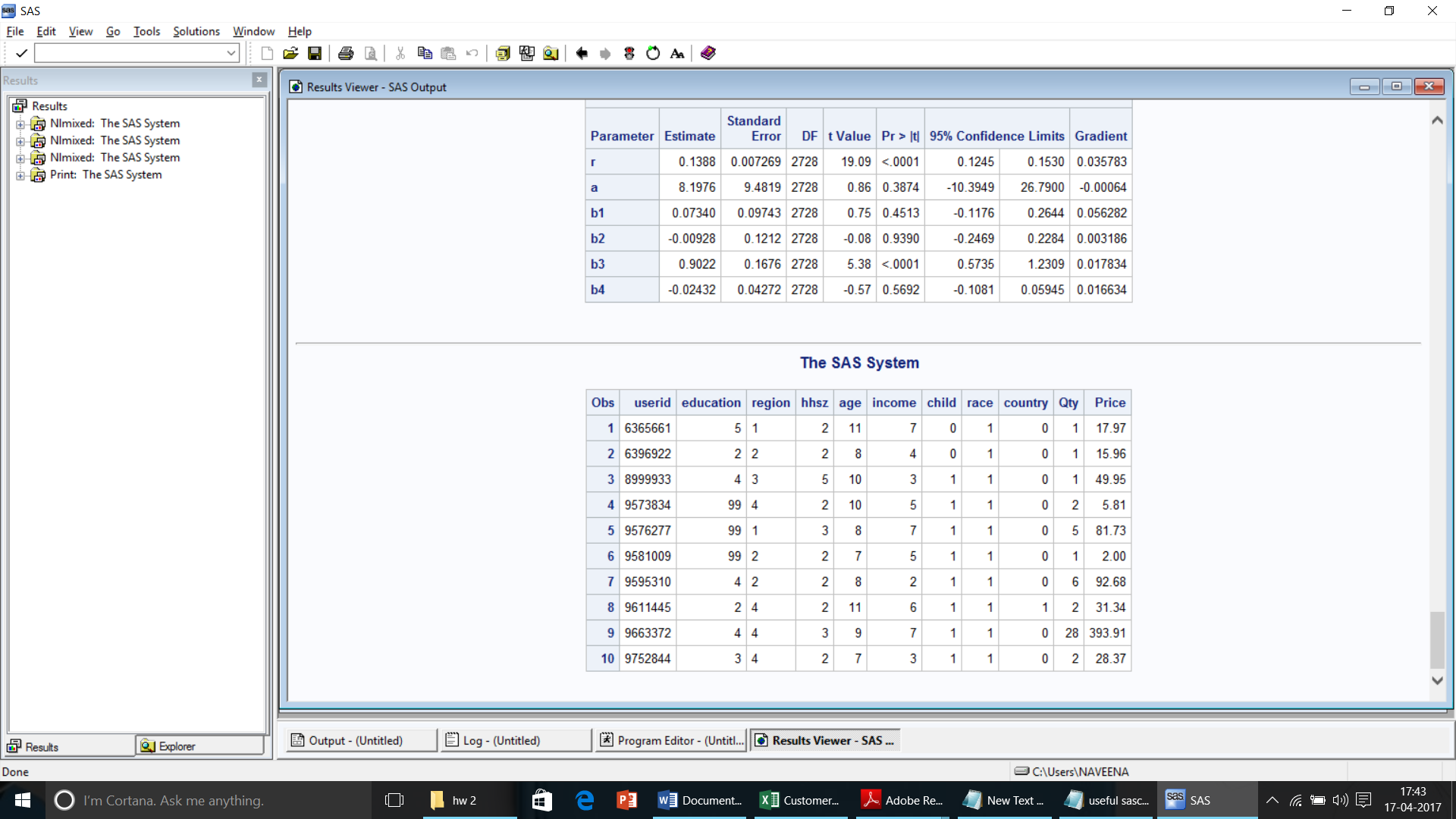
Applying NBD model, the Poisson regression model, and the NBD Regression Models on Customer data

Using these models to analyse the customer purchasing behaviour at two competitor stores, Amazon.com and Barnes & Noble.

