Part I: Examples Integrating SAS and Advanced Modeling

1. The NBD Model

```
proc nlmixed data=mis6334.Billboard;
  parms alpha=.5 r=.5; *Our decision variables;
  Xx =
(gamma(r+exposures)/(gamma(r)*fact(exposures)))*((alpha/(alpha+1))**r)*(1/(al
pha+1))**exposures; *P(X=x|r,alpha);
  ll = peoplecount*log(Xx); *sum of ll is what we are trying to maximize;
  model peoplecount ~ general(ll);
run;
```

alpha	r	Negative Log Likelihood
0.5	0.5	849.509336

Iteration History

Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	11	655.4252	194.0842	145.532	-4876.86
2	14	653.9099	1.515209	41.5461	-29.2737
3	18	649.8283	4.081679	18.8781	-4.12042
4	20	649.7050	0.123291	2.49637	-0.19282
5	22	649.6892	0.015759	0.67906	-0.02841
6	25	649.6888	0.000385	0.13381	-0.00070
7	28	649.6888	3.039E-6	0.003734	-5.34E-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|	95% Confidence Limits		Gradient
alpha	0.2175	0.02978	24	7.30	<.0001	0.1561	0.2790	-0.00373
r	0.9693	0.1135	24	8.54	<.0001	0.7350	1.2035	0.000158

The optimized LL value is -649.688 that we can see from iteration 7 in the Iteration History table.
The alpha value is .2175, and its corresponding p-value is <.0001.
The r value is .9693, and its corresponding p-value is <.0001;

2. **The Poisson Regression Model**

```
proc nlmixed data=mis6334.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;
```

34	124	6291.4996	0.003652	23.6812	-0.00449
35	126	6291.4976	0.002057	11.3724	-0.00271
36	129	6291.4968	0.000794	6.20975	-0.00145
37	132	6291.4967	0.000027	0.54396	-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|	95% Confidence Limits		Gradient
m0	0.04387	0.01834	2728	2.39	0.0168	0.007904	0.07984	-0.54396
b1	0.09385	0.03510	2728	2.67	0.0075	0.02502	0.1627	-0.24554
b2	0.004234	0.04093	2728	0.10	0.9176	-0.07601	0.08448	-0.03167
b3	0.5883	0.05502	2728	10.69	<.0001	0.4804	0.6961	-0.08027
b4	-0.03591	0.01529	2728	-2.35	0.0189	-0.06590	-0.00593	-0.10097

Optimized LL value= -6291.4967
Estimated value of lambda = .04387
Estimated value of income(b1) = .09385
Estimated value of sex(b2) = .004234
Estimated value of age(b3) = .5883
Estimated value of HHsize(b4) = -.03591

At a 5% significance level, all parameters are significant except for the parameter for sex. Age is the most important variable when determining the number of visits. As income and age increases, exposure rate increases, and when hhsize increases, exposure rate decreases. Our model also shows that more females visit websites than males. However, sex is not significant so we cannot be confident about the impact of sex on the number of visits.

3. The NBD Regression Model

```
proc nlmixed data=mis6334.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;
```

32	112	2888.9674	0.001026	0.68012	-0.00301
33	114	2888.9671	0.000267	7.83069	-0.00063
34	118	2888.9663	0.000803	1.69034	-0.00193
35	121	2888.9662	0.000146	0.25909	-0.00018
36	124	2888.9661	0.000048	0.11004	-0.00008
37	127	2888.9661	1.801E-6	0.056282	-3.45E-6

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|	95% Confidence Limits		Gradient
r	0.1388	0.007269	2728	19.09	<.0001	0.1245	0.1530	0.035783
a	8.1976	9.4819	2728	0.86	0.3874	-10.3949	26.7900	-0.00064
b1	0.07340	0.09743	2728	0.75	0.4513	-0.1176	0.2644	0.056282
b2	-0.00928	0.1212	2728	-0.08	0.9390	-0.2469	0.2284	0.003186
b3	0.9022	0.1676	2728	5.38	<.0001	0.5735	1.2309	0.017834
b4	-0.02432	0.04272	2728	-0.57	0.5692	-0.1081	0.05945	0.016634

Optimized LL value= -2888.9661
Estimated value of r = .1388
Estimated value of a = 8.1976
Estimated value of income(b1) = .07340
Estimated value of sex(b2) = -.00928
Estimated value of age(b3) = .9022
Estimated value of HHsize(b4) = -.02432

Only values for r and age are significant. All over values are not significant due to p-values greater than .05. r and age are both positively related to exposure rates.

