UT DALLAS | Naveen Jindal School of Management

MIS 6334.001 – ADVANCED BUSINESS ANALYSTICS WITH SAS

PROF. XIANJUN GENG

PROJECT 2 REPORT

PROJECT MEMBERS:

GROUP NO. 8

VINEET PRADEEP JOSHI

PIYUSH KAMDAR

SANGSANG JIA

ANDREA TAPIA

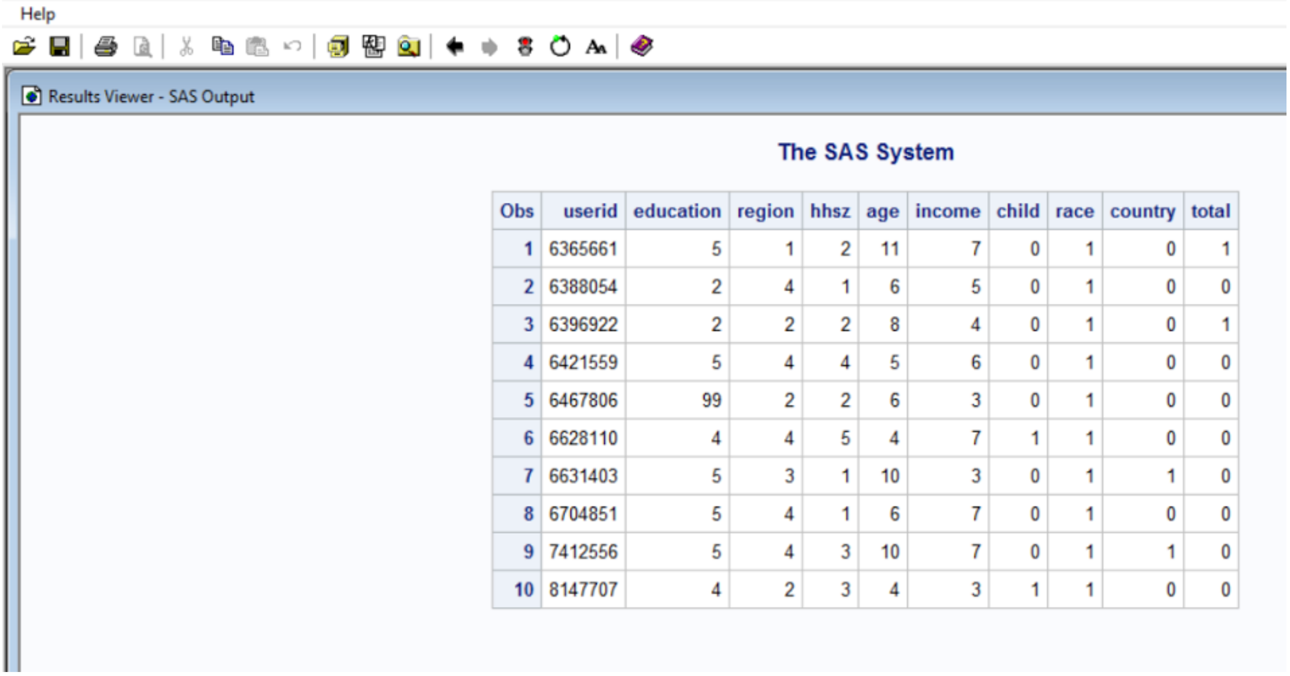
JUN ZHOU

Part I. Modeling Count Data

1. **SAS Code:**

```
proc sql;

create table datanew as

          select a.userid, a.education, a.region, a.hhsz, a.age,
     a.income, a.child, a.race, a.country, sum(a.qty_new) as total
          from
               (select *,CASE WHEN domain = "amazon.com" THEN qty
     = 0
                              Else qty = qty END AS qty_new
               from Mis_6334.Aba_project2_data_books) a
          group by a.userid, a.education, a.region, a.hhsz, a.age,
     a.income, a.child, a.race, a.country;
     quit;

     data Mis_6334.BNN_1;
     set Work.datanew;
     run;

     proc print data = Mis_6334.Bnn_1 (obs=10); run;
```

SAS Results:



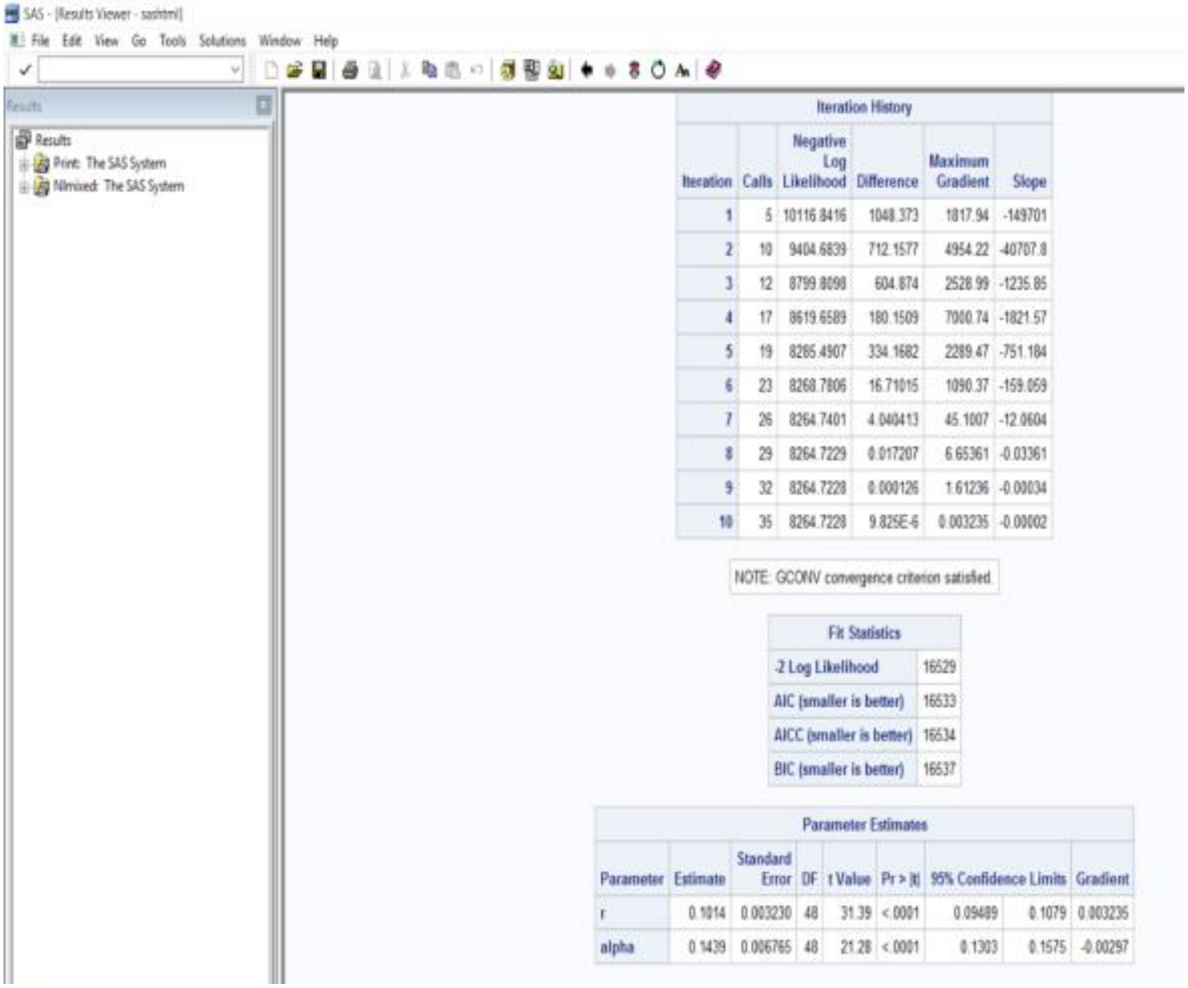
The SAS System

Obs	userid	education	region	hhsz	age	income	child	race	country	total
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	99	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

2. **SAS Code:**

```
                proc sql;
                create table Mis_6334.Q2 as
                select total, count(userid) as total_ppl from
Mis_6334.BNN_1
                group by total;
quit;

PROC NLMIXED DATA = Mis_6334.Q2;
parms r = 1, alpha = 1;
ll = total_ppl*log( (gamma(r + total)/(gamma(r)*perm(total))) *
((alpha/(alpha+1))**r) * ((1/(alpha+1))**total) );
MODEL total_ppl ~ general(ll);
RUN;
```

SAS Results:


SAS - [Results Viewer - sashtml]
File Edit View Go Tools Solutions Window Help
Results
Results
Print: The SAS System
Nlmixed: The SAS System

Iteration History					
Iteration	Calls	Negative Log Likelihood	Difference	Maximum Gradient	Slope
1	5	10116.8416	1048.373	1817.94	-149701
2	10	9404.6839	712.1577	4954.22	-40707.8
3	12	8799.8098	604.874	2528.99	-1235.85
4	17	8619.6589	180.1509	7000.74	-1821.57
5	19	8285.4907	334.1682	2289.47	-751.184
6	23	8268.7806	16.71015	1090.37	-159.059
7	26	8264.7401	4.040413	45.1007	-12.0604
8	29	8264.7229	0.017207	6.65361	-0.03361
9	32	8264.7228	0.000126	1.61236	-0.00034
10	35	8264.7228	9.825E-6	0.003235	-0.00002

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	16529
AIC (smaller is better)	16533
AICC (smaller is better)	16534
BIC (smaller is better)	16537

