Integrating SAS and Advanced Modeling

The NBD Model

Write SAS code and conduct maximum likelihood esti- mation (MLE) for the NBD Model; estimate r and α. Report your code and the estimated values. When reporting MLE results, please provide the optimized LL value, all the estimated parameter val- ues, and the corresponding p-values. Other statistics are optional

```
* NBD Model for billboards;

PROC NLMIXED DATA=abi.billboard;
PARMS a=1 r=1;
ll                                                                     =
peoplecount*log((Gamma(r+exposures)/(Gamma(r)*Fact(exposures))*((a/(a+
1))**r)*((1/(a+1))**exposures)));
MODEL peoplecount ~ general(ll);
RUN;
```

Iteration History

Iter		Calls	NegLogLike	Diff	MaxGrad	Slope
1		5	657.00411	288.4486	114.0546	-2875.41
2		7	653.178148	3.825962	16.38484	-63.113
3		9	651.01791	2.160238	101.077	-1.80625
4		10	649.696636	1.321274	3.757159	-2.77324
5		11	649.691059	0.005578	0.480362	-0.02279
6		13	649.68883	0.002229	0.114922	-0.0044
7		15	649.688827	2.217E-6	0.006761	-4.66E-6

NOTE: GCONV convergence criterion satisfied.

Fit Statistics

-2 Log Likelihood	1299.4
AIC (smaller is better)	1303.4
AICC (smaller is better)	1303.9
BIC (smaller is better)	1305.7

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
a	0.2175	0.02978	24	7.31	<.0001	0.05	0.1561	0.2790	0.006761
r	0.9693	0.1135	24	8.54	<.0001	0.05	0.7350	1.2035	-0.00175

After running the NBD model for the billboards data we get r = 0.969, α= 0.218 and Loglikelihood = -649.69. The values of r and α are significant as their p-values are < 0.05.
Using these values of r and α to calculate number of exposures over 4 weeks for the 250 people:

```
Data abi.billsim (drop = r a p);
Call streaminit(123);
r = 0.969;
a = 0.218;
t = 4;
Do i=0 to 23;
      p                                                                          =
(Gamma(r+i)/(Gamma(r)*Fact(i))*((a/(a+t))**r)*((t/(a+t))**i));
      x = 250*p;
Output;
End;
Run;
```

Running the above code, we get the number of exposures (x) as follows:

	t	i	x
1	4	0	14.163679127
2	4	1	13.015272711
3	4	2	12.151290644
4	4	3	11.40419849
5	4	4	10.730977668
6	4	5	10.113272268
7	4	6	9.5410340081
8	4	7	9.0078528758
9	4	8	8.5091962503
10	4	9	8.0416185837
11	4	10	7.6023608972
12	4	11	7.1891285556
13	4	12	6.7999588812
14	4	13	6.4331375957
15	4	14	6.0871434718
16	4	15	5.7606101235
17	4	16	5.4522986642
18	4	17	5.1610775128
19	4	18	4.8859070559
20	4	19	4.625827703
21	4	20	4.3799503746
22	4	21	4.1474487753
23	4	22	3.9275530042
24	4	23	3.7195441876

The Poisson Regression Model

Write SAS code to estimate parameters (_0 and the vector_) using MLE for the Poisson Regression Model. Report your code and the estimated values. What are some managerial takeaways?

Code:

```
proc nlmixed data=abi.kc;
  /* m stands for lambda */
  parms m0=1 b1=0 b2=0 b3=0 b4=0;
  m=m0*exp(b1*income+b2*sex+b3*age+b4*HHSize);
  ll = total*log(m)-m-log(fact(total));
  model total ~ general(ll);
run;
```

Result:

31		67	6291.51509	0.055528	55.52937	-0.13954
32		69	6291.50579	0.009297	39.31139	-0.01563
33		72	6291.5033	0.00249	104.2272	-0.0048
34		73	6291.49967	0.003631	23.56034	-0.00445
35		74	6291.4975	0.002175	10.79874	-0.00271
36		76	6291.49677	0.000725	5.895111	-0.00133
37		78	6291.49675	0.000025	0.44181	-0.00005

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	12583
AIC (smaller is better)	12593
AICC (smaller is better)	12593
BIC (smaller is better)	12623

