Data Mining 2015

Report Assignment 2
Graphical Models

Ferdinand van Walree - 3874389
Rizkiyanto - 4234634

1. Introduction

This report gives an analysis of an undirected model from data. The data set contains data from a medical study on adult patients admitted to an intensive care unit.

2. Analysis

With the function that we already created, we try to answer the following question.

Question a. How many different graphical models are there for this data set?

Given k nodes there are (k nCr 2) different edges. Every edge can either be included or excluded, which means we have 2^(k nCr 2) different undirected graphical models. The dataset contains 10 attributes so we have 10 nodes. The number of different undirected graphical models equals: 2^(10 ncr 2) = 3.518*10^13

Question b. How many cells does the table of counts for this data set have? How many parameters does the saturated model have?

The number of cells of the table of counts depends on the number of values each attribute can have. This is because each cell will correspond to one value of each attribute. The total number of cells must then be the product of the number of possible values of each attribute. The data set documentation tells us how many different values each attribute can hold. Therefor the number of cells of the table of counts must be:

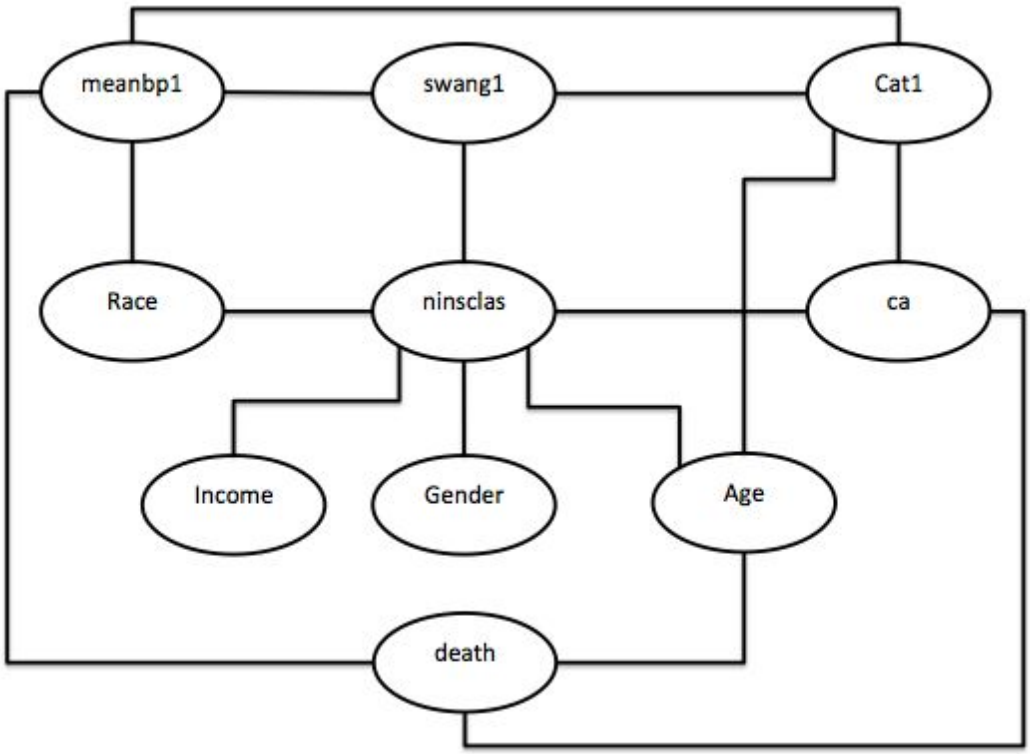
9 * 2 * 2 * 2 * 3 * 6 * 4 * 3 * 5 * 2 = 155520

The saturated model has as many parameters as there are cells in the table: 155520

Question c. Perform a forward-backward search on this data set, starting from the empty graph (independence model). Use the BIC scoring function. List the cliques of the resulting model in a nicely formatted table (don't dump the R output in your report) and draw the independence graph.

