ll for the new dataset = -12,737

LR = -2(-18,806 - -12,737) = -2(-6069) = 12,138

χ2 (.05,k) for df=3 (because we added 3 variables to the analysis) 🡪 7.815

Hence LR> χ2; signifying that the more complicated model I.e. the Poisson Model with the additional Poisson variables is better than the original model.

**We now executed the NBD Regression model on the dataset with the additional 3 variables:-**

***proc nlmixed data= abi.Permdata\_q10\_date;***

***/\* m stands for lamdha \*/***

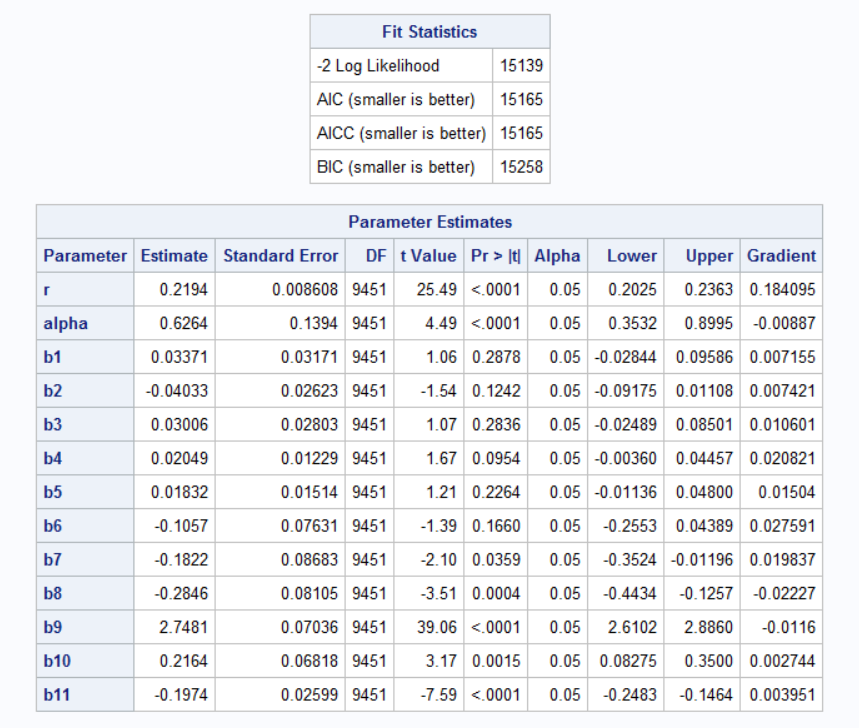
***parms r=1 alpha=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0, b8=0, b9=0, b10=0, b11=0;***

***expon=exp(b1\*education+b2\*region+b3\*hhsz+b4\*age+b5\*income+b6\*child+b7\*race+b8\*country+b9\*LoyalCust+b10\*Weekend+b11\*Season);***

***ll = log((gamma(r+total\_count)/(gamma(r)\*fact(total\_count)))\*((alpha/(alpha+expon))\*\*r)\*((expon/(alpha+expon))\*\*total\_count));***

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

***run;***



**Comparison between the NBD Regression Model on the original dataset and the NBD Regression on the dataset with the 3 additional attributes:-**

ll for the original dataset = -8,368.5

ll for the new dataset = -7,569.5

LR = -2(-8,368.5 - - 7,569.5) = -2(-799) = 1598

χ2 (.05,k) for df=3 (because we added 3 variables to the analysis) 🡪 7.815

Hence LR> χ2; signifying that the more complicated model I.e. the NBD Model with the additional NBD variables is better than the original model.

**Conclusion for constructing additional variables to the dataset:**

We observed that by constructing the 3 variables; loyalty, Season and Weekend, the performance of both Poisson and NBD models increase. We also observed the p-value of these 3 variables from both Poisson and NBD Regression and concluded that the 3 variables especially loyalty and Seasons are significant in determining the total purchase of customers. Hence constructing significant variables have increased the performance of the models.

1. Interaction effects:-

A) Qualitative consideration

We brainstormed on the effect of each input variable on the target variable and also their effect and relationship with each other. We decided on the following 3 interaction effects:-

1. Hhsz\*child

In the age of nuclear families, household size is strongly correlated to having children. If the customer has children his/her household size would generally be more than a customer who has no children.

1. region\*race\*country

Country, region and race are demographic information of a customer and are strongly correlated with each other. The race and region where a customer resides is often related to the country of residence of the consumer.

