Relationship of Living Conditions with Life Expectancy

Probabilistic Modeling Exam

Erik Maresia
963724
March 2022

1. Introduction

The purpose of this study is to study the relationships between life expectancy and other variables related to living conditions. It will try to create graphical models to visualise the conditional independency of the variables. In particular undirected gaussian graphical models are used.

2. Data

The dataset used is a dataset showing life expectancy factors from the World Health Organization (WHO). The dataset is a panel data dataset with 183 countries over the years 2000-2016. This report will only consider the year 2016. The dataset has 32 variables, but 15 are immediately dropped for either having too many missing values, or being unrelated to this analysis. After checking for missing values, the dataset size is 121 observations.

There are 18 variables of interest, of which all are continuous. The most important variable is the variable concerning life expectancy (*life_expect*). A box plot of the variable can be seen in Appendix A.

1.1 Exploratory analysis

A correlation analysis was performed on all the variables expect life expectancy. Due to the relatively high dimension of the dataset, it was analysed if some variables could be dropped. Some variables were indeed very correlated with each other, and therefore omitted. Two correlation plots can be seen in Figure 1. For example it is seen that only *hepatitis* is left from the cluster of diseases *hepatitis, measles, polio* and *diphtheria*. In other words the the omitted variables were not explaining anything new that the other variables weren't. The remaining variables can be seen from Table 1.

Table 1

variable name	description	type
life_expect	Life expectancy at birth (years)	continuous
age1.4mort	Death rate between ages 1 and 4	continuous
alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)	continuous
bmi	Mean BMI (kg/m^2) (18+) (age-standardized estimate)	continuous
age5.19thinness	Prevalence of thinness among children and adolescents	continuous
age5.19obesity	Prevalence of obesity among children and adolescents	continuous
hepatitis	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)	continuous
basic_water	Population using at least basic drinking-water services	continuous
gghe.d	Domestic general government health expenditure	continuous
che_gdp	Current health expenditure (CHE) as percentage of gross domestic product (GDP) (%)	continuous
une_pop	Population (thousands)	continuous
une_hiv	Prevalence of HIV, total (% of population ages 15-49)	continuous
une_gni	GNI per capita	continuous