Score: 15841.66

Cliques Nrs	Cliques
1,3,10	Cat1, Swang1, Meanbp1
1,8	Cat1, Ca
1,9	Cat1, Age
2,8	Death, Ca
2,9	Death, Age
2,10	Death, Meanbp1
3,6	Swang1, Ninsclas
4,6	Gender, Ninsclas
5,6	Race, Ninsclass
5,10	Race, Meanbp1
6,7	Ninsclas, Income
6,8	Ninsclas, Ca
6,9	Ninsclas, Age



Question d. Using the independence graph you found under (c), what can you say about the pairwise conditional independence between income and gender? If we are interested in predicting whether or not someone survives, looking at the independence graph, which variables appear to be sufficient for that purpose?

We can see that income and gender are separated by ninsclas. We have the following independence: income ⊥ gender | ninsclass

For the second part of the question: To predict whether someone survives the following variables must be known: Age, Ca, Meanbp1

Question e. Perform a forward-backward search on this data set, starting from the complete graph (saturated model). Use the BIC scoring function. List the cliques of the resulting model and draw the independence graph. How does the model and its score compare to the result you found under (c)? □

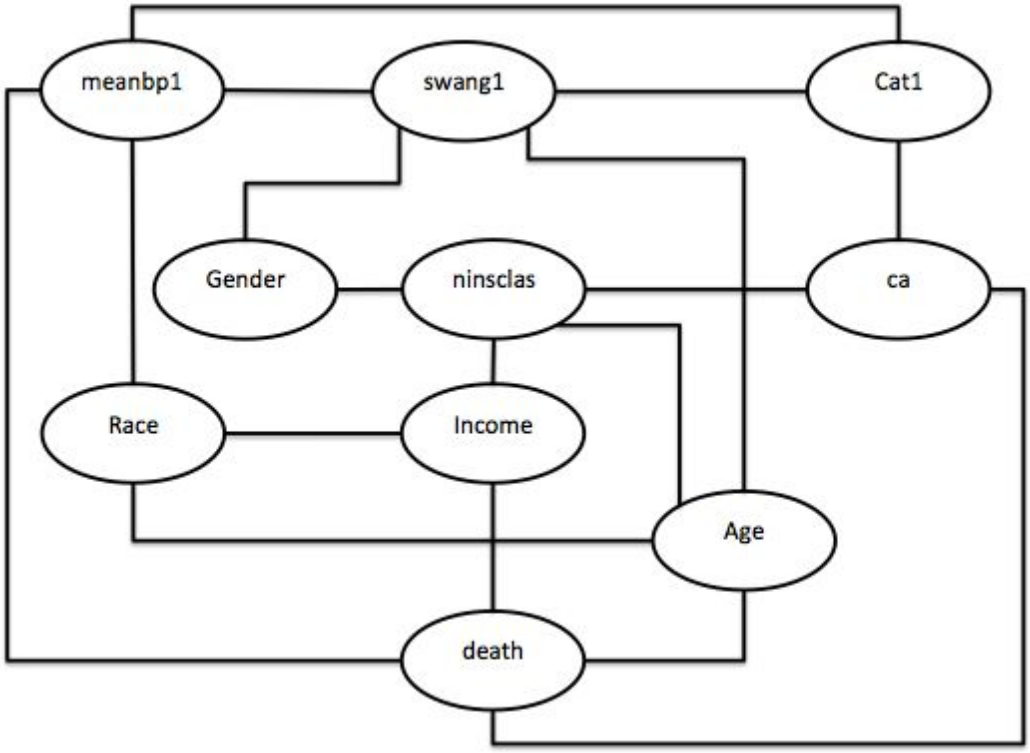
This is the saturated model:

```
> ma
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]    0    1    1    1    1    1    1    1    1     1
 [2,]    1    0    1    1    1    1    1    1    1     1
 [3,]    1    1    0    1    1    1    1    1    1     1
 [4,]    1    1    1    0    1    1    1    1    1     1
 [5,]    1    1    1    1    0    1    1    1    1     1
 [6,]    1    1    1    1    1    0    1    1    1     1
 [7,]    1    1    1    1    1    1    0    1    1     1
 [8,]    1    1    1    1    1    1    1    0    1     1
 [9,]    1    1    1    1    1    1    1    1    0     1
[10,]    1    1    1    1    1    1    1    1    1     0
> 
```