In the Poisson Regression model sex is not significant, but all the other variables are. In the NBD Regression model only age is a significant demographic in determining exposure rates. The AIC, AICC, and BIC values are all lower in the NBD Regression model which means that it is a better model.

Part II: Analysis of New Real Data

Question 1

```
DATA mis6334.books (drop=VAR15);
infile 'C:\mis6334\PROJECT\books1.txt' delimiter='09'x MISSOVER DSD
lrecl=50000 firstobs=2 IGNOREDOSEOF;
informat userid best32. ;
informat education best32. ;
informat region $1. ;
informat hhsz best32. ;
informat age best32. ;
informat income best32. ;
informat child best32. ;
informat race best32. ;
informat country best32. ;
informat domain $20. ;
informat date best32. ;
informat product $132. ;
informat qty best32. ;
informat price best32. ;
informat VAR15 $1. ;
format userid best12. ;
format education best12. ;
format region $1. ;
format hhsz best12. ;
format age best12. ;
format income best12. ;
format child best12. ;
format race best12. ;
format country best12. ;
format domain $20. ;
format date best12. ;
format product $132. ;
format qty best12. ;
format price best12. ;
format VAR15 $1. ;
input
         userid
         education
         region $
         hhsz
         age
         income
         child
         race
         country
         domain $
         date
         product $
         qty
         price
       VAR15 $
                                   ;
RUN;
```

```
data bnbooks;
set mis6334.books;
if domain = "barnesandnoble.com";
run;

proc means data=bnbooks NOPRINT;
class userid;
id education region hhsz income child race country;
output out=bnbookssum
sum(qty) = NumBooks;
run;

Data Bnbookssum;
set Bnbookssum (drop = _TYPE_ _FREQ_);
if userid = . then delete;
run;

PROC PRINT data=Bnbookssum (obs=10);
run;
```

The SAS System

Obs	userid	education	region	hhsz	income	child	race	country	NumBooks
1	6365661	5	1	2	7	0	1	0	1
2	6396922	2	2	2	4	0	1	0	1
3	8999933	4	3	5	3	1	1	0	1
4	9573834	99	4	2	5	1	1	0	2
5	9576277	99	1	3	7	1	1	0	5
6	9581009	99	2	2	5	1	1	0	1
7	9595310	4	2	2	2	1	1	0	6
8	9611445	2	4	2	6	1	1	1	2
9	9663372	4	4	3	7	1	1	0	28
10	9752844	3	4	2	3	1	1	0	2

Question 2

```
/*count number of Amazon books purchased*/

data amazonbooks;
set mis6334.books;
if domain = "amazon.com";
run;

proc means data=amazonbooks NOPRINT;
class userid;
id education region hhsz income child race country;
output out=amazonbookssum
sum(qty) = NumBooksamazon;
run;


Data amazonbookssum;
set amazonbookssum (drop = _TYPE_ _FREQ_);
if userid = . then delete;
run;

/* merge barnesandnoble count dataset with amazon count dataset */

data mergedbooks;
merge amazonbookssum bnbookssum;
by userid;
if NumBooks = . then NumBooks = 0;
run;

/* Find peoplecount for barnesandnobles */

proc means data=mergedbooks NOPRINT;
class NumBooks;
output out=nbdmodel
n(userid) = peoplecount;
run;

data nbdmodel;
set nbdmodel (drop= _TYPE_ _FREQ_);
if NumBooks = . then delete;
run;

/*Print first ten observations */

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

The SAS System

Obs	NumBooks	peoplecount
1	0	7639
2	1	753
3	2	362
4	3	175
5	4	126
6	5	82
7	6	74
8	7	30
9	8	48
10	9	31

