CATEGORICAL DATA ANALYSIS

PROF. ZVI GILULA

TA- Nadav Har-Tuv

Authors: Yoel Graumann, Dylan Vialla, Dekel Karp

INFERENTIAL ANALYSIS

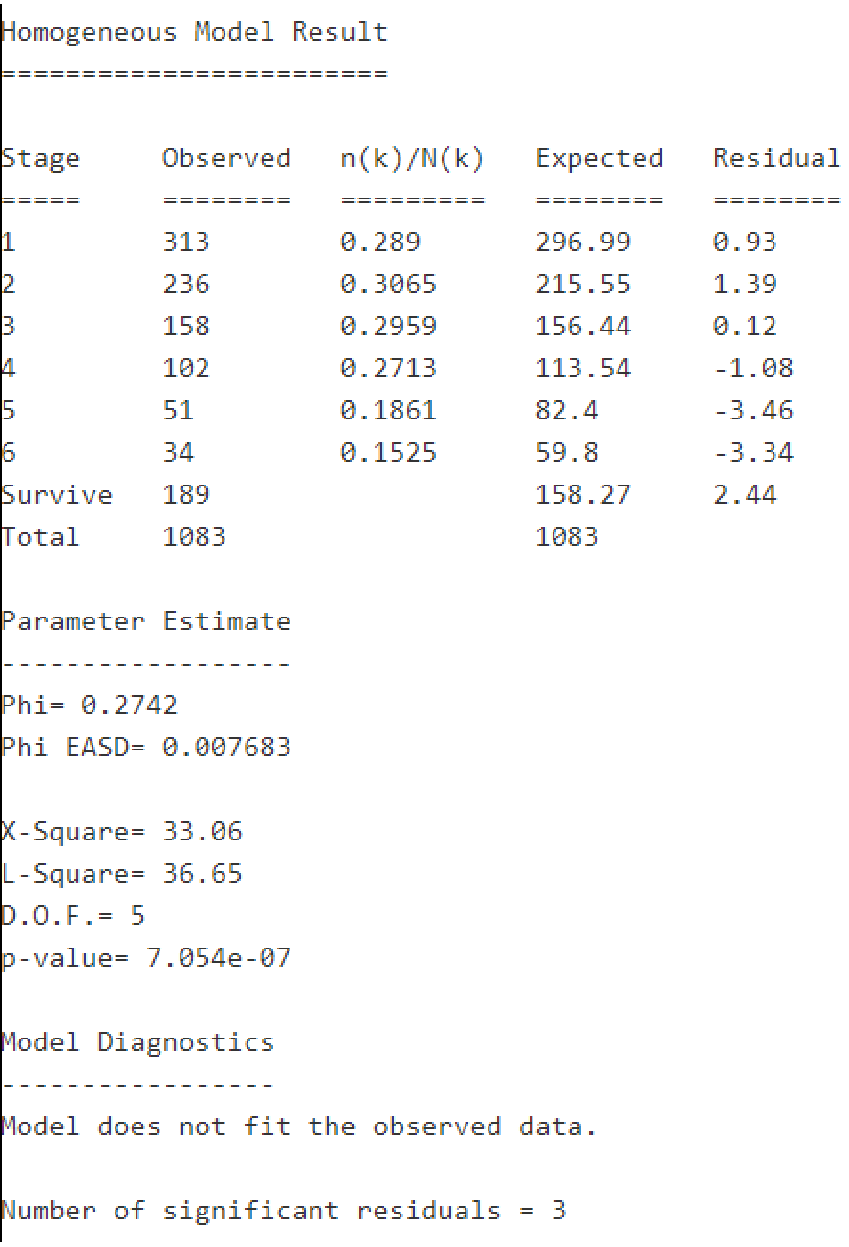
The Following categorical data analysis was done using the Insight Software which was created by Prof. Zvi Gilula using R.

Question 1

This question is obviously about Time Developing Phenomena and hence we will use Column 5 from the data set for our analysis. Column 5 denotes the hour which the buyer left the (1-6) workshop, "-1" is given to buyers who succeeded in all tests.

A: Preliminary Model fitting: investigation and early conclusions

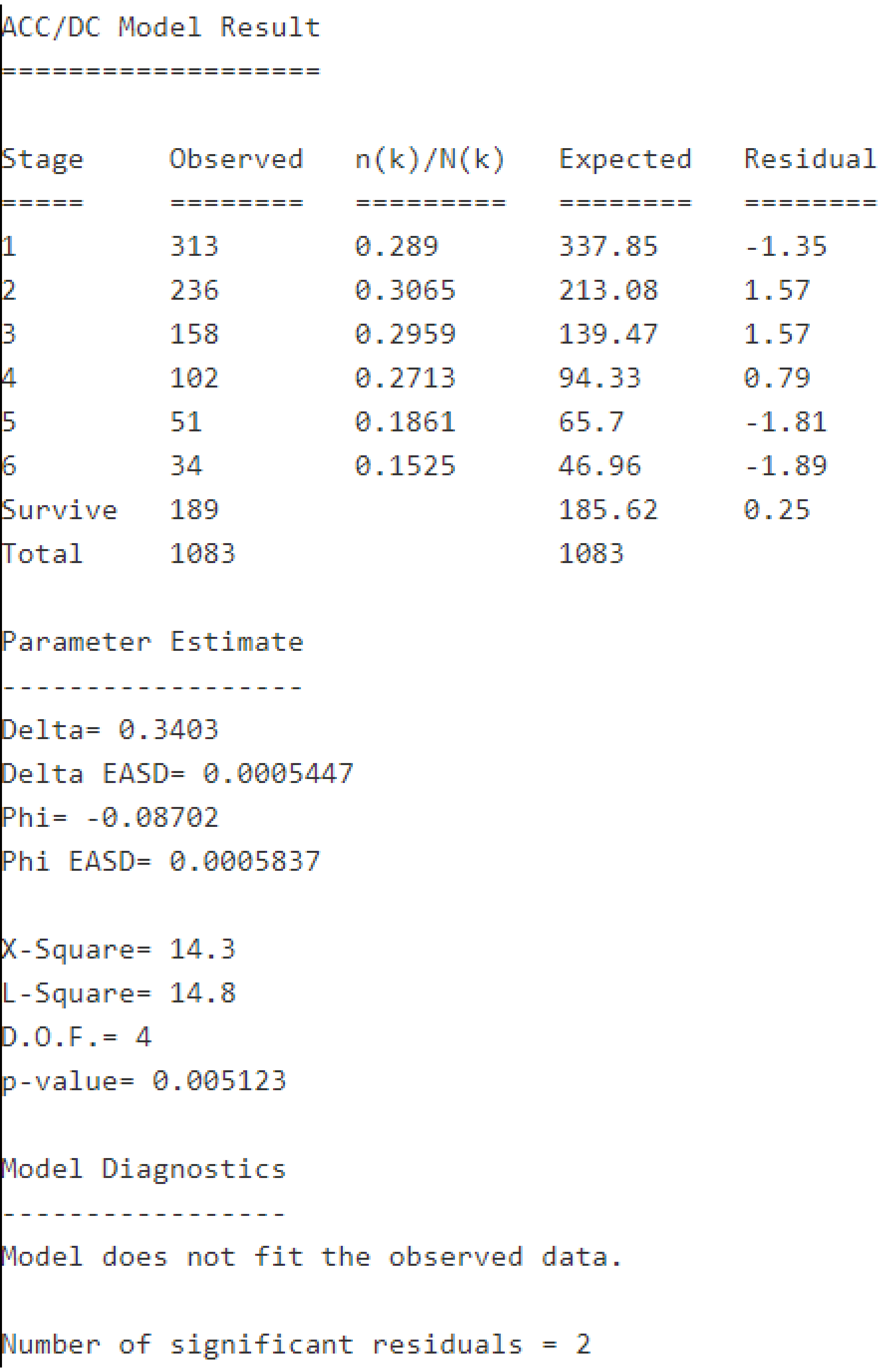
In insight, we choose the Analysis Method to be Survival and Model name to be Homogeneous, ACC/DC. This is the output that we got.



```
Homogeneous Model Result
========================

Stage     Observed   n(k)/N(k)   Expected   Residual
=====     ========   =========   ========   ========
1         313        0.289       296.99     0.93
2         236        0.3065      215.55     1.39
3         158        0.2959      156.44     0.12
4         102        0.2713      113.54     -1.08
5         51         0.1861      82.4       -3.46
6         34         0.1525      59.8       -3.34
Survive   189                    158.27     2.44
Total     1083                   1083

Parameter Estimate
------------------
Phi= 0.2742
Phi EASD= 0.007683

X-Square= 33.06
L-Square= 36.65
D.O.F.= 5
p-value= 7.054e-07

Model Diagnostics
-----------------
Model does not fit the observed data.

Number of significant residuals = 3
```

Let's talk about the obvious things first. We can see that phi has been estimated to be Phi=0.2742 and the asymptotic standard deviation has a value of Phi EASD=0.007683. We notice a super small P_value < 0.05 which means that the homogenous model does Not fit the observed data. We can also see that we have 3 significant residuals. As we have learned in class, significant residuals are standardized residuals which are greater than 1.64 or smaller than -1.64. negative outlying residuals (such as stages 6 and 5) are significant OVER estimates of the model. On the other hand, positive residuals (such as "survive" stage) are significant UNDER estimates of the model.