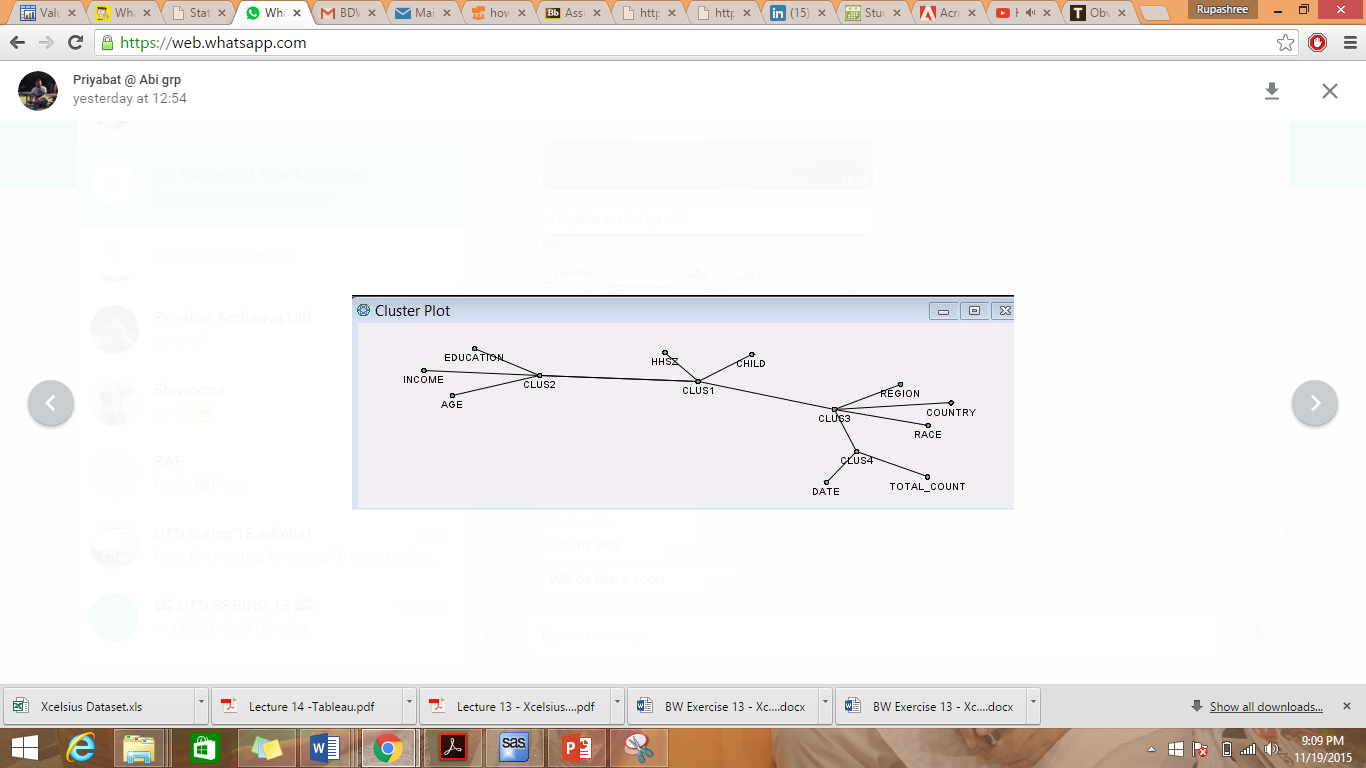
1. age\*income\*education

The older a consumer is, it is more likely that he/she would be more educated and earning better. Also more educated a consumer is, it is more likely that he/she earns more.

B) Variable Clustering On Enterprise Miner

On applying Variable Clustering on SAS EM, we get the following clusters:-

1. Cluster 1- HHSZ and Child
2. Cluster 2- Education, Income and Age
3. Cluster 3 – Region , Country and Race



We thus concluded that HHSZ and Child, Education, Income and Age and Region, Country and Race are strongly related to each other and hence their interaction effects can be considered for the analysis.

***data abi.Permdataq11;***

***set abi.Permdata;***

***parameter1 = hhsz\*child;***

***parameter2 = region\*race\*country;***

***parameter3 = age\*income\*education;***

***run;***

1. Interaction Effect 1- **hhsz\*child:**

***Poisson Regression:-***

***Proc contents data = abi.Permdataq11 POSITION;***

***proc nlmixed data= abi.Permdataq11;***

***/\* m stands for lamdha \*/***

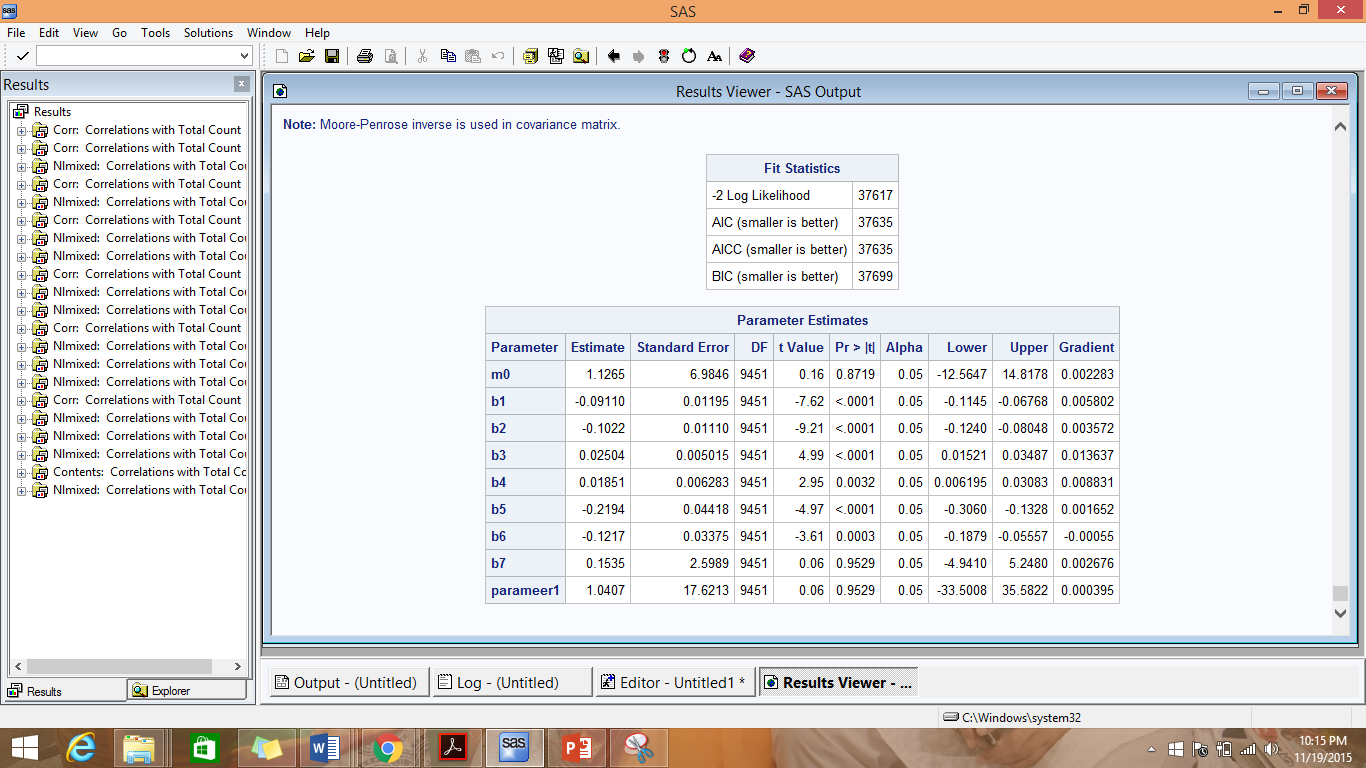
***parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;***