The total number of books purchased from B&N for each of the first 10 customers is given below.



An NBD model was run on the books dataset and the parameters are computed as alpha=0.1426 and r= 0.1014.

PROC SQL;

CREATE TABLE Project.book1 AS

SELECT count(userid) AS users, qty from project.book1 group by qty;

run;

/\* The NBD Model \*/

PROC NLMIXED DATA=project.book1;

PARMS r=1 a=1;

retain Prob;

IF Qty = 0 THEN DO;

Prob1=(a/(a+1))\*\*r;

Prob=Prob1;

LL=users\*Log(Prob);

END;

ELSE IF Qty >0 THEN DO;

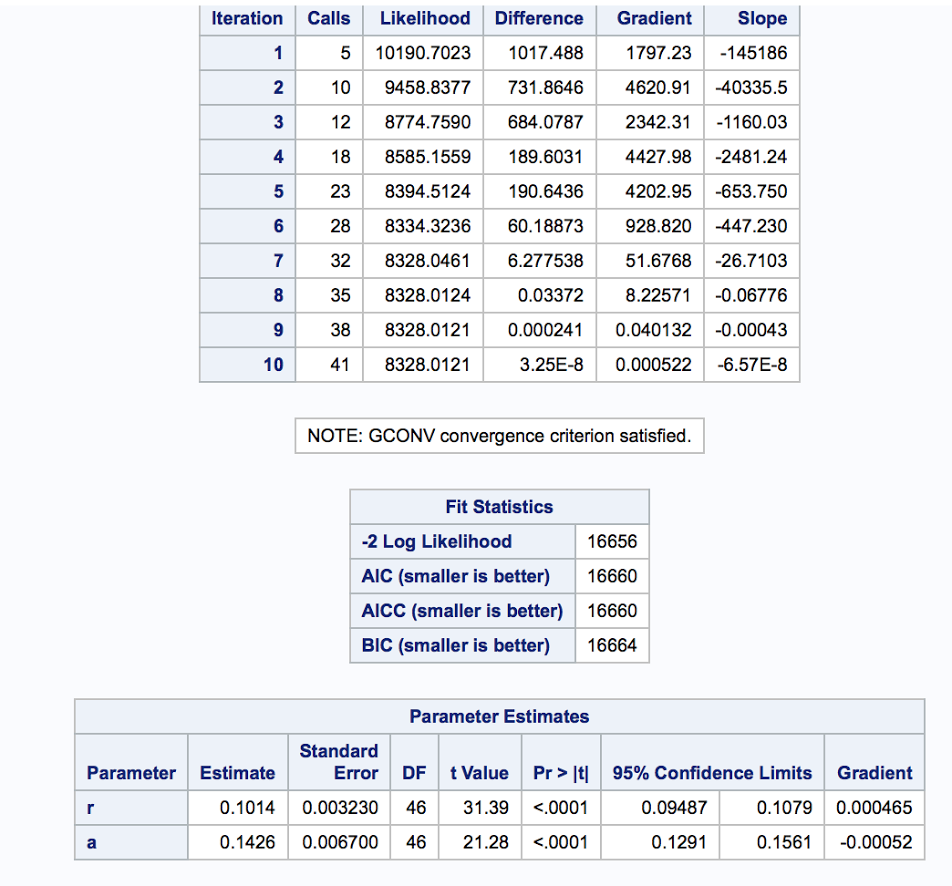
Prob=((r+(Qty-1))/(Qty\*(a+1)))\*Prob;

LL=users\*Log(Prob);

END;

MODEL Qty~general(LL);

RUN;