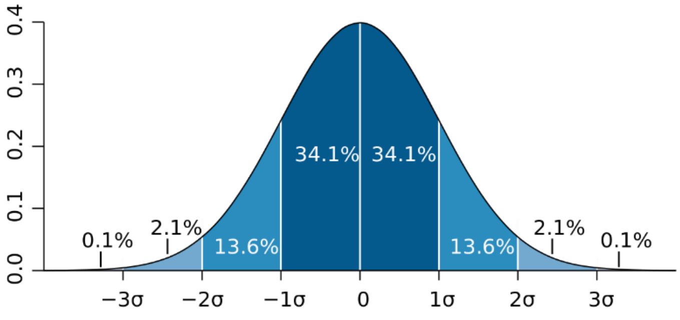
Notice that only 189 out of 1083 succeeded in all of the tests.



```
ACC/DC Model Result
===================

Stage      Observed    n(k)/N(k)    Expected    Residual
=====      ========    =========    ========    ========
1          313         0.289        337.85      -1.35
2          236         0.3065       213.08      1.57
3          158         0.2959       139.47      1.57
4          102         0.2713       94.33       0.79
5          51          0.1861       65.7        -1.81
6          34          0.1525       46.96       -1.89
Survive    189                      185.62      0.25
Total      1083                     1083

Parameter Estimate
------------------
Delta= 0.3403
Delta EASD= 0.0005447
Phi= -0.08702
Phi EASD= 0.0005837

X-Square= 14.3
L-Square= 14.8
D.O.F.= 4
p-value= 0.005123

Model Diagnostics
-----------------
Model does not fit the observed data.

Number of significant residuals = 2
```

Let's get the obvious things out of the way; We can See that Delta has been estimated to be DELTA=0.3404 and the Asymptotic standard deviation, Delta EASD=0.005447. Phi is very small and negative at -0.08702 and the asymptotic standard deviation Phi EASD=0.0005837. we can also see that we have two significant and negative residuals (stages 5 and 6). Finally, we can see that the p_value <0.05 and hence the model does Not fit the data.

We have a very big problem here, both models do not fit the observed data. But We can notice kind of extreme standardized residuals for the homogeneity model. We can see, using this graph from Wikipedia(bottom left), that 2+ standard deviations are kind of extreme. For example, the expected value of stage 5, E[x]=82.4 are 3.46 standard deviations under the mean. This leaves us a very small tail (remember that the tail converges to the x axis very fast). These extreme residuals in the homogeneity model led us to the following conclusion: There isn't a single mathematical function that "rules" over the whole phenomena. We need to do splining.

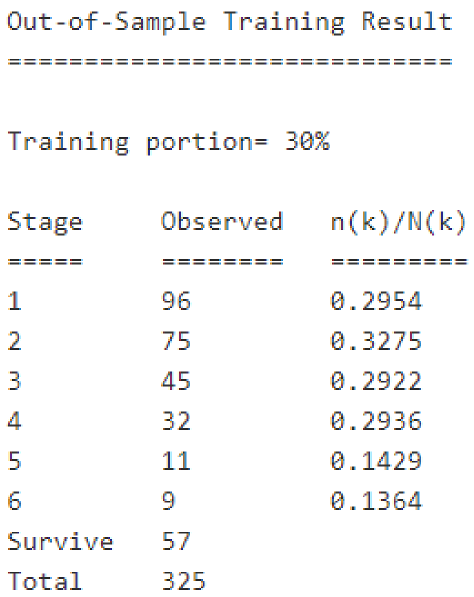


B: Splining investigation and recommendation.

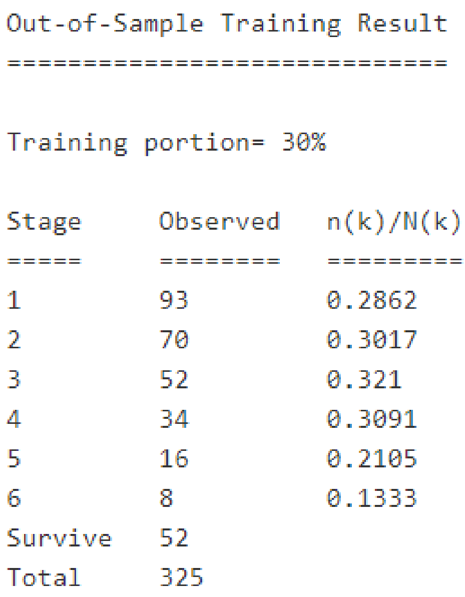
Since none of the models fit the data in view of some extra ordinary residuals from the poor fits, we are encouraged to find ways of splining the data. Let's look at the empirical estimators under the column n(k)/N(k) (They are not dependent on the models, I.E they are the same for both models). We can see that stages 1-4 are about the same in magnitude, while stages 5-6 are decreasing in magnitude with the stages. This leads us to believe that we might want to try to fit a homogeneity model for stages 1-4 and an ACC/DC model for stages 5-6. In other words, we recommend doing splining for stages 1-4;5-6. And since we do not want the current data to dictate the model / splining, the smart thing to do would be to split the data into train and test (also known as cross-validation) and then do splining. This will help us "defend" the uncertainty element of our model so we could report a useful model to our employers. We recommend a training proportion of 30 percent.

C: Splining According to recommendation, Analysis & interpretation of significant estimators of parameters.

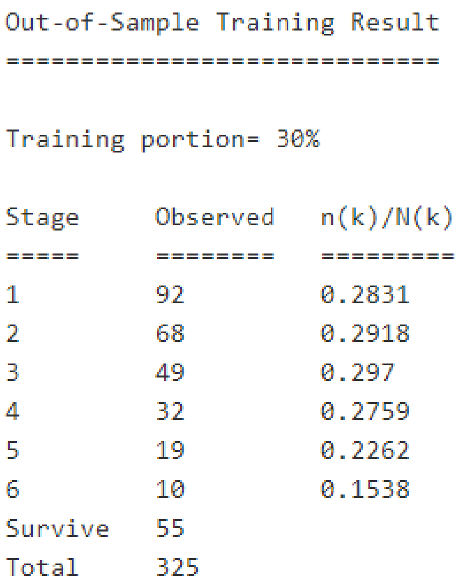
To do that, we will go to insight and under Analysis Method choose Survival. Then, Under Model Name we will choose Out-of-Sample Splining. We will choose a training proportion of 30. This is the output:



```
Out-of-Sample Training Result
=============================

Training portion= 30%

Stage       Observed    n(k)/N(k)
=====       ========    =========
1           96          0.2954
2           75          0.3275
3           45          0.2922
4           32          0.2936
5           11          0.1429
6           9           0.1364
Survive     57
Total       325
```

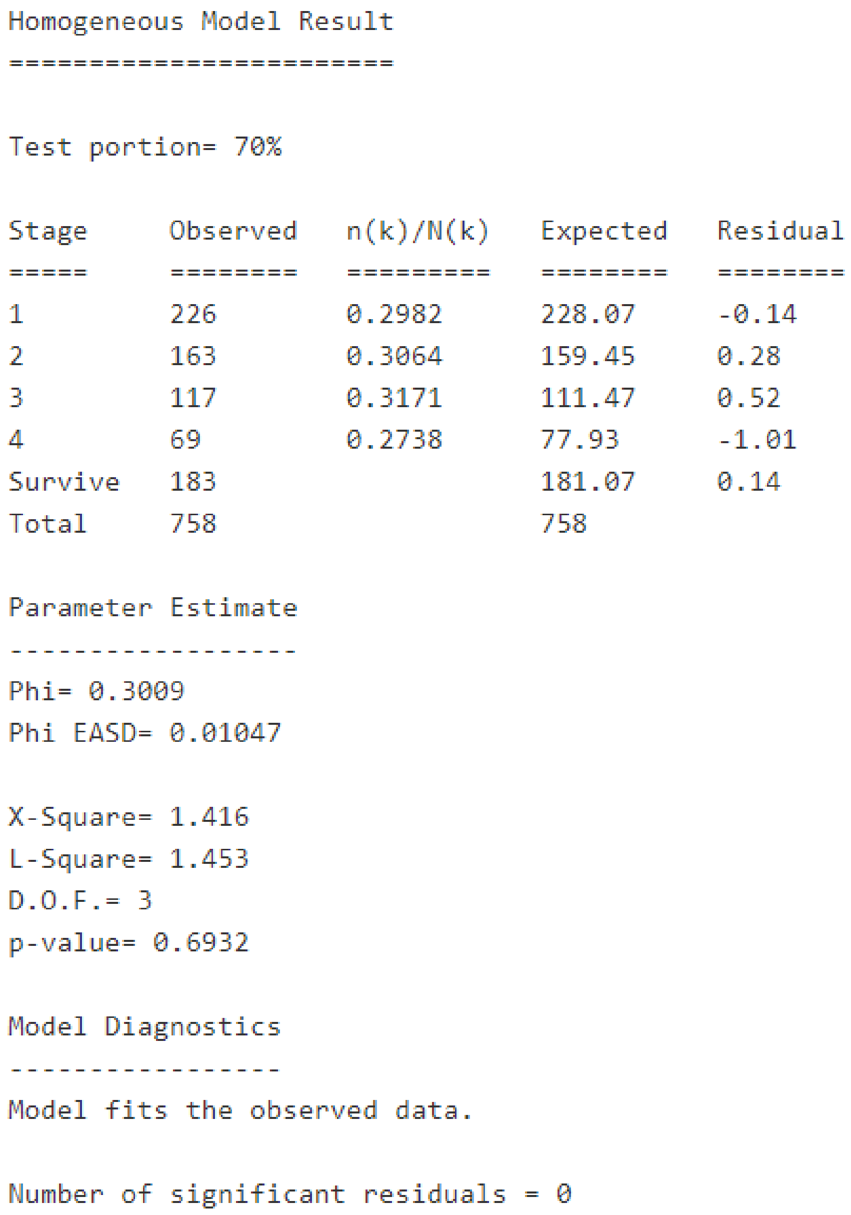


```
Out-of-Sample Training Result
=============================

Training portion= 30%

Stage       Observed    n(k)/N(k)
=====       ========    =========
1           93          0.2862
2           70          0.3017
3           52          0.321
4           34          0.3091
5           16          0.2105
6           8           0.1333
Survive     52
Total       325
```



```
Out-of-Sample Training Result
=============================

Training portion= 30%

Stage       Observed    n(k)/N(k)
=====       ========    =========
1           92          0.2831
2           68          0.2918
3           49          0.297
4           32          0.2759
5           19          0.2262
6           10          0.1538
Survive     55
Total       325
```