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
m0	0.04387	0.01782	2728	2.46	0.0139	0.05	0.008926	0.07882	-0.44181
b1	0.09385	0.03439	2728	2.73	0.0064	0.05	0.02641	0.1613	-0.19574
b2	0.004236	0.04090	2728	0.10	0.9175	0.05	-0.07597	0.08444	-0.02777
b3	0.5883	0.05475	2728	10.74	<.0001	0.05	0.4809	0.6956	-0.06349
b4	-0.03591	0.01529	2728	-2.35	0.0189	0.05	-0.06589	-0.00594	-0.08553

Running Poisson regression model for the khakhichinos data we get – m0(lambda) = .04387, $\beta1$ = 0.09385, $\beta2$ = 0.004236, $\beta3$ = 0.5883, $\beta4$ = -0.03591 and loglikelihood = -6291.4967.

The p-value of $\beta2$ is 0.91 which is not less than 0.05. This implies that $\beta2$ is not significant. So, the managerial takeaway is that $\beta2$ corresponding to income of the consumer is not significant in predicting his visiting rate.

The NBD Regression Model

Write SAS code to estimate parameters (r, _ and the vector _) using MLE for NBD Regression Model. Report your code and the estimated values. What are some managerial takeaways? Explain the difference in results between the NBD and the Poisson Regression Model.

Code:

```
* The NBD Regression Model for Khaki Chinos;

proc nlmixed data=abi.kc;
  parms r=1 a=1 b1=0 b2=0 b3=0 b4=0;
  expBX=exp(b1*income+b2*sex+b3*age+b4*HHSize);
  ll                =                log(gamma(r+total))-log(gamma(r))-
log(fact(total))+r*log(a/(a+expBX))+total*log(expBX/(a+expBX));
  model total ~ general(ll);
run;
```

Results:

31		63	2888.96814	0.004446	9.767101	-0.00733
32		65	2888.96736	0.000786	0.339248	-0.00203
33		66	2888.96657	0.000784	5.078462	-0.00068
34		68	2888.96624	0.000336	0.35606	-0.00058
35		70	2888.96613	0.00011	0.480104	-0.00014
36		72	2888.96611	0.000013	0.012623	-0.00003

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	5777.9
AIC (smaller is better)	5789.9
AICC (smaller is better)	5790.0
BIC (smaller is better)	5825.4

Parameter Estimates									
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
r	0.1388	0.007269	2728	19.09	<.0001	0.05	0.1245	0.1530	-0.00467
a	8.2007	9.5022	2728	0.86	0.3882	0.05	-10.4316	26.8330	-0.00007
b1	0.07348	0.09755	2728	0.75	0.4513	0.05	-0.1178	0.2648	0.012623
b2	-0.00927	0.1212	2728	-0.08	0.9390	0.05	-0.2469	0.2284	-0.00033
b3	0.9020	0.1677	2728	5.38	<.0001	0.05	0.5732	1.2308	-0.00395
b4	-0.02432	0.04272	2728	-0.57	0.5692	0.05	-0.1081	0.05945	0.007999

Running NBD regression model for the khakhichinos data we get – $r = .1388$, $\alpha = 8.2007$, $\beta 1 = 0.07348$, $\beta 2 = -0.00927$, $\beta 3 = 0.9020$, $\beta 4 = -0.02432$ and loglikelihood = 2888.96611.

The p-values of α, $\beta 1$, $\beta 2$ and $\beta 4$ are all > 0.05, implying that all of them are insignificant. B3 corresponding to sex is the only significant variable. So, the managerial takeaway is that Sex is the only factor that significantly explains the visiting rate of customers.

Poisson regression model accounts for only the observed heterogeneity among the customers (lambda constant), whereas NBD regression model takes into account both observed and unobserved heterogeneity (lambda follows gamma distribution).
Poisson regression shows sex, age and household size as significant variables in predicting visiting rate, whereas NBD regression shows only sex is the significant variable.

Analysis of New Real Data

In this part of the project, you will adapt the models you used in Part I and apply them to the dataset books.txt. The dataset records customer purchases at two competitors, Amazon.com and BARNES & NOBLE (B&N) in 2007. Some customer demographic variables | education, household size (hhsz), income, and race | are also in the dataset.

