Virtual Surf or Physical Serve?

Book-buying insights for Amazon and Barnes &Noble

Hua Guo
Yanwei Jia
Yi Liu
Pengfei Liu
Fei Cen
Lu Chen

Advanced Business Intelligence Group 6
Professor Xianjun Geng

The Group 6 of Advanced Business Intelligence for Consumer and Customer Insight applies a unique, integrated approach that combines quantitative and qualitative consumer research with a deep understanding of business strategy and competitive dynamics. The Group's works hit the best results among all the groups, and the group got the highest praise of the professor.

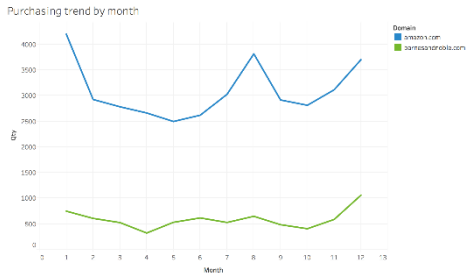
The Advanced Business Intelligence course is for machine learning and business strategy by leading student hands on real world business analysis. The course is the most popular and the hardest course in the Information Technology and Management Department of the University of Texas at Dallas. And the instructor, Professor Geng, is the most popular professor.

Executive Summary

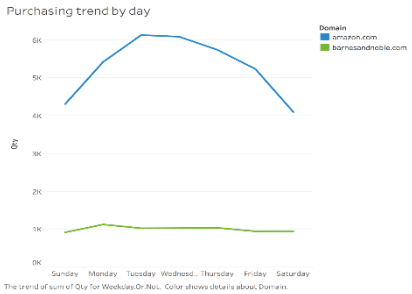
The negative binomial model (NBD) form is more appropriate than the Poisson, for NBD model has a better optimized LL value. NBD model can capture unobserved heterogeneity in variables.

Four variables, education, region, race, and age may have statistical significance for book-buying propensity. Constructed variables, month and day also reflect variance in consumer behavior. Key findings are presented below.

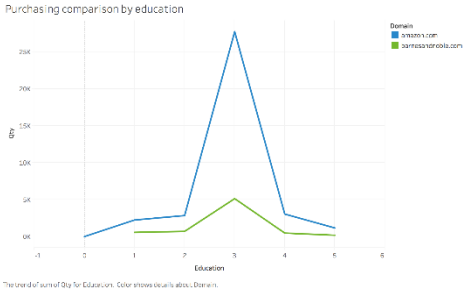
Sales peak in matriculation season



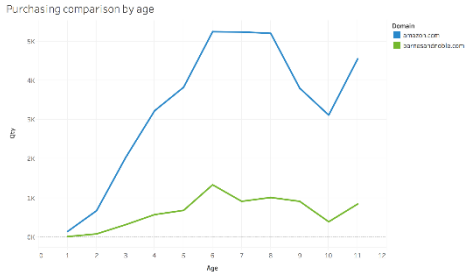
Purchase hikes from Tuesday to Thursday



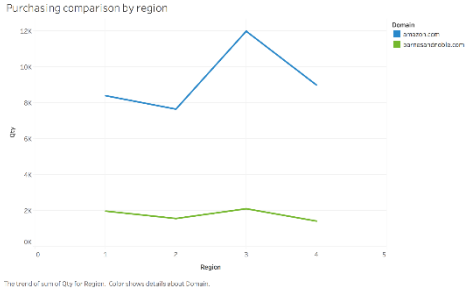
Purchase concentrate in buyers with education level 3



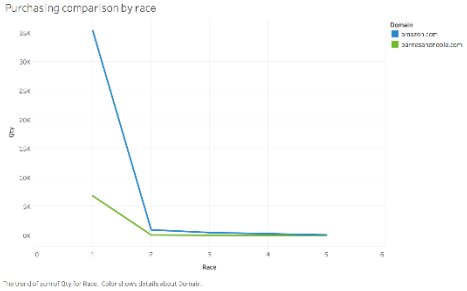
Purchase plummets after age 8 but rallies at age 11



Region 3 counts for the most of the purchase



Overwhelming majority of purchaser is of ethnic group 1



Contents

1. Preprocess the Dataset

Little has been written about the book-buying behaviors. With consumer behavior changing, forward-looking companies need to create effective strategies for wining its business.

Before beginning, we'll check the data quality to establish analysis strategies.

1.1 Check data type

First, we checked the data type of each features to make sure the data be analyzed is right.

```
*Set a libraby and read the dataset in the library;
libname project2 "C:\ABI";
DATA project2.aba;
INFILE "C:\ABI\aba.sas7bat";
RUN;

*Check datatype to make sure each variable's type is right;
PROC CONTENTS data = project2.aba  position;
RUN;
```

Figure 1.1 Data type checking results

Variables in Creation Order						
#	Variable	Type	Len	Format	Informat	Label
1	userid	Num	8			userid
2	education	Num	8			education
3	region	Num	8			region
4	hhsz	Num	8			hhsz
5	age	Num	8			age
6	income	Num	8			income
7	child	Num	8			child
8	race	Num	8			race
9	country	Num	8			country
10	domain	Char	18	$18.	$18.	domain
11	date	Num	8			date
12	product	Char	215	$215.	$215.	product
13	qty	Num	8			qty
14	price	Num	8			price

From the above results, we can find that variable domain and product are Char, all the other variables are Num. Here, we'll change the date's format to 'YYMMDD8.' for future analysis.

```
*Change DATE format from num to date;
DATA project2.aba_date;
SET  project2.aba;
x =  INPUT(PUT(DATE,8.),YYMMDD10.);
FORMAT x YYMMDD10.;
DROP date;
RENAME x = date;
RUN;
```

1.2 Check missing value

```
*Count the nmber of missing value;
```

```
DATA project2.aba_mis (KEEP = totalRecords missing_edu missing_region
missing_age);
SET project2.aba_date END = last;
RETAIN totalRecords missing_edu missing_region missing_age 0;
totalRecords+1;
IF education = 99 THEN DO;
   missing_edu+1;
END;
IF region ^IN (1,2,3,4) THEN DO;
   missing_region+1;
END;
IF age = 99 THEN DO;
   missing_age+1;
END;
IF last;
RUN;
PROC PRINT DATA = project2.aba_mis;
RUN;
```

Figure 1.1 Missing value checking results

Obs	totalRecords	missing_edu	missing_region	missing_age
1	40945	30238	46	3

From above table, we found that variable education contains too many missing value (75%), we can discard this feature because even after imputation, it will not generate accurate estimation. On the other hand, the education variable still has more than 10,000 records. This number is big enough for analyzing. We think any feature cannot be discarded easily, so we may try use the 10,000 + records in the following analysis, and may impute the missing value for comparing analysis results.

1.3 Standardize missing value

We found education, region, and age has missing value in different format, before we go further process, we decided to standardize them for following steps.

```
*Standlize missing value;
DATA project2.aba_stand;
SET project2.aba_date;
IF  education = 99 THEN DO;
   education = .;
END;
IF  region ^IN (1,2,3,4) THEN DO;
   region = .;
END;
IF  age = 99 THEN DO;
   age = .;
END;
RUN;
```

1.4 Check dataset properties

We will check the mean, median, mode and skewness of the data.

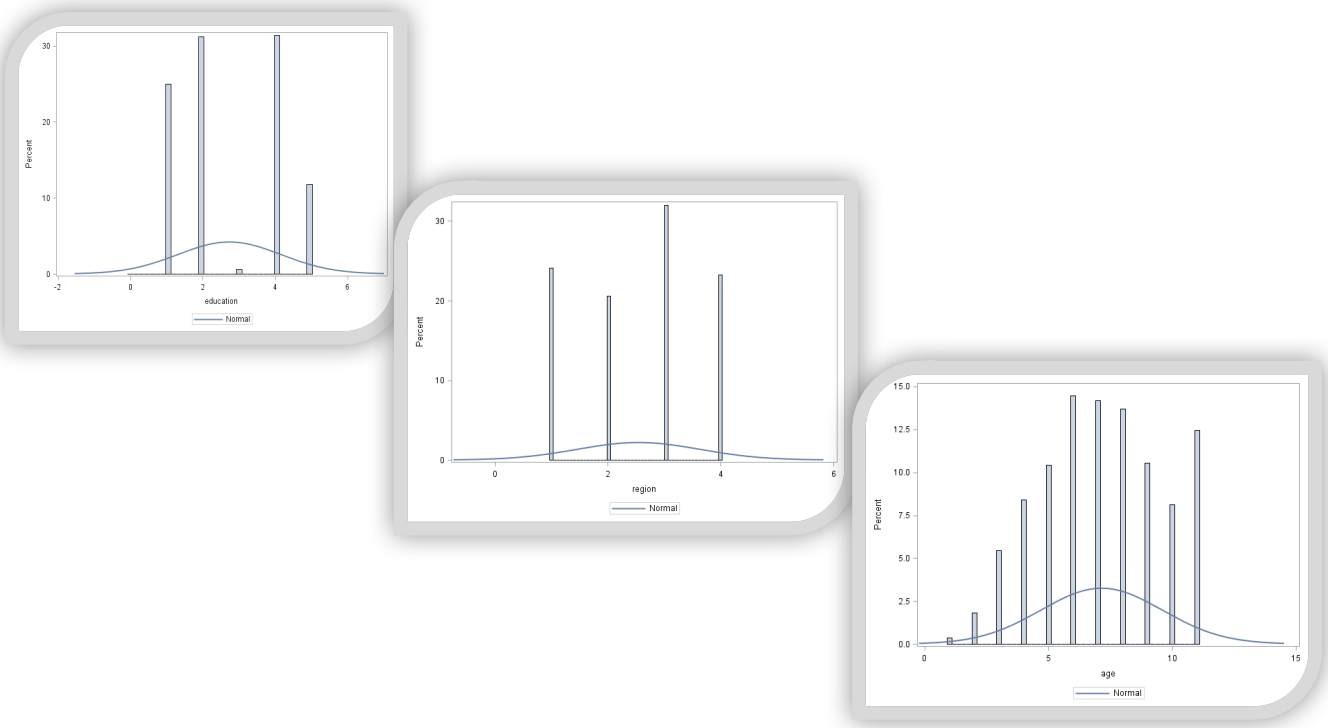
```
PROC MEANS mean median mode data = project2.aba_stand;
RUN;
PROC UNIVARIATE data = project2.aba_stand;
inset skewness;
```

```
RUN;
```