Score : 15850.53

Cliques Nrs	Cliques
1,3,10	Cat1, Swang1, Meanbp1
1,8	Cat1, Ca
2,7	Death, Income
2,8	Death, Ca
2,9	Death, Age
2,10	Death, Meanbp1
3,4	Swang1, Gender
3,9	Swang1, Age
5,7	Race, Income
5,9	Race, Age
5,10	Race, Meanbp1
4,6	Gender, Ninsclas
6,7	Ninsclas, Income
6,8	Ninsclas, Ca
6,9	Ninsclas, Age

Independence graph:



Comparison of the independence model and the saturated model: Both models have nearly the same score, with the independence model having the best score. If we compare the independence graphs of both models we can see that the saturated model has a few more edges, but besides those edges the graphs are very similar, which of course makes sense since the scores are so close to each other.

Question f. Perform a forward-backward search with AIC scoring on this data set, starting from the complete graph and empty graph. Give the cliques of the models you find, and the score of the models. Are they the same?

Empty graph/independence model:

Score : 14278.21

Cliques Nrs	Cliques
1,3,9	Cat1, Swang1, Age
1,3,10	Cat1, Swang1, Meanbp1
1,2,8	Cat1, Death,Ca
1,4,8	Cat1, Gender,Ca

1,2,9	Cat1, Death, Age
1,2,10	Cat1, Death, Meanbp1
1,4,10	Cat1, Gender, Meanbp1
2,7	Death, Income
3,6,9	Race, Income
4,5,6	Swang1, Race, Ninsclas
5,6,7	Race, Meanbp1
5,6,9	Race, Ninsclas, Income
4,5,10	Gender, Race, Meanbp1
4,6,8	Gender, Ninsclas, Ca

Complete graph/saturated model:

Score : 14278.21

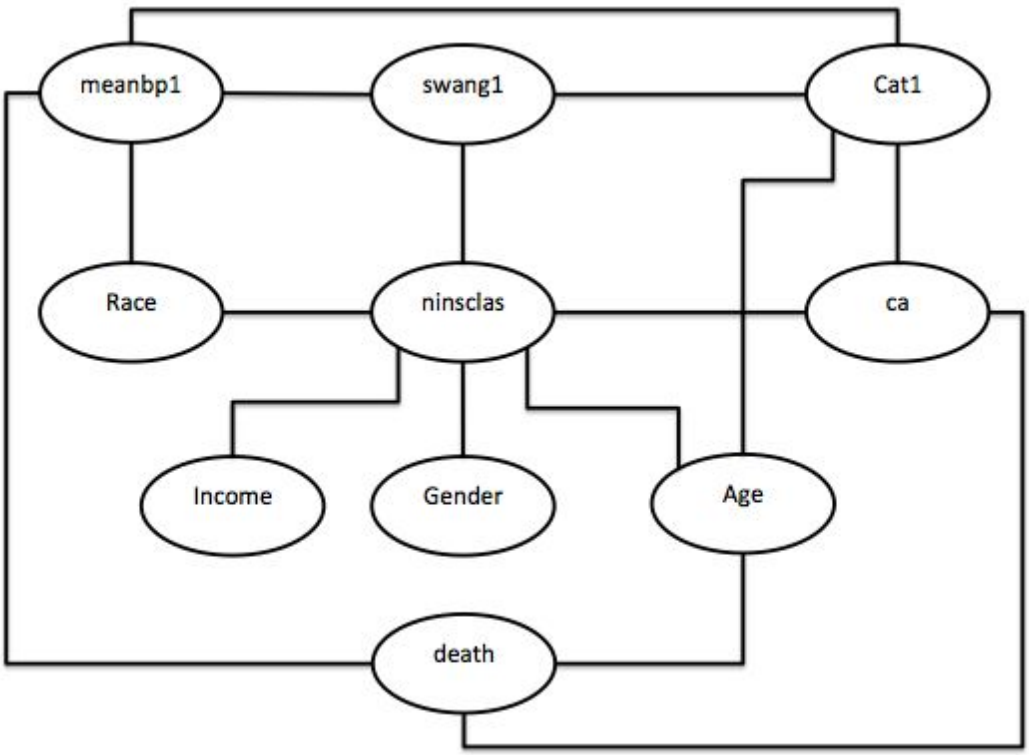
Cliques Nrs	Cliques
1,4,8	Cat1, Gender, Ca
1,4,10	Cat1, Gender, Meanbp1
4,5,6	Gender, Race,Ninsclas
4,5,10	Gender,Race,Meanbp1
4,6,8	Gender, Ninsclas, Ca
2,7	Death, Income
5,6,7	Race, Ninsclas, Income
1,2,8	Cat1, Death, Ca
1,2,9	Cat1, Death, Age
1,3,9	Cat1, Swang1, Age
3,6,9	Swang1, Ninsclas, Age