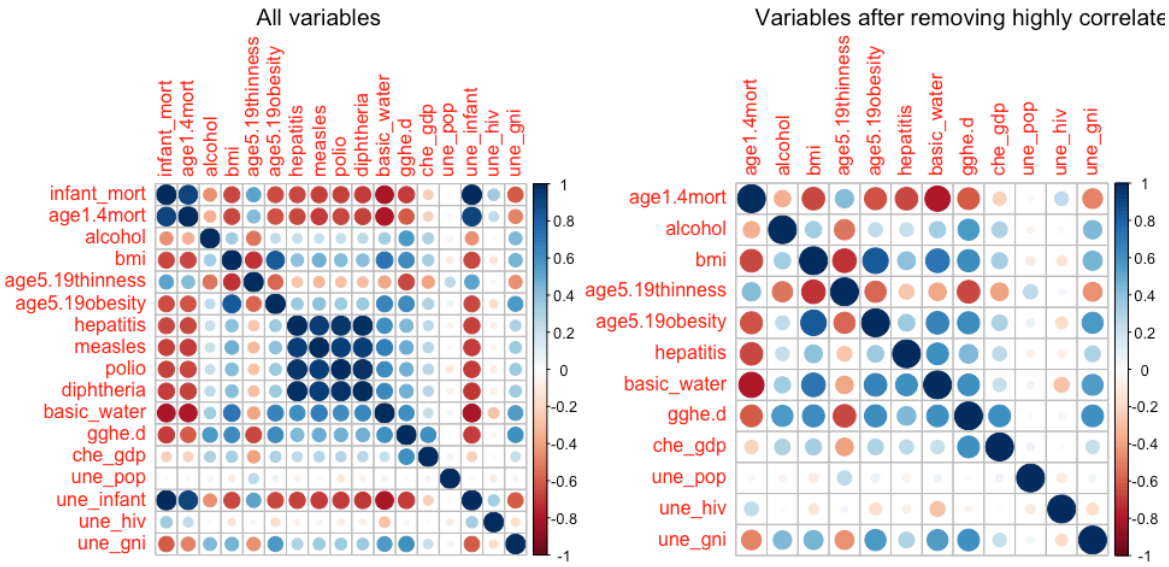


Figure 1

3. Methodology

This report is only concerned in gaussian graphical models, and in particular undirected graphs. Two methods are considered for the model selection. Namely simultaneous p-values and convex optimisation. Due to the relatively high dimension of the data, stepwise model selection is omitted for reasons of expensive computational time.

2.1 Concentration Matrix

In order to model how the variables are related, a concentration matrix must be derived. The concentration matrix K of a dataset is the inverse of the variance-covariance matrix

$$K = \Sigma^{-1}$$

Since the variables are all measured in different scales, a partial correlation matrix is used. This partial correlation matrix roughly measures the dependance of the variables. The partial correlation matrix for the dataset can be found from Figure 2.

	life_expect	age1.4mort	alcohol	bmi	age5.19thinness	age5.19obesity	hepatitis	basic_water	gghe.d	che_gdp	une_pop	une_hiv	une_gni
life_expect	100	-88	40	63	-50	66	57	82	71	27	4	-45	67
age1.4mort	-88	100	-34	-65	43	-62	-65	-81	-61	-23	-3	24	-49
alcohol	40	-34	100	34	-52	24	23	33	57	31	-6	3	43
bmi	63	-65	34	100	-71	83	40	72	62	32	-10	-13	46
age5.19thinness	-50	43	-52	-71	100	-57	-28	-39	-65	-40	25	5	-45
age5.19obesity	66	-62	24	83	-57	100	35	65	63	31	8	-18	57
hepatitis	57	-65	23	40	-28	35	100	61	44	25	-7	-8	31
basic_water	82	-81	33	72	-39	65	61	100	60	22	6	-29	56
gghe.d	71	-61	57	62	-65	63	44	60	100	61	2	-4	60
che_gdp	27	-23	31	32	-40	31	25	22	61	100	7	1	23
une_pop	4	-3	-6	-10	25	8	-7	6	2	7	100	-11	6
une_hiv	-45	24	3	-13	5	-18	-8	-29	-4	1	-11	100	-17
une_gni	67	-49	43	46	-45	57	31	56	60	23	6	-17	100

Figure 2

2.2 Simultaneous p-values

Simultaneous p-values (SIN) is a special type of thresholding technique. It attempts to identify subgraphs within the graph, and include edges that

are definitely present in the true graph. This method has a parameter α, which needs to be determined to partition the simultaneous p-values into three sets; a significant one, an intermediate one and a non-significant one. The probability that a graph G(α) is not a subgraph of the true model is less or equal to α . A plot for these simultaneous p-values van be seen from Figure 3.