/\* The NBD Model \*/

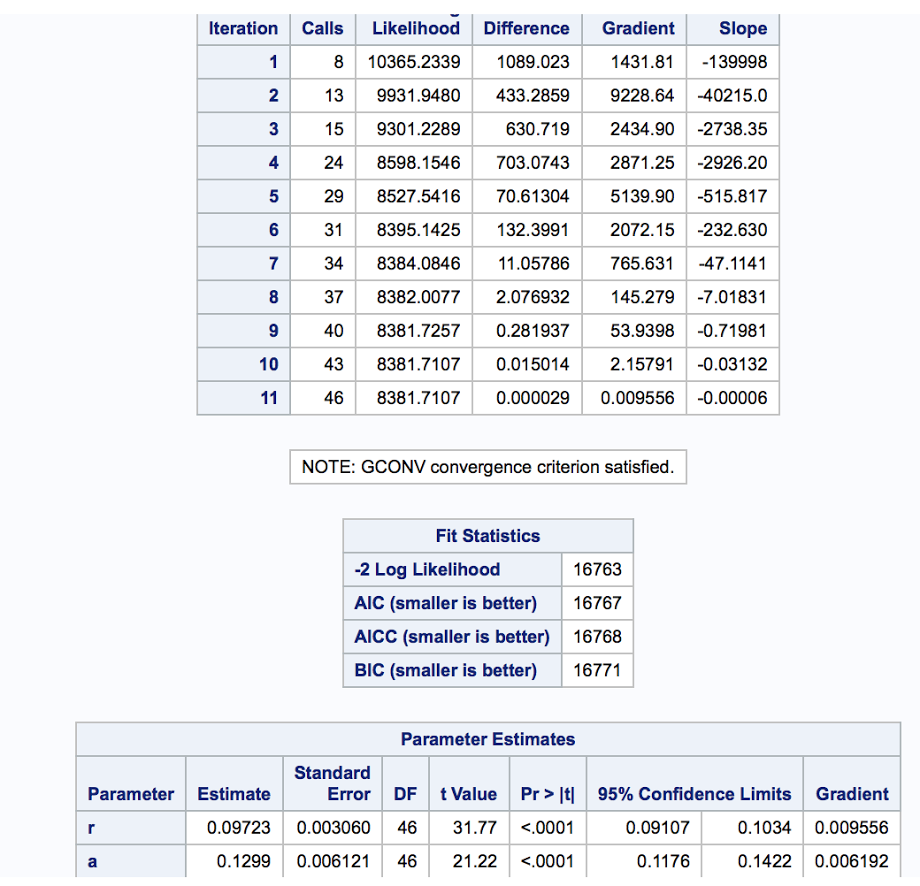
proc nlmixed data=project.book1;

parms r=1 a=1;

ll = users \* ( log(gamma(r+Qty))-log(gamma(r))-log(fact(Qty))+r\*log(a/(a+1))+Qty\*log(1/(a+1)) );

model Qty ~ general(ll);

run;



data a;

alpha=0.09723;

r=0.1299;

/\*Probability of purchasing zero books\*/

P0=(alpha/(alpha+1))\*\*r;

/\*Calculating Reach\*/

Reach=100\*(1-P0);

/\*Calculating Expected Value\*/

Ex=r/alpha;

/\*Calculating Average Frequency\*/

AvgFreq=Ex/(1-P0);

/\*Calculating Gross Rating Point\*/

GRP=Reach\*AvgFreq;

run;

**Reach** = **18.97%**

**Average** **Frequency** = **3.9456**

**Gross ratings Points (GRPs)** = **74.849**

From the above table we get the following values:

**Reach**=18.97 -i.e. 18.97% of population are buying at least one book in the given time period.

**Average Frequency**=3.945 i.e. 3.945 books are bought on an average by each person.

**Gross Rating Points**= 74.849 i.e. the average number of books bought per 100 people is 74.849.

The Poisson regression model was built using household size, age, income, children, country, region, race and education as the demographic variables.