5,6,9	Race, Ninsclas
1,2,10	Cat1, Death, Meanbp1
1,3,10	Cat1, Swang1, Meanbp1

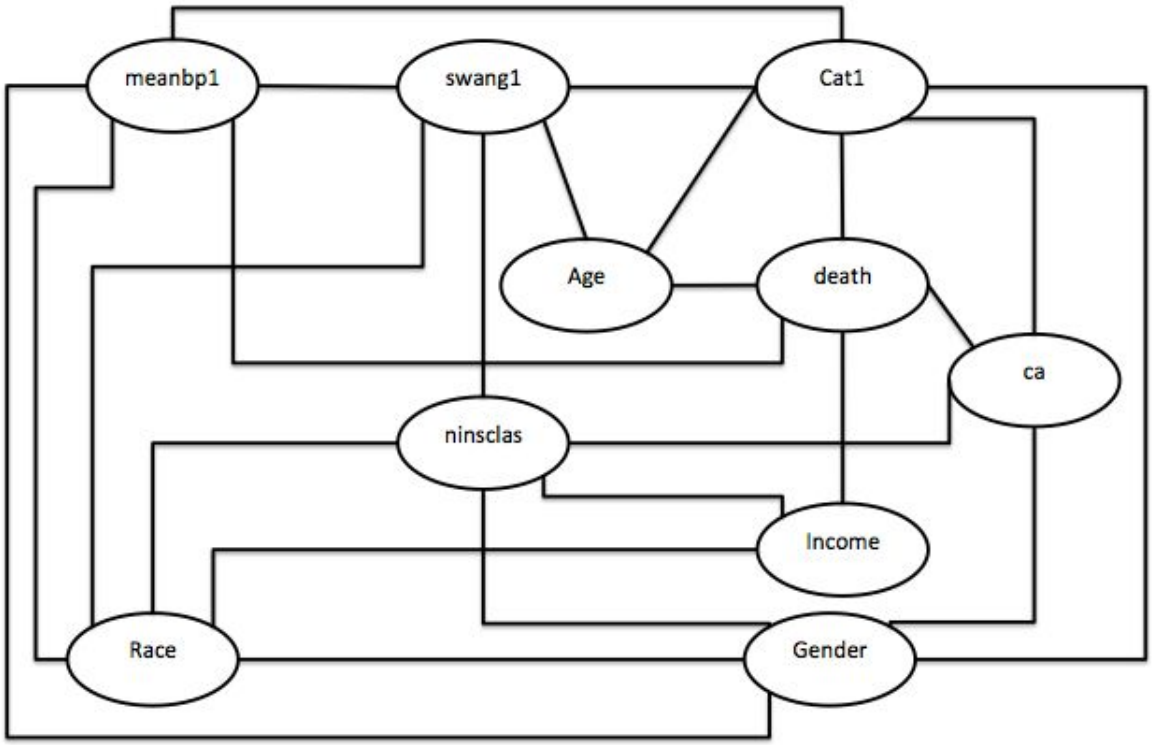
The score is exactly the same with 14278.21 for the search start with the empty graph and complete graph.