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.1014	0.003230	48	31.39	<.0001	0.09489	0.1079	0.003236
alpha	0.1439	0.006765	48	21.28	<.0001	0.1303	0.1575	-0.00297

3. Calculation Findings:

Reach: 1-P(X(t)=0)=1-(**α**/(**α**+1))^r=1-(0.1439/1.1439)^0.1014=0.1896

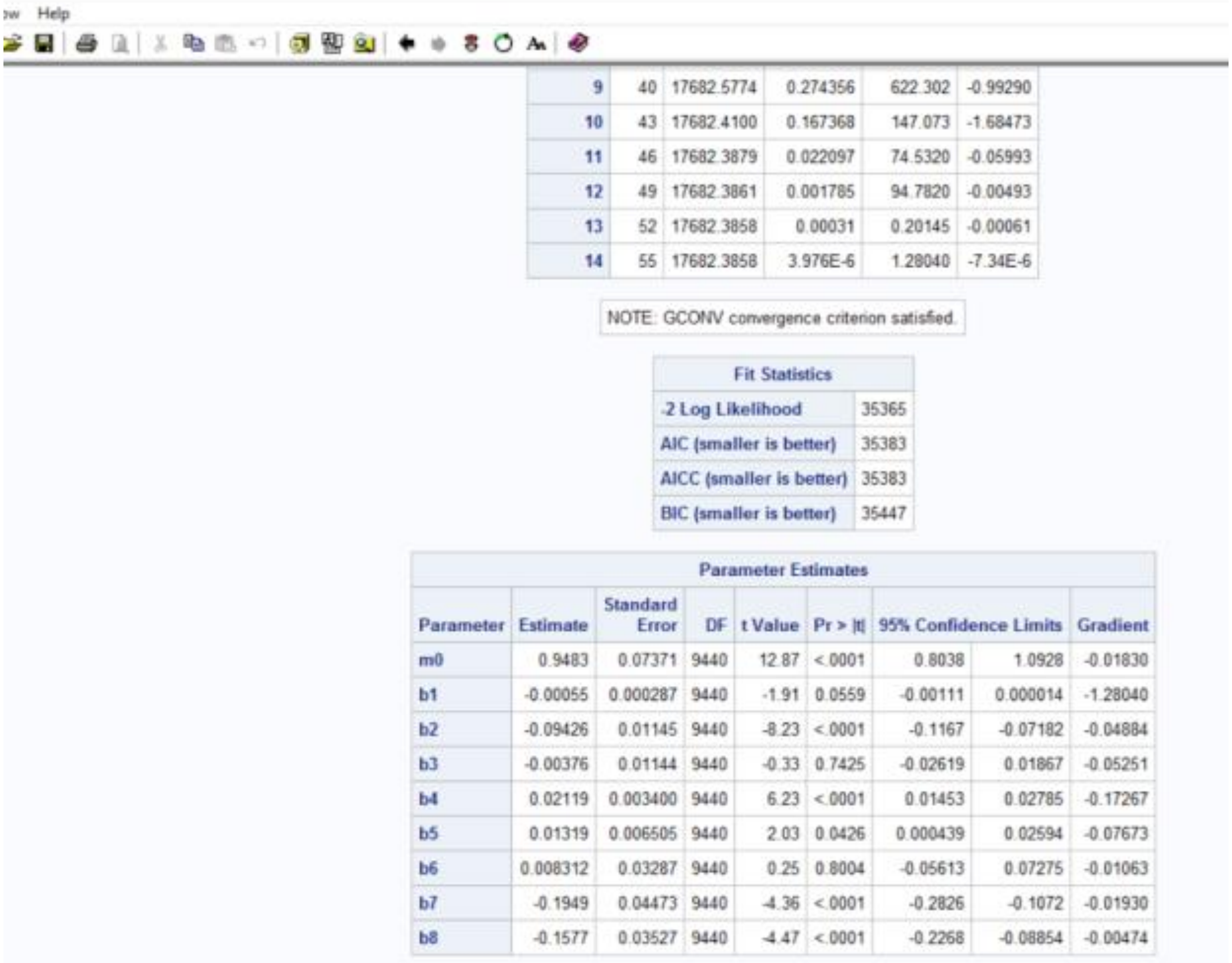
Average Frequency: r/**α**=0.7046

GRP:100*0.7046=70.46

4. SAS Code:

```
PROC NLMIXED DATA=Mis_6334.Bnn_1;
   /* m stands for lamdha */
   parms m0=1 b1=0 b2=0 b3=0 b4=0 b5=0 b6=0 b7=0 b8=0;

m=m0*exp(b1*education+b2*region+b3*hhsz+b4*age+b5*income+b6*chil
d+b7*race+b8*country);
   ll = total*log(m)-m-log(fact(total));
   MODEL total ~ general(ll);
   RUN;
```

SAS Results:



9	40	17682.5774	0.274356	622.302	-0.99290
10	43	17682.4100	0.167368	147.073	-1.68473
11	46	17682.3879	0.022097	74.5320	-0.05993
12	49	17682.3861	0.001785	94.7820	-0.00493
13	52	17682.3858	0.00031	0.20145	-0.00061
14	55	17682.3858	3.976E-6	1.28040	-7.34E-6

NOTE: GCONV convergence criterion satisfied.

Fit Statistics	
-2 Log Likelihood	35365
AIC (smaller is better)	35383
AICC (smaller is better)	35383
BIC (smaller is better)	35447

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
m0	0.9483	0.07371	9440	12.87	<.0001	0.8038	1.0928	-0.01830
b1	-0.00055	0.000287	9440	-1.91	0.0559	-0.00111	0.000014	-1.28040
b2	-0.09426	0.01145	9440	-8.23	<.0001	-0.1167	-0.07182	-0.04884
b3	-0.00376	0.01144	9440	-0.33	0.7425	-0.02619	0.01867	-0.05251
b4	0.02119	0.003400	9440	6.23	<.0001	0.01453	0.02785	-0.17267
b5	0.01319	0.006505	9440	2.03	0.0426	0.000439	0.02594	-0.07673
b6	0.008312	0.03287	9440	0.25	0.8004	-0.05613	0.07275	-0.01063
b7	-0.1949	0.04473	9440	-4.36	<.0001	-0.2826	-0.1072	-0.01930
b8	-0.1577	0.03527	9440	-4.47	<.0001	-0.2268	-0.08854	-0.00474

Managerial Takeaways:

- In terms of using “date” in the regression, we opted not to include as books are year-round purchases and we would need much more data with a timestamp. Additionally, for those customers who purchased once or very little in a time frame, that record would not be meaningful.
- Poisson Regression Model LL: -17,682.5
- Upon comparing with 0.05 alpha-level, the following predictor variables are not significant, and we can infer that these factors do not affect the customer purchasing behavior at Barnes and Noble.
  - b1 (education)
  - b3 (House hold size)
  - b6 (child)
- In contrast, the following factors are significant per a 0.05 alpha-level.
  - b2 (region)
  - b4 (age)
  - b5 (income)
  - b7 (race)
  - b8 (country)

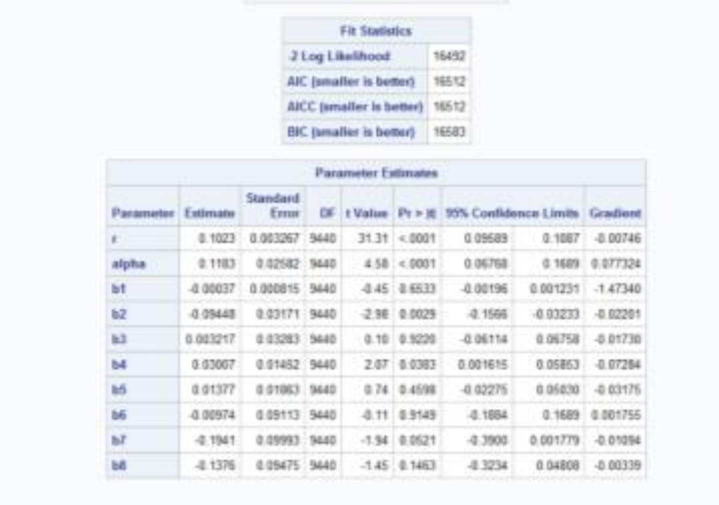
5. **SAS Code:**

```
m =
exp(b1*education+b2*region+b3*hhsz+b4*age+b5*income+b6*child+b7*
race+b8*country);
  ll = log( (gamma(r + total)/(gamma(r)*perm(total))) *
((alpha/(alpha+m))**r) * ((m/(alpha+m))**total) );
```

6. **SAS Code:**

```
PROC NLMIXED DATA=Mis_6334.Bnn_1;

       parms r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0
     b7=0 b8=0;
       /* m gives us the exp(beta*x) values which are then used in
     the formula for calculating the log likelihood */
       m                                                          =
     exp(b1*education+b2*region+b3*hhsz+b4*age+b5*income+b6*chil
     d+b7*race+b8*country);
       ll  =  log(  (gamma(r  +  total)/(gamma(r)*perm(total)))  *
     ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total) );
       MODEL total ~ general(ll);
     RUN;
```

SAS Results:



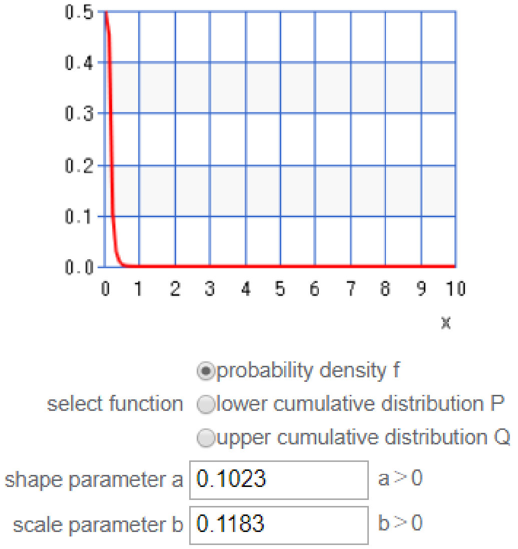
Fit Statistics	
-2 Log Likelihood	16492
AIC (smaller is better)	16512
AICC (smaller is better)	16512
BIC (smaller is better)	16583

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.1023	0.003267	9440	31.31	<.0001	0.09589	0.1087	-0.00746
alpha	0.1183	0.02582	9440	4.58	<.0001	0.06768	0.1689	0.077324
b1	-0.00037	0.000815	9440	-0.45	0.6533	-0.00196	0.001231	-1.47340
b2	-0.09448	0.03171	9440	-2.98	0.0029	-0.1566	-0.03233	-0.02201
b3	0.003217	0.03283	9440	0.10	0.9220	-0.06114	0.06758	-0.01730
b4	0.03007	0.01452	9440	2.07	0.0383	0.001615	0.05853	-0.07284
b5	0.01377	0.01863	9440	0.74	0.4598	-0.02275	0.05030	-0.03175
b6	-0.00974	0.09113	9440	-0.11	0.9149	-0.1884	0.1689	0.001755
b7	-0.1941	0.09993	9440	-1.94	0.0521	-0.3900	0.001779	-0.01094
b8	-0.1376	0.09475	9440	-1.45	0.1463	-0.3234	0.04808	-0.00339

Managerial Takeaways:

- NBD Regression model LL: -8,246
- Upon comparing with 0.05 alpha-level, the following predictor variables are not significant:
    - b1(education)
    - b3(House hold size)
    - b5 (income)
    - b6 (child)
    - b7 (race)
    - b8 (country)
- In contrast, the following factors are significant per a 0.05 alpha-level.
    - b2(region)
    - b4(age)
- Moreover, we plotted the Gamma distribution using shape and scale parameters Source: http://keisan.casio.com/exec/system/1180573216

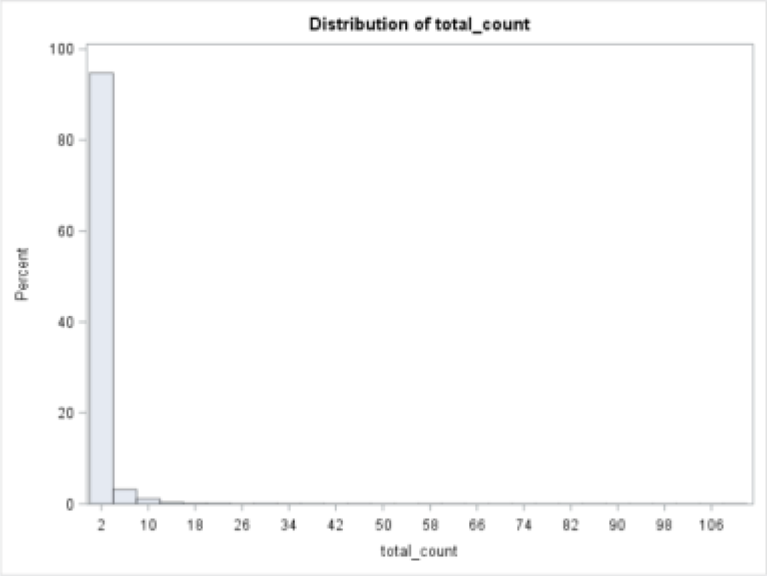


select function ● probability density f
○ lower cumulative distribution P
○ upper cumulative distribution Q

shape parameter a 0.1023 a>0

scale parameter b 0.1183 b>0

When comparing this distribution with the distribution of our data using SAS the distributions are very similar which suggests our NBD model fits the data well.



7.**Noticeable difference regarding the managerial takeaways:**

- In comparing LL values, we find that the NBD Regression model LL value of -8,246 is greater than the Poisson LL value of -17,682.5 which suggests that NBD regression fits the data better.
- The NBD Regression model has less significant variables than the Poisson model.

8. **LR test Takeaways:**

Null Hypothesis - NBD Regression model is not different from the Poisson Regression model.

NBD Regression model has one extra degree of freedom than the Poisson Regression model.

LLB = -2LL for the NBD Regression model is 16,492
LLA = -2LL for the Poisson Regression model is 35,365
LR= -2（LL（for　NBD Regression）- LL(for Poisson Regression)) = 35,365 – 16,492 = 18,873

X^2(0.05, 1) = 3.841

To reject the null hypothesis, LR = -2（LLB- LLA）> X^2(0.05, 1)

To reject the null hypothesis, LR = 18,873 > 3.841

Clearly, we reject the null hypothesis. Our NBD Regression model is different from the Poisson regression model.

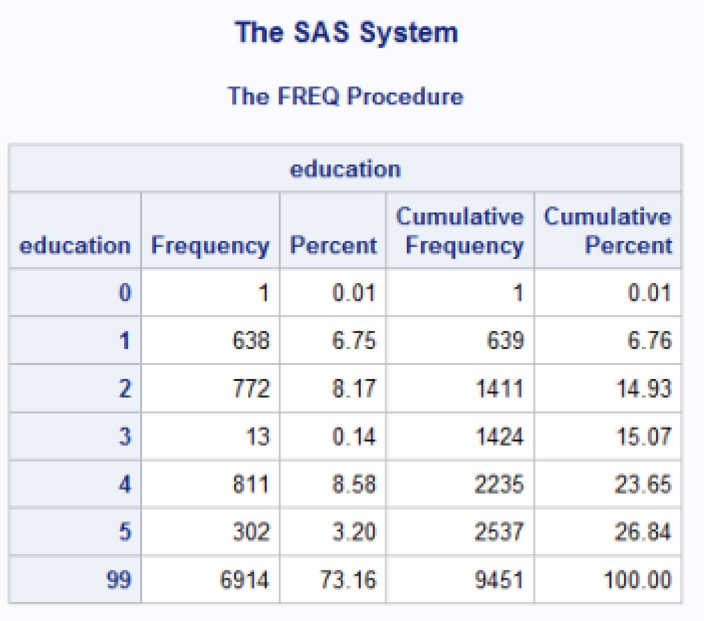
Additionally, after finding that the NBD Regression and Poisson models are different when comparing their LL values, we find that the NBD Regression model LL value of -8,246 is greater than the Poisson LL value of -17,682.5 which suggests that NBD regression fits the data better.

Part II. Improving the model

9. **SAS Code:**

```
proc freq data=mis6334.datanew;
table education;
run;
```

SAS Results:



The SAS System

The FREQ Procedure

education				
education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1	0.01	1	0.01
1	638	6.75	639	6.76
2	772	8.17	1411	14.93
3	13	0.14	1424	15.07
4	811	8.58	2235	23.65
5	302	3.20	2537	26.84
99	6914	73.16	9451	100.00

Per the proc freq output, we find education has 6,914 missing values so we should remove this variable.