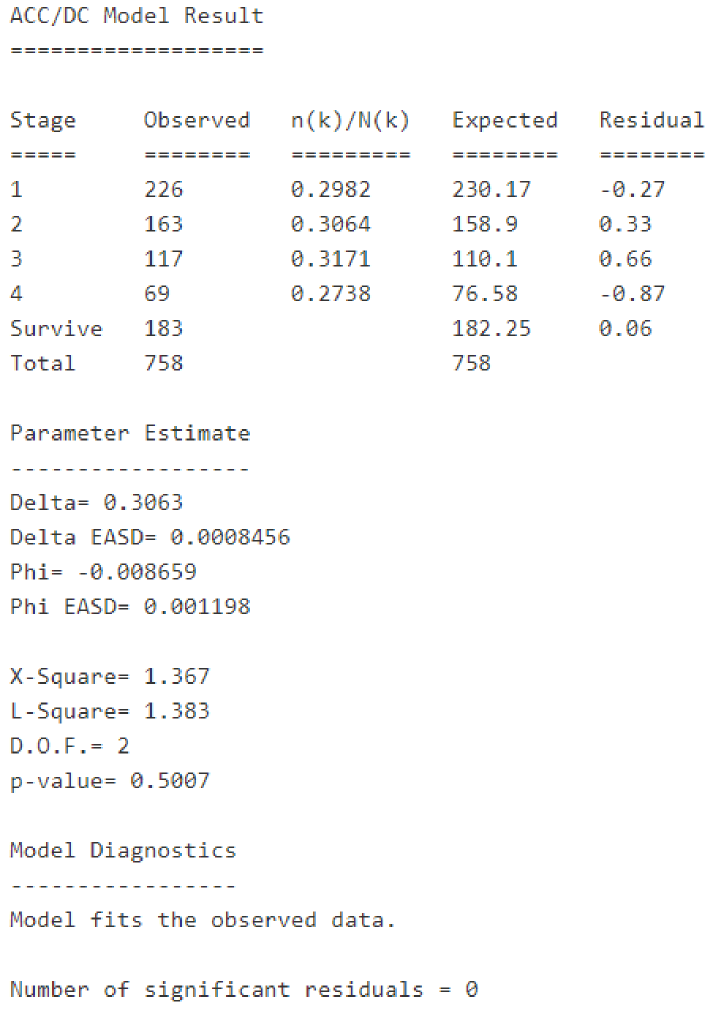
Since the code picks 30 percent of the data randomly, every time we press the "go training!" button, we decided to run the training 38 times, the pictures above are a fair representation of most of what we got from training. We can clearly see that the splining should be 1-4;5-6. We shall indeed input 1-4;5-6 in the "testing splining" method in insight. We got the following output:



```
Homogeneous Model Result
========================

Test portion= 70%

Stage      Observed   n(k)/N(k)   Expected   Residual
=====      ========   =========   ========   ========
1          226        0.2982      228.07     -0.14
2          163        0.3064      159.45     0.28
3          117        0.3171      111.47     0.52
4          69         0.2738      77.93      -1.01
Survive    183                    181.07     0.14
Total      758                    758

Parameter Estimate
------------------
Phi= 0.3009
Phi EASD= 0.01047

X-Square= 1.416
L-Square= 1.453
D.O.F.= 3
p-value= 0.6932

Model Diagnostics
-----------------
Model fits the observed data.

Number of significant residuals = 0
```

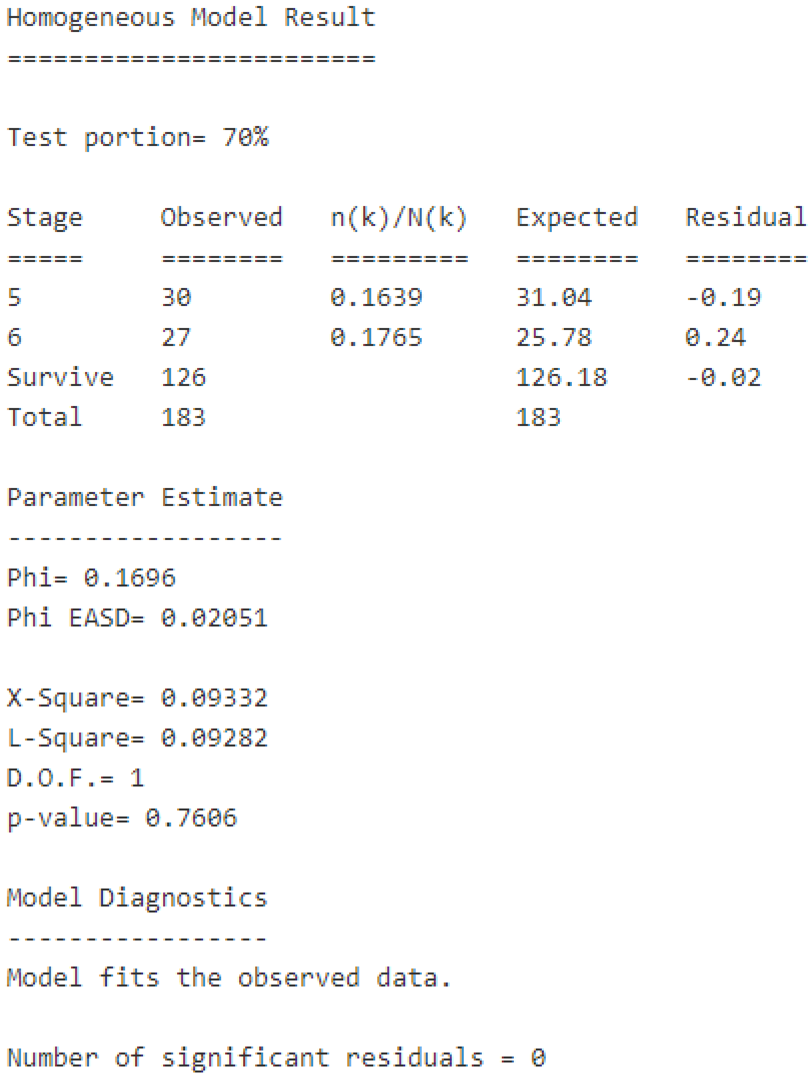
FOR STAGES 1-4

We can see that the pvalue is 0.06932, therefore the model fits the data. We expected it to be high considering the n(k)/N(k) values. We can see that the residuals are not as extreme as before. In fact, there are zero significant residuals. We can see phi=0.3009 and Phi EASD=0.01047. According to this model, the conditional probability of leaving the workshop is the same, and is about 0.3. (0.3 out of the people at current-stage k will leave the workshop).



```
ACC/DC Model Result
===================

Stage      Observed   n(k)/N(k)   Expected   Residual
=====      ========   =========   ========   ========
1          226        0.2982      230.17     -0.27
2          163        0.3064      158.9      0.33
3          117        0.3171      110.1      0.66
4          69         0.2738      76.58      -0.87
Survive    183                    182.25     0.06
Total      758                    758

Parameter Estimate
------------------
Delta= 0.3063
Delta EASD= 0.0008456
Phi= -0.008659
Phi EASD= 0.001198

X-Square= 1.367
L-Square= 1.383
D.O.F.= 2
p-value= 0.5007

Model Diagnostics
-----------------
Model fits the observed data.

Number of significant residuals = 0
```

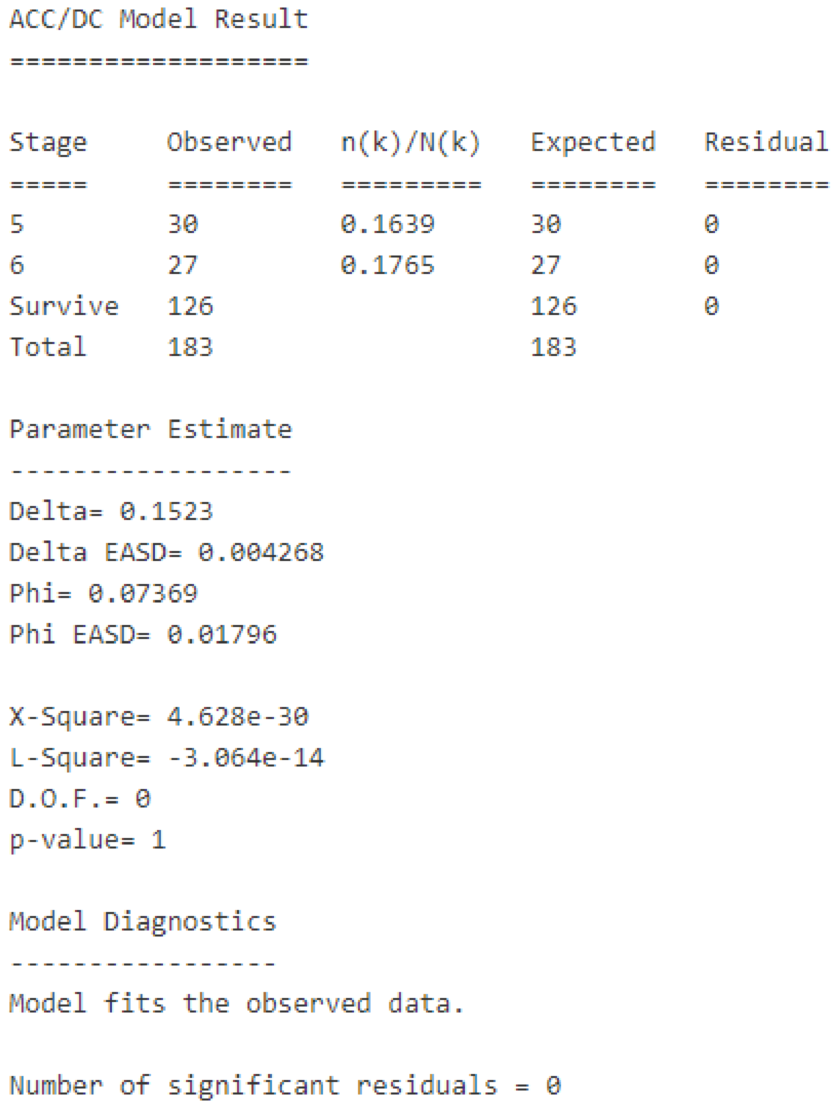
Here we can also see that the residuals are not as extreme, we have 0 significant residuals. The p-value is 0.5007, therefore, the model fits the observed data. We must remember that the Homogeneity model is a sub-model of the ACC/DC model. In other words, if we the the PHI of the ACC/DC model to be zero, we will be left with a constant number, DELTA=0.3063 which is exactly what the Homogeneity model claims. Additionally, we can see that the DELTA of the ACC/DC model is very close to the PHI of the Homogeneity model. We can see that the Phi in the ACC/DC model is very close to zero. This basically sets PHI of the ACC/DC model to be zero, giving us the Homogeneity model.



```
Homogeneous Model Result
========================

Test portion= 70%

Stage     Observed   n(k)/N(k)   Expected   Residual
=====     ========   =========   ========   ========
5         30         0.1639      31.04      -0.19
6         27         0.1765      25.78      0.24
Survive   126                    126.18     -0.02
Total     183                    183

Parameter Estimate
------------------
Phi= 0.1696
Phi EASD= 0.02051

X-Square= 0.09332
L-Square= 0.09282
D.O.F.= 1
p-value= 0.7606

Model Diagnostics
-----------------
Model fits the observed data.

Number of significant residuals = 0
```