Figure 1.4.1 Mean, median and mode

Variable	Label	Mean	Median	Mode
userid	userid	14007010.41	14411385.00	14648977.00
education	education	2.7362473	2.0000000	4.0000000
region	region	2.5436808	3.0000000	3.0000000
hhsz	hhsz	3.1741116	3.0000000	2.0000000
age	age	7.1519955	7.0000000	6.0000000
income	income	4.7151301	5.0000000	7.0000000
child	child	0.7082672	1.0000000	1.0000000
race	race	1.0659665	1.0000000	1.0000000
country	country	0.1559165	0	0
qty	qty	1.0779094	1.0000000	1.0000000
price	price	15.9082052	10.8500000	6.9900000
date		17349.82	17356.00	17176.00

Figure 1.4.2 Distribution plot of education(left), region(central), and age(right)



From the means procedure and distribution plots, we can found that our data is similar normal distributed. Moreover, our data is more than 40,000, based on central limit theorem we can treat our data as normal distributed.

1.5 Impute missing value

Since the missing value of age and value only count for very small fraction of total data, 0.1% or less respectively, we may just drop the data with negligible influence. However, we decide to fill these missing values instead of just drop the data.

Age and Region have every small skewness, and the rounded mean equals to median. The age and region should be integer, so we will fill the missing age by using rounded mean (= median).

Figure 1.5.1 The UNIVARIATE Procedure
Variable: age (age)

Moments			
N	40942	**Sum Weights**	40942
Mean	7.15199551	**Sum Observations**	292817
Std Deviation	2.45728081	**Variance**	6.03822896
Skewness	-0.0962596	**Kurtosis**	-0.8334176
Uncorrected SS	2341437	**Corrected SS**	247211.132
Coeff Variation	34.3579747	**Std Error Mean**	0.01214424

Figure 1.5.2 The UNIVARIATE Procedure

Variable: region (region)

Moments			
N	40899	**Sum Weights**	40899
Mean	2.54368077	**Sum Observations**	104034
Std Deviation	1.09351395	**Variance**	1.19577276
Skewness	-0.1320303	**Kurtosis**	-1.2878173
Uncorrected SS	313534	**Corrected SS**	48904.7143
Coeff Variation	42.9894333	**Std Error Mean**	0.00540714

```
DATA project2.aba_imp_mean;
SET project2.aba_stand END = last;
IF region ^IN (1, 2, 3, 4) THEN DO;
   region = 3;
END;
IF age = . THEN DO;
   age = 7;
END;
RUN;
```

Education has over 75% missing values, we'll keep it as original for Question 1-3, while we'll use bagged trees to fill the missing values for Question 4.

```
##use caret preprocess to impute the missing value##
library("caret")
book <- read.csv("aba_imp_mean.csv")
```

```
book_preproc <- preProcess(book,method = c("bagImpute"))
book_impute <- predict(book_preproc,book)

book_impute$education <- round(book_impute$education)
write.csv(book_impute,"aba_imputed.csv")
```

2. Part I. Modeling Count Data

Q1 Count the number of books purchased from BN

Process the raw data using SAS to generate a count dataset in a format similar to the raw data in thec"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.

```
PROC SORT DATA = project2.aba_imp_mean;
   by userid;*sort the data by userid to make sure it will be okay to
use in data steps later;
RUN;

DATA project2.aba_BN (DROP = domain date product qty price);
SET project2.aba_imp_mean;
by userid;
IF first.userid THEN count = 0; *initialize the number of purchased
book;
IF domain = 'barnesandnoble.com' THEN count+qty; *add the qty purchased
from BN;
IF last.userid; *keep the final sum;
RUN;

PROC PRINT DATA = project2.aba_BN (OBS=10);
    TITLE 'Number of books purchased from B&N by all customers (10
observations)';
RUN;
```

Figure 2.Q1.1 Number of books purchased from B&N by all customers (10 observations)

Obs	userid	education	region	hhsz	age	income	child	race	country	count
1	6365661	5	1	2	11	7	0	1	0	1
2	6388054	2	4	1	6	5	0	1	0	0
3	6396922	2	2	2	8	4	0	1	0	1
4	6421559	5	4	4	5	6	0	1	0	0
5	6467806	.	2	2	6	3	0	1	0	0
6	6628110	4	4	5	4	7	1	1	0	0
7	6631403	5	3	1	10	3	0	1	1	0
8	6704851	5	4	1	6	7	0	1	0	0
9	7412556	5	4	3	10	7	0	1	1	0
10	8147707	4	2	3	4	3	1	1	0	0

To show B&N's consumer purchasing results, we should add one more if condition as following code:

```
PROC PRINT DATA = project2.aba_BN (OBS=10);
   WHERE count > 0 ;
   TITLE 'Number of books purchased from B&N by B&N customer (10
observations)';
RUN;
```

Figure 2.Q1.2 Number of books purchased from B&N by all customers (10 observations)

Obs	userid	education	region	hhsz	age	income	child	race	country	count
1	6365661	5	1	2	11	7	0	1	0	1
3	6396922	2	2	2	8	4	0	1	0	1
12	8999933	4	3	5	10	3	1	1	0	1
19	9573834	.	4	2	10	5	1	1	0	2
20	9576277	.	1	3	8	7	1	1	0	5
22	9581009	.	2	2	7	5	1	1	0	1
24	9595310	4	2	2	8	2	1	1	0	6
31	9611445	2	4	2	11	6	1	1	1	2
34	9663372	4	4	3	9	7	1	1	0	28
36	9752844	3	4	2	7	3	1	1	0	2

Q2 NBD Model

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.)

We start with creating new dataset include frequency of book purchase from B&N, with the number of people in each frequency category.

```
PROC SORT DATA = project2.aba_BN;
BY count;
RUN;
DATA project2.aba_BN_NBD (KEEP = count peoplecount rename = (count =
bookbought));
*bookbought = the numbers of book purchased,
peoplecount = the numbers of customer who bought particular numbers of
books;
SET  project2.aba_BN;
BY count;
IF first.count THEN peoplecount = 0;
peoplecount+1;
IF last.count;
RUN;
```

Now, run the NBD model on aba_BN_NBD dataset.

```
PROC NLMIXED DATA = project2.aba_BN_NBD;
PARMS r=1 a=1;
```

```
ll=peoplecount*log((gamma(r+bookbought)/(gamma(r)*fact(bookbought)))*((
a/(a+1))**r)*(
(1/(a+1))**bookbought));
model peoplecount ~ general(ll);
run;
```

Figure 2.Q2.1 Iteration History

Iteration History						
Iter		**Calls**	**NegLogLike**	**Diff**	**MaxGrad**	**Slope**
1		8	10365.2339	1089.023	1431.815	-139998
2		13	9931.94799	433.2859	9228.643	-40215
3		15	9301.22895	630.719	2434.901	-2738.35
4		24	8598.15462	703.0743	2871.251	-2926.2
5		29	8527.54158	70.61304	5139.898	-515.817
6		31	8395.14247	132.3991	2072.149	-232.63
7		34	8384.08461	11.05786	765.6312	-47.1141
8		37	8382.00768	2.076932	145.2793	-7.01831
9		40	8381.72574	0.281937	53.93981	-0.71981
10		43	8381.71073	0.015014	2.157912	-0.03132
11		46	8381.7107	0.000029	0.009556	-0.00006

Figure 2.Q2.2 Parameter Estimates

Parameter Estimates									
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**Alpha**	**Lower**	**Upper**	**Gradient**
r	0.09723	0.003060	46	31.77	<.0001	0.05	0.09107	0.1034	0.009556
a	0.1299	0.006121	46	21.22	<.0001	0.05	0.1176	0.1422	0.006192

Optimized LL value = -8381.7107 as shown in last row of figure 2.Q2.1.