10. **SAS Code:**

Code for creating the weekend variable

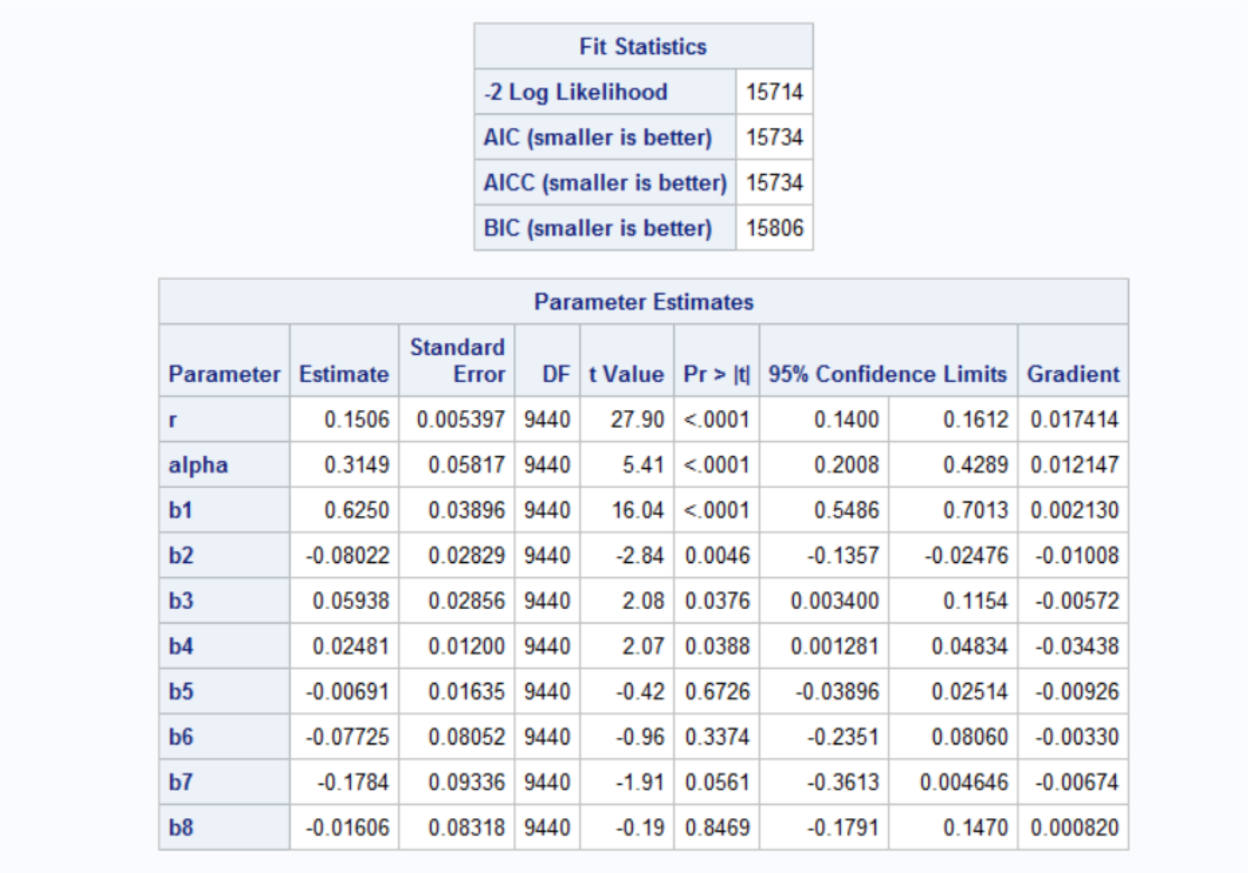
```
/* Extract day of the week */
DATA abasas.dataset;
SET abasas.Aba_project2_data_books;
date = input(put(date,8.),yymmdd8.);
format date yymmdd10. ;
day_of_week = weekday(date);
RUN;
```

```
/* Create a variable for weekend purchases */
DATA abasas.dataset;
SET abasas.dataset;
IF (day_of_week < 2) OR (day_of_week > 6)
THEN weekend = 1;
ELSE weekend = 0;
RUN;

/* Table for purchases made on weekends from B and N */
PROC SQL;
CREATE TABLE dummy_1 AS
SELECT userid, region, hhsz, age, income, child, race, country,
domain,
CASE WHEN domain = "amazon.com" THEN qty = 0
ELSE qty = qty END AS qty_new
FROM abasas.dataset
WHERE weekend = 1;
QUIT;

/* Aggregating weekend purchases per user */
PROC SQL ;
CREATE TABLE dummy_3 AS
SELECT userid, SUM(qty_new) AS total_count_weekend
FROM dummy_1
GROUP BY userid;
QUIT;

/* Merging weekend count data with the original dataset */
DATA abasas.weekend;
MERGE abasas.count_data dummy_3;
BY userid;
IF total_count_weekend = . THEN
total_count_weekend = 0;
RUN;

/* NBD Regression excluding education but including weekend*/
PROC NLMIXED DATA = dummy_7;
PARMS r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0 b7=0
b8=0;
/* m gives us the exp(beta*x) values which are then used in the
formula for calculating the log likelihood */
m =
exp(b1*total_count_weekend+b2*region+b3*hhsz+b4*age+b5*income+b6
*child+b7*race+b8*country);
ll = log( (gamma(r + total_count)/(gamma(r)*perm(total_count)))
* ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total_count) );
MODEL total_count ~ general(ll);
```

```
RUN;
```

SAS Results:

NBD Regression results excluding education and including the total_weekend_count variable



Fit Statistics	
-2 Log Likelihood	15714
AIC (smaller is better)	15734
AICC (smaller is better)	15734
BIC (smaller is better)	15806

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.1506	0.005397	9440	27.90	<.0001	0.1400	0.1612	0.017414
alpha	0.3149	0.05817	9440	5.41	<.0001	0.2008	0.4289	0.012147
b1	0.6250	0.03896	9440	16.04	<.0001	0.5486	0.7013	0.002130
b2	-0.08022	0.02829	9440	-2.84	0.0046	-0.1357	-0.02476	-0.01008
b3	0.05938	0.02856	9440	2.08	0.0376	0.003400	0.1154	-0.00572
b4	0.02481	0.01200	9440	2.07	0.0388	0.001281	0.04834	-0.03438
b5	-0.00691	0.01635	9440	-0.42	0.6726	-0.03896	0.02514	-0.00926
b6	-0.07725	0.08052	9440	-0.96	0.3374	-0.2351	0.08060	-0.00330
b7	-0.1784	0.09336	9440	-1.91	0.0561	-0.3613	0.004646	-0.00674
b8	-0.01606	0.08318	9440	-0.19	0.8469	-0.1791	0.1470	0.000820

Creating variable for loyalty (BN Purchases/Total Purchases)

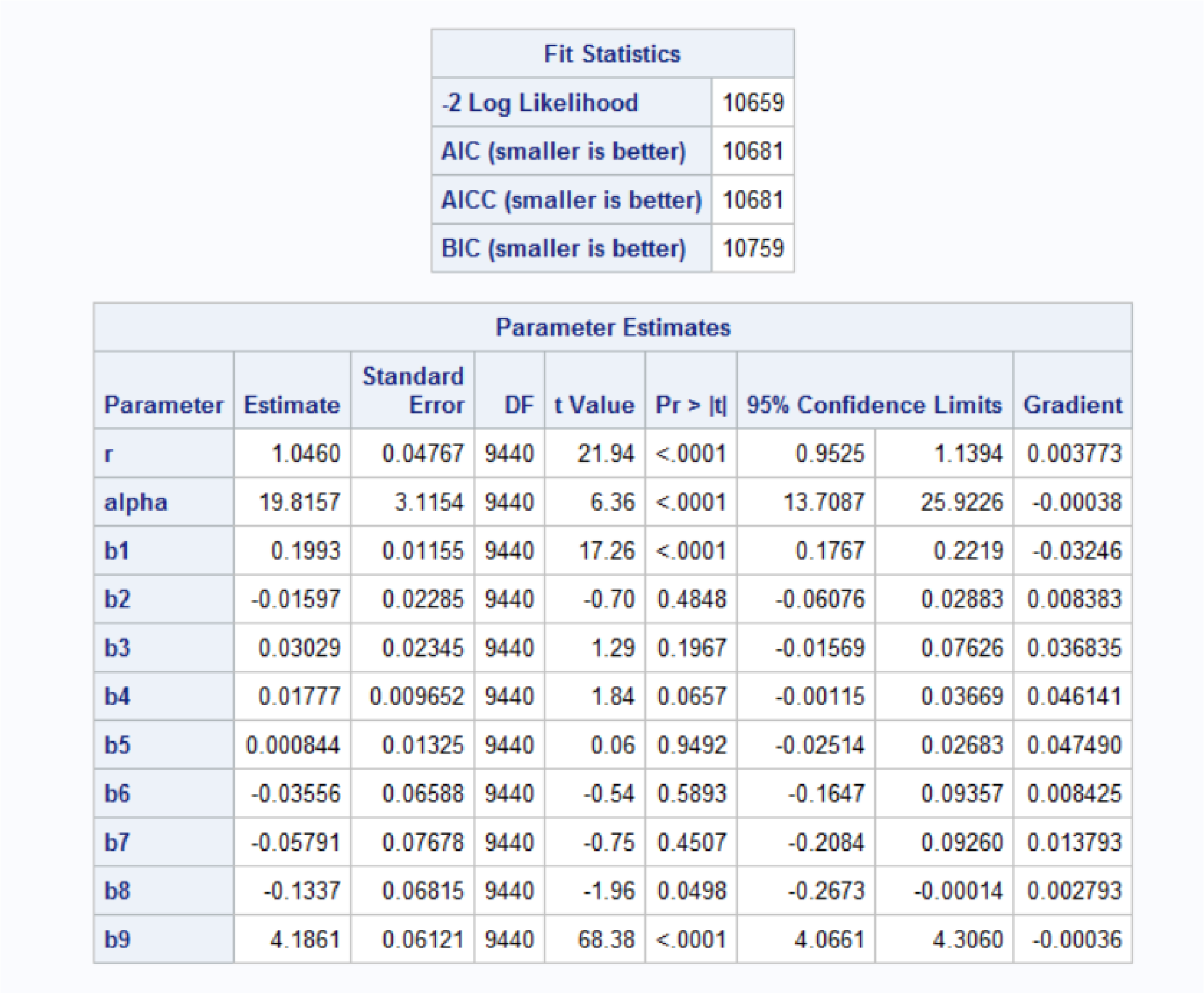
```
/* Amazon purchases */
data amazon;
set abasas.Aba_project2_data_books;
if domain='amazon.com';
run;

/* B and N purchases */
data bandn;
set abasas.Aba_project2_data_books;
if domain~='amazon.com';
run;

/* Total Amazon  */
proc sql;
create table amazon_total as
```

```
select userid, sum(qty) as total_amazon
from amazon
group by userid;
quit;

/* Total Barnes and Nobles */
proc sql;
create table bandn_total as
select userid, sum(qty) as total_bandn
from bandn
group by userid;
quit;

/* Total Amazon and Barnes and Nobles */
data all;
merge amazon_total bandn_total;
by userid;
run;

/* Loyalty table */
data all;
set all;
if total_amazon=. then loyalty=1;
else if total_bandn=. then loyalty=0;
else loyalty =total_bandn/(total_bandn+ total_amazon);
run;

/* Merging the table with the loyalty variable with the original
dataset */
data abasas.weekend;
merge abasas.weekend all;
by userid;
run;

/* NBD Regression excluding education but including weekend and
loyalty*/
PROC NLMIXED DATA = abasas.weekend;
PARMS r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0 b7 =
0 b8 = 0 b9 = 0;
/* m gives us the exp(beta*x) values which are then used in the
formula for calculating the log likelihood */
m =
exp(b1*total_count_weekend+b2*region+b3*hhsz+b4*age+b5*income+b6
*child+b7*race+b8*country + b9*loyalty);

ll = log( (gamma(r + total_count)/(gamma(r)*perm(total_count)))
* ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total_count) );
```

```
MODEL total_count ~ general(ll);
RUN;
```