1. Write a SAS program that reads the data in books.txt and generates a count dataset. That is, for each customer count the number of books purchased from B&N in 2007, while keeping the demographic variables. Print the first 10 records of this dataset.

```
proc import datafile='/folders/myfolders/books.txt'   out=work.nt
(drop=VAR15);
getnames=yes;
run;

proc sql;
create table regnt as
select unique(userid), education, region,hhsz, age, income, child,
race, country, avg(price) as ppbook, avg(qty) as qtyperv, sum(qty)
as qty, avg(wend) as wend
from nt
where domain = 'barnesandn'
group by userid;
quit;

proc print data=nbd_bn (obs=10);
run;
```

Obs	userid	education	region	hhsz	age	income	child	race	country	ppbook	qtyperv	qty	wend
1	6365661	5	1	2	11	7	0	1	0	17.9700	1.00000	1	0.00
2	6396922	2	2	2	8	4	0	1	0	15.9600	1.00000	1	0.00
3	8999933	4	3	5	10	3	1	1	0	49.9500	1.00000	1	0.00
4	9573834	99	4	2	10	5	1	1	0	2.9050	1.00000	2	0.00
5	9576277	99	1	3	8	7	1	1	0	16.3460	1.00000	5	0.00
6	9581009	99	2	2	7	5	1	1	0	2.0000	1.00000	1	1.00
7	9595310	4	2	2	8	2	1	1	0	23.1700	1.50000	6	0.25

Obs	userid	education	region	hhsz	age	income	child	race	country	ppbook	qtyperv	qty	wend
8	9611445	2	4	2	11	6	1	1	1	15.6700	1.00000	2	1.00
9	9663372	4	4	3	9	7	1	1	0	43.7678	3.11111	28	0.00
10	9752844	3	4	2	7	3	1	1	0	14.1850	1.00000	2	1.00

Here qty is the count variable.

2. Build an NBD model, ignoring the demographic variables. Report your results. (Hint: you will need to create a data set similar to that used in the billboard exposures example.)

```
proc sql;
create table bn as
select unique(userid), sum(qty) as qty
from nt
where domain in ('barnesandn')
group by userid;
quit;

proc sql;
create table nbd_bn as
select   unique(qty)   as   exposures,count(unique(userid))   as
peoplecount
from bn
group by qty
order by qty;
quit;

proc print data=nbd_bn (obs=10);
run;
```

Obs	exposures	peoplecount
1	1	753
2	2	362
3	3	175
4	4	126
5	5	82
6	6	74
7	7	30
8	8	48

Obs	exposures	peoplecount
9	9	31
10	10	20

```
proc NLMIXED data= nbd_bn;
parms r=2 alpha=2;
m=              ((gamma(r+exposures))/(gamma(r)*fact(exposures)))*
((alpha/(alpha+1))**r)*((1/(alpha+1))**exposures);
ll=peoplecount*log(m);
model peoplecount ~ general(ll);
run;
```

The NLMIXED Procedure

Specifications	
Data Set	WORK.NBD_BN
Dependent Variable	peoplecount
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	45
Observations Not Used	0
Total Observations	45
Parameters	2

Initial Parameters		
r	**alpha**	**Negative Log Likelihood**
2	2	6958.29235

Iteration History					
Iteration	**Calls**	**Negative Log Likelihood**	**Difference**	**Maximum Gradient**	**Slope**
1	7	4781.1500	2177.142	854.922	-42140.4
2	13	4649.7965	131.3535	202.887	-4181.07
3	17	4614.7250	35.07154	2937.60	-75.5577
4	21	4487.3001	127.4249	322.732	-520.923
5	24	4485.4744	1.82569	293.072	-10.1766
6	26	4483.1792	2.295144	7.49085	-4.25434
7	29	4483.1725	0.006758	0.10297	-0.01168
8	32	4483.1725	9.624E-6	0.002305	-0.00002

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	8966.3
AIC (smaller is better)	8970.3
AICC (smaller is better)	8970.6
BIC (smaller is better)	8974.0

Parameter Estimates								
Parameter	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**	**95% Confidence Limits**		**Gradient**
r	1.2024	0.04687	45	25.65	<.0001	1.1080	1.2968	0.000706
alpha	0.3080	0.01418	45	21.72	<.0001	0.2794	0.3366	-0.00231

3. Calculate the values of (i) Reach, (ii) Average Frequency, and (iii) Gross Ratings Points (GRPs) based on the NBD Model. Show your work.