Estimated parameter value for *r* and *a* is 0.09723 and 0.1299, respectively.

Each of the parameter has a p-value less than 0.001, meaning they are robust.

Q3 Reach, Average Frequency and GRPs

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

Given r = 0.09723 and a = 0.1299, we have

P(X(t) =0)|r, a) = (a/(a+r))^r = (0.1299/1.1299)^0.09723 = 0.8103

E(x) = r/a = 0.09723/0.1299 = 0.7485

Reach: 1 – P(X = 0) = 1 – 0.8103 = 0.1897

Average Frequency: E(x) / Reach = 0.7485 / 0.1897 = 3.9457

GRPs: 100 * E(x) = 74.85

Q4 Poisson Regression Model

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?)

In this step, we'll compare the results of keep original education variable, drop education variable, and imputed education. First, let's see the results.

```
*Keep original education;
PROC NLMIXED DATA = project2.aba_BN;
*m stands for lamdha 1.region 2.hhsz 3.age 4,income 5.child 6.race
7.country 8.education;
PARMS m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;
m=m0*exp(b1*region + b2*hhsz + b3*age +b4*income +b5*child + b6*race +
b7*country + b8*education);
ll = count * log(m) - m - log(fact(count));
model count ~ general(ll);
RUN;
```

Figure 2.Q4.1 Iteration History (keep education)

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	7	5335.93007	68.52481	1010.059	-329719
2	10	5236.49234	99.43773	258.1952	-8654.18
3	13	5229.61869	6.873653	316.2926	-518.021
4	16	5223.53299	6.085698	205.7796	-284.827
5	18	5207.38998	16.14301	196.6932	-238.166
6	21	5206.03328	1.356697	164.7983	-83.0562
7	25	5200.72074	5.312543	51.93542	-13.7424
8	28	5200.34785	0.37289	12.51784	-2.52062
9	31	5200.1984	0.149456	41.07219	-0.54748
10	34	5200.13719	0.06121	42.95756	-0.14437
11	37	5200.10726	0.029925	24.49258	-0.0239
12	40	5200.09918	0.008077	0.36936	-0.01882
13	43	5200.09907	0.000119	0.358829	-0.00023
14	46	5200.09906	6.817E-7	0.001534	-1.24E-6

Figure 2.Q4.2 Parameter Estimates (keep original education)

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
m0	1.2246	0.1694	2537	7.23	<.0001	0.05	0.8924	1.5567	-0.00018
b1	-0.1886	0.02152	2537	-8.76	<.0001	0.05	-0.2307	-0.1464	0.001534
b2	-0.06638	0.02110	2537	-3.15	0.0017	0.05	-0.1078	-0.02500	-0.00141
b3	0.03102	0.01048	2537	2.96	0.0031	0.05	0.01046	0.05158	0.001411
b4	0.05632	0.01285	2537	4.38	<.0001	0.05	0.03112	0.08151	-0.00007
b5	0.2922	0.06630	2537	4.41	<.0001	0.05	0.1622	0.4222	-0.00019

b6	-0.08510	0.05723	2537	-1.49	0.1371	0.05	-0.1973	0.02711	0.000277
b7	-0.3970	0.06624	2537	-5.99	<.0001	0.05	-0.5269	-0.2671	0.000382
b8	-0.1292	0.01648	2537	-7.84	<.0001	0.05	-0.1616	-0.09693	0.000195

From figure 2.Q4.2, we can see education matters (p-value < 0.0001). Let's see the results of drop education:

```
*Drop education;
DATA project2.aba_BN_DropEdu (DROP = education);
SET project2.aba_BN;
PROC NLMIXED DATA = project2.aba_BN;
*m stands for lamdha 1.region 2.hhsz 3.age 4,income 5.child 6.race
7.country;
PARMS m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;
m=m0*exp(b1*region + b2*hhsz + b3*age +b4*income +b5*child + b6*race +
b7*country);
ll = count * log(m) - m - log(fact(count));
model count ~ general(ll);
                                        RUN;
```

Figure 2.Q4.3 Iteration History (drop education)

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	7	19131.3561	118.2637	11148.89	-4732921
2	10	19019.6203	111.7358	9266.612	-24720.9
3	13	19007.6722	11.94811	9587.303	-4739.42
4	16	18993.5093	14.16297	9221.457	-1209.33
5	18	18965.1286	28.38068	6604.476	-1653.16
6	22	18895.8788	69.24975	4059.086	-384.994
7	24	18866.5994	29.27939	3510.019	-467.328
8	27	18851.5934	15.00599	3529.117	-159.313
9	29	18834.0776	17.51585	55.73741	-49.6731
10	33	18833.433	0.644568	175.3987	-11.1432
11	36	18833.3092	0.123785	17.08166	-0.22293
12	39	18833.2797	0.029559	18.53352	-0.02442
13	42	18833.2751	0.004634	3.19224	-0.00867
14	45	18833.275	0.000037	0.546293	-0.00009

Figure 2.Q4.4 Parameter Estimates (drop education)

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
m0	0.9533	0.07213	9451	13.22	<.0001	0.05	0.8119	1.0947	0.047886
b1	-0.1029	0.01110	9451	-9.27	<.0001	0.05	-0.1246	-0.08111	0.115971
b2	-0.01572	0.01108	9451	-1.42	0.1561	0.05	-0.03744	0.006006	0.117042
b3	0.02478	0.005009	9451	4.95	<.0001	0.05	0.01496	0.03459	0.546293
b4	0.01522	0.006325	9451	2.41	0.0161	0.05	0.002819	0.02762	0.300087
b5	0.07428	0.03202	9451	2.32	0.0204	0.05	0.01151	0.1371	0.044482
b6	-0.2081	0.04424	9451	-4.70	<.0001	0.05	-0.2948	-0.1214	0.016519
b7	-0.1176	0.03374	9451	-3.49	0.0005	0.05	-0.1837	-0.05145	-0.01508

```
*With imputed education;
PROC IMPORT OUT= WORK.aba_imputed DATAFILE= "C:\ABI\aba_imputed.xls"
            DBMS=xls REPLACE;
     GETNAMES=YES;
RUN;

PROC NLMIXED DATA = work.aba_BN_imputed;
*m stands for lamdha 1.region 2.hhsz 3.age 4,income 5.child 6.race
7.country 8.education;
PARMS m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;
m=m0*exp(b1*region + b2*hhsz + b3*age +b4*income +b5*child + b6*race +
b7*country + b8*education);
ll = count * log(m) - m - log(fact(count));
model count ~ general(ll);
RUN;
```

Figure 2.Q4.5 Iteration History (imputed education)

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	8	18952.0361	297.5836	1481.135	-5294971
2	11	18856.761	95.27512	1853.863	-34468
3	15	18841.84	14.92101	1464.537	-5853.71
4	18	18836.2425	5.597472	1535.224	-974.663
5	21	18830.0917	6.150884	1464.523	-398.86
6	24	18820.734	9.357664	1287.688	-355.756
7	26	18805.0467	15.68733	445.7959	-67.8507
8	29	18801.3277	3.718918	124.9747	-12.4967
9	32	18800.641	0.686751	91.5355	-8.98713
10	34	18800.3098	0.331161	786.7841	-1.46298
11	36	18800.0986	0.211233	485.7869	-1.29813
12	38	18799.7361	0.362475	209.9448	-1.93594
13	40	18799.487	0.249122	32.01242	-0.41708
14	43	18799.3338	0.153207	38.31138	-0.23223
15	46	18799.3131	0.020649	13.0633	-0.02257
16	49	18799.3127	0.000426	3.675301	-0.00106
17	52	18799.3127	0.000019	0.006934	-0.00004

Figure 2.Q4.6 Parameter Estimates (imputed education)

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
m0	1.3683	0.1192	9451	11.48	<.0001	0.05	1.1347	1.6019	0.001742
b1	-0.1015	0.01111	9451	-9.14	<.0001	0.05	-0.1233	-0.07977	0.006433
b2	-0.01489	0.01109	9451	-1.34	0.1796	0.05	-0.03663	0.006856	0.005605
b3	0.02480	0.005014	9451	4.95	<.0001	0.05	0.01497	0.03463	0.00517
b4	0.01754	0.006341	9451	2.77	0.0057	0.05	0.005113	0.02997	0.006934
b5	0.07153	0.03207	9451	2.23	0.0257	0.05	0.008675	0.1344	0.000716
b6	-0.2091	0.04418	9451	-4.73	<.0001	0.05	-0.2957	-0.1225	0.002349
b7	-0.1151	0.03374	9451	-3.41	0.0007	0.05	-0.1812	-0.04892	0.000282
b8	-0.1298	0.01561	9451	-8.32	<.0001	0.05	-0.1604	-0.09924	0.005824

From figure 2.4.6, we can also found education matters (p-value < 0.0001). So, we cannot drop variable education in granted.

optimized LL value

comparing the 3 results, we can found that the optimized LL value (-18799) is very similar between drop education and imputed education (-18833), while the optimized LL value of keep education (-5200) is far different.