/\*To convert region from character to numeric\*/

Data new;

set project.book1;

reg=input(region,2.);

run;

DATA project.book2;

set work.new;

drop region;

run;

/\*Poisson Regression\*/

proc nlmixed data=project.book2;

/\* m stands for lambda \*/

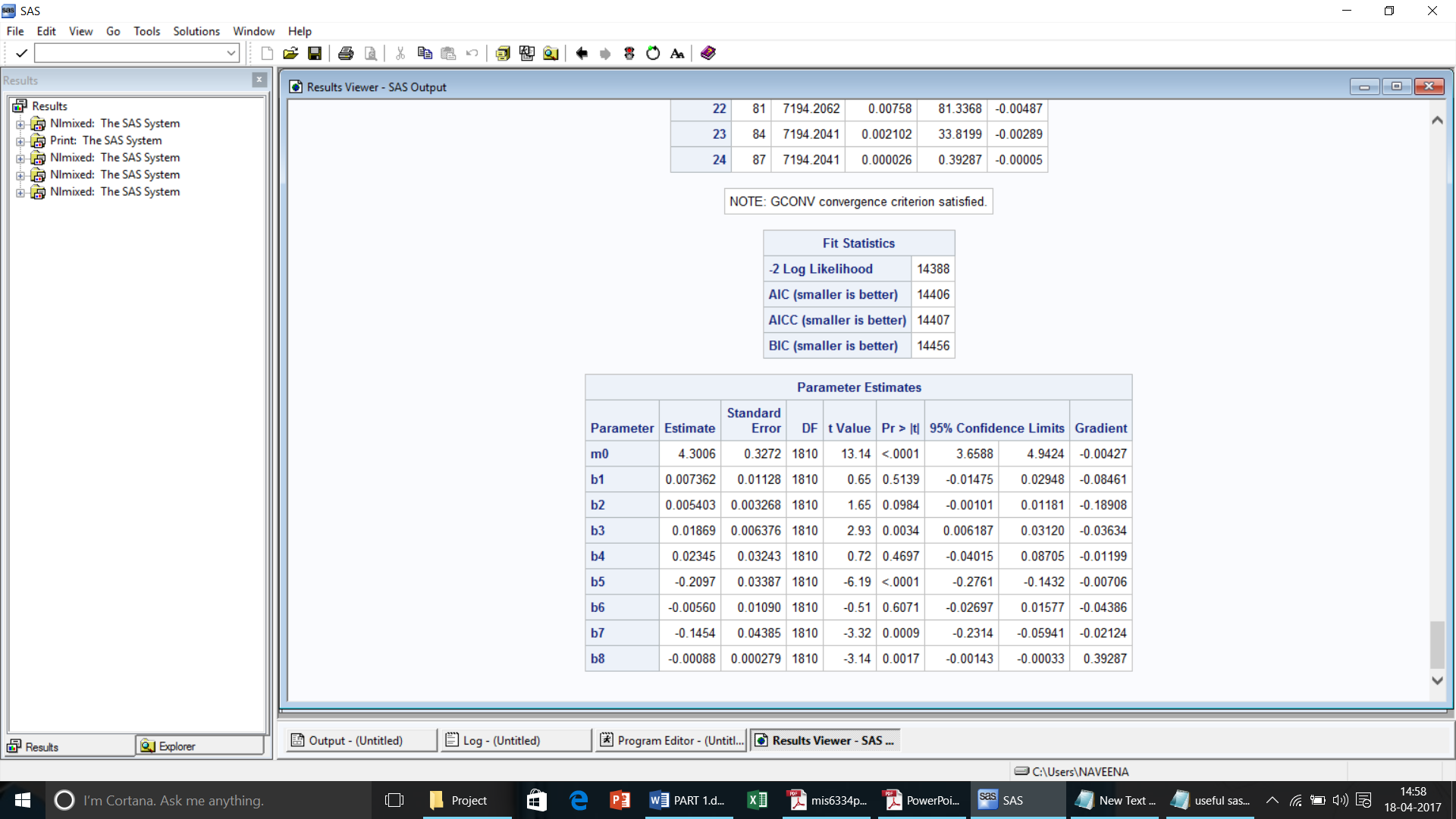
parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;