$$P(X(t) = 0|r,\ \alpha) = \left(\frac{\alpha}{\alpha + t}\right)^{r} = \left(\frac{0.3080}{0.3080 + 1}\right)^{1.2024} = 0.1757$$

$$E(X(t)) = \frac{rt}{\alpha} = \frac{1.2024 \times 1}{0.3080} = 3.9039$$

(i) Reach = $100 \times \left(1 - \mathrm{P}(X(t) = 0)\right) = 82.43\%$

(ii) Average Frequency = $\frac{E(X(1))}{\left(1 - \mathrm{P}(X(t)=0)\right)} = 4.736$

(iii) GRP = $100 \times E(X(1)) = 390.39$

4. Build a Poisson regression model using the demographic information (customer characteristics) provided. Report your results. What are the managerial takeaways | which customer characteristics seem to be important?
Optional: You have exibility in choosing the variables to include | if you wish to do so, you can choose to
eliminate some (via feature selection, for example) or create new ones (from the variables you have available - for example, fraction of weekend purchases). This is optional for this project, but if you do anything along these lines, please provide your justification.

```
/*Poisson Regression*/
data nt;
set nt;
  date_new = input(put(date, 8.), yymmdd8.);
run;

data nt;
set nt;
wend=0;
if weekday(date_new) = 1 then wend = 1;
if weekday(date_new) = 7 then wend = 1;
run;
```

```
proc sql;
create table regnt as
select unique(userid), education, region,hhsz, age, income, child,
race, country, avg(price) as ppbook, avg(qty) as qtyperv, sum(qty)
as qty, avg(wend) as wend
from nt
where domain = 'barnesandn'
group by userid;
quit;
```

The NLMIXED Procedure

Specifications	
Data Set	WORK.REGNT
Dependent Variable	qty
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	1812
Observations Not Used	0
Total Observations	1812
Parameters	10

Initial Parameters										
m0	**b1**	**b2**	**b3**	**b4**	**b5**	**b6**	**b7**	**b8**	**b9**	**Negative Log Likelihood**
1	0	0	0	0	0	0	0	0	0	11610.6198

Iteration History					
Iteration	**Calls**	**Negative Log Likelihood**	**Difference**	**Maximum Gradient**	**Slope**
1	11	8482.2441	3128.376	87700.2	-1.545E9
2	17	8403.6288	78.61527	86992.3	-2636435
3	25	7987.7527	415.8761	79176.0	-4903066
4	29	7564.2104	423.5423	80267.0	-645866
5	32	7352.7800	211.4304	77240.6	-28663.5
6	35	7313.1244	39.65562	72793.7	-6893.09
7	37	7230.5197	82.60471	73236.3	-1866.58
8	39	7170.3245	60.19517	17127.2	-576.996
9	41	7079.0083	91.3162	9143.52	-632.772
10	44	7064.8560	14.15228	12889.5	-157.720
11	46	7041.9189	22.93715	11976.5	-102.924
12	49	7035.6473	6.271555	14561.4	-25.9087
13	51	7025.1150	10.53235	19589.9	-56.2726
14	53	7012.3028	12.8122	5217.41	-15.5383
15	56	7003.9174	8.385376	4010.26	-10.5739
16	59	6999.4945	4.422915	9173.32	-7.83933
17	62	6998.0458	1.44873	810.360	-3.73147
18	65	6997.6360	0.409723	1311.62	-0.72265
19	68	6997.5519	0.084177	572.436	-0.18623
20	71	6997.5191	0.032728	76.7433	-0.05086
21	74	6997.5177	0.001436	23.3419	-0.00267
22	77	6997.5176	0.000119	14.8538	-0.00012
23	80	6997.5175	0.000032	1.69826	-0.00004

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	13995
AIC (smaller is better)	14015
AICC (smaller is better)	14015
BIC (smaller is better)	14070

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
m0	2.8829	0.1574	1812	18.32	<.0001	2.5743	3.1915	-0.00762
b1	-0.00099	0.000279	1812	-3.55	0.0004	-0.00154	-0.00044	-1.69826
b2	0.009962	0.01133	1812	0.88	0.3793	-0.01226	0.03218	-0.05393
b3	0.007292	0.003146	1812	2.32	0.0206	0.001123	0.01346	-0.15363
b4	0.03014	0.006408	1812	4.70	<.0001	0.01757	0.04270	-0.05850
b5	0.001840	0.03252	1812	0.06	0.9549	-0.06194	0.06562	-0.01816
b6	-0.2469	0.03428	1812	-7.20	<.0001	-0.3141	-0.1796	-0.00531
b7	-0.00738	0.000728	1812	-10.13	<.0001	-0.00880	-0.00595	-0.92927
b8	0.2446	0.009802	1812	24.95	<.0001	0.2254	0.2638	-0.02823
b9	0.1702	0.02865	1812	5.94	<.0001	0.1140	0.2264	0.000419

Derived Variables:

3 variables have been derived from the available set of variables. Their description and justification for selection are as follows:

- Price per book (ppbook): This variable explains the type of customer, whether they buy expensive or cheap books. It would be interesting to see if the customers who buy expensive books tend to buy more or not.

