Data Analysis

Mathijs Baaijens (3542068)
Jarno Le Conté (3725154)

Question (a) - How many different models are there for this data?

possible edges: n*(n-1)/2 = 10*(10-1)/2 = 45
possible models: $2^{edges} = 2^{45} = 35184372088832$

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

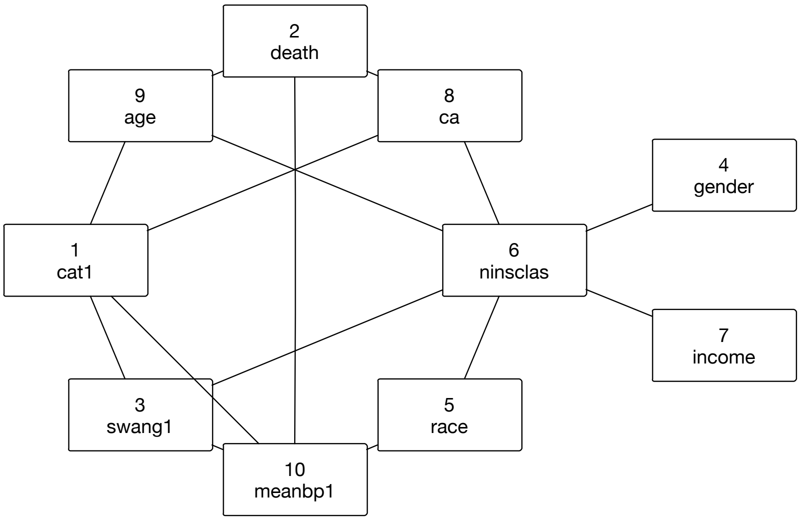
	cat1	death	swang1	gender	race	ninsclas	income	ca	age	meanbp1
values	9	2	2	2	3	6	4	3	5	2

9*2*2*2*3*6*4*3*5*2 = 155520 number of cells
8*1*1*1*2*5*3*2*4*1 = 1920 number of parameters

Question (c) - Independence model using BIC score

13 cliques, BIC score 15841.66 *(independence model)*												
1 8	**1 9**	1 3 10	2 8	2 9	2 10	**3 6**	4 6	**5 6**	6 7	6 8	6 9	5 10

TRACE
1. Add: 9 − 6   (score= 20139.73)
2. Add: 8 − 1   (score= 18561.18)
3. Add: 7 − 6   (score= 17313.90)
4. Add: 3 − 1   (score= 16947.05)
5. Add: 8 − 2   (score= 16706.60)
6. Add: 6 − 5   (score= 16466.82)
7. Add: 10 − 3   (score= 16227.81)
8. Add: 9 − 2   (score= 16024.58)
9. Add: 9 − 1   (score= 15936.67)
10. Add: 10 − 2   (score= 15899.67)
11. Add: 10 − 1   (score= 15878.31)
12. Add: 10 − 5   (score= 15857.81)
13. Add: 6 − 4   (score= 15847.04)
14. Add: 6 − 3   (score= 15843.36)
15. Add: 8 − 6   (score= 15841.66)