Estimated λ

Estimated λ(keep education) = 1.2246, with a p-value < 0.001, is very robust.
Estimated λ(drop education) = 0.9533, with a p-value < 0.001, is very robust.
Estimated λ(imputed education) = 1.3683, with a p-value < 0.001, is very robust.

Variables' parameter estimate

We can also discovered that b6 (race) has no statistical significance when we keep education because p-value= 0.1371, greater than 0.05, while b2 (hhsz) has no significant effect when drop education or impute education for their p-value > 0.05 respectively (p(drop education) = 0.1561, p(imputed education)=0.1796).

For SAS is friendly with missing value, we may conclude that by the number of records increasing race increasing its statistical significance while hhsz losing its statistical significance.

More data, more persuasive. **In the following analysis, we'll only display the imputed education version.**

All the other variables' p-value are less than 0.05, meaning those valuables are statistically significant.

Takeaway

The Poisson regression model assume numbers of books bought by individual from B&N in 2007 is distributed follow a Poisson distribution with a parameter λ (mean number of book individual purchased)

The demographic information will influence the λ.

As regression results suggest, household size of customer has no influence on the λ. For region (b1), race (b6), country (b7), and education (b8), move from lowest category to largest category has **negative impact** on the λ. Other variables, include age, income, and child, when move from lowest category to largest category, has **positive impact** on the λ.

We didn't take care of date in this step for the following reasons:

a) Because in Poisson Regression Model, Y_i denote the number of times individual I visits the site in a unit time period, we set the whole year as the time period. So, we dropped data during analysis in this step.
b) In this step, we only concern the number of books bought by customers. So, each book-buyer should only have one record that indicates how many books she bought in the given time period and other buyer related information. We cannot group all records that have the same userID together as one record by using date.

Q5 Formula LL for NBD Regression Model

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.

LL(r, α, β) = Sum of Ln $(P(Y_i = y))$ Given that:

$$P(Y_i = y) = \frac{\Gamma(r+y)}{\Gamma(r)y!}\left(\frac{\alpha}{\alpha+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^r\left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\alpha+\exp(\boldsymbol{\beta}'\mathbf{x}_i)}\right)^y$$

In SAS, we can write:

```
ll=log(gamma(r+count))-log(gamma(r))-
log(fact(count))+r*log(a/(a+expBX))+count*log(expBX/(a+expBX));
```

Q6 NBD Regression Model

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?

```
PROC NLMIXED DATA = work.aba_BN_imputed;
*1.region 2.hhsz 3.age 4,income 5.child 6.race 7.country 8.education;
PARMS r=1 a=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;
expBX = exp(b1*region + b2*hhsz + b3*age +b4*income +b5*child + b6*race
+ b7*country +b8*education);
ll = log(gamma(r+count)) - log(gamma(r))- log(fact(count)) +
r*log(a/(a+expBX)) + count*log(expBX/(a+expBX));
model count ~ general(ll);
RUN;
```

Figure 2.Q6.2 Parameter Estimates (NBD Regression Model)

Iteration History						
Iter		Calls	NegLogLike	Diff	MaxGrad	Slope
1		7	11280.0445	174.2126	2239.299	-1449616
2		15	9470.57747	1809.467	8119.057	-70575.3
3		21	9404.66519	65.91228	10537.66	-61622.6
4		24	9375.1607	29.50449	10475.44	-977.834
5		26	9277.28341	97.87729	9313.871	-1940.08
6		37	8712.25538	565.028	3380.621	-7909.4
7		41	8702.58939	9.665992	4019.83	-1817.74
8		45	8490.91706	211.6723	2823.482	-1003.81
9		48	8433.64351	57.27355	4903.967	-627.017
10		50	8393.47299	40.17052	1148.343	-325.508
11		53	8377.19649	16.2765	294.6432	-27.232
12		56	8375.37006	1.826432	155.1778	-3.86248
13		59	8374.44537	0.924686	130.4013	-1.48627
14		62	8374.33078	0.114592	151.7908	-0.11173
15		66	8372.40972	1.921064	104.3921	-0.10961
16		71	8371.14381	1.265905	323.9334	-3.32194
17		73	8369.46952	1.674288	560.9624	-3.56255
18		76	8368.60254	0.866984	243.6148	-3.8597

19		78	8368.23779	0.364751	420.2127	-0.79783
20		80	8367.69208	0.54571	151.6247	-2.3958
21		83	8367.48616	0.205919	15.40133	-0.29931
22		86	8367.44178	0.044378	20.10581	-0.0383
23		89	8367.42891	0.012869	5.211488	-0.02646
24		92	8367.42629	0.002627	4.948225	-0.00499
25		95	8367.42614	0.000146	1.045264	-0.00021
26		98	8367.42612	0.000016	0.256857	-0.00002

Optimized LL value = -8367.42612 as shown in the last row of figure above.

Figure 2.Q6.2 Parameter Estimates (NBD Regression Model)

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
r	0.09845	0.003107	9451	31.68	<.0001	0.05	0.09236	0.1045	-0.25686
a	0.07642	0.01877	9451	4.07	<.0001	0.05	0.03962	0.1132	0.032923
b1	-0.09987	0.03214	9451	-3.11	0.0019	0.05	-0.1629	-0.03686	-0.01015
b2	-0.00446	0.03334	9451	-0.13	0.8935	0.05	-0.06982	0.06089	-0.03466
b3	0.02927	0.01497	9451	1.95	0.0506	0.05	-0.00008	0.05862	-0.0328
b4	0.01764	0.01874	9451	0.94	0.3467	0.05	-0.01910	0.05438	-0.0649
b5	0.05681	0.09215	9451	0.62	0.5375	0.05	-0.1238	0.2374	-0.00523
b6	-0.2222	0.1009	9451	-2.20	0.0278	0.05	-0.4200	-0.02430	-0.00108
b7	-0.06879	0.09646	9451	-0.71	0.4758	0.05	-0.2579	0.1203	-0.00031
b8	-0.1271	0.04766	9451	-2.67	0.0077	0.05	-0.2205	-0.03368	-0.0226

The estimated r = 0.09845 and α = 0.07642, both with a p-value < 0.0001, are very robust.

For parameter estimate, the only 3 variables that help explain λ is region (b1), race (b6) and education (b8).

region (b1) = -0.09987@ p-value = 0.0019;
race (b6) = -0.2222@p-value = 0.0278;
education (b8) = -0.1271@p-value = 0.0077

Since both p-value increase compare to what in the Poisson regression model (recall, p-value < 0.0001), their explaining power decrease.

All other demographic variables have very little influence or even irrelevant with book purchase decision, due to their large p-value, which makes the parameter estimate statistically insignificant.

Takeaway:

We believe the Poisson regression model in Question 4 only capture the **observed heterogeneity** among individuals (use the observed demographic information), yet there is unobserved heterogeneity not captured by Poisson regression model and lower its explanatory power. We use NBD regression model to capture the **unobserved heterogeneity**.

In NBD regression model, we assume λ vary across population by follow a gamma distribution, with parameter r and α. Demographic information also influence λ.

Compare with the Poisson regression model, the optimized LL value increase dramatically from –18799.3127 to –8367.42612. This suggest NBD regression model fits the given dataset much better than Poisson regression model.

As regression result suggest, given r = 0.09845 and α = 0.07642, **region, race and education**, move from lowest category to largest category has a **negative impact** on the λ.

Other demographic variables have **no influence** on the λ.

Q7 Difference between Poisson Regression and NBD Regression

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?)

There is significant difference between the results of the two models.

The differences are analysis below:
In Poisson regression model, the optimized LL value is -18799.3127. And the estimated parameter value of λ_0 is 1.3683, with a p-value less than 0.0001. At a 1% level, we found this estimator robust.

In NBD regression model, the optimized LL value is -8367.42612. And the estimated parameter value of r and α is 0.07642 and 0.09845. With the P-value less than 0.001. The estimators are very robust.

Since the LL increased in NBD regression model compared with Poisson, this suggest NBD regression model fits better for the data.

There are five demographic variables help explain the λ in Poisson regression model.

region (b1) = –0.1015@ p-value < 0.0001;
age (b3) = 0.02480@p-value = < 0.0001;
income (b4) = 0.01754@p-vlaue = 0.0057;
child (b5) = 0.07153@p-value = 0.0257;
race (b6) = –0.2091@p-value < 0.0001;
country (b7) = –0.1151@p-value = 0.0007;
education(b8) = –0.129 @p-value <0.0001

While in NBD regression model, there are only 3 left, and their explaining power decrease.

region (b1) = –0.09987@ p-value = 0.0019;
race (b6) = –0.2222@p-value = 0.0278;
education (b8) = –0.1271@p-value = 0.0077

The cause of the difference is due to existence of unobserved heterogeneity in the variables (for instance, people has different buying frequencies). In the Poisson regression, the explanatory variables may not fully capture differences among individuals. But the NBD regression is able to capture unobserved component of differences among the variable.