- Average quantity per visit (qtyperv): Do bulk buyers buy more quantity? This variable was created to answer this question.

- Fraction of weekend purchases (wend): Do more purchases happen on the weekend rather than on weekdays ? The date that was provided in the dataset was converted into a SAS Date type and the weekday() function was used to determine whether a day is a weekend or not. Further the variable indicating weekend (binary, 1 standing for is a weekend) was averaged to obtain fraction of weekend purchases.

Managerial Takeaways:

- The average number of books a person buys at B&N is 2.8829. Interestingly the average m0 is lesser than the NBD model.
- Household size and number of children do not have any significant explanatory power for the number of books bought. Understandably they are consistent.
- Education has a negative parameter estimate meaning as the label for education increases the quantity of books sold decreases.
- Country has a negative parameter estimate, if the variable is encoded as 0-domestic and 1-foreign, then it means that B&N domestic customers buy more than their foreign customers.
- ppbook has a negative intercept meaning people who buy expensive books buy lesser quantities.
- Age and income have slight positive parameter estimates meaning as the age and income increase the number of books bought also increase.
- qtyperv: As the quantity per purchase increases the number of books bought also increases.
- wend: The quantity of books purchased increases for a person who buys most of his books on weekends!

5. Next, we start the setup for developing an NBD regression model. What is the formula for the log-liklihood expression, LL?

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

To make NBD Regression Model we have got the below formula from above:

```
Where y=qty
prob=exp(b1*education+b2*hhsz+b3*age+b4*income+b5*child+b6*country+b7*
avg_visits)
m=((gamma(r+qty))/(gamma(r)*fact(qty)))*      ((alpha/(alpha+prob))^r)*
(prob/(alpha+prob))^qty;
LL = log(m);
```

6. Build a NBD regression model using the demographic information provided. Report your results. What are the managerial takeaways | which customer characteristics seem to be important?
Optional: As with the Poisson regression, you have exibility in choosing the variables to include | if you
wish to do so, you can choose to eliminate some (via feature selection, for example) or create new ones (from the variables you have available - for example, fraction of weekend purchases).

This is optional for this project, but if you do anything along these lines, please provide your justification.

```
proc nlmixed data=regnt;
parms r=1 alpha=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0 b9=0;
prob=exp(b1*education+b2*hhsz+b3*age+b4*income+b5*child+b6*countr
y+b7*ppbook+b8*qtyperv+b9*wend);
m=((gamma(r+qty))/(gamma(r)*fact(qty)))*
((alpha/(alpha+prob))**r)* (prob/(alpha+prob))**qty;
ll = log(m);
model qty ~ general(ll);
run;
```

The NLMIXED Procedure

Specifications	
Data Set	WORK.REGNT
Dependent Variable	qty
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions	
Observations Used	1812
Observations Not Used	0
Total Observations	1812
Parameters	11

Initial Parameters											
r	**alpha**	**b1**	**b2**	**b3**	**b4**	**b5**	**b6**	**b7**	**b8**	**b9**	**Negative Log Likelihood**
1	1	0	0	0	0	0	0	0	0	0	6159.30585