SAS Results:

NBD Regression results excluding education and including the total_weekend_count and loyalty



Fit Statistics	
-2 Log Likelihood	10659
AIC (smaller is better)	10681
AICC (smaller is better)	10681
BIC (smaller is better)	10759

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.0460	0.04767	9440	21.94	<.0001	0.9525	1.1394	0.003773
alpha	19.8157	3.1154	9440	6.36	<.0001	13.7087	25.9226	-0.00038
b1	0.1993	0.01155	9440	17.26	<.0001	0.1767	0.2219	-0.03246
b2	-0.01597	0.02285	9440	-0.70	0.4848	-0.06076	0.02883	0.008383
b3	0.03029	0.02345	9440	1.29	0.1967	-0.01569	0.07626	0.036835
b4	0.01777	0.009652	9440	1.84	0.0657	-0.00115	0.03669	0.046141
b5	0.000844	0.01325	9440	0.06	0.9492	-0.02514	0.02683	0.047490
b6	-0.03556	0.06588	9440	-0.54	0.5893	-0.1647	0.09357	0.008425
b7	-0.05791	0.07678	9440	-0.75	0.4507	-0.2084	0.09260	0.013793
b8	-0.1337	0.06815	9440	-1.96	0.0498	-0.2673	-0.00014	0.002793
b9	4.1861	0.06121	9440	68.38	<.0001	4.0661	4.3060	-0.00036

Managerial Takeaways:

- NBD Regression model LL: -5,329.5 which is a much more improved model.
- Upon comparing with 0.05 alpha-level, the following predictor variables are not significant:
  - b2(region)
  - b3(House hold size)
  - b4(age)
  - b5 (income)
  - b6 (child)
  - b7 (race)

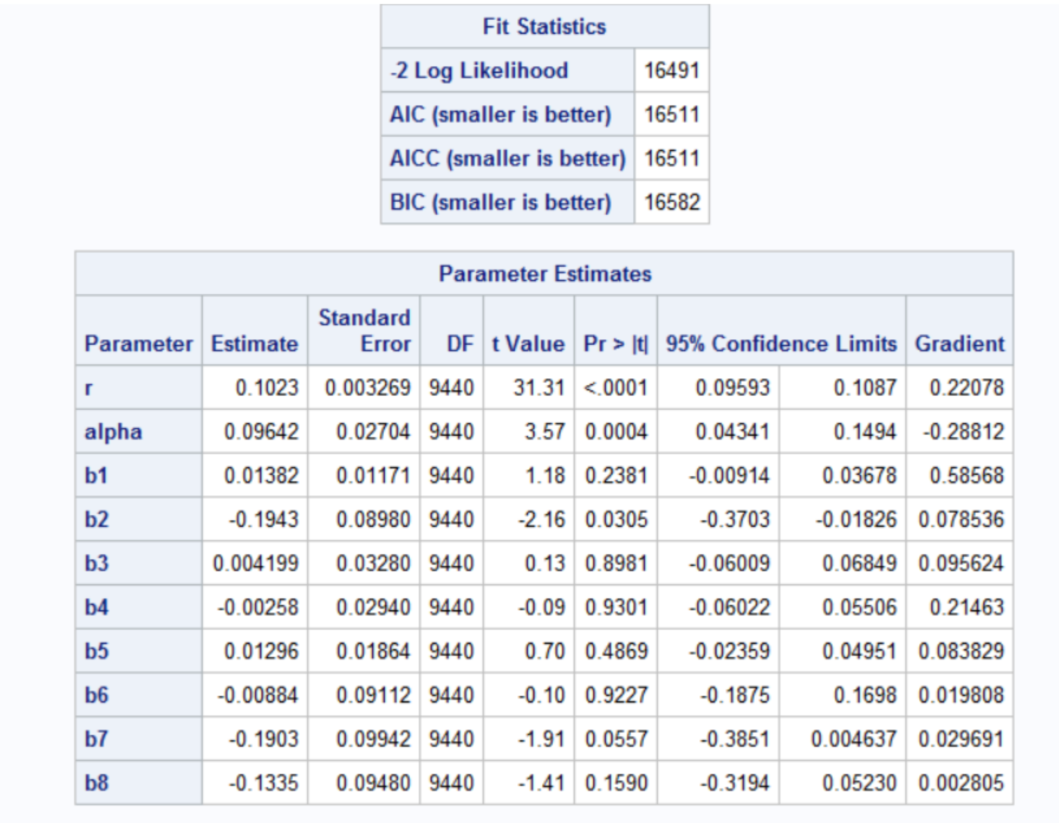
- In contrast, the following factors are significant per a 0.05 alpha-level. Where b1 and b9 were the variables that we created.
    - b1(weekend)
    - b8(country)
    - b9(loyalty)

11. **SAS Code:**

```
/* NBD Regression with an interaction variable between region
and age*/

PROC NLMIXED DATA = abasas.weekend;

PARMS r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0 b7 =
0 b8 = 0 ;

/* m gives us the exp(beta*x) values which are then used in the
formula for calculating the log likelihood */

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

ll = log( (gamma(r + total_count)/(gamma(r)*perm(total_count)))
* ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total_count) );

MODEL total_count ~ general(ll);

RUN;
```

SAS Results:

NBD Regression Results

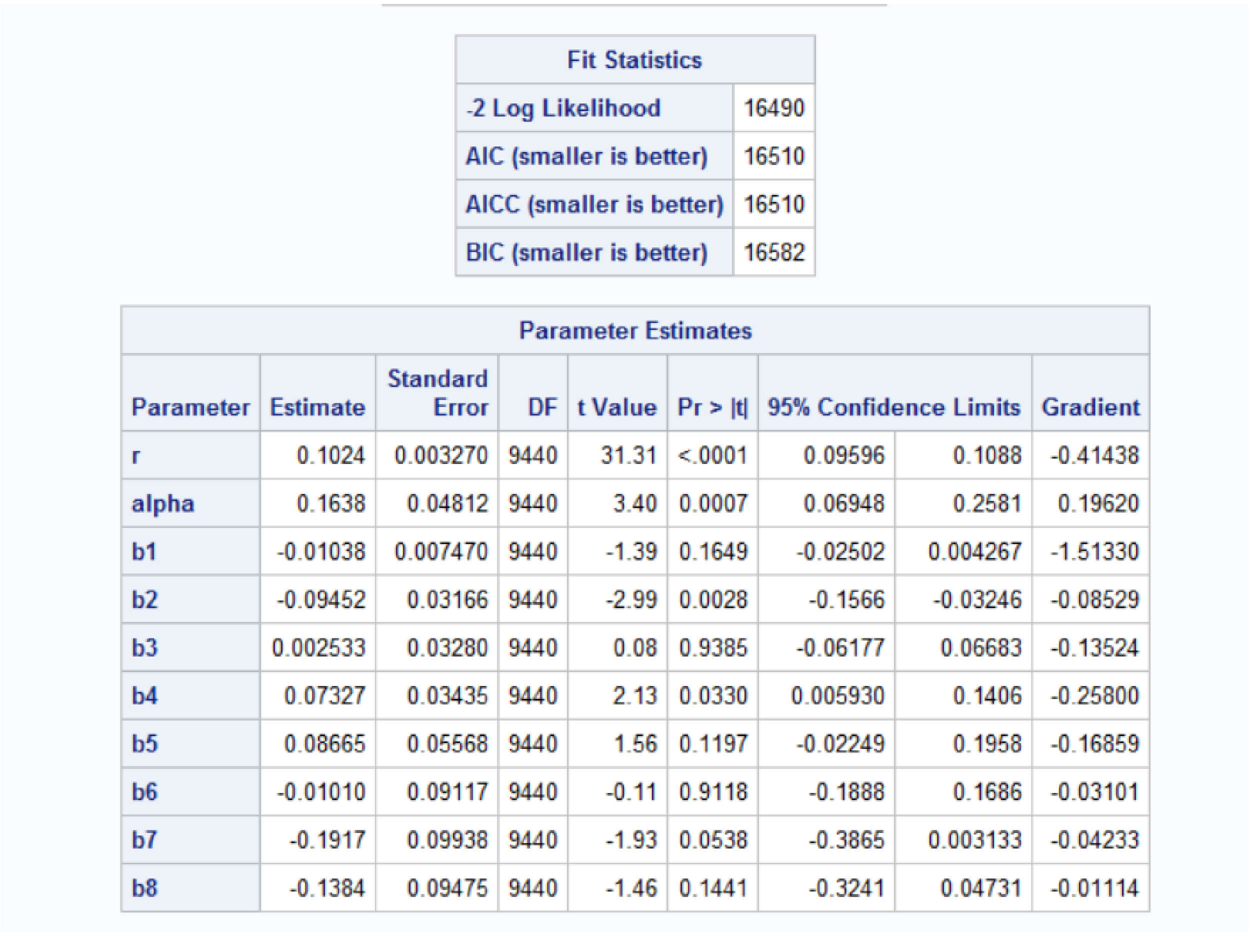


Fit Statistics	
-2 Log Likelihood	16491
AIC (smaller is better)	16511
AICC (smaller is better)	16511
BIC (smaller is better)	16582