Q8 Compare Poisson Regression and NBD Regression

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.)

LR = –2 (LLNBD – LLPoisson)

If LR > Chi-Squire (0.05, k) NBD Regression performance is not the same with Poisson Regression

Here K = 1 since NBD Regression has 1 more constraints than Poisson Regression model.

LR= –2 (– 8367.42612 – (18799.3127))= 20863.77316

Chi-Squire (0.05, 1) =3.8415;

We can found that the ***NBD Regression model is superior to the Poisson Regression model for fitting data much better.***

3. Part II. Improving the Model Compare

Please try to improve the **NBD Regression Model** using the following three methods. Two hints:

- Note that not all things we try can improve the model – in case of no improvement, concisely write down why you think it didn't work.
- Since we are not using any validation dataset, the correct way to compare models is to use the LR test, which you can easily do manually with the LL values reported by SAS for each model and a Chi-squares table (see, e.g., https://www.medcalc.org/manual/chi-square-table.php, the "0.05" column)
- The following questions regarding model improvement are all open questions. For each question, just give a few tries and report your results and your thoughts

Q9 Feature selection

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.)

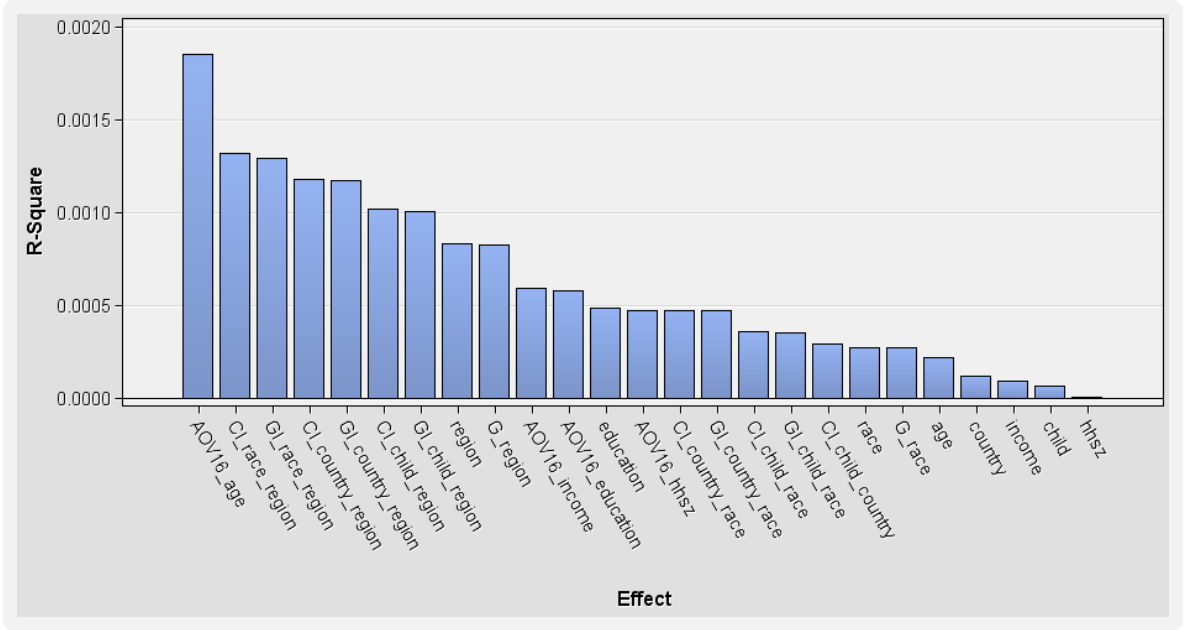
In this question, we'll use the dataset generated in **Question 1,** `project2.aba_BN`. Quick recap what we have in question one:

1. Filled all the missing values in region variables
2. Kept the education variable as original

3. Generated a new variable named count, representing the number of books purchased from BN

From feature selection results from SAS Enterprise Miner, we can found that variable selection method is based on R square. The higher R square, the more impact of the variable. The results show that Region contributes the most to the number of book purchased. Followed by Education, Race, Age, Country, Income, Child, and Hhsz in respectively.

Figure 3.Q9.1 Feature selection results (Histogram) from SAS Enterprise Miner



SAS Enterprise Miner **rejected all the original variables** for those R-Square are too small to have enough impact on count.

Figure 3.Q9.3 Feature selection results from SAS Enterprise Miner

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
AOV16_age	Input	Ordinal	Numeric		
AOV16_education	Input	Ordinal	Numeric		
G_region	Input	Nominal	Numeric	Grouped Levels for region	
age	Rejected	Interval	Numeric	age	Varsel2:Small R-square val...
child	Rejected	Binary	Numeric	child	Varsel2:Small R-square val...
country	Rejected	Nominal	Numeric	country	Varsel2:Small R-square val...
education	Rejected	Interval	Numeric	education	Varsel2:Small R-square val...
hhsz	Rejected	Interval	Numeric	hhsz	Varsel2:Small R-square val...
income	Rejected	Interval	Numeric	income	Varsel2:Small R-square val...
race	Rejected	Nominal	Numeric	race	Varsel2:Small R-square val...
region	Rejected	Binary	Numeric	region	Varsel2:Small R-square val...

Figure 3.Q9.3 Feature selection results (R-Square) from SAS Enterprise Miner

The DMINE Procedure

R-Squares for Target Variable: count

Effect	DF	R-Square	
AOV16: age	10	0.001851	
Class: race*region	15	0.001317	
Group: race*region	4	0.001293	
Class: country*region	7	0.001175	
Group: country*region	4	0.001168	
Class: child*region	7	0.001019	
Group: child*region	4	0.001005	
Class: region	3	0.000828	
Group: region	2	0.000824	
AOV16: income	6	0.000590	
AOV16: education	5	0.000576	
Var: education	1	0.000482	R2 < MINR2
AOV16: hhsz	5	0.000472	R2 < MINR2
Class: country*race	7	0.000471	R2 < MINR2
Group: country*race	3	0.000468	R2 < MINR2
Class: child*race	7	0.000354	R2 < MINR2
Group: child*race	4	0.000351	R2 < MINR2
Class: child*country	3	0.000288	R2 < MINR2
Class: race	3	0.000267	R2 < MINR2
Group: race	2	0.000267	R2 < MINR2
Var: age	1	0.000217	R2 < MINR2
Class: country	1	0.000118	R2 < MINR2
Var: income	1	0.000086757	R2 < MINR2
Class: child	1	0.000060629	R2 < MINR2
Var: hhsz	1	0.000004015	R2 < MINR2

Since SAS Enterprise Miner suggested reject all the original variables. We'll try to check LL and log likelihood by dropping variables.

Figure 3.Q9.3 Optimized LL values by dropping variables

Dropped Variable	Optimized LL value	-2 Log Likelihood
Region	-8372.26175	16745
Hhsz	-8367.435	16735
Age	-8369.34099	16739
Income	-8367.86781	16736
Child	-8367.61562	16735
Race	-8369.60011	16739
Country	-8367.67755	16735
Education	-8371.02347	16742
Income, Child, and hhsz	-8367.91261	16736
NO drop	-8367.42612	16735

From the results above, we can found that

1) keep all the 8 original variables has the best result.
2) Drop any variable(s) the results only get worse very slightly

Here, we conclude that we could not improve the performance of NBD regression model by using variable selection.

Q10 Construct new variables

10. You can also construct some variables on your own (e.g. convert date to weekday/weekend, or to holiday/non-holiday, or to seasons, construct percentage of weekend purchases, degree of

loyalty to BN etc. -- totally your call and just try 2-3 ideas). Report your code (including the code for constructing the new variables) and the MLE results. Which newly constructed variables matter, i.e., what is your new managerial takeaway?

