CS 584-04: Machine Learning

Fall 2019: Assignment 4

Amitdeb Prasad Bhattacharya A20402789

Question 1

a) (5 points) List the aliased parameters that you found in your model.

	Deviance	Degrees of Freedom	Value
Intercept + group_size	987.576	6	4.34e-210
Intercept + group_size + homeowner	5867.781	2	0.0
Intercept + group_size + homeowner + married_couple	84.578	2	4.30e-19
Intercept + group_size + homeowner + married_couple + group_size * homeowner	254.078	6	5.51e-52
Intercept + group_size + homeowner + married_couple + group_size *	70.842	2	4.138e-16

b) (5 points) How many degrees of freedom do you have in your model?

```
Degrees of Freedom = 20
```

c) (10 points) After entering a model effect, calculate the Deviance test statistic, its degrees of freedom, and its significance value between the current model and the previous model. List your Deviance test results by the model effects in a table.

```
Column Numbers of the Non-redundant Columns:
(0, 1, 2, 3, 5, 7, 9, 11, 13, 17)
Optimization terminated successfully.
         Current function value: 0.889553
         Iterations 5
                          MNLogit Regression Results
==========================================================================
==========
Dep. Variable:                          A   No. Observations:
665249
Model:                          MNLogit   Df Residuals:
665229
Method:                           MLE   Df Model:
18
```

```
Date:                Sat, 02 Nov 2019   Pseudo R-squ.:                0
.006101
   Time:                        02:13:18   Log-Likelihood:            -
   5.9177e+05
   converged:                       True   LL-Null:                   -
   5.9541e+05
   Covariance Type:            nonrobust   LLR p-value:
   0.000
==============================================================================
============================
                             A=1       coef    std err          z
   P>|z|      [0.025      0.975]
   ------------------------------------------------------------------------
   ------------------------------
   const                             0.4396      0.091      4.822
   0.000       0.261       0.618
   group_size_1                      1.0885      0.093     11.763
   0.000       0.907       1.270
   group_size_2                      0.9573      0.092     10.454
   0.000       0.778       1.137
   group_size_3                      0.3439      0.095      3.610
   0.000       0.157       0.531
   homeowner_0                       0.8002      0.259      3.093
   0.002       0.293       1.307
   married_couple_0                 -0.2157      0.017    -12.873
   0.000      -0.249      -0.183
   group_size_1 * homeowner_0       -1.5056      0.260     -5.793
   0.000      -2.015      -0.996
   group_size_2 * homeowner_0       -1.1646      0.259     -4.493
   0.000      -1.673      -0.657
   group_size_3 * homeowner_0       -0.6546      0.267     -2.450
   0.014      -1.178      -0.131
   homeowner_0 * married_couple_0    0.2125      0.026      8.224
   0.000       0.162       0.263
   ------------------------------------------------------------------------
   ------------------------------
                             A=2       coef    std err          z
   P>|z|      [0.025      0.975]
   ------------------------------------------------------------------------
   ------------------------------
   const                            -0.9255      0.134     -6.927
   0.000      -1.187      -0.664
   group_size_1                      0.8015      0.135      5.923
   0.000       0.536       1.067
   group_size_2                      0.7281      0.134      5.429
   0.000       0.465       0.991
   group_size_3                      0.5275      0.138      3.810
   0.000       0.256       0.799
   homeowner_0                       0.5423      0.361      1.504
   0.133      -0.164       1.249
   married_couple_0                 -0.1882      0.023     -8.327
   0.000      -0.232      -0.144
group_size_1 * homeowner_0        -0.9834      0.362     -2.716      0.
007      -1.693      -0.274
```

```
group_size_2 * homeowner_0        -0.7156      0.361     -1.981
0.048      -1.423      -0.008
group_size_3 * homeowner_0        -0.5987      0.372     -1.611
0.107      -1.327       0.130
homeowner_0 * married_couple_0     0.2124      0.035      6.065
0.000       0.144       0.281
==========================================================================
==============================
Model Parameter Estimates:
                                          0         1
const                              0.439563 -0.925506
group_size_1                       1.088485  0.801493
group_size_2                       0.957293  0.728103
group_size_3                       0.343931  0.527471
homeowner_0                        0.800157  0.542297
married_couple_0                  -0.215748 -0.188178
group_size_1 * homeowner_0        -1.505554 -0.983441
group_size_2 * homeowner_0        -1.164638 -0.715556
group_size_3 * homeowner_0        -0.654639 -0.598700
homeowner_0 * married_couple_0     0.212483  0.212433
Model Log-Likelihood Value = -591774.333631724
Number of Free Parameters = 20
Deviance Chi=Square Test
Chi-Square Statistic =  70.84227676969022
  Degrees of Freedom =  2
     Significance =  4.138043547449837e-16
```