<u>FOR STAGES 5-6</u>

We can see that the p-value is 0.7606 which means that the model fits the data. We can see that the constant conditional probability PHI is 0.1696, but the expected values are not an exact match according to the residuals.



```
ACC/DC Model Result
===================

Stage     Observed   n(k)/N(k)   Expected   Residual
=====     ========   =========   ========   ========
5         30         0.1639      30         0
6         27         0.1765      27         0
Survive   126                    126        0
Total     183                    183

Parameter Estimate
------------------
Delta= 0.1523
Delta EASD= 0.004268
Phi= 0.07369
Phi EASD= 0.01796

X-Square= 4.628e-30
L-Square= -3.064e-14
D.O.F.= 0
p-value= 1

Model Diagnostics
-----------------
Model fits the observed data.

Number of significant residuals = 0
```

According to the ACC/DC model we can see that the delta and the phi together come with an exact match for observed=expected, and hence the residual values of 0. The p-value of the model is 1 and that means the model fits the observed data.

We can compare the p-values(thanks to cross validation) for each model and each spline, and come to the following conclusion: For stages 1-4 the appropriate model would be the homogeneity model. And for stages 5-6 the appropriate model would be the ACC/DC model.

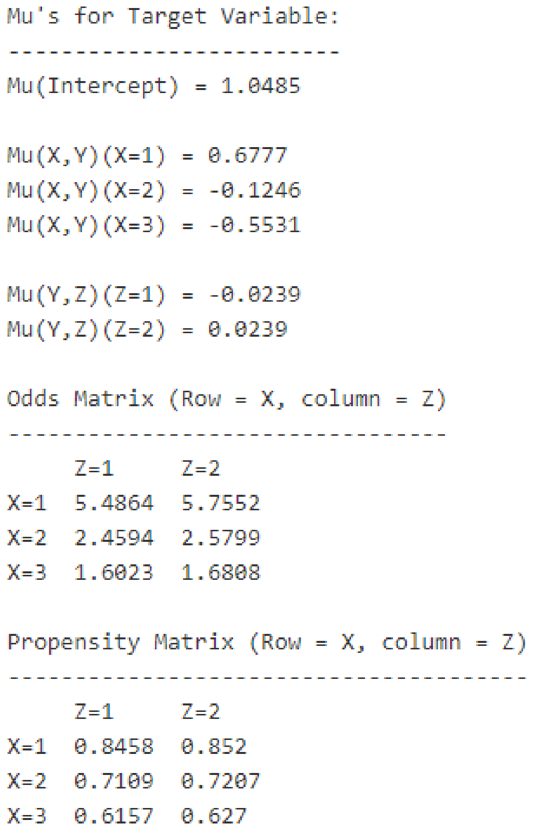
D: Characterization With 2 Best Explanatory Variables & Interpretation of Propensity Matrix.

For this question we will Do Survival Analysis, and then pick the model's name Explanatory Variable. We will examine both splines.

FOR STAGES 1-4

```
Sorted by descriptiveness: 6, 8, 4, 9, 3, 2, 10, 11, 7
```

According to insight, Education Level and Gender are the 2 most descriptive explanatory variables when it comes to stages 1-4. Now we can use the log-linear model with the file that insight created for us. We got a p-value of 0.79. The model fits the data matrix odds for target variable Y. Let Z be the Gender column and X be the Education column:



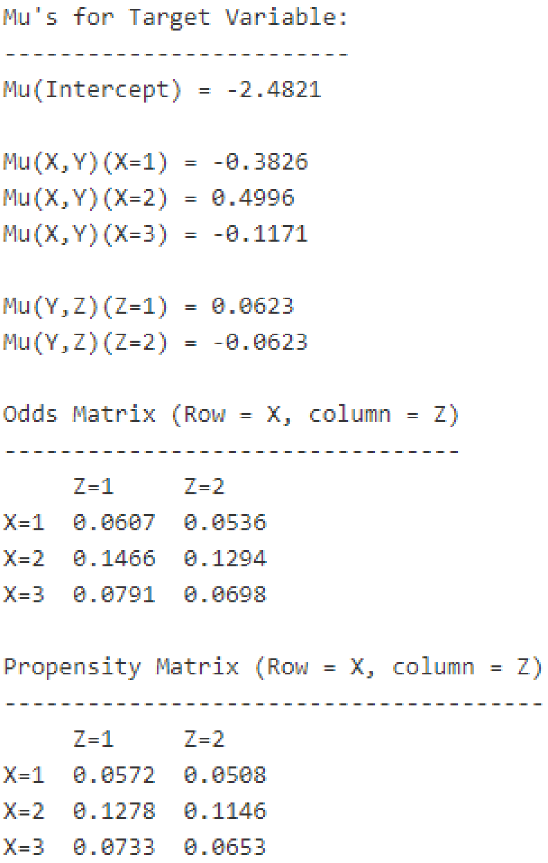
```
Mu's for Target Variable:
-------------------------
Mu(Intercept) = 1.0485

Mu(X,Y)(X=1) = 0.6777
Mu(X,Y)(X=2) = -0.1246
Mu(X,Y)(X=3) = -0.5531

Mu(Y,Z)(Z=1) = -0.0239
Mu(Y,Z)(Z=2) = 0.0239

Odds Matrix (Row = X, column = Z)
--------------------------------
     Z=1     Z=2
X=1  5.4864  5.7552
X=2  2.4594  2.5799
X=3  1.6023  1.6808

Propensity Matrix (Row = X, column = Z)
--------------------------------------
     Z=1     Z=2
X=1  0.8458  0.852
X=2  0.7109  0.7207
X=3  0.6157  0.627
```

According to the odds matrix, we can see that X=1,Z=2 or in words: males with no bagrut were 5.7 more likely to drop out of the workshop by the end of this stage time when compared to people with any other profile. According to the propensity matrix 85.2% of the males with no bagrut left the workshop at stages 1-4. On the other hand Females with Above-Highschool education had a conditional probability of 0.6157 to leave the workshop/fail the exams at stages 1-4. In summary, for stages 1-4 males with no bagrut were the worst.

FOR STAGES 5-6

```
Sorted by descriptiveness: 6, 7, 3, 2, 4, 11, 10, 8, 9
```

According to insight, Education level and TV size are the 2 most descriptive explanatory variables when it comes to stages 5-6. Let's use the log linear model again. Let X=education column, Z=TV size column. Using the log-linear model we get: p-value of 0.8732 so the model fits the observed data. Also, the following output:



```
Mu's for Target Variable:
-------------------------
Mu(Intercept) = -2.4821

Mu(X,Y)(X=1) = -0.3826
Mu(X,Y)(X=2) = 0.4996
Mu(X,Y)(X=3) = -0.1171

Mu(Y,Z)(Z=1) = 0.0623
Mu(Y,Z)(Z=2) = -0.0623

Odds Matrix (Row = X, column = Z)
--------------------------------
     Z=1     Z=2
X=1  0.0607  0.0536
X=2  0.1466  0.1294
X=3  0.0791  0.0698

Propensity Matrix (Row = X, column = Z)
--------------------------------------
     Z=1     Z=2
X=1  0.0572  0.0508
X=2  0.1278  0.1146
X=3  0.0733  0.0653
```

According to the propensity matrix, X=2,Z=1, or in words, people with High School Education and a TV bigger than 40 inches have a conditional probability of 0.1278 to leave the workshop at stages 5-6. 12.78% of the people with this profile will leave the workshop.

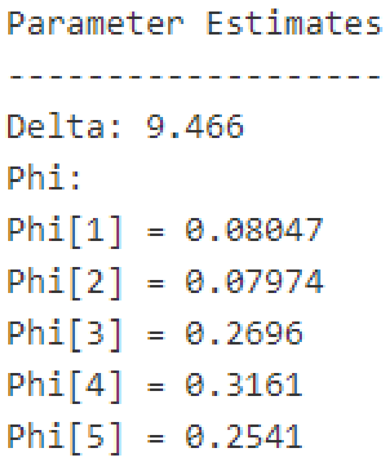
Question 2

Analysis of Brand Loyalty.

A: Comparing Q & M models and interpretation.