***m=m0\*exp(b1\*education+b2\*region+b3\*age+b4\*income+b5\*race+b6\*country+b7\* parameer1);***

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

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

***run;***



**Comparison of Poisson Regression model on the dataset with the Interaction effect with Poisson Regression model on the original dataset:-**

ll of Poisson Regression model on the dataset with the Interaction effect = -18,808.5

ll of Poisson Regression model on the original dataset = -18,806

LR = -2(-18,808.5 - -18,806) = 5

χ2 (.05,k) for df=1 (because we have drooped hhsz and child and added hhsz\*child to the analysis) 🡪 3.841

Thus, LR > χ2; signifying that the more complicated model I.e. the Poisson Model on the original dataset is better than the model on the new dataset.

**NBD Regression:-**

***proc nlmixed data= abi.Permdataq11;***

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

***m=exp(b1\*education+b2\*region+b3\*age+b4\*income+b5\*race+b6\*country+ b7\*parameer1);***

***num1 = gamma(r+total\_count);***

***denom1 = gamma(r) \* fact(total\_count);***

***term1 = num1 / denom1;***

***denom2 = a+m;***

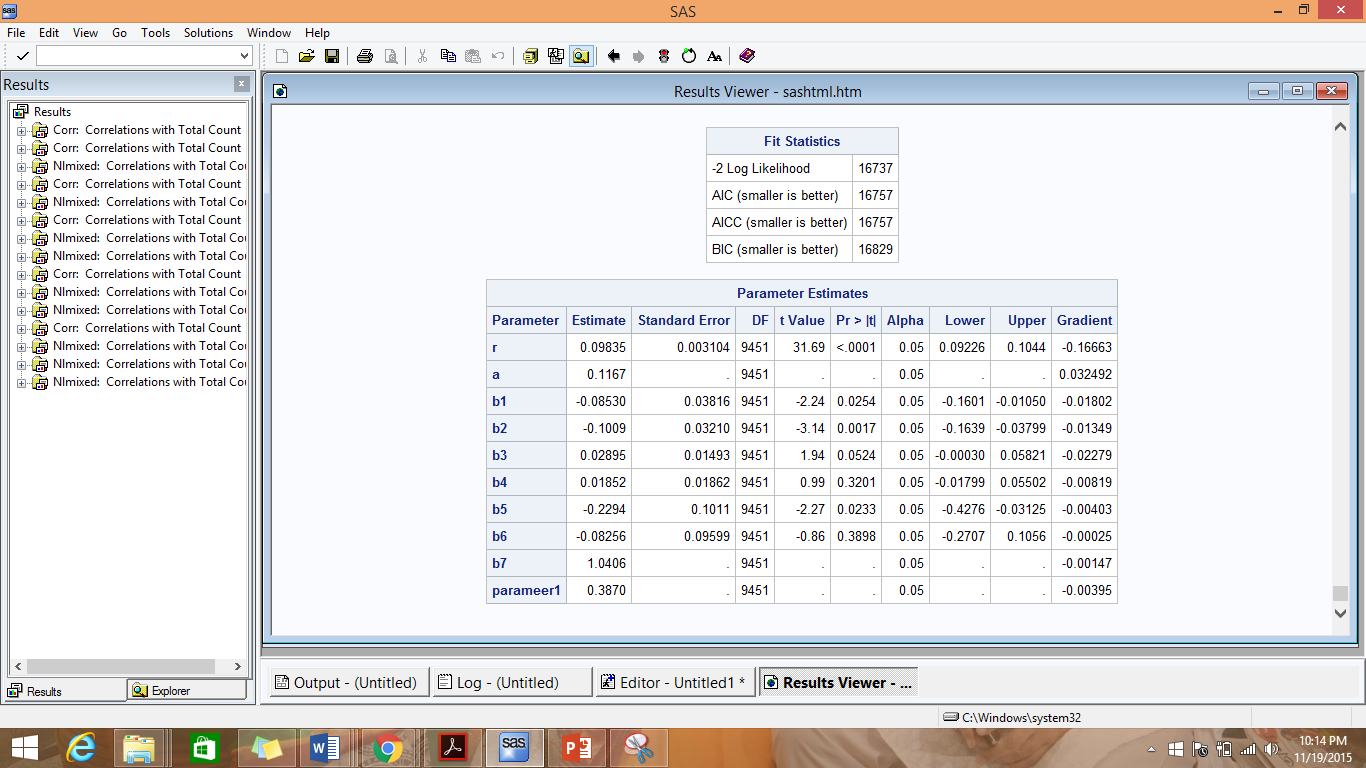
***term2 = (a / denom2)\*\*r;***

***term3 = (m/denom2)\*\*total\_count;***

***ll = log(term1\*term2\*term3);***

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

***run;***



**Comparison of NBD Regression model on the dataset with the Interaction effect with NBD Regression model on the original dataset:-**

ll of NBD Regression model on the dataset with the Interaction effect = -8368.5

ll of NBD Regression model on the original dataset = -8368.5

LR = -2(-8368.5 - - 8368.5) = 0

χ2 (.05,k) for df=1 (because we have drooped hhsz and child and added hhsz\*child to the analysis) 🡪 3.841

Thus, LR < χ2; signifying that the more complicated model I.e. the NBD Model on the original dataset is not different from the model on the new dataset. Hence the new, simpler model can be used.