In this question, we'll construct 3 new variables: BuyDay, Holiday, and Month

BuyDay: 1 to 7 (represent Monday to Sunday)
Holiday:

1 : 4 days before and 4 days after NEWYEAR
2 : 4 days before and 4 days after USINDEPENDENCE
3 : 4 days before and 4 days after LABOR
4 : 4 days before and 4 days after THANKSGIVING
5 : 4 days before and 4 days after CHRISTMAS

Month: 1,2,3,4,5,6,7,8,9,10,11,12 (calendar month)

In this step, we'll use the `aba_imputed` dataset, because we already imputed missing values, and change the date type.

```
DATA work.aba_constructed (DROP = product price dday);
set work.aba_imputed;

*Set the Holiday variable;
DATA work.aba_constructed (DROP = domain product qty price dday sum_qty
avg_price day holidays_or_not weekday_or_not );
set work.aba_imputed2;
*Set the Holiday variable;
if (DATE => (holiday('NEWYEAR', 2007)-4) and DATE <=
(holiday('NEWYEAR', 2007)+4))
   then Holiday=1;/*NEWYEAR 4 days before and 4 days after*/
   else if (DATE => (holiday('USINDEPENDENCE', 2007)-4) and DATE <=
(holiday('USINDEPENDENCE', 2007)+4))
      then Holiday=2;/*USINDEPENDENCE 4 days before and 4 days after */
   else if (DATE => (holiday('LABOR', 2007)-4) and DATE <=
(holiday('LABOR', 2007)+4))
      then Holiday=3;/*LABOR 4 days before and 4 days after*/
   else if (DATE => (holiday('THANKSGIVING', 2007)-4) and DATE <=
(holiday('THANKSGIVING', 2007)+4))
      then Holiday=4;/*thanksgiving 4 days before and 4 days after*/
   else if (DATE => (holiday('CHRISTMAS', 2007)-4) and DATE <=
(holiday('CHRISTMAS', 2007)+4))
      then Holiday=5;/*CHRISTMAS 4 days before and 4 days after*/
   else Holiday=0;

*Set the variable of Month;
   if DATE => "01Jan2007"d and DATE <= "31Jan2007"d then Month=1;
   else if DATE => "01Feb2007"d and DATE <= "28Feb2007"d then Month=2;
   else if DATE => "01Mar2007"d and DATE <= "31Mar2007"d then Month=3;
   else if DATE => "01Apr2007"d and DATE <= "30Apr2007"d then Month=4;
   else if DATE => "01May2007"d and DATE <= "31May2007"d then Month=5;
   else if DATE => "01Jun2007"d and DATE <= "30Jun2007"d then Month=6;
   else if DATE => "01Jul2007"d and DATE <= "31Jul2007"d then Month=7;
   else if DATE => "01Aug2007"d and DATE <= "31Aug2007"d then Month=8;
   else if DATE => "01Sep2007"d and DATE <= "30Sep2007"d then Month=9;
   else if DATE => "01Oct2007"d and DATE <= "31Oct2007"d then Month=10;
```

```
   else if DATE => "01Nov2007"d and DATE <= "30Nov2007"d then Month=11;
   else if DATE => "01Dec2007"d and DATE <= "31Dec2007"d then Month=12;

/*Set the variable of Weekend*/
   dday=weekday(DATE);
   if dday =1 then BuyDay=7;
   if dday =2 then BuyDay=1;
   if dday =3 then BuyDay=2;
   if dday =4 then BuyDay=3;
   if dday =5 then BuyDay=4;
   if dday =6 then BuyDay=5;
   if dday =7 then BuyDay=6;
Run;
```

Now, we'll use we constructed dataset with new variables to check consumers book-buying behavior. If the quantity of books a customer bought from holiday, weekend or a given day more than what they bought from other days, we'll treat them as holiday/weekend customer.

```
Proc Sort Data = work.aba_constructed;
   By userid;
Run;

data work.aba_constructed_Hcount(drop=Holiday BuyDay Month
total_H_times count_BN_times  domain qty);
   set work.aba_constructed;
   by userid;
   if first.userid then do;
      total_H_times=0; count_BN_times=0; count_BN_qty=0;
      /*count_BN_qty: the total quantity buy from BN;
      count_times: the total times buy from BN;
      total_H_times: total times buy from holiday*/
end;
count_BN_times+1;
   if holiday^=0 then total_H_times+1;
   if domain='barnesandnoble.com' then count_BN_qty + qty;
   if last.userid then averageho=round(total_H_times/count_BN_times);
   if last.userid;
run;

proc genmod data = work.aba_constructed_count;
   model  count_BN_qty = region hhsz age income child race country
education averageho
   /dist=NB  link=log type1 type3;
run;
```

The results show below:

Figure 3.Q10.1 Criteria For Assessing Goodness of Fit-Holiday

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	9441	4237.8366	0.4489
Scaled Deviance	9441	4237.8366	0.4489
Pearson Chi-Square	9441	15587.4210	1.6510
Scaled Pearson X2	9441	15587.4210	1.6510
Log Likelihood		1431.9806	
Full Log Likelihood		-8366.6392	
AIC (smaller is better)		16755.2784	

AICC (smaller is better)		16755.3064	
BIC (smaller is better)		16833.9710	

Figure 3.Q10.2 LR statistics for type 1-Holiday

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	2833.8182			
region	2844.2912	1	10.47	0.0012
hhsz	2844.3699	1	0.08	0.7791
age	2848.5274	1	4.16	0.0415
income	2849.5897	1	1.06	0.3027
child	2850.0252	1	0.44	0.5093
race	2854.0955	1	4.07	0.0436
country	2855.1926	1	1.10	0.2949
education	2862.3873	1	7.19	0.0073
averageho	2863.9612	1	1.57	0.2096

Form type 1 results, we can found region, age, race and education have stats significance for their p-value <.05. Averageho has no statistical significance. This means the new variable we constructed, holiday, is not appropriate by the data.

Figure 3.Q10.3 LR statistics for type 3-Holiday

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
region	1	9.28	0.0023
hhsz	1	0.03	0.8689
age	1	3.87	0.0492
income	1	0.84	0.3589
child	1	0.35	0.5550
race	1	4.21	0.0403
country	1	0.53	0.4677
education	1	7.22	0.0072
averageho	1	1.57	0.2096

Form type 3 results, we can found region, age, race and education have stats significance for their p-value >.05. Type 3 also demonstrated that the new variable, holiday, is not fit better.

Code for BuyDay

```
*BuyDay;
data work.aba_constructed_Dcount(drop=Holiday BuyDay Month
total_H_times count_BN_times  domain qty);
   set work.aba_constructed;
   by userid;
   if first.userid then do;
      total_D_times=0; count_BN_times=0; count_BN_qty=0;
      /*count_BN_qty: the total quantity buy from BN;
      count_times: the total times buy from BN;
      total_H_times: total times buy from BuyDay*/
end;
count_BN_times+1;
total_D_times+BuyDay;
   if domain='barnesandnoble.com' then count_BN_qty + qty;
```

```
   if last.userid then averageBDay =
round(total_D_times/count_BN_times);
   if last.userid;
run;

proc genmod data = work.aba_constructed_Dcount;
   class averageBDay;
   model  count_BN_qty = region hhsz age income child race country
education averageBDay
   /dist=NB  link=log type1 type3;
run;
```

Results as below:

Figure 3.Q10.4 Criteria For Assessing Goodness of Fit-BuyDay

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	9436	4249.3712	0.4503
Scaled Deviance	9436	4249.3712	0.4503
Pearson Chi-Square	9436	14526.8998	1.5395
Scaled Pearson X2	9436	14526.8998	1.5395
Log Likelihood		1456.8895	
Full Log Likelihood		-8341.7303	
AIC (smaller is better)		16715.4605	
AICC (smaller is better)		16715.5182	
BIC (smaller is better)		16829.9226	

Figure 3.Q10.5 LR statistics for type 1-BuyDay

LR Statistics For Type 1 Analysis				
Source	**2*LogLikelihood**	**DF**	**Chi-Square**	**Pr > ChiSq**
Intercept	2833.8182			
region	2844.2912	1	10.47	0.0012
hhsz	2844.3699	1	0.08	0.7791
age	2848.5274	1	4.16	0.0415
income	2849.5897	1	1.06	0.3027
child	2850.0252	1	0.44	0.5093
race	2854.0955	1	4.07	0.0436
country	2855.1926	1	1.10	0.2949
education	2862.3873	1	7.19	0.0073
averageBDay	2913.7790	6	51.39	<.0001

Form type 1 results, we can found region, age, race, education, and averageBDay have stats significance for their p-value <.05. This means the new variable we constructed, BuyDay, would be appropriated by the data.

Figure 3.Q10.6 LR statistics for type 3-BuyDay

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
region	1	9.97	0.0016
hhsz	1	0.01	0.9316
age	1	3.88	0.0489
income	1	0.92	0.3384
child	1	0.14	0.7060

race	1	5.52	0.0188
country	1	0.08	0.7708
education	1	4.16	0.0415
averageBDay	6	51.39	<.0001

Form type 3 results, we can found region, age, race, education, and averageBDay have stats significance for their p-value <.05. Type 3 result also demonstrates that the new variable we constructed, BuyDay, would be appreciated by the data.