Question g. Compare the complexity of the models you found with BIC to those you found with AIC. Can you explain the difference?

BIC from the empty graph:



AIC from the empty graph:



The difference is the model from AIC is more complex than the model from BIC. This happens because BIC penalizes complex models more heavily than AIC. If we look at the number of parameters of a given model (model complexity), we can see that AIC multiplies that by two and adds that to its quality score. BIC, however multiples the number of parameters by the log of the number of counts of the table of observed counts. So the table only needs to be of decent size for BIC to penalize complex models more heavily than AIC.

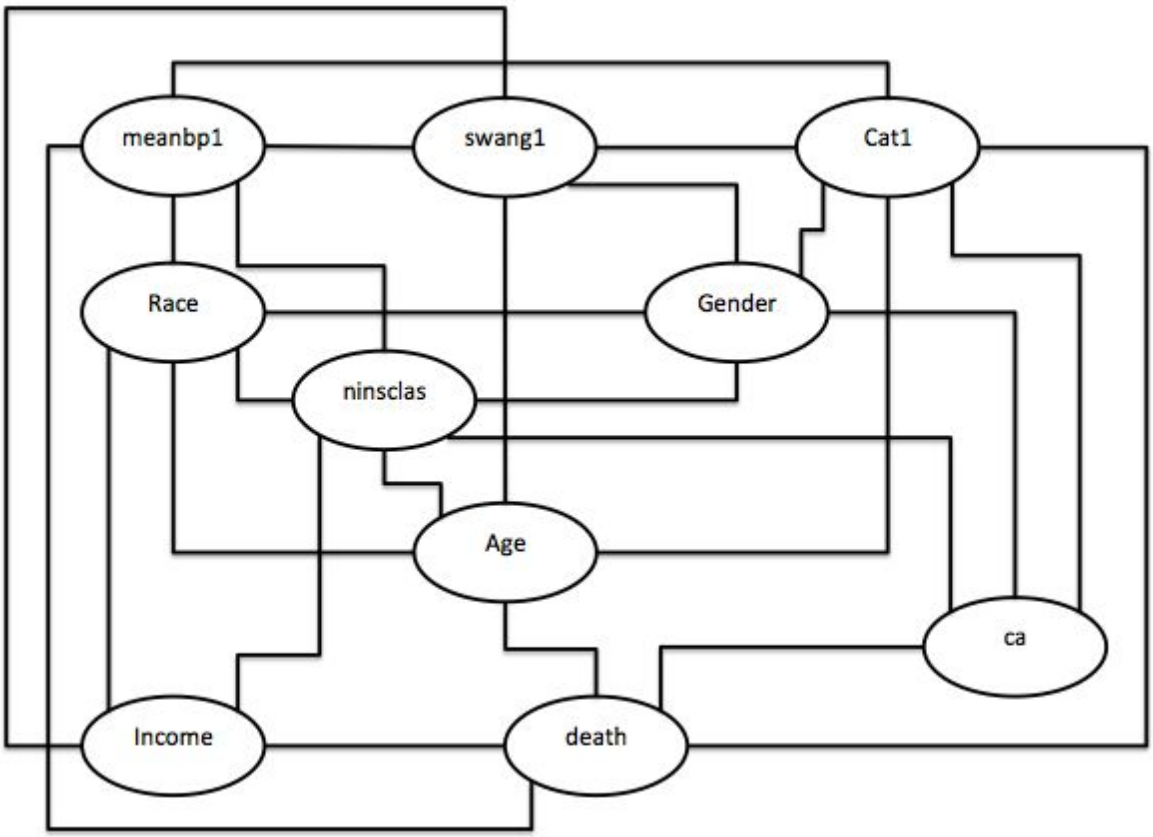
Question h. Use random restarts to see whether you can find better scoring models than you have found so far (both for AIC and for BIC scoring). Report how many restarts you performed, which values of prob you tried, and the best model that was found.