```
/* NBD MODEL */

proc nlmixed data=nbdmodel;
  parms alpha=.5 r=.5; *Our decision variables;
  Xx =
(gamma(r+NumBooks)/(gamma(r)*fact(NumBooks)))*((alpha/(alpha+1))**r)*(1/(alph
a+1))**NumBooks; *P(X=x|r,alpha);
  ll = peoplecount*log(Xx); *sum of ll is what we are trying to maximize;
  model peoplecount ~ general(ll);
run;
```

Iteration History					
Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	10	9161.2705	810.6365	1588.86	-306461
2	13	8839.6058	321.6647	4714.58	-52120.8
3	15	8721.5997	118.0061	14409.1	-787.050
4	17	8481.4675	240.1322	4653.64	-4308.92
5	20	8463.3040	18.16348	3128.07	-94.2939
6	24	8389.8814	73.42265	957.757	-96.2154
7	27	8382.8463	7.035079	306.726	-8.78148
8	30	8381.7612	1.085078	66.6914	-1.81031
9	33	8381.7110	0.050228	5.19184	-0.10809
10	36	8381.7107	0.000291	0.080373	-0.00059
11	39	8381.7107	1.413E-7	0.049863	-2.38E-7

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	16763
AIC (smaller is better)	16767
AICC (smaller is better)	16768
BIC (smaller is better)	16771

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
alpha	0.1299	0.006121	46	21.22	<.0001	0.1176	0.1422	-0.01909
r	0.09723	0.003060	46	31.77	<.0001	0.09107	0.1034	0.049863

Optimized LL value: -8381.7107
Estimated alpha value: .1299
Estimated r value: .09723

<u>Question 3</u>

P(X(0)) = (.1299/.1299+1)^.09723 = .810324813

E(X(1)) = (.09723*1)/.1299 = .748498845

Reach

100 * (1-P(X(t)=0) = 100 * (1-.810324813) =**<u>18.97%</u>**

Average Frequency

E(X(1))/(1-P(X(t)=0)) = .748498845/(10.810324813) = **<u>3.94621</u>**

GRPs

100*E(X(1)) = 100 * .748498845 = **<u>74.8498845</u>**

Question 4

```
Data Pbooks;
set Mergedbooks (drop=NumBooksamazon);
run;

/* checking for missing values */

Proc means data=Pbooks N;
class education;
var education;
run;
```

The MEANS Procedure

Analysis Variable : education		
education	N Obs	N
0	1	1
1	638	638
2	772	772
3	13	13
4	811	811
5	302	302
99	6914	6914

We can assume that 99 are missing values. Since there are so many missing values in the education variable, we will not use it in our Poisson model.

```
/* fix missing region value */

data Pbooks;
set Pbooks;
if region='*' then region=.;
run;

/* Poisson Regression Model */

proc nlmixed data=Pbooks;
  /* m stands for lambda */
  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 = NumBooks*log(m)-m-log(fact(NumBooks));
  model NumBooks ~ general(ll);
run;
```

14	44	18821.9064	0.001616	1.25182	-0.00253
15	47	18821.9064	0.000022	0.036143	-0.00005

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	37644
AIC (smaller is better)	37660
AICC (smaller is better)	37660
BIC (smaller is better)	37717

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
m0	0.9836	0.07107	9440	13.84	<.0001	0.8443	1.1229	0.000780
b1	-0.1020	0.01110	9440	-9.18	<.0001	-0.1237	-0.08021	-0.00992
b2	-0.01454	0.01108	9440	-1.31	0.1895	-0.03627	0.007181	0.036143
b3	0.01931	0.003410	9440	5.66	<.0001	0.01263	0.02600	-0.01438
b4	0.01574	0.006312	9440	2.49	0.0126	0.003370	0.02812	-0.01917
b5	0.07322	0.03205	9440	2.28	0.0224	0.01039	0.1360	0.006430
b6	-0.2081	0.04423	9440	-4.70	<.0001	-0.2948	-0.1214	0.003172
b7	-0.1180	0.03375	9440	-3.50	0.0005	-0.1842	-0.05187	0.000525

Takeaways:

Optimized LL value: -18821.9064
All variables are significant except for hhsize due its p-value being > .05.
Race holds the highest significance, but not by much.
Region, race, and country have a negative relationship with a customer's numbers of purchases at barnesandnoble.com.
Age, income, and child have a positive relationship with a customer's numbers of purchases at barnesandnoble.com

Question 5

LL=log((gamma(r+NumBooks)/(gamma(r)fact(NumBooks)))*((alpha/(alpha+e^{Bx}))**r)*((e^{Bx}/(alpha+e^{Bx}))** NumBooks))

Where e^{Bx} = exp((b1*region)+(b2*hhsz)+(b3*age)+(b4*income)+(b5*child)+(b6*race)+(b7*country))

Question 6

Just like in question 4, we are not including education in our model due to too many missing values.