1. Interaction Effect 2- **region\*race\*country**

***Poisson Regression:-***

***Proc contents data = abi.Permdataq11 POSITION;***

***proc nlmixed data= abi.Permdataq11;***

***/\* m stands for lamdha \*/***

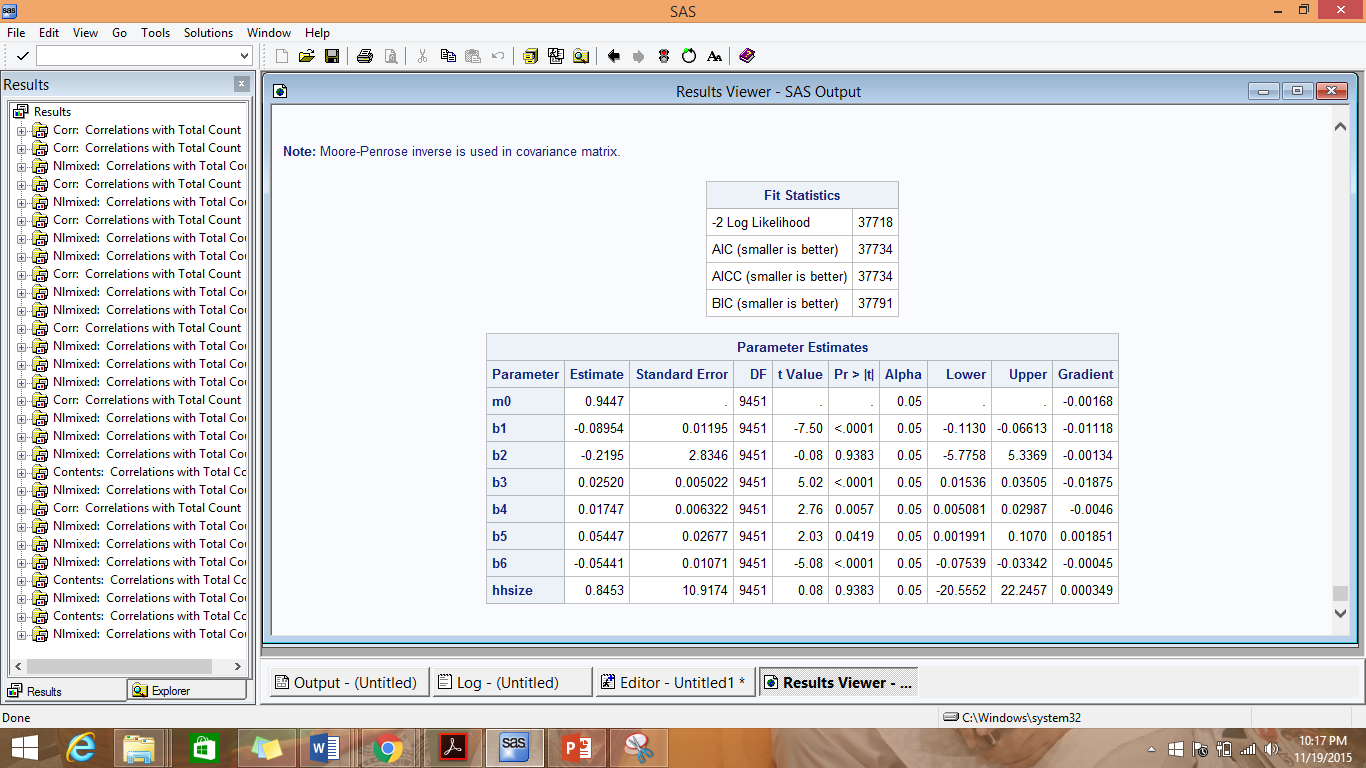
***parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0;***

***m=m0\*exp(b1\*education+b2\*hhsize+b3\*age+b4\*income+b5\*child+b6\* parameter2);***

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

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

***run;***



**Comparison of Poisson Regression model on the dataset with the Interaction effect with Poisson Regression model on the original dataset:-**

ll of Poisson Regression model on the dataset with the Interaction effect = -18,859

ll of Poisson Regression model on the original dataset = -18,806

LR = -2(-18,859 - -18,806) = 106

χ2 (.05,k) for df=2 (because we have drooped region, race and country and added region\*race\*country to the analysis) 🡪 5.991

Thus, LR > χ2; signifying that the more complicated model I.e. the Poisson Model on the original dataset is better than the model on the new dataset.