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.1023	0.003269	9440	31.31	<.0001	0.09593	0.1087	0.22078
alpha	0.09642	0.02704	9440	3.57	0.0004	0.04341	0.1494	-0.28812
b1	0.01382	0.01171	9440	1.18	0.2381	-0.00914	0.03678	0.58568
b2	-0.1943	0.08980	9440	-2.16	0.0305	-0.3703	-0.01826	0.078536
b3	0.004199	0.03280	9440	0.13	0.8981	-0.06009	0.06849	0.095624
b4	-0.00258	0.02940	9440	-0.09	0.9301	-0.06022	0.05506	0.21463
b5	0.01296	0.01864	9440	0.70	0.4869	-0.02359	0.04951	0.083829
b6	-0.00884	0.09112	9440	-0.10	0.9227	-0.1875	0.1698	0.019808
b7	-0.1903	0.09942	9440	-1.91	0.0557	-0.3851	0.004637	0.029691
b8	-0.1335	0.09480	9440	-1.41	0.1590	-0.3194	0.05230	0.002805

```
/* NBD Regression with an interaction variable between income
and age*/

PROC NLMIXED DATA = abasas.weekend;

PARMS r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0 b7 =
0 b8 = 0;

/* m gives us the exp(beta*x) values which are then used in the
formula for calculating the log likelihood */

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

ll = log( (gamma(r + total_count)/(gamma(r)*perm(total_count)))
* ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total_count) );

MODEL total_count ~ general(ll);

RUN;
```



Fit Statistics	
-2 Log Likelihood	16490
AIC (smaller is better)	16510
AICC (smaller is better)	16510
BIC (smaller is better)	16582

Parameter Estimates

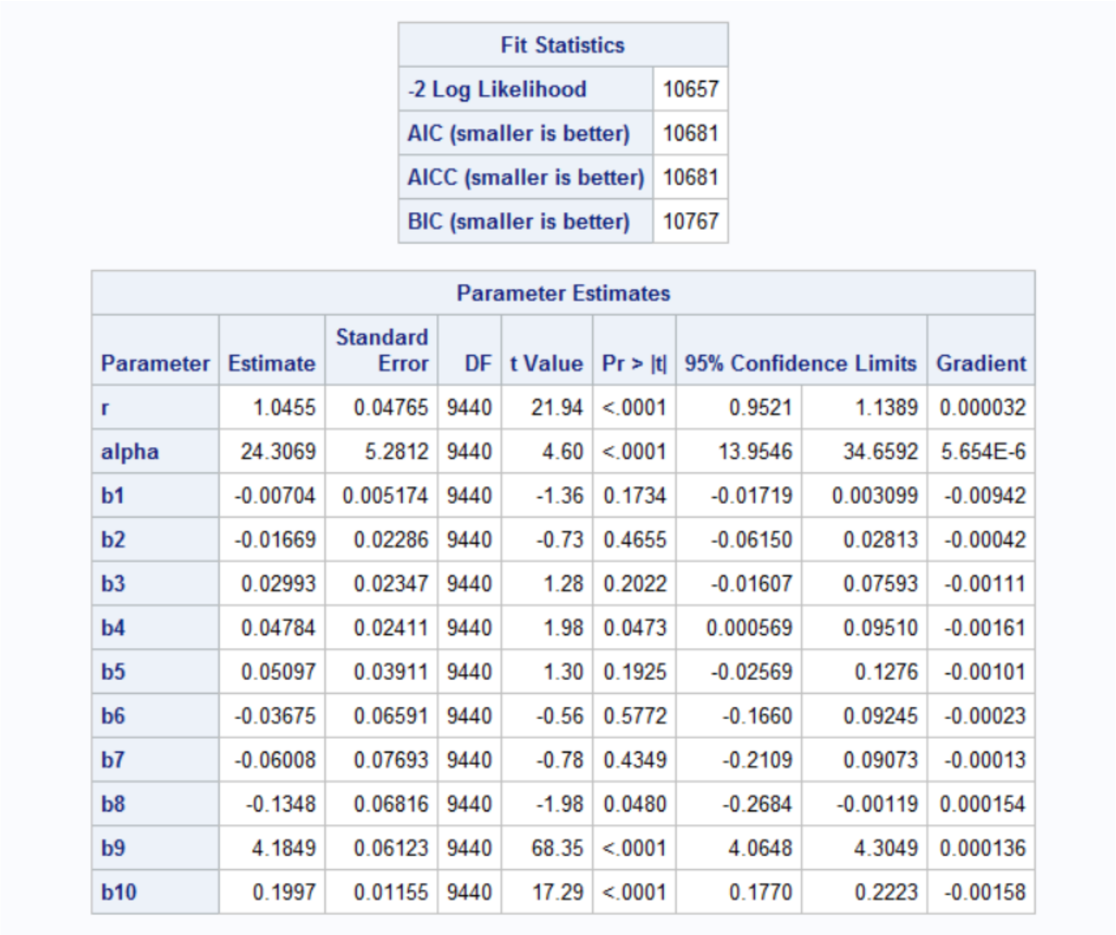
Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	0.1024	0.003270	9440	31.31	<.0001	0.09596	0.1088	-0.41438
alpha	0.1638	0.04812	9440	3.40	0.0007	0.06948	0.2581	0.19620
b1	-0.01038	0.007470	9440	-1.39	0.1649	-0.02502	0.004267	-1.51330
b2	-0.09452	0.03166	9440	-2.99	0.0028	-0.1566	-0.03246	-0.08529
b3	0.002533	0.03280	9440	0.08	0.9385	-0.06177	0.06683	-0.13524
b4	0.07327	0.03435	9440	2.13	0.0330	0.005930	0.1406	-0.25800
b5	0.08665	0.05568	9440	1.56	0.1197	-0.02249	0.1958	-0.16859
b6	-0.01010	0.09117	9440	-0.11	0.9118	-0.1888	0.1686	-0.03101
b7	-0.1917	0.09938	9440	-1.93	0.0538	-0.3865	0.003133	-0.04233
b8	-0.1384	0.09475	9440	-1.46	0.1441	-0.3241	0.04731	-0.01114

We also tried running models by including interaction variables, weekend, and loyalty but we noticed that whenever we included weekend and loyalty, all other variables were always insignificant.

Some examples below:

```
/* NBD Regression with demographic variables, interaction
variable between income and age, weekend, and loyalty*/

PROC NLMIXED DATA = abasas.weekend;

PARMS r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0 b7 =
0 b8 = 0 b9=0 b10=0;

/* m gives us the exp(beta*x) values which are then used in the
formula for calculating the log likelihood */

m =
exp(b1*income*age+b2*region+b3*hhsz+b4*age+b5*income+b6*child+b7
*race+b8*country+b9*loyalty+b10*total_count_weekend);

ll = log( (gamma(r + total_count)/(gamma(r)*perm(total_count)))
* ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total_count) );

MODEL total_count ~ general(ll);

RUN;
```

SAS Results:



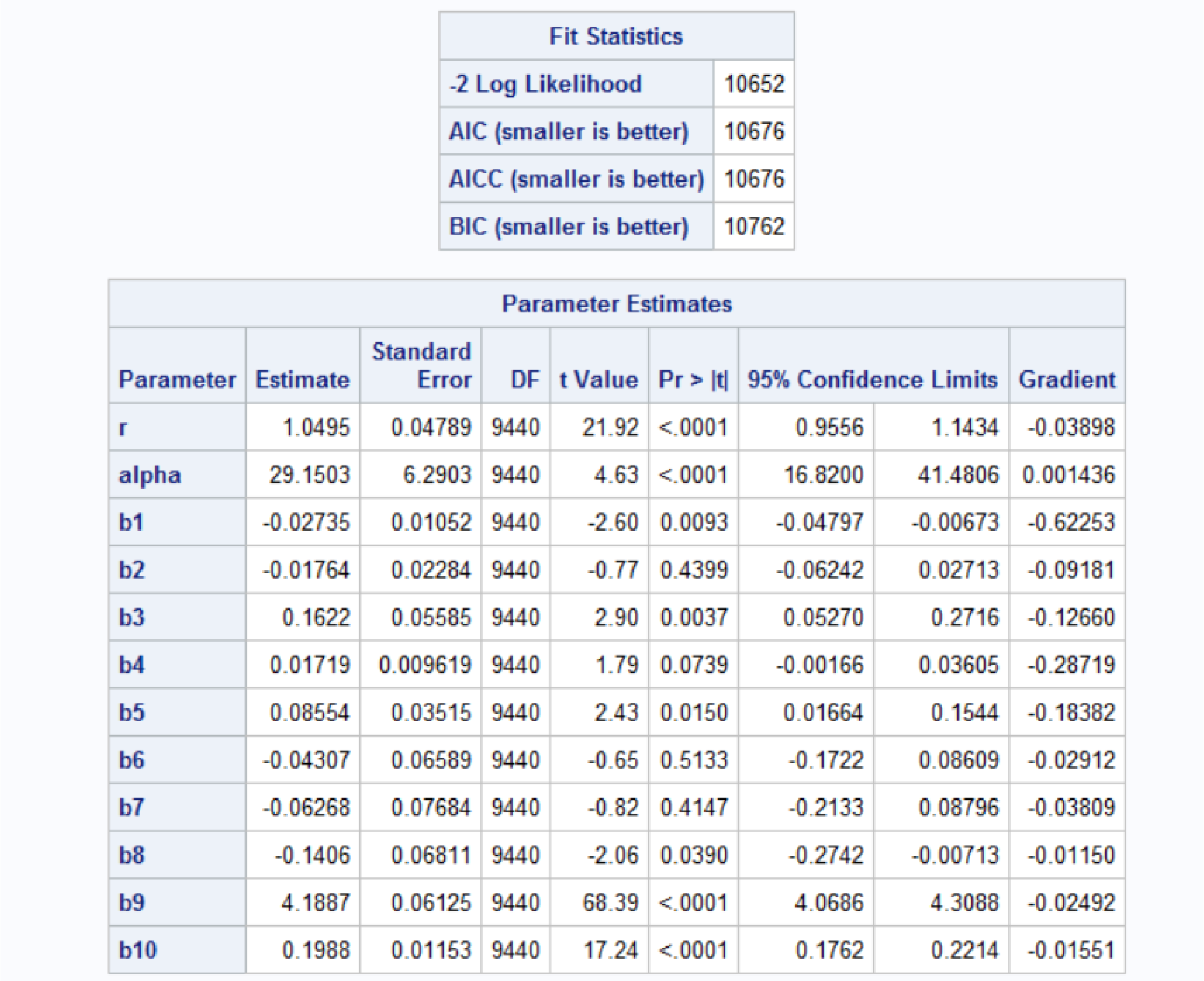
Fit Statistics	
-2 Log Likelihood	10657
AIC (smaller is better)	10681
AICC (smaller is better)	10681
BIC (smaller is better)	10767