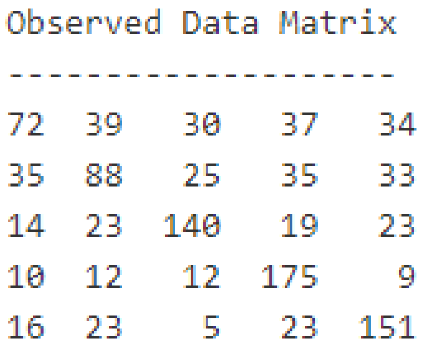
MODEL M

We can see that the model fits the data, with a p-value of 0.3217 and just one significant residual. We can see that the loyalty parameter delta is estimated to be 9.466. which is bigger than 2.5. According to what we have learned in lecture, this means that we can interpret a very strong loyalty to the floors, which is consistent with the Observed Data Matrix. We know that the bigger delta is, the more it shows conservatism, or concentration on the diagonal.



```
Parameter Estimates
--------------------
Delta: 9.466
Phi:
Phi[1] = 0.08047
Phi[2] = 0.07974
Phi[3] = 0.2696
Phi[4] = 0.3161
Phi[5] = 0.2541
```

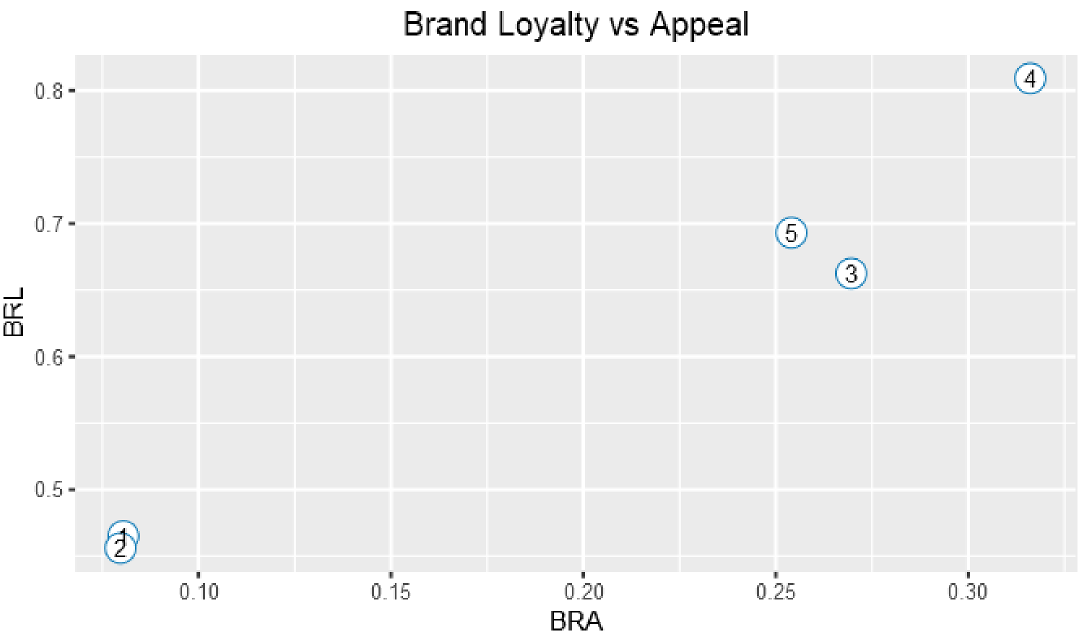
The Phis have a "relative" meaning. Floor 4 has the biggest phi=0.3161. floor 2 has the smallest phi=0.07974



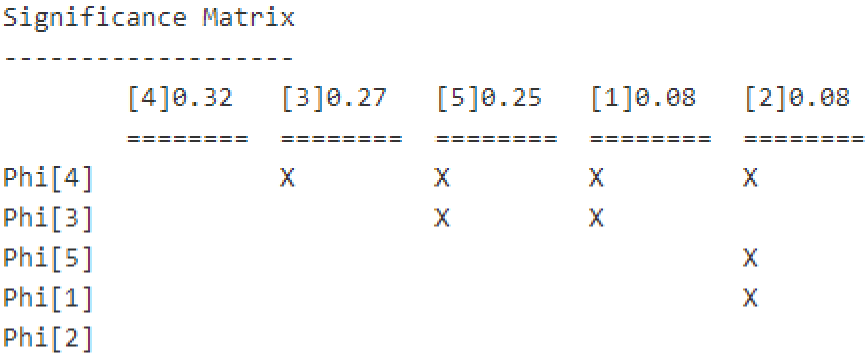
```
Observed Data Matrix
--------------------
72  39   30   37   34
35  88   25   35   33
14  23  140   19   23
10  12   12  175    9
16  23    5   23  151
```

Let's move on to Brand Loyalty vs. Appeal graph: In our case, the brand is the floor of the mall. The graph shows the phis on the x-axis and brand relative loyalty. BRL measures the proportion of going back to the same floor relative the penetration of each visit to the floor.

We can see that floor 4 is superior to all other floors, with the highest BRL and Phi values. (From the people who visited floor 4 at the start, a lot of the people stayed loyal to the floor). We can also see that floor two is in the worst position relative to all other floors.



But can this information be generalized to the whole population, and not only to the data? We must analyze the significance matrix.



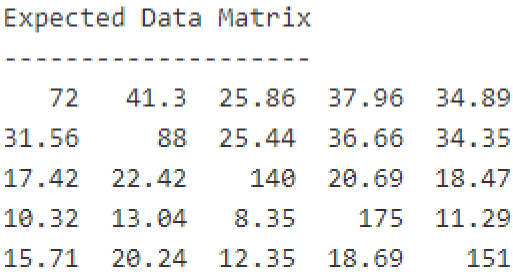
```
Significance Matrix
-------------------
        [4]0.32   [3]0.27   [5]0.25   [1]0.08   [2]0.08
        ========  ========  ========  ========  ========
Phi[4]            X         X         X         X
Phi[3]                      X         X
Phi[5]                                          X
Phi[1]                                          X
Phi[2]
```

Let's look at the cells above the diagonal. Cells that have an X means that the outcome is significant. We can see that 4 is indeed significantly different from all the other floors. In other words, floor 4 is indeed more attractive than the other floors when it comes to the general population. There are hints for transitive relation problem, which we will talk about in B. In summary, according to the M model, floor 4 is superior to all other floors. To gain better understanding, let's study the pure relative appeal of the floors via model Q

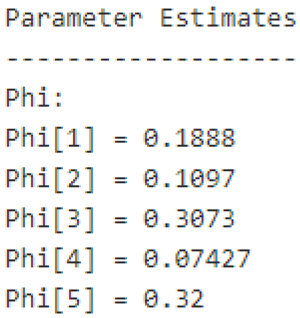
<u>MODEL Q</u>

This model will only look at the "switching probabilities". Mathematically, this model applies only to the off-diagonal elements. More clearly, it estimates the probabilities of switching from floor to floor. The observed values of the diagonal are equal to the expected of this model, they are seen as constant.

We can see that the Q-Model fits the data, as the p-value is 0.1445. We can see that the expected data on the diagonal is the same as the observed data, according to the Q Model. This gives this model less Degrees of freedom (M model has D.O.F=12 and Q Model has D.O.F=8)



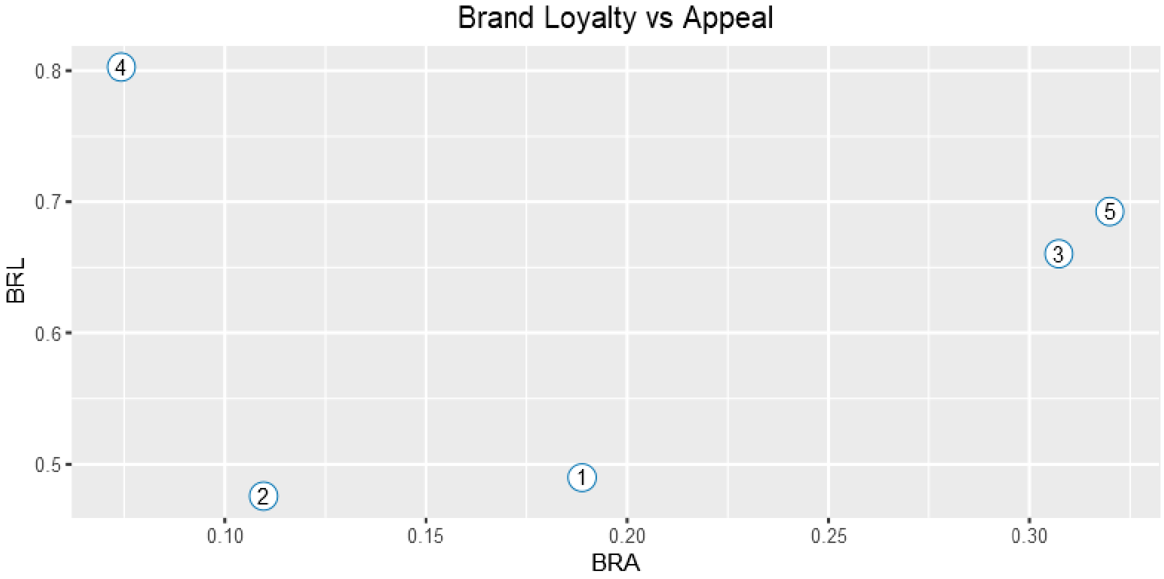
```
Expected Data Matrix
--------------------
   72   41.3  25.86  37.96  34.89
31.56     88  25.44  36.66  34.35
17.42  22.42    140  20.69  18.47
10.32  13.04   8.35    175  11.29
15.71  20.24  12.35  18.69    151
```



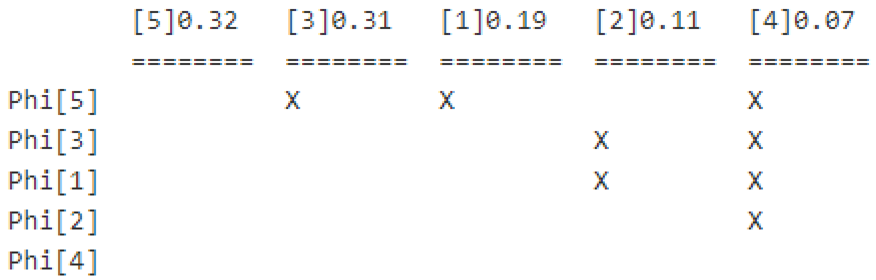
```
Parameter Estimates
-------------------
Phi:
Phi[1] = 0.1888
Phi[2] = 0.1097
Phi[3] = 0.3073
Phi[4] = 0.07427
Phi[5] = 0.32
```

We can see that we don’t have a delta parameter here, because it is used to model the diagonal, which is assumed to be constant under Q-model's assumptions. We can see that this time the phis are not in the same magnitudes as in M model. We can see that the maxim this time is phi[5], and phi[4] is the smallest. This difference between models is because model Q is learning about the population which has not evolved loyalty (off diagonal).