Question (d) - Independencies

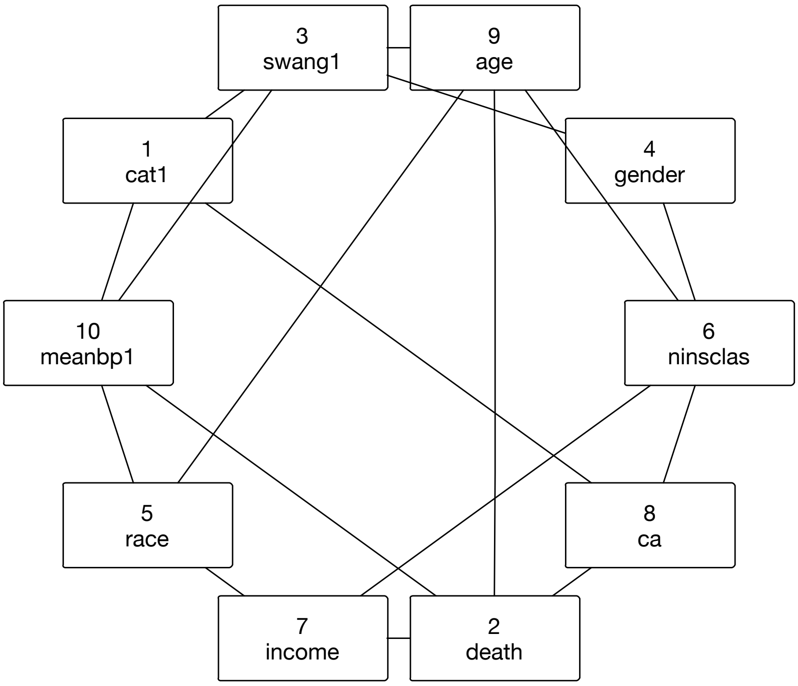
Gender and income are independent given ninsclas (**gender ⊥ income | ninsclas**). This means that given some value for ninsclas gender and income are independent. Another formulation is that gender and income look related but this is only because they are both dependant on ninsclas.
If you are interested in predicting whether or not someone survives, you have look at the direct

dependencies of the variable 'death', which means that the value depends on cancer status (ca), age and blood pressure (meanbp1).

Question (e) - Saturated model using BIC score

15 cliques, BIC score 15850.53 *(saturated model)*														
1 3 10	1 8	**2 7**	2 8	2 9	2 10	**3 4**	4 6	**5 7**	**5 9**	5 10	6 7	6 8	**3 9**	6 9

```
TRACE
1. Remove:    6 − 1   (score= 352418.67 )
2. Remove:    9 − 7   (score= 145996.05 )
3. Remove:    8 − 5   (score=  85022.56 )
4. Remove:    2 − 1   (score=  64150.58 )
5. Remove:    4 − 3   (score=  51172.69 )
6. Remove:    6 − 2   (score=  41682.01 )
7. Remove:   10 − 1   (score=  34213.22 )
8. Remove:   10 − 6   (score=  29598.23 )
9. Remove:    9 − 1   (score=  26137.86 )
10. Remove:   7 − 1   (score=  23410.33 )
11. Remove:   8 − 6   (score=  21604.37 )
12. Remove:   6 − 5   (score=  19937.64 )
13. Remove:  10 − 9   (score=  19036.88 )
14. Remove:  10 − 7   (score=  18358.74 )
15. Remove:   5 − 2   (score=  17953.09 )
16. Remove:   8 − 3   (score=  17582.91 )
17. Remove:   8 − 4   (score=  17231.89 )
18. Remove:   5 − 1   (score=  16916.77 )
19. Remove:   6 − 3   (score=  16633.86 )
20. Remove:   6 − 4   (score=  16404.68 )
21. Remove:   5 − 3   (score=  16269.59 )
22. Remove:   5 − 4   (score=  16145.29 )
23. Remove:   7 − 2   (score=  16062.76 )
24. Remove:   9 − 8   (score=  16011.30 )
25. Remove:   4 − 2   (score=  15969.11 )
26. Remove:   3 − 2   (score=  15935.30 )
27. Remove:   8 − 7   (score=  15913.10 )
28. Remove:   4 − 1   (score=  15903.07 )
29. Remove:   7 − 3   (score=  15897.47 )
30. Remove:   7 − 4   (score=  15892.60 )
31. Add:      8 − 6   (score=  15889.27 )
32. Remove:   9 − 4   (score=  15886.83 )
33. Add:      6 − 4   (score=  15876.29 )
34. Remove:  10 − 8   (score=  15873.92 )
35. Add:     10 − 1   (score=  15856.19 )
36. Remove:  10 − 4   (score=  15854.62 )
37. Add:      4 − 3   (score=  15851.79 )
38. Add:      7 − 2   (score=  15850.53 )
```