d) (5 points) Calculate the Feature Importance Index as the negative base-10 logarithm of the significance value. List your indices by the model effects.

```
Feature Importance Index for (Intercept + group_size) = 209.36172341080683
Feature Importance Index for (Intercept + group_size + homeowner) = inf
Feature Importance Index for (Intercept + group_size + homeowner + married
_couple) = 18.365879862820417
Feature Importance Index for (Intercept + group_size + homeowner + married
_couple + group_size * homeowner) = 51.25868244189017
Feature Importance Index for (Intercept + group_size + homeowner + married
_couple + group_size * homeowner + homeowner * married_couple) = 15.383204
943269693
```

e) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for A = 0, 1, 2 based on the multinomial logistic model. List your answers in a table with proper labelling.

	group_size	homeowner	married_couple	0	1	2
0	1	0	0	0.259651	0.589175	0.151174
1	1	0	1	0.260092	0.592106	0.147802
2	1	1	0	0.183602	0.682030	0.134368
3	1	1	1	0.154023	0.709918	0.136059
4	2	0	0	0.221936	0.621105	0.156959
5	2	0	1	0.222321	0.624216	0.153463
6	2	1	0	0.202510	0.659773	0.137718
7	2	1	1	0.170552	0.689450	0.139999
8	3	0	0	0.239570	0.604616	0.155814
9	3	0	1	0.239992	0.607660	0.152348
10	3	1	0	0.301140	0.531297	0.167563
11	3	1	1	0.259017	0.567017	0.173966
12	4	0	0	0.194485	0.669686	0.135829
13	4	0	1	0.194692	0.672592	0.132716
14	4	1	0	0.387719	0.484974	0.127306
15	4	1	1	0.339172	0.526404	0.134424

f) (5 points) Based on your model, what values of group_size, homeowner, and married_couple will maximize the odds value Prob(A=1) / Prob(A = 0)? What is that maximum odd value?

g) (5 points) Based on your model, what is the odds ratio for group_size = 3 versus group_size = 1, and A = 2 versus A = 0? Mathematically, the odds ratio is (Prob(A=2)/Prob(A=0) | group_size = 3) / ((Prob(A=2)/Prob(A=0) | group_size = 1).

```
1.0249543364157785
```

h) (5 points) Based on your model, what is the odds ratio for homeowner = 1 versus homeowner = 0, and A = 0 versus A = 1? Mathematically, the odds ratio is (Prob(A=0)/Prob(A=1) | homeowner = 1) / ((Prob(A=0)/Prob(A=1) | homeowner = 0).

```
0.6232245044401726
```

Question 2

a) (5 points) Show in a table the frequency counts and the Class Probabilities of the target variable.

A	count	class probability
0	143691	0.215996
1	426067	0.640462
2	95491	0.143542

b) (5 points) Show the crosstabulation table of the target variable by the feature group_size. The table contains the frequency counts.

Frequency Table:

A	group_size 1	2	3	4
0	115460	25728	2282	221
1	329552	91065	5069	381
2	74293	19600	1505	93

Row Fraction Table:

A	group_size 1	2	3	4
0	0.803530	0.179051	0.015881	0.001538
1	0.773475	0.213734	0.011897	0.000894
2	0.778010	0.205255	0.015761	0.000974

c) (5 points) Show the crosstabulation table of the target variable by the feature homeowner. The table contains the frequency counts.

```
Frequency Table:
 homeowner       0       1
A
0           78659   65032
1          183130  242937
2           46734   48757

Row Fraction Table:
 homeowner         0         1
A
0          0.547418  0.452582
1          0.429815  0.570185
2          0.489407  0.510593
```

d) (5 points) Show the crosstabulation table of the target variable by the feature married_couple. The table contains the frequency counts.

```
Frequency Table:
 married_couple       0     1
A
0              117110  26581
1              333272  92795
2               75310  20181

Row Fraction Table:
 married_couple         0         1
A
0              0.815013  0.184987
1              0.782206  0.217794
2              0.788661  0.211339
```