In words, Floor number 4 is the most attractive to people are already loyal to that floor. BUT, when it comes to people who have not developed any loyalty yet, floor 4 is very unattractive. We recommend an investigation to get NEW people. On the other hand, we can see that floor 3 is attractive to their already loyal people and also attractive (in fact the most attractive) to people who have not developed any loyalty yet. We can conclude from the Q and M plots that floor 2 is really worst off. We can also see that floor 4 has a high BRL relative to the BRA, according to this model. As we explained earlier, this is because floor 4 has a high loyalty when it comes to people who already developed it, and low appeal when it comes to people who have not developed loyalty yet.



significance matrix:



```
        [5]0.32   [3]0.31   [1]0.19   [2]0.11   [4]0.07
        ========  ========  ========  ========  ========
Phi[5]            X         X                   X
Phi[3]                                X         X
Phi[1]                                X         X
Phi[2]                                          X
Phi[4]
```

We can see that this time floor 5 is significantly different from the other floors. Which hints that it is very attractive to people who haven't developed loyalty. There are hints for transitive relation problem, which we will talk about in B.

<u>In summary:</u>

Model M has a reasonable fit.

Floor 4 which under model M has the highest appeal, has under model Q the lowest appeal.

Floor 2 which under model M has the lowest appeal, under model Q is just in the middle when it comes to appeal.

Floor 5 which has third highest appeal under model M, has the highest appeal under model Q.

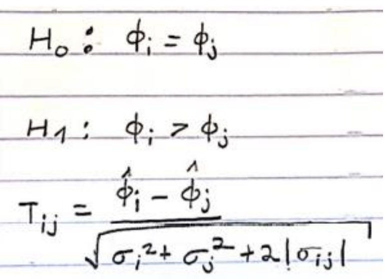
It seems like model 4 has successfully created its dedicated fans, but is performing very badly when it comes to attracting newcomers.

Floor 2 might be gaining slowly but surely some popularity among newcomers.

Floor 5 is doing decently when it comes to people already loyal, and is thriving in finding new customers.

B: Identification of Transitive Relation Problems.

When looking at the significance matrix for model M, we can see that there is no significant difference between floor 3 and floor 2. Despite the fact that floors 2 and 1 have the about the same phi values, and floor 3 is significantly different from floor 1. To understand why that is we must explain how any two phis are being compared. We have to compare them <u>Statistically</u>, that is by the means of hypothesis testing. Given by this process:



$$H_0 : \phi_i = \phi_j$$

$$H_1 : \phi_i > \phi_j$$

$$T_{ij} = \frac{\hat{\phi}_i - \hat{\phi}_j}{\sqrt{\sigma_i^2 + \sigma_j^2 + 2|\sigma_{ij}|}}$$

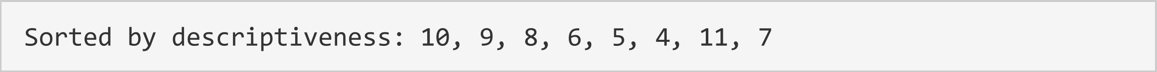
Where T_ij is the test statistic. The main point here is that the comparison is not only dependent on the difference between the pure values of phis, the difference is also given by what's in the denominator of the test statistic T_ij. The asymptotic variance plays a big role here. More specifically, the bigger the covariance is in absolute value, the smaller the quotient. And if the quotient is not bigger than the Critical value, it means that the difference is not significant and hence there will not be an X in the cell intersecting two phis in the significance matrix. We know that high absolute covariance hints at a serious competition between the phis. Therefore, lack of transitive relation, or a lack of x in a cell intersection, means that there is serious competition between two floors. To easily find lack of transitive relation, we must look at each row and see if there is a cell that is empty, but cells before it, to the left, have an "x".

According to model M significance matrix, there is a strong (more extreme) competition between floor 3 and 2.

According to model Q significance matrix, there is a strong (more extreme) competition between floor 2 and 5.

C: Characterization of loyalists to floors 1 & 2 with 2 best variables.

When we created the target variable to characterize the floors 1,2 loyalists. We created target variables according to people who purchased at floor 1,2 at the start and who purchased at floor 1,2 at the end (we used "with 1,2" for both 1st and 2nd purchase in insight). We got the following variables sorted by descriptiveness.

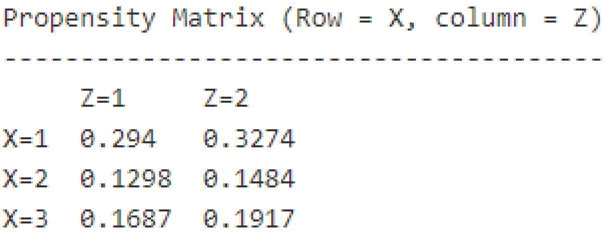


```
Sorted by descriptiveness: 10, 9, 8, 6, 5, 4, 11, 7
```

In words, Age Group & "is there a partner?" are the two best explanatory variables when it comes to loyalists to floors 1,2.

Log linear model

Let Y=loyalty to floors 1,2 (column 12) , X=age group(column 10), Z=is there a partner (column 9). We get a p-value of 0.7541, the model fits the data. We have 0 significant residuals. We have the following propensity matrix:



```
Propensity Matrix (Row = X, column = Z)
----------------------------------------
     Z=1     Z=2
X=1  0.294   0.3274
X=2  0.1298  0.1484
X=3  0.1687  0.1917
```

We can see that the maximal proportion in the matrix is given when X=1 and Z=2. In words, the most loyal(out of the people who are already loyal to floors 1,2) people who purchased at floors 1,2 at the start bought at floors 1,2 at the end are people who are ages 1-30, and do not have a partner. Floors 1,2 are attractive to young bachelors and bachelorettes. The lowest proportion in the matrix is when X=2,Z=1. In words, from the already loyal people, the least loyal of them are the people who are aged 31-50 and have a partner.

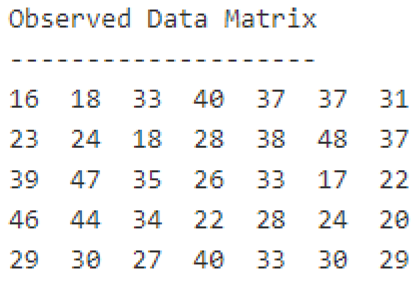
Question 3

Ranking of brands.

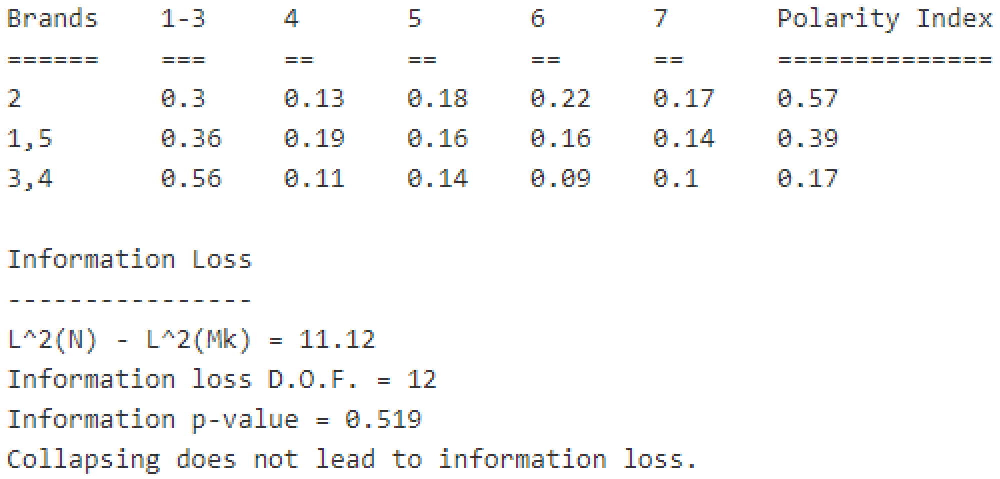
A: Ranking of floors at the start of the time-line by satisfaction.

Let's start with the Exploratory model first.

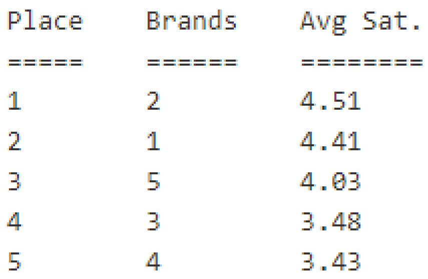
We get a very small p-value of 2.467e-08 which means that there is no homogeneity. and the following observed data matrix to the right.



```
Observed Data Matrix
--------------------
16   18   33   40   37   37   31
23   24   18   28   38   48   37
39   47   35   26   33   17   22
46   44   34   22   28   24   20
29   30   27   40   33   30   29
```



```
Brands    1-3     4       5       6       7       Polarity Index
======    ===     ==      ==      ==      ==      ==============
2         0.3     0.13    0.18    0.22    0.17    0.57
1,5       0.36    0.19    0.16    0.16    0.14    0.39
3,4       0.56    0.11    0.14    0.09    0.1     0.17

Information Loss
----------------
L^2(N) - L^2(Mk) = 11.12
Information loss D.O.F. = 12
Information p-value = 0.519
Collapsing does not lead to information loss.
```



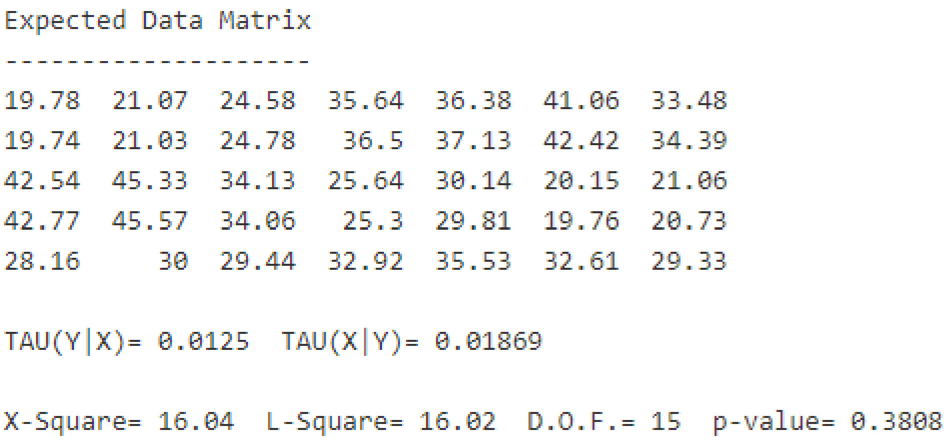
```
Place    Brands    Avg Sat.
=====    ======    ========
1        2         4.51
2        1         4.41
3        5         4.03
4        3         3.48
5        4         3.43
```

We can see the brands ordered stochastically, with the collapsed rankings of 1,2,3 into one column. Only after collapsing rankings 1-3 we got a stochastic ordering. Additionally, we can see that we did not lose any information by using the Yakir-Gilula technique. Another thing we have to notice is the polarity index, we can see that for all floors, the PI is smaller than 1. This means that the dis-satisfaction is greater than the satisfaction for all floors. In other words, floor 2, despite being ranked 1st, is still unsatisfactory to the customers.