**NBD Regression:-**

***proc nlmixed data= abi.Permdataq11;***

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

***m=exp(b1\*education+b2\*hhsz+b3\*age+b4\*income+b5\*child+b6\* parameter2);***

***num1 = gamma(r+total\_count);***

***denom1 = gamma(r) \* fact(total\_count);***

***term1 = num1 / denom1;***

***denom2 = a+m;***

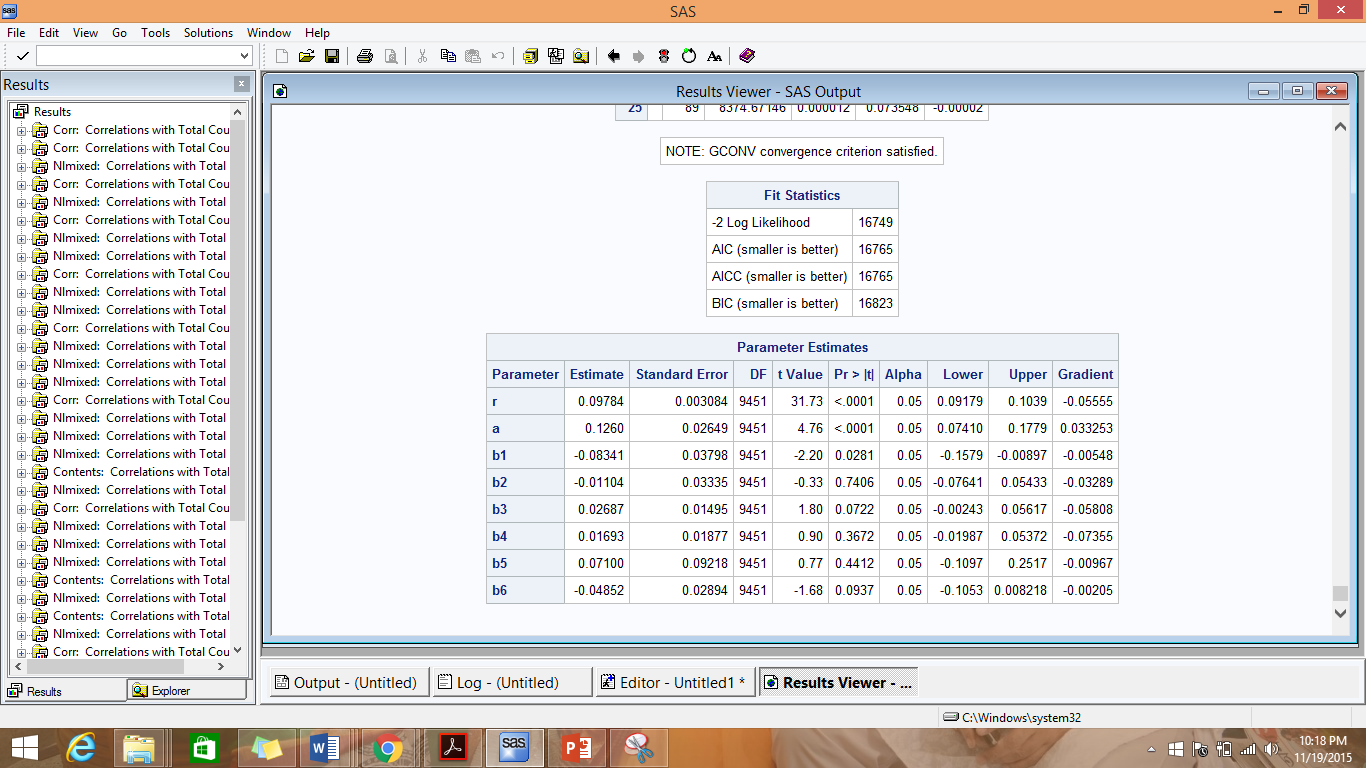
***term2 = (a / denom2)\*\*r;***

***term3 = (m/denom2)\*\*total\_count;***

***ll = log(term1\*term2\*term3);***

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

***run;***



**Comparison of NBD Regression model on the dataset with the Interaction effect with NBD Regression model on the original dataset:-**

ll of NBD Regression model on the dataset with the Interaction effect = -8374.5

ll of NBD Regression model on the original dataset = -8368.5

LR = -2(-8374.5 - -8368.5) = 12

χ2 (.05,k) for df=2 (because we have drooped region, race and country and added region\*race\*country to the analysis) 🡪 5.991

Thus, LR > χ2; signifying that the more complicated model I.e. the NBD Model on the original dataset is better than the model on the new dataset.