AIC

Number of restart	Probability	Seed	Score
5	0.2	1	14263.97
5	0.5	1	14263.97
5	0.35	1	14263.97
5	0.85	1	14288.22

The best model for AIC using the gm.restart function is with the probability of

0.2 and the result of the score 14263.97

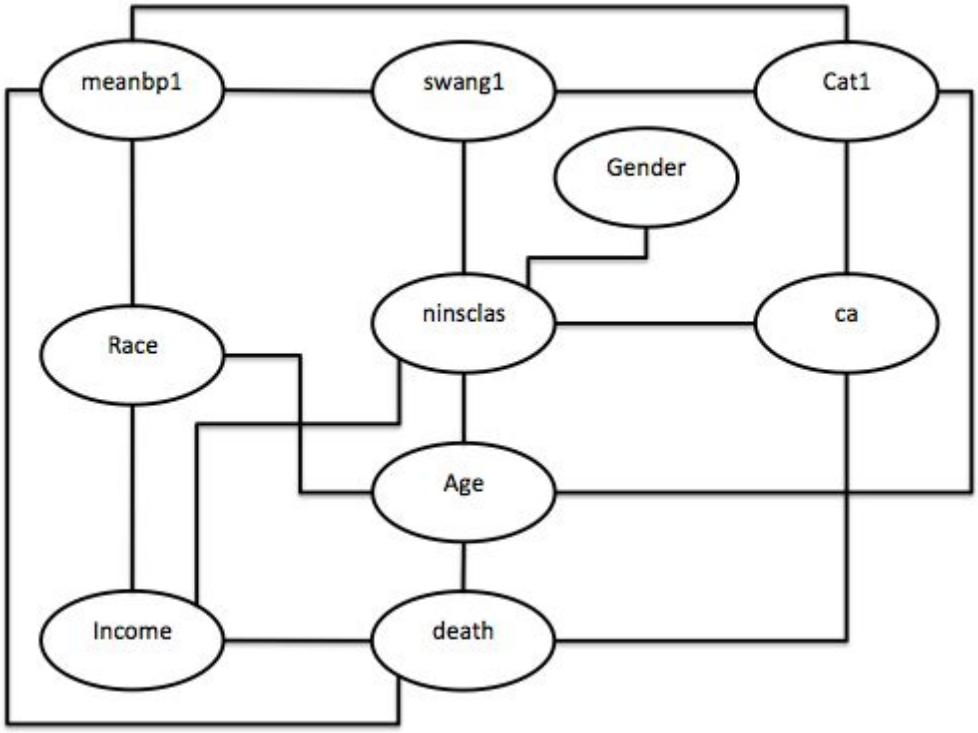


The best AIC model with 5 times restart and probability of 0.2

BIC

Number of restart	Probability	Seed	Score
5	0.2	1	15783.74
5	0.5	1	15787.08
5	0.4	1	15843.32
5	0.3	1	15787.08
5	0.25	1	15843.32
5	0.1	1	15807.95

The best model for BIC using the gm.restart function is with the probability of 0.2 and the result of the score 15783.74.



The best BIC model with 5 times restart and probability of 0.2