e) (10 points) Calculate the Cramer's V statistics for the above three crosstabulations tables. Based on these Cramer's V statistics, which feature has the largest association with the target A?

```
Observed Count:
 A               0       1      2
group_size
1           115460  329552  74293
2            25728   91065  19600
3             2282    5069   1505
4              221     381     93
Column Total:
 group_size
1    519305
2    136393
3      8856
4       695
dtype: int64
Row Total:
 A
0    143691
1    426067
2     95491
dtype: int64
Overall Total:
 665249
Expected Count:
 [[1.12167707e+05 3.32595349e+05 7.45419441e+04]
 [2.94603172e+04 8.73545940e+04 1.95780888e+04]
 [1.91285894e+03 5.67193540e+03 1.27120566e+03]
 [1.50117091e+02 4.45121398e+02 9.97615104e+01]]


Observed Count:
 A              0       1      2
homeowner
0          78659  183130  46734
1          65032  242937  48757
Column Total:
 homeowner
0    308523
1    356726
dtype: int64
Row Total:
 A
0    143691
1    426067
2     95491
dtype: int64
Overall Total:
 665249
Expected Count:
 [[ 66639.67686235 197597.39442074  44285.92871692]
 [ 77051.32313765 228469.60557926  51205.07128308]]


Observed Count:
 A                    0       1      2
married_couple
0               117110  333272  75310
1                26581   92795  20181
Column Total:
 married_couple
0    525692
1    139557
dtype: int64
Row Total:
 A
0    143691
1    426067
2     95491
dtype: int64
Overall Total:
 665249
Expected Count:
 [[113547.27203198 336685.9827884   75458.74517962]
 [ 30143.72796802  89381.0172116   20032.25482038]]


                    Test Statistic DF  Significance Association    Measure
homeowner       Chi-square   6270.49  2             0     CramerV  0.0970864
married_couple  Chi-square   699.285  2  1.41953e-152     CramerV  0.0324216
group_size      Chi-square   977.276  6  7.34301e-208     CramerV   0.027102
```

homeowner

f) (5 points) Based on the assumptions of the Naïve Bayes model, express the joint probability Prob(A = a, group_size = g, homeowner = h, married_couple = m) as a product of the appropriate probabilities.

```
Probability of each class
[0.21599582 0.64046244 0.14354174]
Empirical probability of features given a class, P(x_i|y)
[[0.65592525 0.24424339 0.09983137]
 [0.6114865  0.2811299  0.1073836 ]
 [0.63197344 0.26028999 0.10773658]]
Number of samples encountered for each class during fitting
[143691. 426067.  95491.]
Number of samples encountered for each (class, feature) during fitting
[[174646.  65032.  26581.]
 [528413. 242937.  92795.]
 [118380.  48757.  20181.]]
```

g) (10 points) For each of the sixteen possible value combinations of the three features, calculate the predicted probabilities for A = 0, 1, 2 based on the Naïve Bayes model. List your answers in a table with proper labelling.

	group_size	homeowner	married_couple	p_a_0	p_a_1	p_a_2
0	1	0	0	0.227037	0.627593	0.145370
1	1	0	1	0.214391	0.637467	0.148142
2	1	1	0	0.205588	0.654128	0.140284
3	1	1	1	0.193842	0.663414	0.142744
4	2	0	0	0.238441	0.614462	0.147097
5	2	0	1	0.225342	0.624635	0.150024
6	2	1	0	0.216281	0.641528	0.142192
7	2	1	1	0.204079	0.651128	0.144794
8	3	0	0	0.250201	0.601084	0.148715
9	3	0	1	0.236653	0.611546	0.151801
10	3	1	0	0.227342	0.628652	0.144006
11	3	1	1	0.214684	0.638559	0.146756
12	4	0	0	0.262308	0.587475	0.150218
13	4	0	1	0.248318	0.598215	0.153467
14	4	1	0	0.238767	0.615513	0.145720
15	4	1	1	0.225656	0.625720	0.148624

h) (5 points) Based on your model, what values of group_size, homeowner, and married_couple will maximize the odds value Prob(A=1) / Prob(A = 0)? What is that maximum odd value?

```
   group_size homeowner married_couple  odd value(p_a_1/p_a_0)
0           1         0              0                2.764273
1           1         0              1                2.973389
2           1         1              0                3.181743
3           1         1              1                3.422441
4           2         0              0                2.576994
5           2         0              1                2.771943
6           2         1              0                2.966181
7           2         1              1                3.190572
8           3         0              0                2.402403
9           3         0              1                2.584145
10          3         1              0                2.765223
11          3         1              1                2.974412
12          4         0              0                2.239641
13          4         0              1                2.409070
14          4         1              0                2.577880
15          4         1              1                2.772896
group_size                          1
homeowner                           1
married_couple                      1
p_a_0                        0.193842
p_a_1                        0.663414
p_a_2                        0.142744
odd value(p_a_1/p_a_0)        3.42244
Name: 3, dtype: object
```

3.42244