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.0455	0.04765	9440	21.94	<.0001	0.9521	1.1389	0.000032
alpha	24.3069	5.2812	9440	4.60	<.0001	13.9546	34.6592	5.654E-6
b1	-0.00704	0.005174	9440	-1.36	0.1734	-0.01719	0.003099	-0.00942
b2	-0.01669	0.02286	9440	-0.73	0.4655	-0.06150	0.02813	-0.00042
b3	0.02993	0.02347	9440	1.28	0.2022	-0.01607	0.07593	-0.00111
b4	0.04784	0.02411	9440	1.98	0.0473	0.000569	0.09510	-0.00161
b5	0.05097	0.03911	9440	1.30	0.1925	-0.02569	0.1276	-0.00101
b6	-0.03675	0.06591	9440	-0.56	0.5772	-0.1660	0.09245	-0.00023
b7	-0.06008	0.07693	9440	-0.78	0.4349	-0.2109	0.09073	-0.00013
b8	-0.1348	0.06816	9440	-1.98	0.0480	-0.2684	-0.00119	0.000154
b9	4.1849	0.06123	9440	68.35	<.0001	4.0648	4.3049	0.000136
b10	0.1997	0.01155	9440	17.29	<.0001	0.1770	0.2223	-0.00158

```
/* NBD Regression with demographic variables, interaction
variable between income and hhsz, weekend, and loyalty*/

PROC NLMIXED DATA = abasas.weekend;

PARMS r = 1 alpha = 1 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5=0 b6=0 b7 =
0 b8 = 0 b9=0 b10=0;

/* m gives us the exp(beta*x) values which are then used in the
formula for calculating the log likelihood */

m =
exp(b1*income*hhsz+b2*region+b3*hhsz+b4*age+b5*income+b6*child+b
7*race+b8*country+b9*loyalty+b10*total_count_weekend);

ll = log( (gamma(r + total_count)/(gamma(r)*perm(total_count)))
* ((alpha/(alpha+m))**r) * ((m/(alpha+m))**total_count) );

MODEL total_count ~ general(ll);

RUN;
```

SAS Results:



Fit Statistics	
-2 Log Likelihood	10652
AIC (smaller is better)	10676
AICC (smaller is better)	10676
BIC (smaller is better)	10762

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits		Gradient
r	1.0495	0.04789	9440	21.92	<.0001	0.9556	1.1434	-0.03898
alpha	29.1503	6.2903	9440	4.63	<.0001	16.8200	41.4806	0.001436
b1	-0.02735	0.01052	9440	-2.60	0.0093	-0.04797	-0.00673	-0.62253
b2	-0.01764	0.02284	9440	-0.77	0.4399	-0.06242	0.02713	-0.09181
b3	0.1622	0.05585	9440	2.90	0.0037	0.05270	0.2716	-0.12660
b4	0.01719	0.009619	9440	1.79	0.0739	-0.00166	0.03605	-0.28719
b5	0.08554	0.03515	9440	2.43	0.0150	0.01664	0.1544	-0.18382
b6	-0.04307	0.06589	9440	-0.65	0.5133	-0.1722	0.08609	-0.02912
b7	-0.06268	0.07684	9440	-0.82	0.4147	-0.2133	0.08796	-0.03809
b8	-0.1406	0.06811	9440	-2.06	0.0390	-0.2742	-0.00713	-0.01150
b9	4.1887	0.06125	9440	68.39	<.0001	4.0686	4.3088	-0.02492
b10	0.1988	0.01153	9440	17.24	<.0001	0.1762	0.2214	-0.01551

Managerial Takeaways:

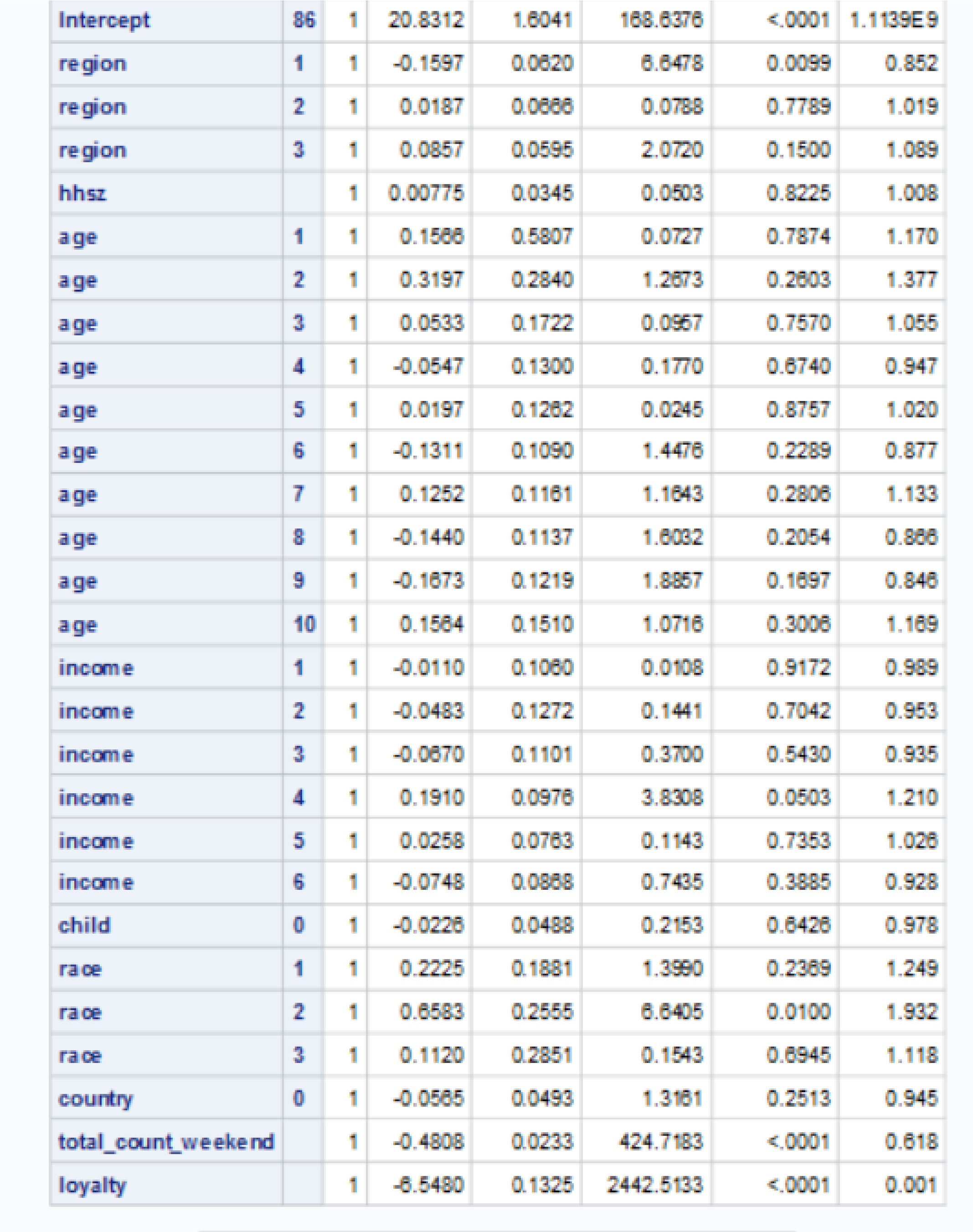
- NBD Regression model LL: -5,326 which is a very small improvement.
- After the addition of two new variables that are weekend and loyalty all the other variables turned out to be insignificant.
- We tried the below two interactions and saw no effect.
  - income*age
  - income*hhsz

Part III. Why Certain Customers Prefer Amazon over BN?