Figure 3.Q10.7 Maximum Likelihood Parameter Estimates-BuyDay

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2476	0.2698	-0.7763	0.2811	0.84	0.3587
region		1	-0.1015	0.0322	-0.1645	-0.0384	9.95	0.0016
hhsz		1	0.0028	0.0332	-0.0623	0.0680	0.01	0.9317
age		1	0.0293	0.0149	0.0001	0.0584	3.87	0.0490
income		1	0.0179	0.0187	-0.0187	0.0545	0.92	0.3372
child		1	0.0348	0.0922	-0.1458	0.2154	0.14	0.7058
race		1	-0.2513	0.1007	-0.4486	-0.0540	6.23	0.0126
country		1	-0.0281	0.0963	-0.2169	0.1606	0.09	0.7703
education		1	-0.0969	0.0477	-0.1905	-0.0034	4.13	0.0422
averageBDay	1	1	-0.0275	0.1656	-0.3522	0.2971	0.03	0.8679
averageBDay	2	1	0.1523	0.1539	-0.1493	0.4539	0.98	0.3224
averageBDay	3	1	0.6194	0.1472	0.3309	0.9079	17.70	<.0001
averageBDay	4	1	0.5902	0.1467	0.3027	0.8776	16.19	<.0001
averageBDay	5	1	0.4588	0.1522	0.1605	0.7572	9.09	0.0026
averageBDay	6	1	0.6809	0.1640	0.3594	1.0023	17.24	<.0001
averageBDay	7	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	9.9325	0.3151	9.3338	10.5697		

We can find more detailed information in the above table: Wednesday, Thursday, Friday and Saturday have statistical significance.

Code for Month:

```
*Month;
data work.aba_constructed_Mcount(drop=Holiday BuyDay Month
total_H_times count_BN_times  domain qty);
   set work.aba_constructed;
   by userid;
   if first.userid then do;
      total_M_times=0; count_BN_times=0; count_BN_qty=0;
      /*count_BN_qty: the total quantity buy from BN;
      count_times: the total times buy from BN;
      total_M_times: total times buy from Month*/
end;
count_BN_times+1;
total_M_times+Month;
   if domain='barnesandnoble.com' then count_BN_qty + qty;
   if last.userid then
averageMonth=round(total_M_times/count_BN_times);
   if last.userid;
run;
```

```
proc genmod data = work.aba_constructed_Mcount;
   class averageMonth;
   model  count_BN_qty = region hhsz age income child race country
averageMonth
   /dist=NB  link=log type1 type3;
run;
```

Figure 3.Q10.8 Criteria For Assessing Goodness of Fit-Month

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	9431	4251.4780	0.4508
Scaled Deviance	9431	4251.4780	0.4508
Pearson Chi-Square	9431	13823.9496	1.4658
Scaled Pearson X2	9431	13823.9496	1.4658
Log Likelihood		1458.7577	
Full Log Likelihood		-8339.8621	
AIC (smaller is better)		16721.7241	
AICC (smaller is better)		16721.8221	
BIC (smaller is better)		16871.9555	

Compared with the base NBD regression model:
LR = –2 (LLNBD – LLPoisson)
LLNBD = – 8367.42612
LR (Holiday) = –2(– 8367.42612 – (– 8366.6392)) = 1.57384
LR (BuyDay) = –2(– 8367.42612 – (– 8341.7303)) = 51.39164
LR (Month) = –2(– 8367.42612 – (– 8339.8621)) = 55.12804
Chi-Square (0.05, 1) = 3.8415

From the results above, we can find that the new variable BuyDay and Month fits better than the original model for their LRs> Chi-Square, while Holiday fit worse than the original model for LR < Chi-Square.

Figure 3.Q10.9 LR statistics for type 1-Month

LR Statistics For Type 1 Analysis				
Source	**2*LogLikelihood**	**DF**	**Chi-Square**	**Pr > ChiSq**
Intercept	2833.8182			
region	2844.2912	1	10.47	0.0012
hhsz	2844.3699	1	0.08	0.7791
age	2848.5274	1	4.16	0.0415
income	2849.5897	1	1.06	0.3027
child	2850.0252	1	0.44	0.5093
race	2854.0955	1	4.07	0.0436
country	2855.1926	1	1.10	0.2949
education	2862.3873	1	7.19	0.0073
averageMonth	2917.5154	11	55.13	<.0001

Form type 1 results, we can found region, age, race, education, and averageMonth have stats significance for their p-value <.05. This means the new variable we constructed, Month, is appreciated by the data.

Figure 3.Q10.9 LR statistics for type 3-Month

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
region	1	9.81	0.0017
hhsz	1	0.06	0.8023
age	1	2.76	0.0967
income	1	0.14	0.7094
child	1	0.42	0.5149
race	1	3.89	0.0486
country	1	0.10	0.7561
education	1	7.17	0.0074
averageMonth	11	55.13	<.0001

Form type 1 results, we can found region, race, education, and averageMonth have stats significance for their p-value <.05. This means the new variable we constructed, BuyDay, is appreciated by the data.

Figure 3.Q10.10 Maximum Likelihood Parameter Estimates-Month

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1579	0.2677	-0.3668	0.6826	0.35	0.5552
region		1	-0.1003	0.0321	-0.1631	-0.0375	9.79	0.0018
hhsz		1	0.0083	0.0333	-0.0569	0.0735	0.06	0.8024
age		1	0.0249	0.0150	-0.0045	0.0542	2.76	0.0969
income		1	0.0070	0.0188	-0.0298	0.0438	0.14	0.7091
child		1	0.0601	0.0922	-0.1205	0.2407	0.43	0.5144
race		1	-0.2103	0.1014	-0.4090	-0.0115	4.30	0.0381
country		1	-0.0300	0.0963	-0.2188	0.1588	0.10	0.7555
education		1	-0.1259	0.0473	-0.2186	-0.0333	7.10	0.0077
averageMonth	1	1	-0.3053	0.1728	-0.6441	0.0334	3.12	0.0773
averageMonth	2	1	-0.3295	0.1756	-0.6736	0.0146	3.52	0.0606
averageMonth	3	1	-0.0659	0.1786	-0.4160	0.2842	0.14	0.7120
averageMonth	4	1	0.2301	0.1761	-0.1151	0.5753	1.71	0.1914
averageMonth	5	1	0.3867	0.1700	0.0535	0.7198	5.18	0.0229
averageMonth	6	1	0.2075	0.1627	-0.1114	0.5264	1.63	0.2021
averageMonth	7	1	0.5659	0.1604	0.2516	0.8803	12.45	0.0004
averageMonth	8	1	0.1453	0.1607	-0.1697	0.4602	0.82	0.3660
averageMonth	9	1	-0.0310	0.1662	-0.3568	0.2948	0.03	0.8521
averageMonth	10	1	0.1079	0.1761	-0.2372	0.4530	0.38	0.5399
averageMonth	11	1	-0.1468	0.1788	-0.4973	0.2038	0.67	0.4119
averageMonth	12	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	9.9120	0.3147	9.3141	10.5484		

We can find that May and July have statistical significance.

Q11 Interaction effects

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

From the results of SAS Enterprise Miner in question 9 (**Figure 3.Q9.1, Figure 3.Q9.1**), we can find the R-Square and suggested interaction effects that are (GI_Region* Race), (CI_Country * Race), and (GI_Child * Region). In this question, we'll try to simple combine those variables to see their

effects. The 3 interaction effects we are trying are (Region* Race), (Country * Race), and (Child * Region).

In this step, we'll use the `Work.aba_bn_imputed` dataset we used in question 4. This dataset has imputed education, region and age, and a generated count variable.

```
*Region*Race;
proc genmod data=Project2.aba_bn_imputed;
model count= region hhsz age income child race country region*race
/dist=NB  link=log type1 type3;
Run;
```

Figure 3.Q11.1 Criteria For Assessing Goodness of Fit- Region* Race

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	9442	4236.2570	0.4487
Scaled Deviance	9442	4236.2570	0.4487
Pearson Chi-Square	9442	16266.4252	1.7228
Scaled Pearson X2	9442	16266.4252	1.7228
Log Likelihood		1429.2047	
Full Log Likelihood		-8369.4150	
AIC (smaller is better)		16758.8301	
AICC (smaller is better)		16758.8534	
BIC (smaller is better)		16830.3689	

LR (Month) = –2(– 8367.42612 – (– 8369.4150)) = – 3.97776
Chi-Square (0.05, 1) = 3.8415
We can find that with the new variable, Region * Race, the log likelihood decreased slightly from – 8367.42612 to – 8369.4150. But get a worse LR. This shows the new variable fit data not better than the original NBD regression model.

Figure 3.Q11.2 LR statistics for type 1- Region* Race

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	2833.8182			
region	2844.2912	1	10.47	0.0012
hhsz	2844.3699	1	0.08	0.7791
age	2848.5274	1	4.16	0.0415
income	2849.5897	1	1.06	0.3027
child	2850.0252	1	0.44	0.5093
race	2854.0955	1	4.07	0.0436
country	2855.1926	1	1.10	0.2949
region*race	2858.4095	1	3.22	0.0729

Figure 3.Q11.3 LR statistics for type 3- Region* Race

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
region	1	0.36	0.5466
hhsz	1	0.07	0.7925
age	1	3.49	0.0618
income	1	0.91	0.3396
child	1	0.47	0.4940
race	1	0.51	0.4741
country	1	1.21	0.2714
region*race	1	3.22	0.0729

From the LR results of type 1 and type 3, we can see the p-value of region*race are greater than .05, demonstrating the region* race variable has no statistical significance. We can conclude that the region*race variable is not appropriate to the dataset.

```
*Country*Region;
proc genmod data=Project2.aba_bn_imputed;
model count= region hhsz age income child race country country*region
/dist=NB  link=log type1 type3;
Run;
```

Figure 3.Q11.4 Criteria For Assessing Goodness of Fit- Country * Region

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	9442	4235.5637	0.4486
Scaled Deviance	9442	4235.5637	0.4486
Pearson Chi-Square	9442	16276.6536	1.7239
Scaled Pearson X2	9442	16276.6536	1.7239
Log Likelihood		1427.5991	
Full Log Likelihood		-8371.0206	
AIC (smaller is better)		16762.0413	
AICC (smaller is better)		16762.0646	
BIC (smaller is better)		16833.5801	

LR (Month) = –2(– 8367.42612 – (– 8371.0206)) = – 7.18896

Chi-Square (0.05, 1) = 3.8415

We can find that with the new variable, Region * Race, the log likelihood decreased slightly from – 8367.42612 to – 8371.0206. But get a worse LR. This shows the new variable fit data not better than the original NBD regression model.