Iteration History					
Iteration	**Calls**	**Negative Log Likelihood**	**Difference**	**Maximum Gradient**	**Slope**
1	10	5233.3728	925.933	69187.0	-3.862E8
2	13	4976.2887	257.0842	38433.3	-1257.06
3	15	4589.6271	386.6616	2037.41	-546.216
4	18	4567.3667	22.2604	5087.77	-110.330
5	20	4540.6663	26.70034	6776.40	-43.5605
6	23	4533.8394	6.826951	2572.99	-10.7735
7	27	4518.4513	15.38809	9669.72	-2.57532
8	29	4492.0606	26.39064	5546.04	-22.6744
9	32	4483.9520	8.108652	583.480	-14.1128
10	36	4463.1953	20.7567	10807.1	-2.56695
11	38	4429.8971	33.29814	3035.02	-28.1878
12	41	4425.7615	4.135632	1373.56	-8.02139
13	43	4421.6651	4.096467	4530.65	-1.60008
14	47	4403.3587	18.30639	7443.37	-7.90063
15	50	4398.5242	4.834452	435.115	-7.30059
16	53	4398.4008	0.123374	37.3168	-0.22804
17	56	4398.3656	0.03519	198.081	-0.01734
18	62	4396.8703	1.495361	1313.57	-0.05566
19	64	4396.0880	0.782291	327.321	-1.34793
20	67	4395.6612	0.426813	29.2740	-0.81919
21	70	4395.6590	0.002221	12.8319	-0.00283
22	76	4395.5919	0.067069	459.887	-0.00149
23	80	4395.2958	0.296045	492.643	-0.11278
24	83	4395.2800	0.015868	84.5584	-0.04314
25	86	4395.2783	0.001694	8.17659	-0.00291
26	88	4395.2757	0.002532	22.3782	-0.00039
27	92	4395.2563	0.019444	52.4267	-0.00542
28	96	4395.2138	0.042537	181.090	-0.02451
29	99	4395.2116	0.002173	12.4155	-0.00414

Iteration History					
Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
30	102	4395.2116	0.000024	0.37289	-0.00004
31	104	4395.2116	8.545E-6	6.57253	-2.33E-6

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	8790.4
AIC (smaller is better)	8812.4
AICC (smaller is better)	8812.6
BIC (smaller is better)	8872.9

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.3234	0.05278	1812	25.07	<.0001	1.2199	1.4269	-0.10439
alpha	0.6714	0.08915	1812	7.53	<.0001	0.4965	0.8462	-0.15326
b1	-0.00068	0.000564	1812	-1.21	0.2278	-0.00179	0.000426	6.57253
b2	0.02035	0.02284	1812	0.89	0.3729	-0.02444	0.06515	0.34008
b3	0.008830	0.008509	1812	1.04	0.2996	-0.00786	0.02552	0.53877
b4	0.02222	0.01281	1812	1.73	0.0829	-0.00290	0.04735	0.47031
b5	-0.02312	0.06361	1812	-0.36	0.7163	-0.1479	0.1016	-0.28004
b6	-0.2078	0.06459	1812	-3.22	0.0013	-0.3345	-0.08110	0.19551
b7	-0.01312	0.001372	1812	-9.57	<.0001	-0.01581	-0.01043	2.93975
b8	0.6506	0.06316	1812	10.30	<.0001	0.5268	0.7745	0.17724
b9	0.1977	0.06384	1812	3.10	0.0020	0.07243	0.3229	0.089618

The derived variables are the same as the Poisson Regression.

Managerial Takeaways:

- The average rate of purchase is far lesser than both the NBD model and Poisson regression at 1.9711.

- The variables education and age alongwith the previous variables household size and child are insignificant. As expected incorporating unobserved variability decreases the significance of variables with lesser explanatory power.

- Country is significant and has the same interpretation as earlier, i.e. , if the variable is encoded as 0-domestic and 1-foreign, then it means that B&N domestic customers buy more than their foreign customers.

- ppbook is significant and negative as earlier meaning as the average price per book increases the qty decreases.

- qtyperv and wend are both positive & significant and their parameter estimates are higher than in Poisson regression. This means that as the quantity per purchase increases and the fraction of buy on the weekend increases the customer buys more quantity of books.

7. Are there any signi_cant differences between the results from the Poisson and NBD regressions? If so, what exactly is the difference? Discuss what you believe about the cause(s) of the difference.

Yes, there is a significant difference.
The NBD regression has a LL of -4395.2116 whereas the Poisson regression has a LL of -6997.5175 making the NBD regression model better.

The NBD regression incorporates unobserved heterogeneity improving its explanatory power hence the lower log likelihood.

8. Briey summarize what you learned from this project. This is an open-ended question, so please include
anything you found worthwhile - relating to the modeling tool (SAS), the modeling process, insights from the modeling, any managerial takeaways that were insightful to you, and so on.

The NBD regression model may have overfitted the data. I would have tested it on a validation dataset before concluding it better than the Poisson regression model. Maybe, because of the way the variables are encoded the parameter estimates of the demographic variables are very small and hence sometimes insignificant.

The target for higher number of book buys are domestic customers who are older, have a higher income, don’t buy expensive books, buy a lot of books on an average per buy and who make most of their purchases on the weekends.