1. Interaction Effect 3- **age\*income\*education**

**Poisson Regression:-**

***Proc contents data = abi.Permdataq11 POSITION;***

***proc nlmixed data= abi.Permdataq11;***

***/\* m stands for lamdha \*/***

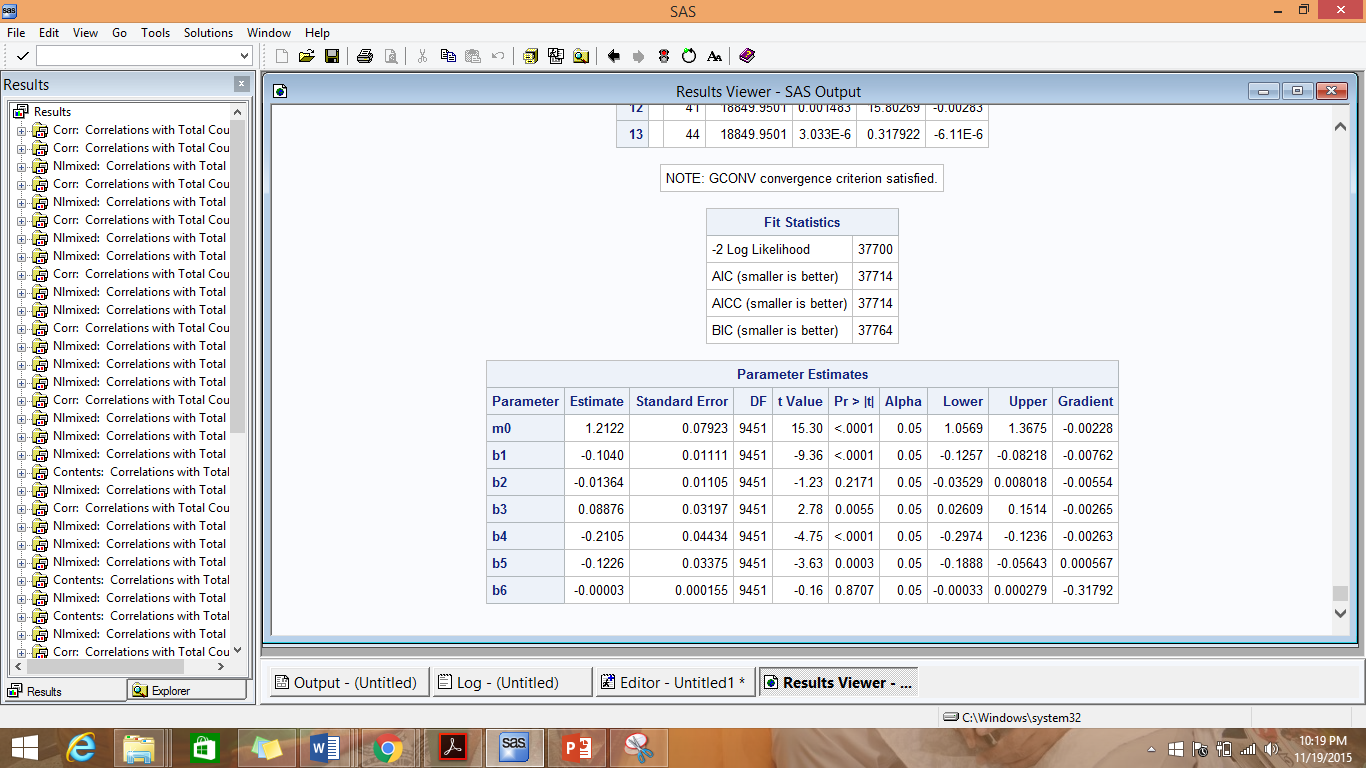
***parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0;***

***m=m0\*exp(b1\*region+b2\*hhsz+b3\*child+b4\*race+b5\*country+ b6\* parameter3);***

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

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

***run;***



**Comparison of Poisson Regression model on the dataset with the Interaction effect with Poisson Regression model on the original dataset:-**

ll of Poisson Regression model on the dataset with the Interaction effect = -18,850

ll of Poisson Regression model on the original dataset = -18,806

LR = -2(-18,850 - - 18,806) = 88

χ2 (.05,k) for df=2 (because we have drooped age, income and education and added age\*income\*education to the analysis) 🡪 5.991

Thus, LR > χ2; signifying that the more complicated model I.e. the Poisson Model on the original dataset is better than the model on the new dataset.