Figure 3.Q11.5 LR statistics for type 1- Country * Region

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	2833.8182			
region	2844.2912	1	10.47	0.0012
hhsz	2844.3699	1	0.08	0.7791
age	2848.5274	1	4.16	0.0415
income	2849.5897	1	1.06	0.3027
child	2850.0252	1	0.44	0.5093
race	2854.0955	1	4.07	0.0436
country	2855.1926	1	1.10	0.2949
region*country	2855.1983	1	0.01	0.9402

Figure 3.Q11.6 LR statistics for type 1- Country * Region

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
region	1	9.07	0.0026
hhsz	1	0.07	0.7847
age	1	3.52	0.0605
income	1	0.87	0.3504
child	1	0.41	0.5228
race	1	3.90	0.0484
country	1	0.21	0.6439
region*country	1	0.01	0.9402

Both the p-value of region * country in type 1 results and in type 3 results demonstrated a very large p-value, which made us unable to conclude the region*country variable appropriate for the dataset.

```
*Child*Region;
proc genmod data=Project2.aba_bn_imputed;
model count= region hhsz age income child race country child*region
/dist=NB  link=log type1 type3;
Run;
```

Figure 3.Q11.7 Criteria For Assessing Goodness of Fit- Child * Region

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	9442	4235.7611	0.4486
Scaled Deviance	9442	4235.7611	0.4486
Pearson Chi-Square	9442	16187.2549	1.7144
Scaled Pearson X2	9442	16187.2549	1.7144
Log Likelihood		1427.9878	
Full Log Likelihood		-8370.6320	
AIC (smaller is better)		16761.2639	
AICC (smaller is better)		16761.2872	
BIC (smaller is better)		16832.8027	

LR (Month) = –2(– 8367.42612 – (– 8370.6320)) = – 6.41176

Chi-Square (0.05, 1) = 3.8415

We can find that with the new variable, Region * Race, the log likelihood decreased slightly from – 8367.42612 to – 8370.6320. But get a worse LR. This shows the new variable fit data not better than the original NBD regression model.

Figure 3.Q11.8 LR statistics for type 1- Child * Region

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	2833.8182			
region	2844.2912	1	10.47	0.0012
hhsz	2844.3699	1	0.08	0.7791
age	2848.5274	1	4.16	0.0415
income	2849.5897	1	1.06	0.3027
child	2850.0252	1	0.44	0.5093
race	2854.0955	1	4.07	0.0436
country	2855.1926	1	1.10	0.2949
region*child	2855.9756	1	0.78	0.3762

Figure 3.Q11.9 LR statistics for type 1- Child * Region

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
region	1	1.09	0.2961
hhsz	1	0.08	0.7768
age	1	3.51	0.0609
income	1	0.81	0.3695
child	1	1.16	0.2804
race	1	3.78	0.0519
country	1	1.10	0.2944
region*child	1	0.78	0.3762

Both the p-value of region * country in type 1 results and in type 3 results show a large enough p-value to conclude the region*country variable appropriate for the dataset.

All in all, we constructed the 3 new combined variables are not appropriate for our dataset.

Part III. Why Certain Customers Prefer Amazon Over BN?

Q12 Consumer purchasing propensity

12. Now let's study why certain customers prefer Amazon over BN and vice versa. We will apply the concepts of a choice model – **logistic regression**. For each customer, you need to generate a binary 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?)

For this question, we'll apply the logistic regression model. We still use the `WORK.aba_imputed` dataset in this step. Quick recap: we imputed missing values of education, region, and age variables in this dataset.

Because we only concern certain consumer prefer Amazon or BN, so we'll drop domain, date, product, qty, price and count variables.

Generate a new variable named NB, representing whether a certain customer purchased book from BN or not.

```
DATA WORK.logistic (DROP = domain date product qty price count);
SET WORK.aba_imputed;
by userid;
IF first.userid THEN count = 0; *initialize the number of purchased
book;
IF domain = 'barnesandnoble.com' THEN count+qty; *add the qty purchased
from BN;
   IF count > 0 THEN BN=1;
   else BN = 0;
IF last.userid; *keep the final sum;
RUN;
```

Now, run the logistic regression.

```
PROC LOGISTIC DATA = WORK.logistic;
Class education region race country hhsz age income child;
model BN = education region race country hhsz age income child/expb;
RUN;
```

Figure 3.Q12.1 Results of logistic regression

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
education	5	11.0999	0.0494
region	3	31.2289	<.0001
race	3	2.1303	0.5458
country	1	2.0492	0.1523

hhsz	5	6.2252	0.2849
age	10	28.7236	0.0014
income	6	2.3366	0.8863
child	1	0.8492	0.3568

From the logistic results, we can find that the education, region, and age variables has a p-value less than .05 respectively. This demonstrates the 3 variables have statistical significance. Other 5 variables would not be appropriate.

Figure 3.Q12.2 Results of logistic regression

Effect	Point Estimate	95% Wald Confidence Limits	
education 0 vs 5	>999.999	<0.001	>999.999
education 1 vs 5	1.144	0.812	1.611
education 2 vs 5	1.000	0.719	1.390
education 3 vs 5	1.120	0.840	1.494
education 4 vs 5	1.506	1.070	2.120
region 1 vs 4	0.706	0.606	0.823
region 2 vs 4	0.893	0.762	1.047
region 3 vs 4	1.008	0.870	1.168
race 1 vs 5	0.619	0.182	2.111
race 2 vs 5	0.759	0.213	2.701
race 3 vs 5	0.685	0.186	2.524
country 0 vs 1	1.106	0.963	1.270
hhsz 1 vs 6	0.725	0.510	1.031
hhsz 2 vs 6	0.824	0.630	1.078
hhsz 3 vs 6	0.872	0.674	1.127
hhsz 4 vs 6	0.780	0.601	1.010
hhsz 5 vs 6	0.899	0.681	1.186
age 1 vs 11	1.442	0.596	3.487
age 2 vs 11	1.543	0.994	2.394
age 3 vs 11	1.379	1.031	1.843
age 4 vs 11	1.031	0.821	1.295
age 5 vs 11	1.237	0.989	1.548
age 6 vs 11	0.849	0.694	1.039
age 7 vs 11	1.152	0.935	1.420
age 8 vs 11	1.069	0.867	1.319
age 9 vs 11	0.982	0.787	1.225
age 10 vs 11	1.246	0.967	1.606
income 1 vs 7	0.944	0.772	1.154
income 2 vs 7	1.001	0.791	1.265
income 3 vs 7	0.960	0.780	1.181
income 4 vs 7	0.989	0.823	1.188
income 5 vs 7	0.900	0.772	1.049
income 6 vs 7	0.973	0.822	1.151
child 0 vs 1	1.076	0.921	1.257

The change in the probability of the event as X changes can be answered by Odds ratio. For example, in the above table, the Point Estimate of education 1 VS 5 is 1.144, which means the odds ratio of preferring

BN is expected to 14.4% (= (1.144 – 1) *100%), given the other variables in the model are held constant. While the Point Estimate of race 3 vs 5 is 31.5% (= (1-0.685) *100%), which means the odds ratio of preferring **Amazon** is expected to 14%, given the other variables in the model are held constant.

From the above table, we can find that people belonging to education 1, 3, 4, region 3, country 0 prefer BN, age 1, 2, 3, 4, 5, ,7, 8, 10, income 2, and child 0 prefer BN, while others prefer Amazon.

Part IV. Summary

Q13 Summary

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.

First, read data description is very important. Raw data always imperfect for analysis. We should read data description very carefully to check the mistakes in the raw data, such as date type is not the right type for analysis.

Second, data preprocessing is very important. Mistakes and noisy data are obstacle to generate knowledge. Clean data may not generate knowledge, but dirty data is hard to generate valuable knowledge.

Third, clean data may not generate knowledge. Our project 2 for example, the data is clean, however, a large part of variables is not appropriate to generate knowledge. At such scenario, data analyst may try to fitting models, and construct new variable to explore the value concealed in the data.

All in all, data analysis is something like digging gems in a quarry. It's exciting and need to pay more attention.