```
proc nlmixed data=Pbooks;
  parms r=1 a=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0;
  expBX=exp(b1*region+b2*hhsz+b3*age+b4*income+b5*child+b6*race+b7*country);
  ll = log(gamma(r+NumBooks))-log(gamma(r))-
log(fact(NumBooks))+r*log(a/(a+expBX))+NumBooks*log(expBX/(a+expBX));
  model NumBooks ~ general(ll);
run;
```

Optimal LL value: -8362.5473

-r, alpha, region, and race are our significant variables due to their p-values being less than .05.

-region and race have a negative relationship with the number of books purchased.

-Race is a more important variable than region when predicting how many books a customer purchases.

24	89	8362.5474	0.000416	1.70336	-0.00055
25	92	8362.5473	0.000051	0.087522	-0.00008

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	16725
AIC (smaller is better)	16743
AICC (smaller is better)	16743
BIC (smaller is better)	16807

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.09812	0.003096	9440	31.69	<.0001	0.09205	0.1042	0.087522
a	0.1070	0.02227	9440	4.80	<.0001	0.06331	0.1506	-0.08536
b1	-0.1033	0.03219	9440	-3.21	0.0013	-0.1664	-0.04017	0.024944
b2	-0.00841	0.03332	9440	-0.25	0.8006	-0.07372	0.05690	0.031233
b3	0.02706	0.01466	9440	1.85	0.0649	-0.00167	0.05579	0.068486
b4	0.01747	0.01876	9440	0.93	0.3518	-0.01930	0.05424	0.016861
b5	0.05834	0.09236	9440	0.63	0.5276	-0.1227	0.2394	0.005722
b6	-0.2098	0.1007	9440	-2.08	0.0373	-0.4072	-0.01232	0.015062
b7	-0.1015	0.09572	9440	-1.06	0.2889	-0.2892	0.08611	0.007374

Question 7

The biggest difference between the Poisson and NBD Regression models are the variables that are significant in each model. In the Poisson regression model, almost all the variables are significant while in the NBD regression model only race and region are significant. However, race is the most significant variable in both models, and race and region negatively impact the number of books purchased in both models. The NBD regression model also has a better optimized LL value of -8362.5473 compared to the Poisson model's optimized LL value of -18821.9064.

In the Poisson regression model, we assume that all the demographics explain the model and use a common lambda for every individual. In the NBD regression model we capture the unobserved component of differences among the individuals that are not included in the demographics given. Some of the demographics that were significant in the Poisson regression model become no longer significant in the NBD model because we capture the unobserved component of differences among individuals which explain the model better than the original demographics given.

Question 8

We learned from this project how to effectively build on different models to come up with the best solution to predict customer purchasing behavior. We can see how our optimized LL went down from the NBD model and the Poisson Regression model when we used the NBD Regression model showing that if we account for explanatory variables and unobserved heterogeneity, we can improve our prediction model. We learned from this project that by not accounting for the unobserved heterogeneity, we may think some variables are significant in predicting the number of books purchased when in reality they may not be significant.

We really enjoyed learning more SAS procedures in this project. Many of the questions could have been done with using PROC SQL, but since we did not cover much PROC SQL in ABI, we challenged ourselves to use different types of code to do the same thing. SAS has so many different options to run code.

Now that we have a prediction model, it would be interesting to do further analysis to see what regions and races prefer Amazon.com over Barnesandnoble.com, and if we could market those demographics differently to optimize profit for Barnes & Noble.