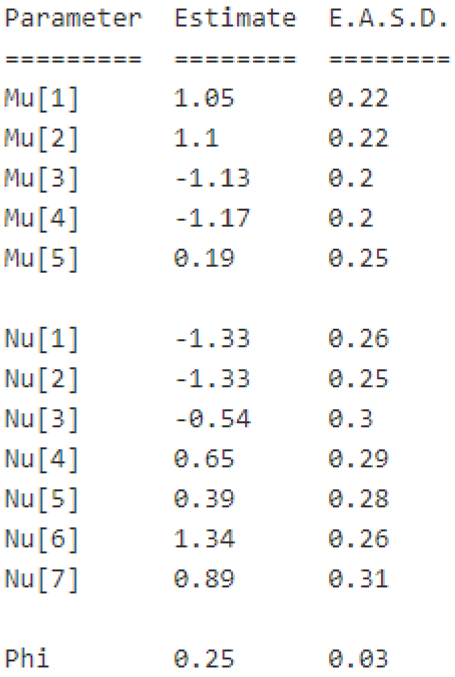
We get the following ranking by averages. We can see that floor 2 is in first place, floor 1 is in second place. And floor 4 is in last place. Interestingly enough, we can see that the brands ranking here are ordered exactly like their stochastic order.

Confirmatory approach.

First of all, we have a p-value of 0.3808. The p-value shows us that this model fits the data and therefore this is preferred to the exploratory approach because it is a model-based approach.

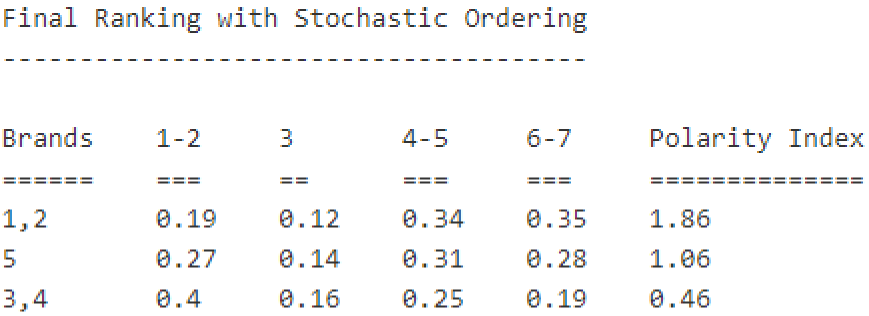


```
Expected Data Matrix
--------------------
19.78  21.07  24.58  35.64  36.38  41.06  33.48
19.74  21.03  24.78   36.5  37.13  42.42  34.39
42.54  45.33  34.13  25.64  30.14  20.15  21.06
42.77  45.57  34.06   25.3  29.81  19.76  20.73
28.16     30  29.44  32.92  35.53  32.61  29.33

TAU(Y|X)= 0.0125  TAU(X|Y)= 0.01869

X-Square= 16.04  L-Square= 16.02  D.O.F.= 15  p-value= 0.3808
```



```
Parameter  Estimate  E.A.S.D.
=========  ========  ========
Mu[1]      1.05      0.22
Mu[2]      1.1       0.22
Mu[3]      -1.13     0.2
Mu[4]      -1.17     0.2
Mu[5]      0.19      0.25

Nu[1]      -1.33     0.26
Nu[2]      -1.33     0.25
Nu[3]      -0.54     0.3
Nu[4]      0.65      0.29
Nu[5]      0.39      0.28
Nu[6]      1.34      0.26
Nu[7]      0.89      0.31

Phi        0.25      0.03
```

We can see the parameters, the estimated and their asymptotic standard deviation. By looking at the Nu's estimates we can see that they are not ALL monotonic, so the model will collapse the satisfaction levels (nu's) accordingly.



```
Final Ranking with Stochastic Ordering
--------------------------------------

Brands   1-2    3      4-5    6-7    Polarity Index
======   ===    ==     ===    ===    ==============
1,2      0.19   0.12   0.34   0.35   1.86
5        0.27   0.14   0.31   0.28   1.06
3,4      0.4    0.16   0.25   0.19   0.46
```

After the collapsing, we can notice that, in contrast to the exploratory approach, the confirmatory approach ranks floors 1 and 2 together in the 1st spot. Floors 3 and 4 are still ranked last. We can also notice that according to the confirmatory analysis, most of the mass is around the higher satisfaction levels. While on the exploratory analysis, most of the mass is on the lower satisfaction levels.

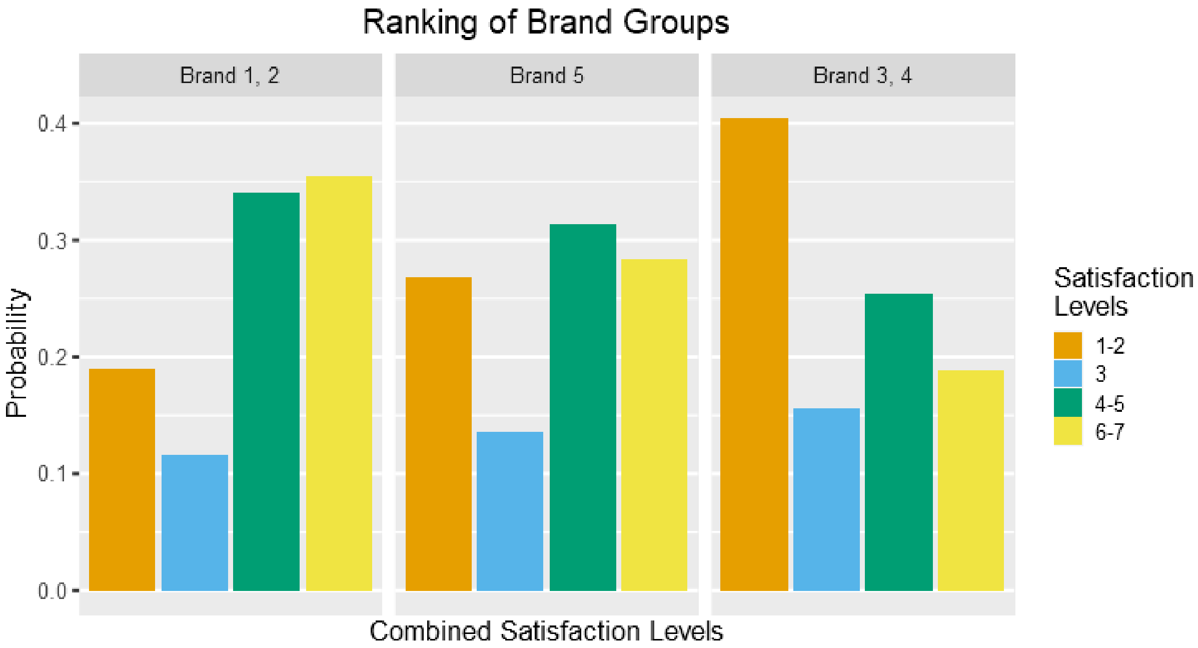
In summary

Since the exponential spacing model fits the data, we will prefer it over the results of the exploratory approach. Here we can see that we must combine the satisfactions (6,7) (4,5) and (1,2). Where in the exploratory we just had to combine 1-3. Since the original scale is not preserved in either model, we will let the p-value of the confirmatory approach lead the way. When looking at the floors, we can see that the main difference between the models is where to rank floor 1. The confirmatory approach ranks floor 1 to be first, along with floor 2. While the exploratory approach ranks floor 2, first and alone. And floor 1 second along with floor 5. Both models rank floors 3,4 last. One of the reasons why we choose the confirmatory model is because it is based on less uncertainty (we use the expected matrix instead of the observed).

B: Differences Between 1st group and last group.

We will compare according to the confirmatory model.

As we can see from the confirmatory model, the group of floors ranked 1st are (1,2) while the group of floors ranked last are (3,4). We can easily compare the densities using a plot:



We can see many differences between floors (1,2) and floors (3,4)

We can see that for floors 1,2 that the lower satisfaction levels both (1-2) and 3, have lower probabilities when compared to floors 3,4. We can see that in floors 1,2 the lower satisfactions in orange, for example, barely reaches 0.2 probability. In floors 3,4 the probability for satisfaction level of 1-2(orange) is more than double of that of floors 1,2. On the other hand, we can see that the colors green and yellow (which represent higher satisfaction) are more dominant in floors 1,2 than in floors 3,4. We can see that a satisfaction of 6-7 (in yellow) is about 1.8 times more likely to happen for floors 1,2 when compared with the worst ranked floors (3,4).

C: Characterization of Customers who had lower satisfactions for floors 1,2.

For this question we will try to characterize the customers who gave a satisfaction level of 1,2,3,4 for the 1st ranked group of floors (1,2).