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

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

model Qty ~ general(ll);

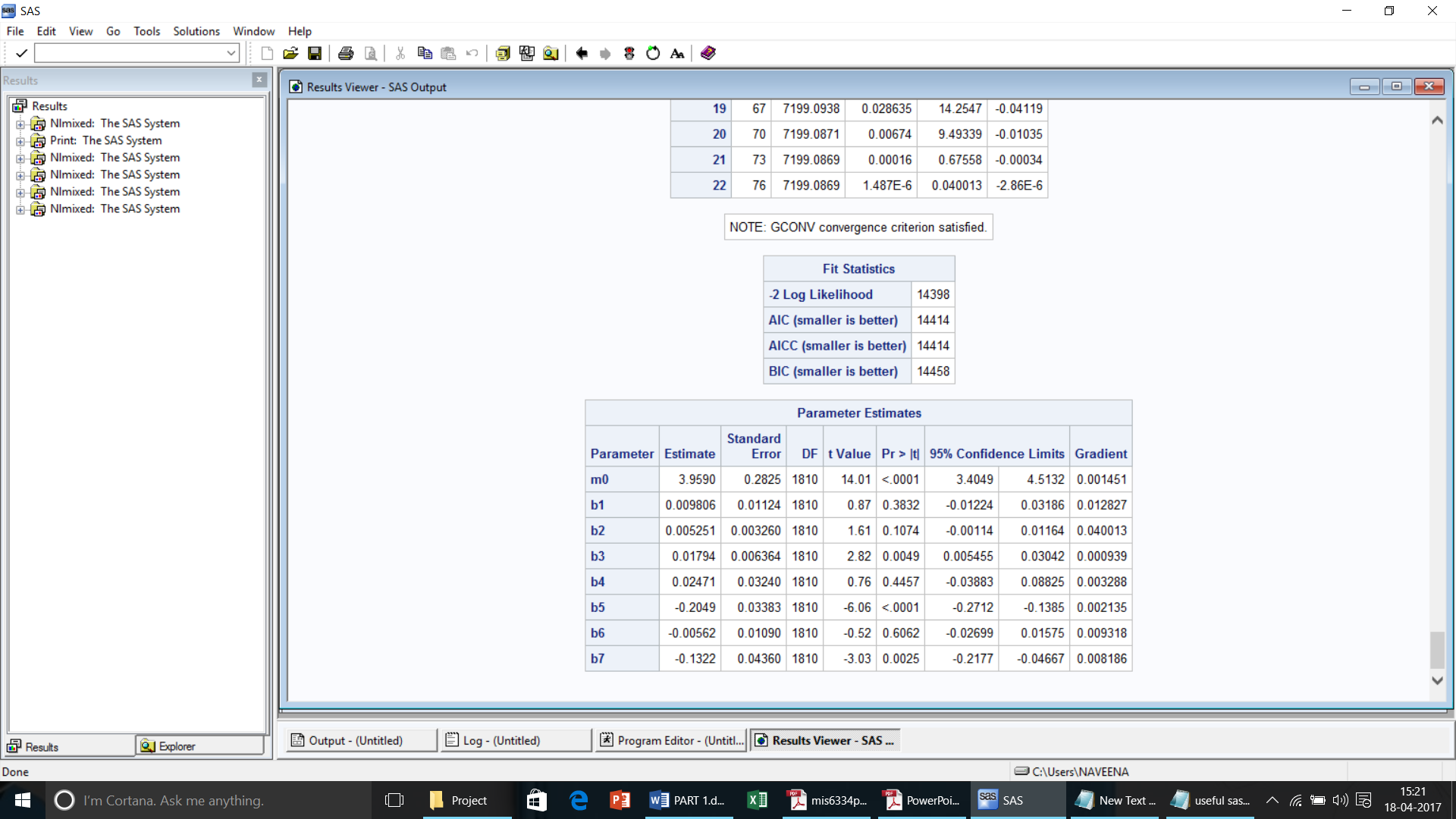
run;