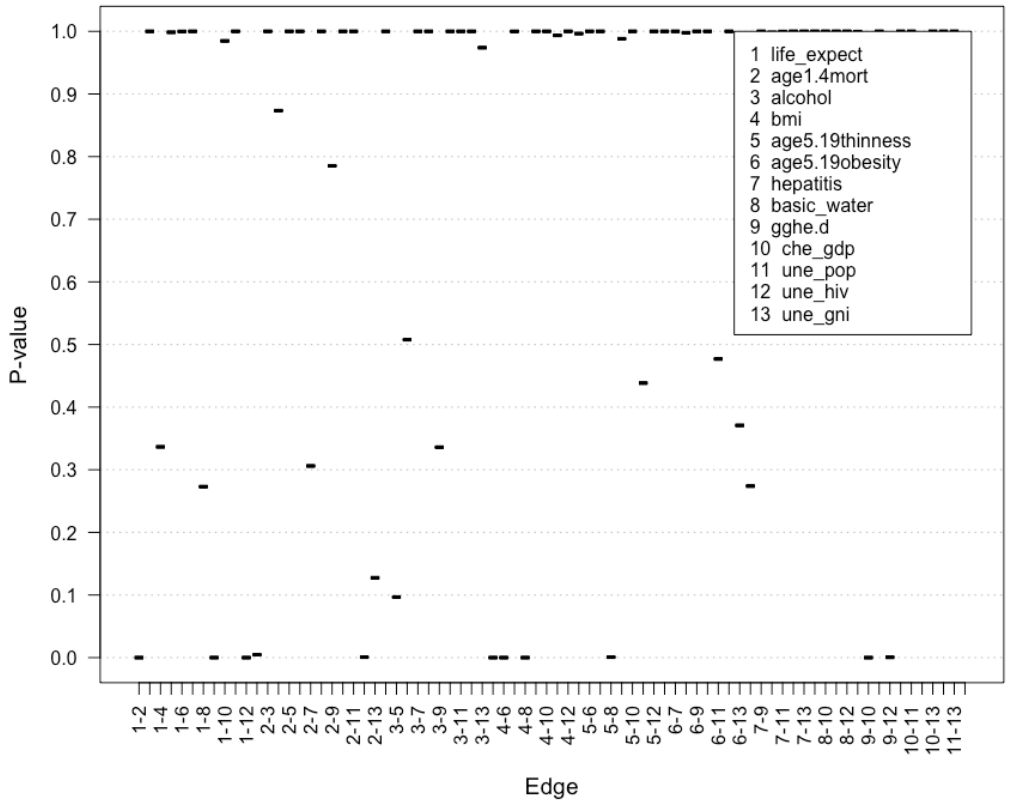


Figure 3

From Figure 4 it is seen what the different models look like visually, when different values for α are selected. The model with $\alpha = 0.2$ is selected for further analysis. The model output can be found from Appendix B.

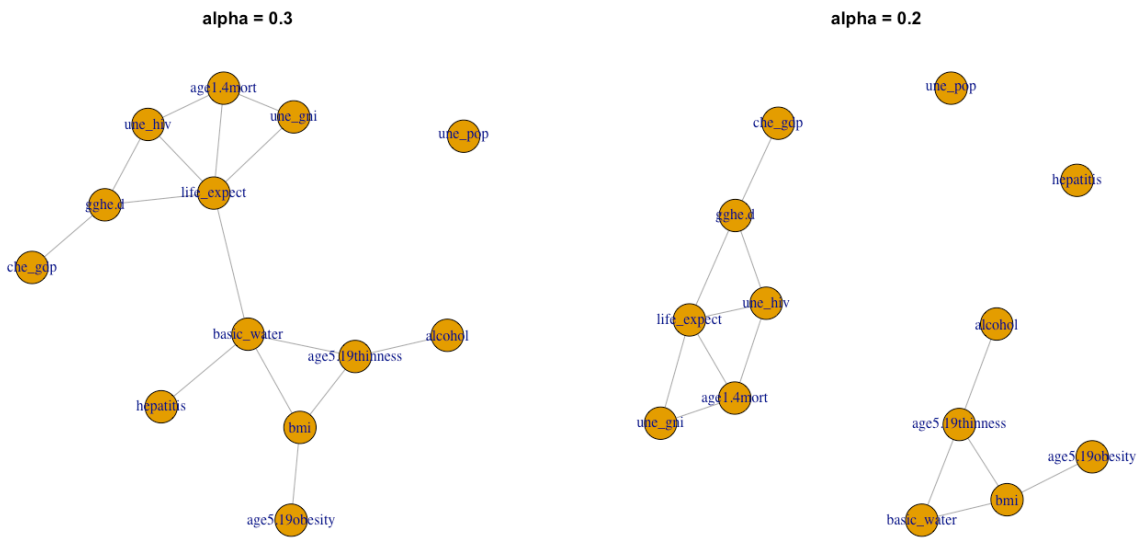


Figure 4

2.3 Convex Optimisation

Convex optimisation is a model selection method, which tries to maximise the log-likelihood for the concentration matrix, K. This technique is penalised by a non-negative parameter ρ. Convex optimisation can be done in R with the package *glasso*.