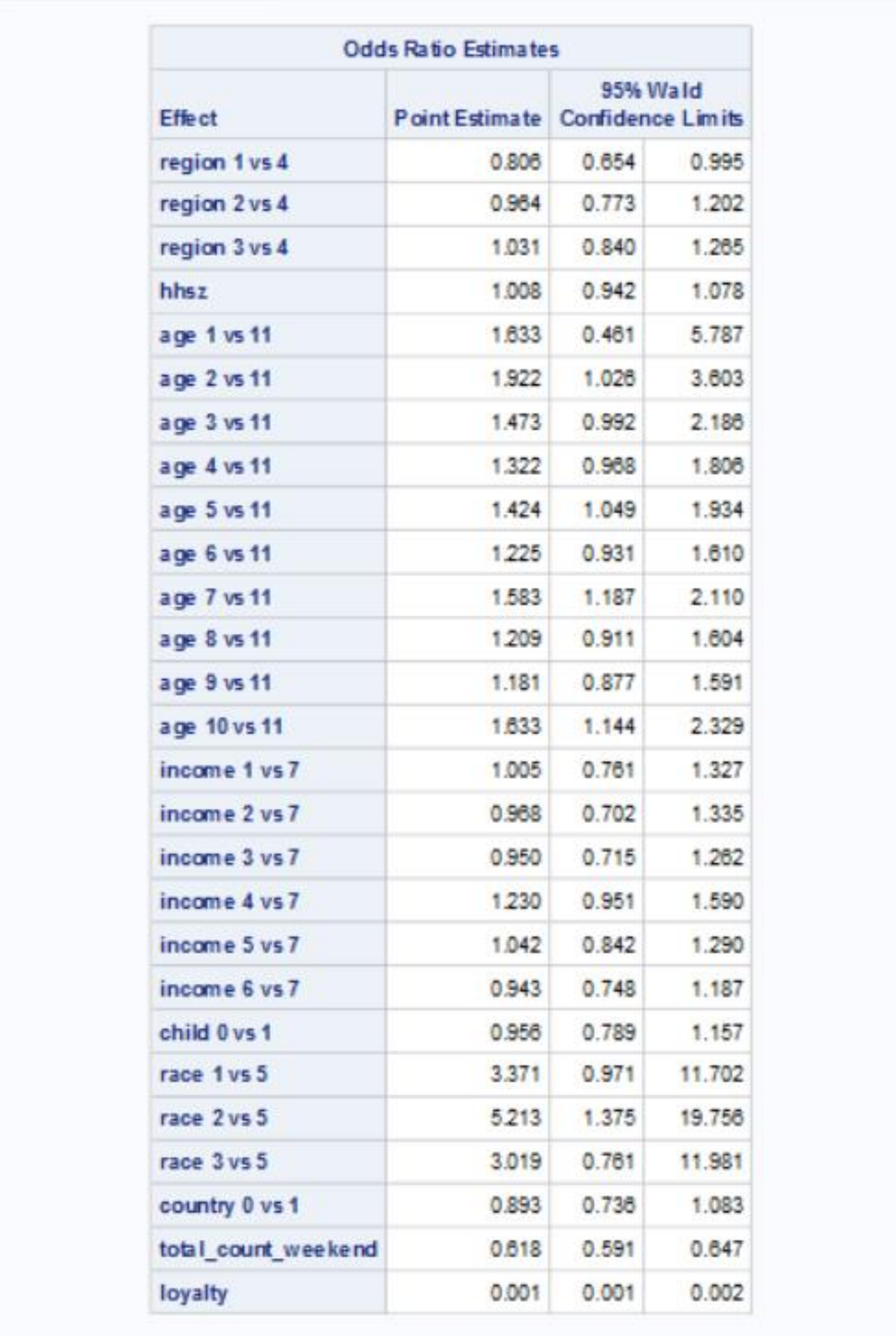
12. SAS Code:

```
/* Creating dummy variable for users who have made a purchase at
B and N */

data abasas.count_data;

set abasas.count_data;

if total_count>0 then purchase_at_bn=1;

else purchase_at_bn=0;

run;

/* Logistic Regression with loyalty and weekend */

proc logistic data= abasas.weekend;

class region age income child race country;

model total_count = region hhsz age income child race country
total_count_weekend loyalty / expb;

run;
```

SAS Results:



Intercept	86	1	20.8312	1.6041	168.6376	<.0001	1.1139E9
region	1	1	-0.1597	0.0620	6.6478	0.0099	0.852
region	2	1	0.0187	0.0666	0.0788	0.7789	1.019
region	3	1	0.0857	0.0595	2.0720	0.1500	1.089
hhsz		1	0.00775	0.0345	0.0503	0.8225	1.008
age	1	1	0.1566	0.5807	0.0727	0.7874	1.170
age	2	1	0.3197	0.2840	1.2673	0.2603	1.377
age	3	1	0.0533	0.1722	0.0957	0.7570	1.055
age	4	1	-0.0547	0.1300	0.1770	0.6740	0.947
age	5	1	0.0197	0.1262	0.0245	0.8757	1.020
age	6	1	-0.1311	0.1090	1.4476	0.2289	0.877
age	7	1	0.1252	0.1161	1.1643	0.2806	1.133
age	8	1	-0.1440	0.1137	1.6032	0.2054	0.866
age	9	1	-0.1673	0.1219	1.8857	0.1697	0.846
age	10	1	0.1564	0.1510	1.0716	0.3006	1.169
income	1	1	-0.0110	0.1060	0.0108	0.9172	0.989
income	2	1	-0.0483	0.1272	0.1441	0.7042	0.953
income	3	1	-0.0670	0.1101	0.3700	0.5430	0.935
income	4	1	0.1910	0.0976	3.8308	0.0503	1.210
income	5	1	0.0258	0.0763	0.1143	0.7353	1.026
income	6	1	-0.0748	0.0868	0.7435	0.3885	0.928
child	0	1	-0.0226	0.0488	0.2153	0.6426	0.978
race	1	1	0.2225	0.1881	1.3990	0.2369	1.249
race	2	1	0.6583	0.2555	6.6405	0.0100	1.932
race	3	1	0.1120	0.2851	0.1543	0.6945	1.118
country	0	1	-0.0565	0.0493	1.3161	0.2513	0.945
total_count_weekend		1	-0.4808	0.0233	424.7183	<.0001	0.618
loyalty		1	-6.5480	0.1325	2442.5133	<.0001	0.001



Odds Ratio Estimates			
		95% Wald	
Effect	Point Estimate	Confidence Limits	
region 1 vs 4	0.806	0.654	0.995
region 2 vs 4	0.964	0.773	1.202
region 3 vs 4	1.031	0.840	1.265
hhsz	1.008	0.942	1.078
age 1 vs 11	1.633	0.461	5.787
age 2 vs 11	1.922	1.026	3.603
age 3 vs 11	1.473	0.992	2.186
age 4 vs 11	1.322	0.968	1.806
age 5 vs 11	1.424	1.049	1.934
age 6 vs 11	1.225	0.931	1.610
age 7 vs 11	1.583	1.187	2.110
age 8 vs 11	1.209	0.911	1.604
age 9 vs 11	1.181	0.877	1.591
age 10 vs 11	1.633	1.144	2.329
income 1 vs 7	1.005	0.761	1.327
income 2 vs 7	0.968	0.702	1.335
income 3 vs 7	0.950	0.715	1.262
income 4 vs 7	1.230	0.951	1.590
income 5 vs 7	1.042	0.842	1.290
income 6 vs 7	0.943	0.748	1.187
child 0 vs 1	0.956	0.789	1.157
race 1 vs 5	3.371	0.971	11.702
race 2 vs 5	5.213	1.375	19.756
race 3 vs 5	3.019	0.761	11.981
country 0 vs 1	0.893	0.736	1.083
total_count_weekend	0.618	0.591	0.647
loyalty	0.001	0.001	0.002

Managerial Takeaways:

- In our SAS results we find the odds ratio estimates which represent how the odds of the event, someone making a purchase from Barnes and Noble, change with a 1 unit increase in that variable, all other things being equal.
    - For example, for the “child 0 vs 1” effect, the odds ratio of a person without a child making a purchase is 0.956 which is less than the odds ratio of a person with a child. Which suggests a person with a child is more likely to buy a book from Barnes and Noble.
        - This key takeaway would be beneficial for Barnes and Noble marketing and sales team to target customers who have kids.
    - Similarly, for the “country 0 vs 1” effect, the odds ratio of a person inside of the U.S. is 0.893 times the odds ratio of a person outside of the U.S. Which suggests a person in the U.S. is less likely to buy a book from Barnes and Noble.
        - This key takeaway would be beneficial for Barnes and Noble marketing and sales team to consider as it relates to attracting the appropriate customer.
- Moreover, we found that most of this customer demographic information is not useful as most variables were not significant per a 0.05 alpha level.

Part IV. Summary

13.

- **Key Managerial Takeaways**
    - **Part I-** In comparing LL values, we find that the NBD Regression model LL value of -8,246 is greater than the Poisson LL value of -17,682.5 which suggests that NBD regression fits the data better.
    - **Part II-** NBD Regression results excluding education and including the total_weekend_count and loyalty, variables we created, is an improved model with an LL value of -5,329.5.
    - **Part III-**Apart from two key findings discussed in the managerial takeaways of this section, we found that most of this customer demographic information is not useful as most variables were not significant per a 0.05 alpha level.
- **BA techniques**
    - In addition to strengthening our data analytics techniques, this project helped us practice identifying the type of distribution our data resembles and as a result the type of business analytics problem we may want to tackle. Thus, in considering the type of distribution our data resembles, we can add this activity as an additional step in the process of applying our human intuitions as well as statistical knowledge when looking at the data to solve a problem.
- **SAS skills**
    - We appreciate the hands-on opportunity provided to complete the project using SAS code. As most of our team is composed of beginner SAS coders, it was extremely helpful to have SAS code examples included in the lecture materials as a guide, enabling us to spend most of our time understanding the components of the code and interpreting the consequent output.
- **New perspectives of BA**

- Furthermore, upon review our managerial takeaways, it was interesting to find that most of our variables were not significant. As novice data miners we may be prone to remove variables that are not significant. Yet, as a consequence of this action, we can cause earlier significant variables to now be insignificant and/or we then introduce a non-intended bias into the model.