**NBD Regression:-**

***proc nlmixed data= abi.Permdataq11;***

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

***m=exp(b1\*region+b2\*hhsz+b3\*child+b4\*race+b5\*country+ b6\*parameter3);***

***num1 = gamma(r+total\_count);***

***denom1 = gamma(r) \* fact(total\_count);***

***term1 = num1 / denom1;***

***denom2 = a+m;***

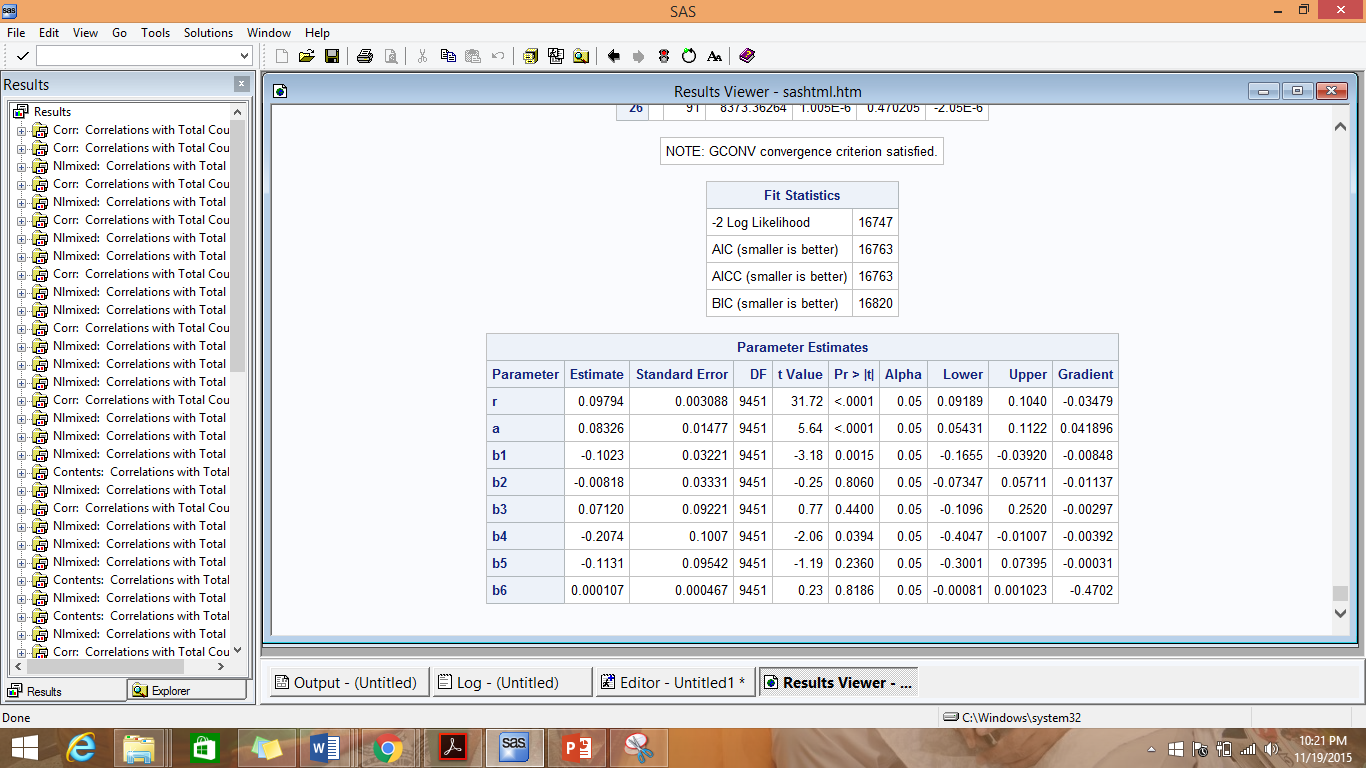
***term2 = (a / denom2)\*\*r;***

***term3 = (m/denom2)\*\*total\_count;***

***ll = log(term1\*term2\*term3);***

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

***run;***



**Comparison of NBD Regression model on the dataset with the Interaction effect with NBD Regression model on the original dataset:-**

ll of NBD Regression model on the dataset with the Interaction effect = -8373.5

ll of NBD Regression model on the original dataset = -8368.5

LR = -2(-8373.5 - -8368.5) = 10

χ2 (.05,k) for df=2 (because we have drooped age, income and education and added age\*income\*education to the analysis) 🡪 5.991

Thus, LR > χ2; signifying that the more complicated model I.e. the NBD Model on the original dataset is better than the model on the new dataset.

**Conclusion on adding the Interaction Effects:**

We thus observed that adding the interaction effects and removing the individual variables did not result in performance enhancement of the models.

**Conclusion drawn from using the Model Enhancement Techniques:-**

We observed a trend that the performance of a model on a dataset with more number of attributes is better than its performance on a dataset with lesser number of variables. As the dataset does not have a large number of attributes and none of the attributes are redundant and they provide new information about the customers, removing any of them is detrimental to the predictive power of the models. Techniques like Attribute Reduction and adding Interaction Effects did not produce the expected performance enhancement of the models probably because the effects were overshadowed by the loss of predictive power of the models due to removal of attributes.

# Part III: Why certain customers prefer Amazon over BN

1. Logistic Regression:-

We executed the logistic regression and results are reported

DATA Permdata\_q12;

SET ABI.Permdata;

/\* we created another column BN which is set to 1 for all customers who have bought from BN and 0 otherwise\*/

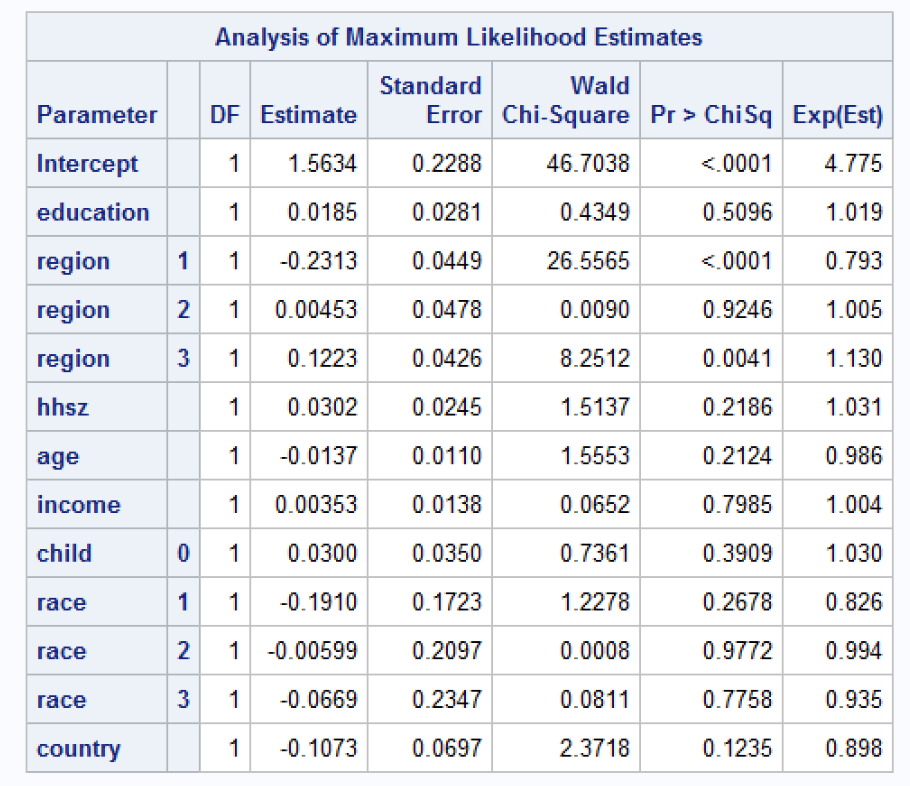
if total\_count > 0 then BN = 1; else BN = 0; run;

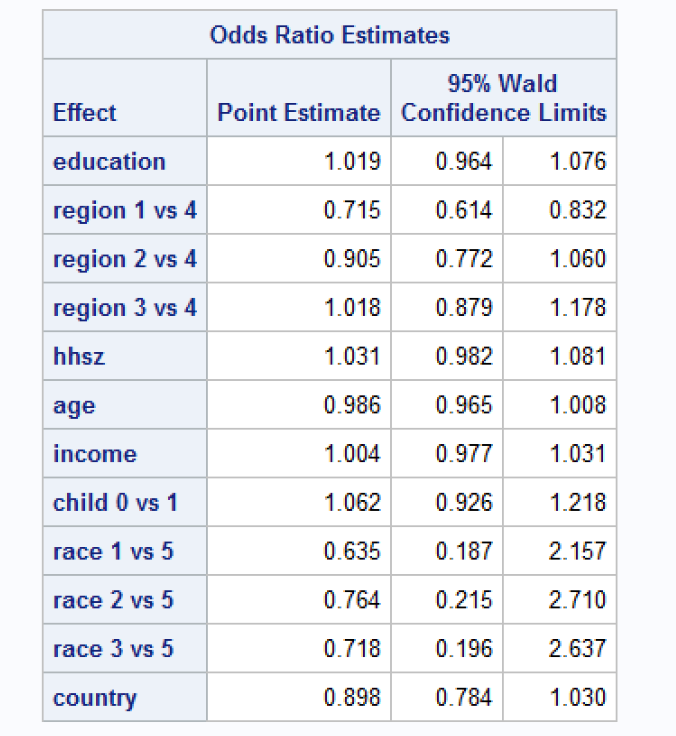
proc logistic data=Permdata\_q12;

class Region child race;