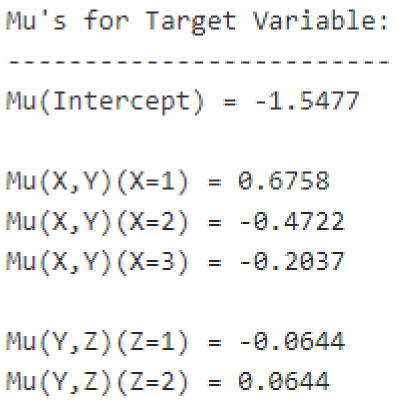
Note that we decided to keep out column 5 because our superiors told us that they are not interested in finding out when the people left the workshop, as they claim it is not useful for them. Our superiors at the company rather know more specific things such as education level / age / gender / political affiliation etc. This is feature engineering forced upon us by the leaders of the company. We got the following columns sorted by descriptiveness by using the Explanatory Variable model:

```
Sorted by descriptiveness: 11, 9, 8, 6, 10, 7
```

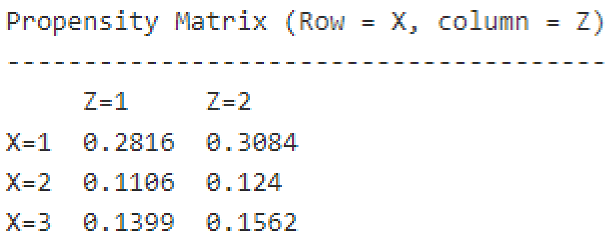
We can see that column 11 is ranked 1st when it comes to describing the individuals we are trying to characterize, and column 9 is ranked 2nd. Column 11 is political affiliation (1-left,2-center,3-right) and column 9 is having a partner (1-yes,2-no). Let's use 3-Dimensional Analysis.

Let X=column 11, Z=column 9, Y=column 12. We get the following result:

P-value of 0.6683 and therefore the model fits the data. Zero significant residuals. The following Mu's for the target variable and the propensity matrix:



```
Mu's for Target Variable:
-------------------------
Mu(Intercept) = -1.5477

Mu(X,Y)(X=1) = 0.6758
Mu(X,Y)(X=2) = -0.4722
Mu(X,Y)(X=3) = -0.2037

Mu(Y,Z)(Z=1) = -0.0644
Mu(Y,Z)(Z=2) = 0.0644
```



```
Propensity Matrix (Row = X, column = Z)
---------------------------------------
     Z=1     Z=2
X=1  0.2816  0.3084
X=2  0.1106  0.124
X=3  0.1399  0.1562
```

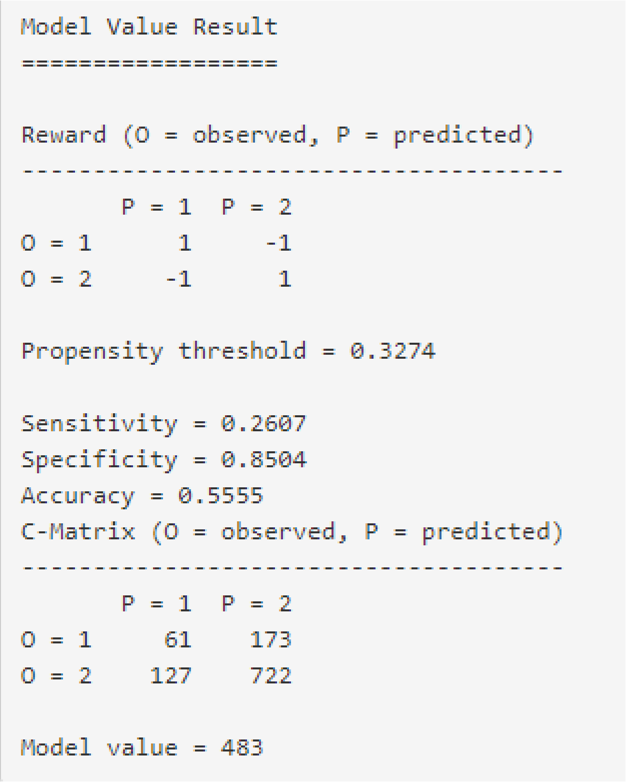
We can see that Mu(X,Y)(x=1)=0.6758, or in words, people who have a left-wing political affiliation, support the target variable. Same logic goes for people without a partner.

From the propensity matrix we can see maxim at {X=1,Z=2}. In other words, we can say that people of the left political spectrum who are bachelors or bachelorettes, were the most likely to rank floors 1,2 on the lower end of the spectrum of satisfaction (1-4). We recommend looking into ways to make these customers. We can see a minimum when {X=2,Z=1}. In words, customers who are on the Center of the political spectrum, with a partner, were the least likely to give floors (1,2) a satisfaction level of (1-4), out of the people who actually did that.

Question 4

A: Sensitivity, Specificity & Model Value.

According to the propensity matrix that we got in Question 2-C. We got a maximum propensity value of 0.3274 when {X=1, Z=2}. In words, people ages 1-30 who do not have a partner have the maximum propensity. We get the following output for this propensity:



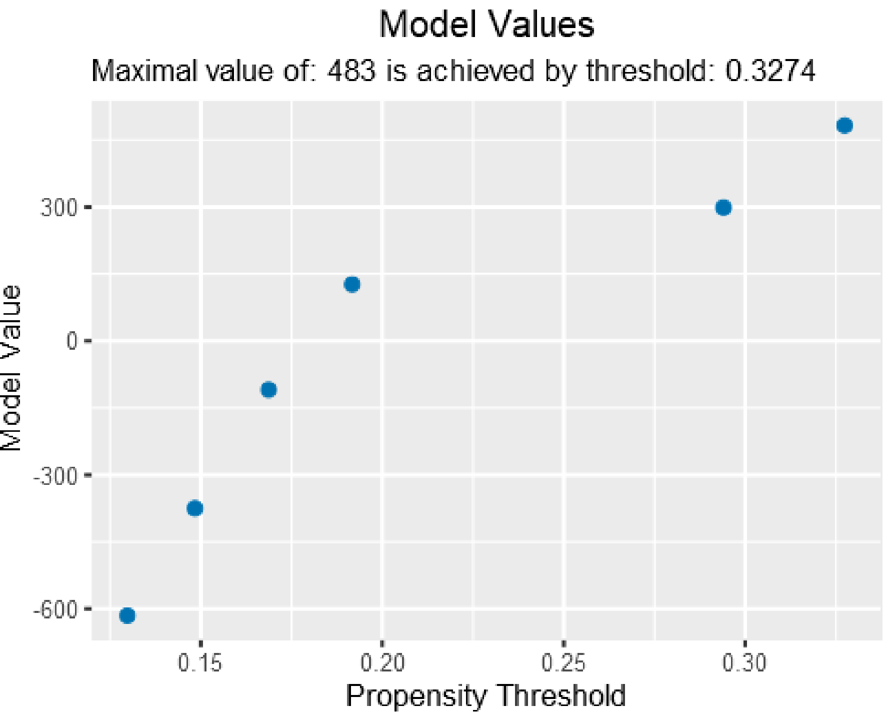
```
Model Value Result
==================

Reward (O = observed, P = predicted)
--------------------------------------
       P = 1  P = 2
O = 1      1     -1
O = 2     -1      1

Propensity threshold = 0.3274

Sensitivity = 0.2607
Specificity = 0.8504
Accuracy = 0.5555
C-Matrix (O = observed, P = predicted)
--------------------------------------
       P = 1  P = 2
O = 1     61    173
O = 2    127    722

Model value = 483
```

When we put a propensity threshold of the maximum propensity, We have a very low sensitivity of 61/(61+173). Which is pretty bad for finding the people who are floor 1 and 2 loyalists out of the observed. On the other hand our model is pretty good in finding the people who are NOT floor 1 and 2 loyalists out of the observed, with a specificity value of 722/(127+722).

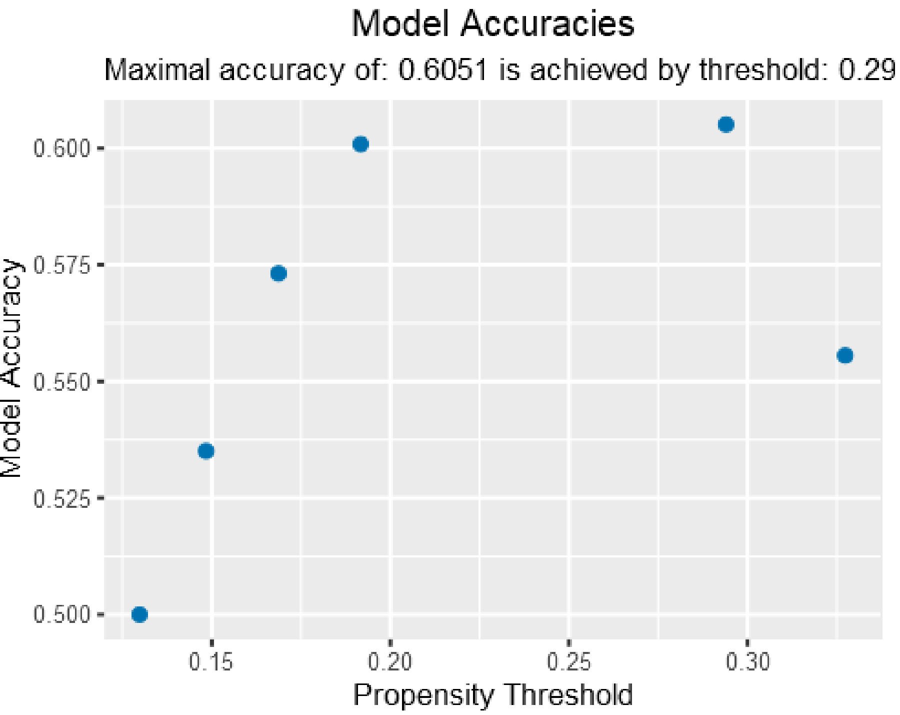
We also have an accuracy value of (sensitivity+specificity)/2 as defined in lecture.

Assuming a uniform reward-penalty matrix, we get a Model Value of 483.

B: Model Values-Model Accuracies tradeoff.



Thanks to Insight, we do not have to manually enter all the propensities. We can easily see that the maximal value of the model 483. Which is given when we use the maximum propensity, we got from Question 2-C. We clearly see that as the propensity increases, the model value also increases.



According to this plot, we can see that the model accuracy actually increases with the propensity threshold, up until a local maximum of 0.6051. The maximum accuracy is given when we use the propensity threshold of 0.29. After reaching the maximum accuracy, the accuracy decreases with the propensity.

There is clearly an accuracy-value threshold which our superiors will have to decide when analyzing this classification problem. In other words, the maximum accuracy and the maximum model value are not achieved with the same characterization of the customers, as we have to use different propensity values to reach both of these maxima. (Different propensities are given to different characterizations of individuals)