From the output we observe the following:

* From the p values we observe that the parameters b3, b5, b7 and b8 are significant i.e. the variables income, country, race and education are highly significant compared to other parameters in determining the quantity of books purchased.
* Negative coefficient of the country variable indicates that customers from country 1 are less likely to purchase than customers from country 0 assuming all other variables as constant.
* Negative coefficient of region variable indicates that with increase in the value in region variable customers become less likely to purchase books while assuming other variables constant.
* Negative coefficient of race variable indicates that with increase in the value in race variable customers become less likely to purchase books while assuming other variables constant.
* Negative coefficient of education variable indicates that with increase in the value in education variable customers become less likely to purchase books while assuming other variables constant.
* Positive coefficients of household size, age, income and child variables imply that increase in these values customers become more likely to purchase books while assuming other variables constant.
* Child variable has the highest coefficient value among all variables.

Since, education has more number of missing values i.e. 73.7%, we are eliminating education variable from the data and running the regression again.



We can observe that there is slight variation in the coefficients of the remaining variables.

As the missing data is eliminated from the analysis we can now infer the results to be more accurate compared to the previous model.

The likelihood expression for the NBD regression model is given below:

ll = log(gamma(r+Qty))-log(gamma(r))-log(fact(Qty))+r\*log(a/(a+expBX))+Qty\*log(expBX/(a+expBX))

The NBD regression model was built using household size, age, income, children, country, region, race and education as the demographic variables.

/\*NBD Regression\*/

proc nlmixed data=project.book2;

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

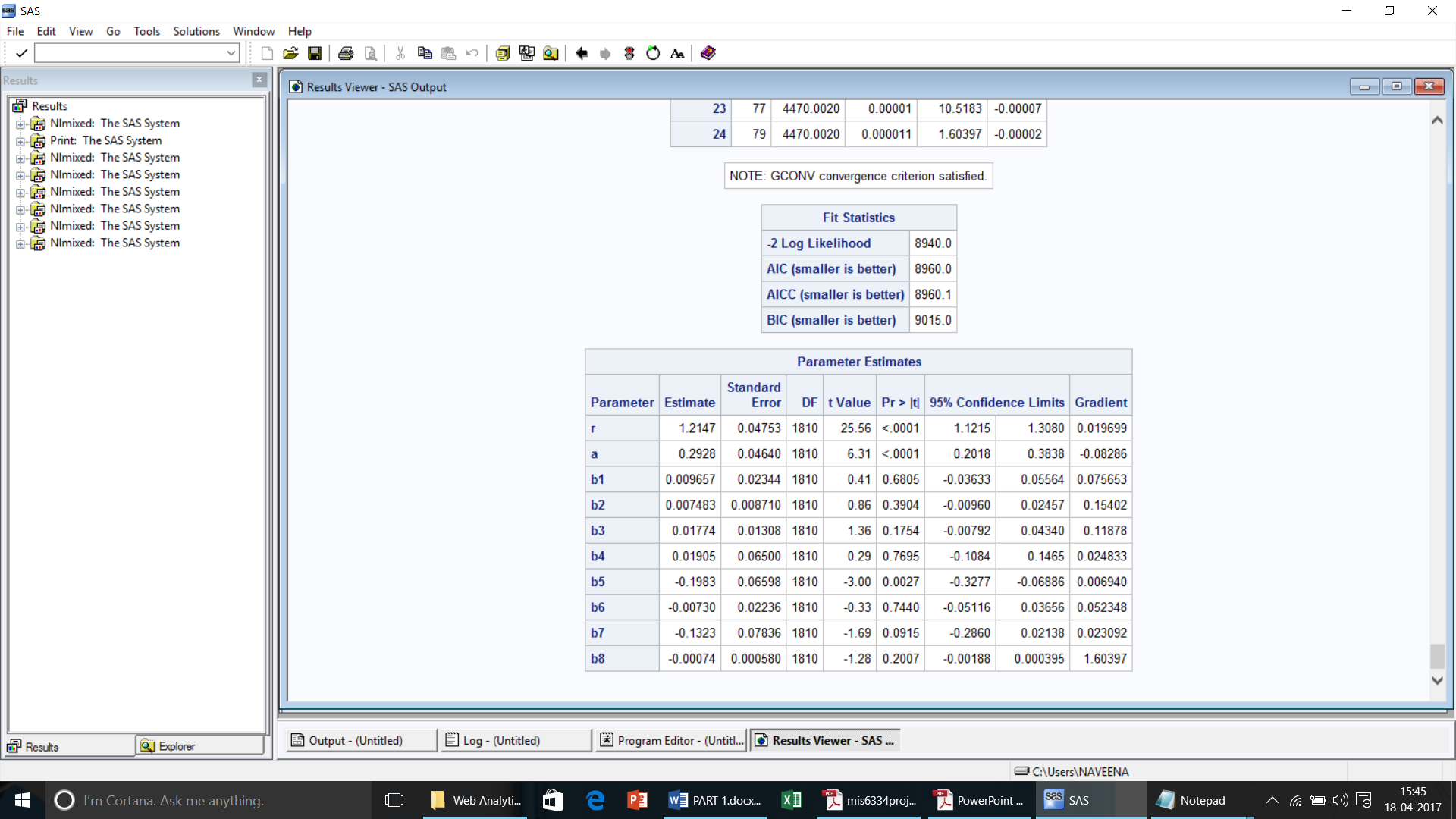
expBX=exp(b1\*hhsz+b2\*age+b3\*income+b4\*child+b5\*country++b6\*reg+b7\*race+b8\*education);