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

run;





From results, we observed region is the only independent variable that is significant for the regression.

**Managerial Takeaways:**

1. Living in region 1 vs region 4, increases the odds of purchasing from Barnes and Nobles (versus not purchasing from Barnes and Nobles) by a factor of 0.715
2. Living in region 2 vs region 4, increases the odds of purchasing from Barnes and Nobles (versus not purchasing from Barnes and Nobles) by a factor of 0.905
3. Living in region 3 vs region 4, increases the odds of purchasing from Barnes and Nobles (versus not purchasing from Barnes and Nobles) by a factor of 1.018

# Part IV: Summary

**Managerial Insights**

We used the NBD model results to calculate the ‘Reach’ or the proportion of customers who have bought books from BN atleast once in 2007 and ‘Average Frequency’ or the average number of books a Barnes and Noble customer has bought in 2007 which was 18.97% and 3.946 respectively. We thus concluded that both the reach and the frequency is quite low for BN and it needs to strengthen it marketing efforts in order to boost sale.

On studying and analyzing the data using various models, we formulated the following insights which can be used to make marketing strategies:

1. We observed that the effect of ‘Race’ is most pronounced in determining lambda. A change in the customer’s race (from 1 to 5) causes around 20% decrease in his/her purchase from BN. Hence BN should concentrate its customer engagement strategies on customers of Race 1 and 2. Churn of customers from Race 1 and 2 can prove to be detrimental for the sale, as a customer from Race 2 would buy 20% less books from BN than Race 1 every other factor keeping constant, Race 3 would buy 20% less than Race 2 and so on. Hence it would be profitable to maintain and attract customers of Race 1 and 2 as such customers tend to buy more books from BN. BN can target the community centers of Race 1 and 2 for their promotional activities and can also start discount offers for the festivals of these 2 races.
2. We have also observed that an Increase in customer’s education by 1 level causes a 8% decrease in their total purchase from BN. This might be due to the fact that a more educated population has more access to internet and thus may prefer online reading. We recommend BN to capitalize more on their slightly less educated customer base by stocking on more books of their choice.
3. We have observed from the Logistic Regression analysis that the probability of purchasing from BN (versus not purchasing from BN) is highest in Region 3 out of the 4 regions. BN in order to maintain the cash cow should keep its stores in Region 3 well stocked. Customers in Region 1 has the least probability of purchasing from BN (versus not purchasing from BN) and thus BN must focus its customer attraction and retention strategies in Region 1.
4. When we added the 3 additional attributes in the analysis, we observed that loyalty to BN and Seasons attributes were significant in determining the total purchase from BN. Hence, a customer who buys only from BN and not from Amazon tend to purchase more books than a customer who buys from both BN and Amazon. BN can introduce some loyalty points programs in order engage the loyal customer base. We also observed that the total purchase is maximum at Season =1 i.e Winters which is expected, it being the holiday season. BN can have ‘Winter special Sale’ to further boost the sale. BN can also have some discounts in Fall in order to promote sale, as Fall is the most unfavorable season for books purchase.
5. An increase in age of a customer by 1 level causes around 25.5% increase in their total purchase from BN, which is expected as we often observe that older people tend to prefer books to online reading. BN can retain their customer base by marketing the kind of books preferred by such age groups.
6. We also observed from the variable clusters that the variables Region, Race and Country are most important in determining total purchase from BN. Thus combing this analysis with that from the Regression models, we can create a portrait of an ideal BN customer who would buy the maximum number of books from BN. Such a customer is a person from US, from Region 1 and is from Race 1. BN can implement target marketing on these kind of customers who are more likely to spend on BN products.

**Learnings:**

This project gave us an opportunity to work with a real life dataset. We built customized BA models by coding in Base SAS and determined the business implications of the results of each our analysis; thus extracting business intelligence from raw data.

We gained immense learnings from working on the project, which can be categorized under the following-

1. Exploration of the raw data, performing various preprocessing techniques and building the customized models gave us substantial experience in coding in Base SAS.
2. The Regression Models strengthened our grasp of Statistics; by increasing our understanding of each statistical parameter and its effect on the business implications
3. We utilized our learning from Project 1, in the variable selection part. We deliberated on the suitability of various techniques for attribute reduction and brainstormed on the significance of each variable in determining the target variable and whether they are provide any additional information. This exercise made us appreciate the need to thoroughly understand the data before applying any kind of analysis on it.
4. We tried many techniques to improve the performance of the models and saw in accordance with the ‘No Free Lunch’ theorem that no single technique enhances model performance. How well a model performs and what techniques enhance model performance depends on the nature of the data. Thus, in our project some enhancement techniques gave expected results, whereas some did not.

Lastly, this project gave us an opportunity to consider the ‘business’ aspect of Business Analytics. Instead of mechanically building models, we tried to understand the data first, deliberated on the questions that we were trying to answer through our analysis, built model and determined the effect of each new information found in the analysis on the business.