An adjacency matrix is built with the values of the estimates of the inverse of the covariance matrix. This adjacency matrix is used to construct the graph, with a value 1 representing an edge between vertices.

The value of ρ, which maximises the log-likelihood is chosen, which is chosen by cross-validation. For this dataset, a value $\rho = 0.17$ is chosen. The chosen graph can be seen from Figure 5. The model output can be found from Appendix C.

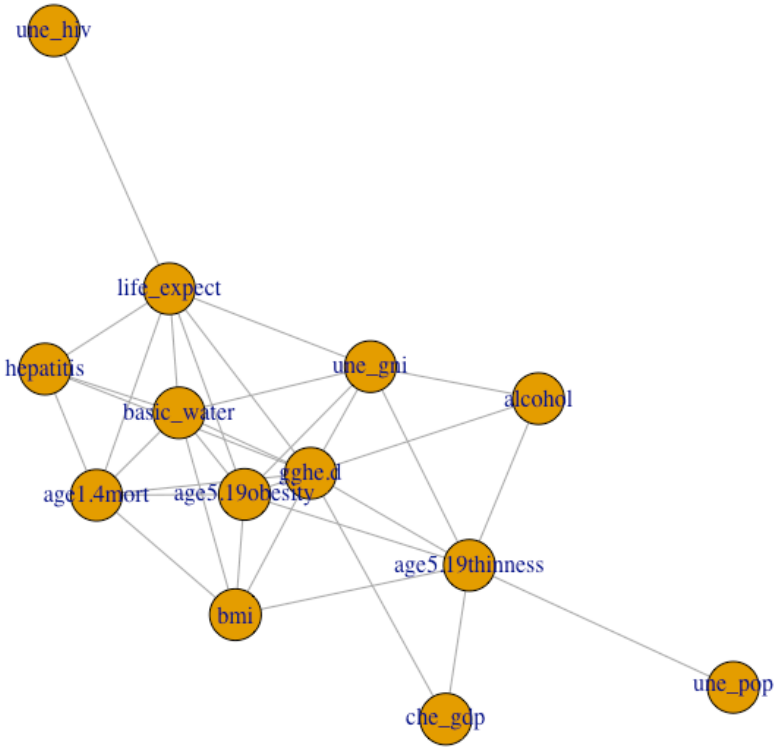


Figure 5

4. Results

This chapter will discuss the findings and the results of the analysis.

3.1 Model Selection

This section will discuss the model chosen to represent the relationships of the variables. As discussed in the previous chapter, two methods were used to select a model. The two models can be compared by using a likelihood ratio test, which shows if one model is a subgraph of the other. The deviance of a model M is

$$D = dev = 2(\hat{l}_{sat} - \hat{l}) = nlog\{\frac{det(S^{-1})}{det(\hat{K})}\} = -nlogdet(S\hat{K}),$$

while the likelihood ratio test, to test M_1 under M_0 where $M_1 \subseteq M_0$ is the difference of the deviance of the models is

$$lrt = 2(\hat{l}_0 - \hat{l}_1) = nlog\{\frac{det(\hat{K}_0)}{det(\hat{K}_1))}\}.$$

Testing the SIN model and the model attained by convex optimisation, where the SIN model is M_1 and the convex optimisation model is M_0, a likelihood ratio test result of 28.21435 is gotten with a degree of freedom of 6. Since the null hypothesis of the likelihood ratio test is that the model M_1 is false, the null is rejected and the model attained by convex optimisation is chosen.

3.2 Hypothesis Testing

Single conditional independence tests can be carried out with the data. This is one way additional way to analyse if the selected model is performing well. Since the dataset size in this report is relatively small, single conditional independence is measured by the F statistic,

$$F = (n - d)(e^{dev/n} - 1),$$

where n is the sample size, d is the degrees of freedom and dev the deviance. The null hypothesis is that there is a conditional independence between the two variables.