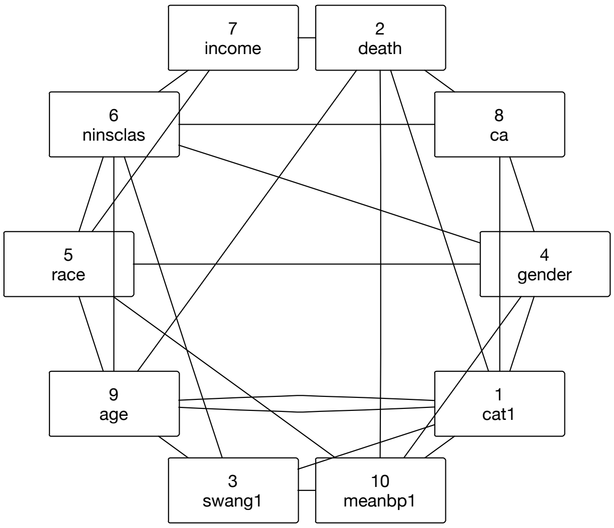


The main difference between this saturated model and the model we found in (c) is that this model contains more dependencies and is therefore less independent. There are some new relations such as income-death, swang1-gender, race-income, race-age and swang1-age, while other relations are omitted such as cat1-age, swang1-ninsclas and race-ninsclas. Ninsclas is less dependent in the saturated model, while death, age, income and gender have now more dependencies. The saturated model have a slightly lower BIC score. The model ‘fits’ better than the model found in (c), but due to the penalty for the additional u-terms for the additional dependencies the BIC score is lower.

Question (f) - Independence and saturated model using AIC score
The models are the same.

14 cliques, AIC score 14278.21 *(independence model)*													
1 2 8	1 2 9	1 2 10	1 3 9	1 3 10	1 4 8	1 4 10	4 5 6	4 5 10	5 6 7	5 6 9	3 6 9	4 6 8	2 7

14 cliques, AIC score 14278.21 *(saturated model)*													
1 2 8	1 2 9	1 2 10	1 3 9	1 3 10	1 4 8	1 4 10	4 5 6	4 5 10	5 6 7	5 6 9	3 6 9	4 6 8	2 7



Question (g) - Difference between AIC and BIC

There are more dependencies in the model found using AIC scoring than in the model found using BIC scoring. This is due to BIC penalising additional parameters (dependencies) more heavily than AIC for datasets with n > 7.

Question (h) - Random restarts

We used the following configurations:

	score type	**prob**	**nstart**	**result score**
A-0.25 (1)	AIC	0.25	1	14278.21
A-0.25 (3)	AIC	0.25	3	14278.21
A-0.25 (9)	AIC	0.25	9	14263.97
A-0.5 (1)	AIC	0.5	1	14341.65
A-0.5 (3)	AIC	0.5	3	14341.65
A-0.5 (9)	AIC	0.5	9	14263.97

	score type	prob	nstart	result score
A-0.75 (1)	AIC	0.75	1	14344.02
A-0.75 (3)	AIC	0.75	3	14344.02
A-0.75 (9)	AIC	0.75	9	14263.97
B-0.25 (1)	BIC	0.25	1	15783.74
B-0.25 (3)	BIC	0.25	3	15783.74
B-0.25 (9)	BIC	0.25	9	15783.74
B-0.5 (1)	BIC	0.5	1	16020.66
B-0.5 (3)	BIC	0.5	3	15783.74
B-0.5 (9)	BIC	0.5	9	15783.74
B-0.75 (1)	BIC	0.75	1	15905.88
B-0.75 (3)	BIC	0.75	3	15850.53
B-0.75 (9)	BIC	0.75	9	15783.74

For both AIC and BIC we measured the best results when doing many restarts.

Furthermore we saw that a probability of 0.25 will give this same best result for all runs of BIC and almost the best result for AIC. So a probability around 0.25 is likely a preferable for this dataset.

We also conclude that BIC will converge more easily, because we found the same result several times, while in the AIC case we measured slightly different values for most runs.