/\*loglikelihood expression\*/

ll = log(gamma(r+Qty))-log(gamma(r))-log(fact(Qty))+r\*log(a/(a+expBX))+Qty\*log(expBX/(a+expBX));

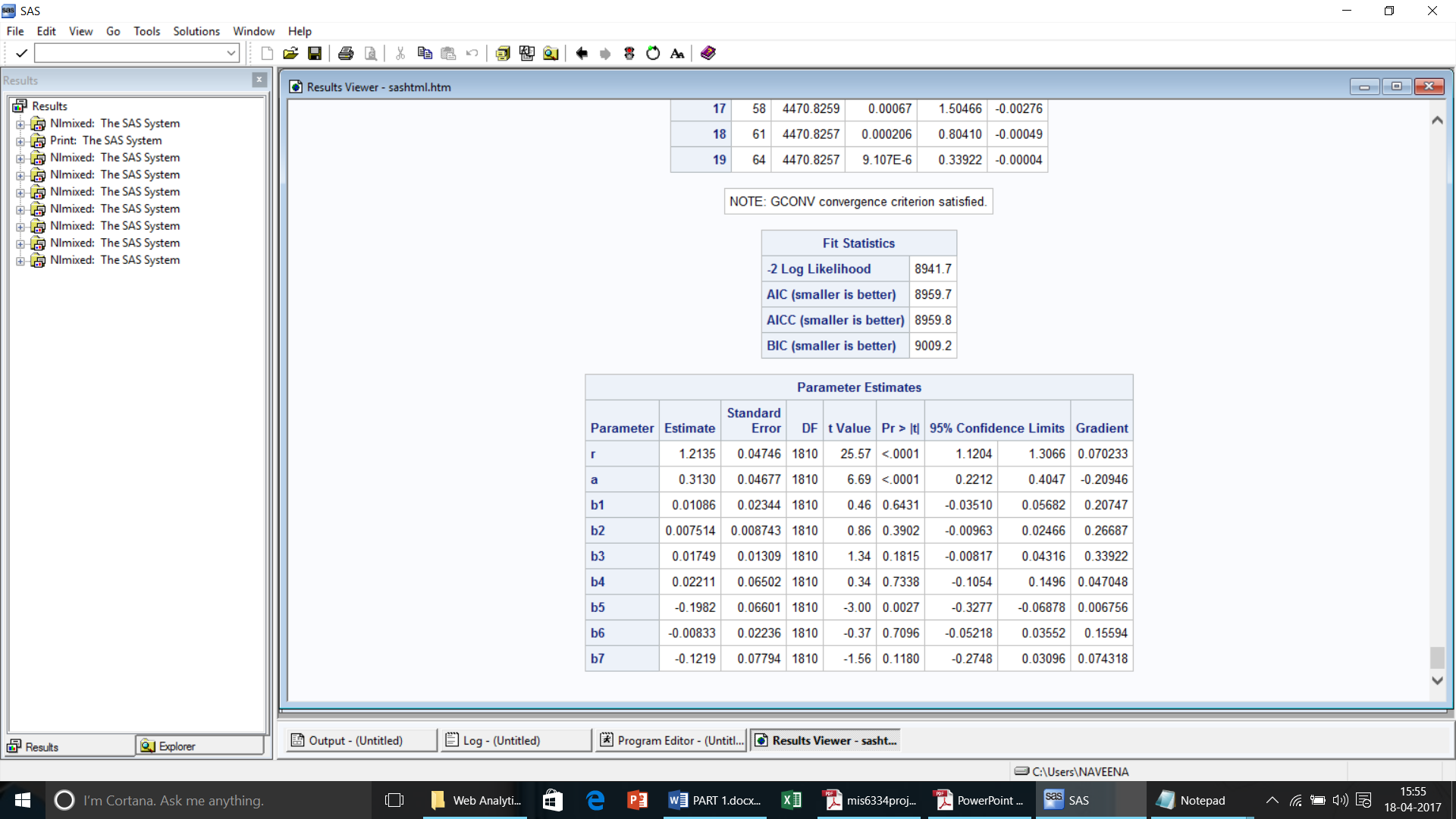
model Qty ~ general(ll);

run;



* From the p values we observe that the parameters b5 is significant i.e. the variable country is highly significant compared to other parameters in determining the quantity of books purchased.
* Negative coefficients of country, region, race and education variables indicate that with increase in the value of these variables customers become less likely to purchase books while assuming other variables constant.
* Positive coefficients of household size, age, income and child variables imply that increase in these values, customers become more likely to purchase books while assuming other variables constant.
* Child variable has the highest coefficient value among all variables.

Since, education has more number of missing values i.e. 73.7%, we are eliminating education variable from the data and running the regression again.



We can observe that there is slight variation in the coefficients of the remaining variables.

As the missing data is eliminated from the analysis we can now infer the results to be more accurate compared to the previous model.

We see that there are significant differences between the results of Poisson regression and NBD regression. From the Poisson regression results b3, b5, b7 and b8 are significant whereas in NBD regression only b5 is observed to be significant. This difference can be attributed to the fact that the gamma distribution in the NBD regression model captures the differences among the individuals i.e. the variables which were significant in the Poisson regression are not significant in NBD regression because the gamma distribution might have captured the variation among individuals for these variables.

Thus, from the analysis, we observe that

* Based on the type of data, we are now able to determine which model would be the best fit for the data to make predictions.
* When there are unobserved differences among the records (for instance people in our case) we use NBD model whereas when there are variables that explain the individual characteristics we use regression model to determine the effect of these characteristics on the outcome variable.
* When the given explanatory variables don’t fully capture the differences among the individuals we use NBD regression model where the gamma distribution in the likelihood function that captures the differences among the individuals. This model gives us a better fit.
* When we are given data that exhibits individual characteristics and demographics we can use regression models to identify the important characteristics that differentiate the outcome variable.