The first test is to test the conditional independence of life expectancy and the population (*une_pop*). The selected model shows that the variables should be conditionally independent from each other, given all other variables. This can be seen from FIGURE. The conditional independence of the two is given by:

$$life\_expectancy \perp\!\!\!\perp une\_pop \mid age1.4mort, alcohol, bmi, age5.19thinness, age5.19obesity, hepatitis, basic\_water, gghe.d, che\_gdp, une\_hiv, une\_gni$$

where the associated p-value is 0.2964, so the null hypothesis of conditional independence cannot be rejected.

The next test is to test the conditional independence of life expectancy (*life_expect*) and mortality between the ages 1-4 (*age1.4mort*). The selected model shows that there is a dependence between the two variables. The conditional independence of the two variables, with the rest of the variables expressed as $\ldots rest$, is given by:

$$life\_expectancy \perp\!\!\!\perp age1.4mort \mid \ldots rest,$$

where the associated p-value is less than 0.0005 and so the null hypothesis of conditional independence can be rejected.

3.4 Further Validation

To validate the findings even further, the dataset is fit with a random forest to understand the importance of the variables. The findings of the random forest correspond to those found in the graphical model. The population of the respective country has the least importance in predicting

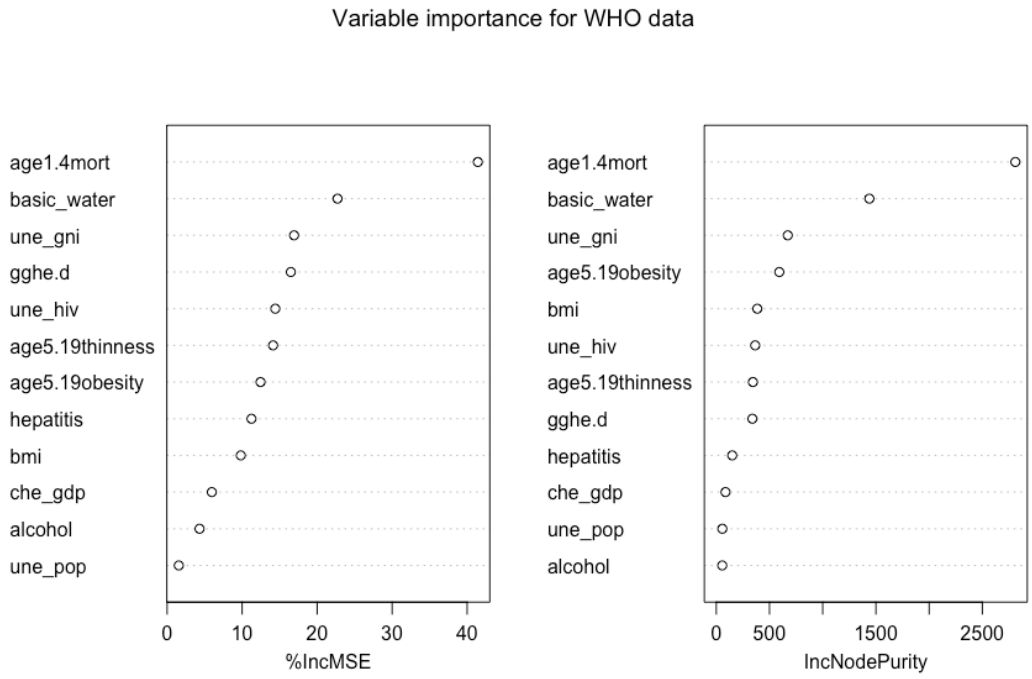


Figure 6

values of life expectancy, while mortality of ages 1-4, basic water, and GNI/ per capita have the most. These findings are similar to the conditional independences discussed above. The variable importance can be seen from Figure 6.

4. Conclusion

Graphical models is a statistical tool to represent conditional independence between variables in a dataset. It is especially helpful, when the dimensionality of the dataset is large.

The model helped to analyse the conditional independence of life expectancy with respect to the other variables in the dataset. In particular, the graphical model helped visualise that variables such as population (*une_pop*) and current health expenditure (*che_gdp*), are conditionally independent of life expectancy, given all the other variables.

Further analysis could be done on the variables and the conditional independence of them. A larger sample size could be more robust, and provide a more significant result. Future analyses could include modelling directed graphs as well as undirected ones.

5. References

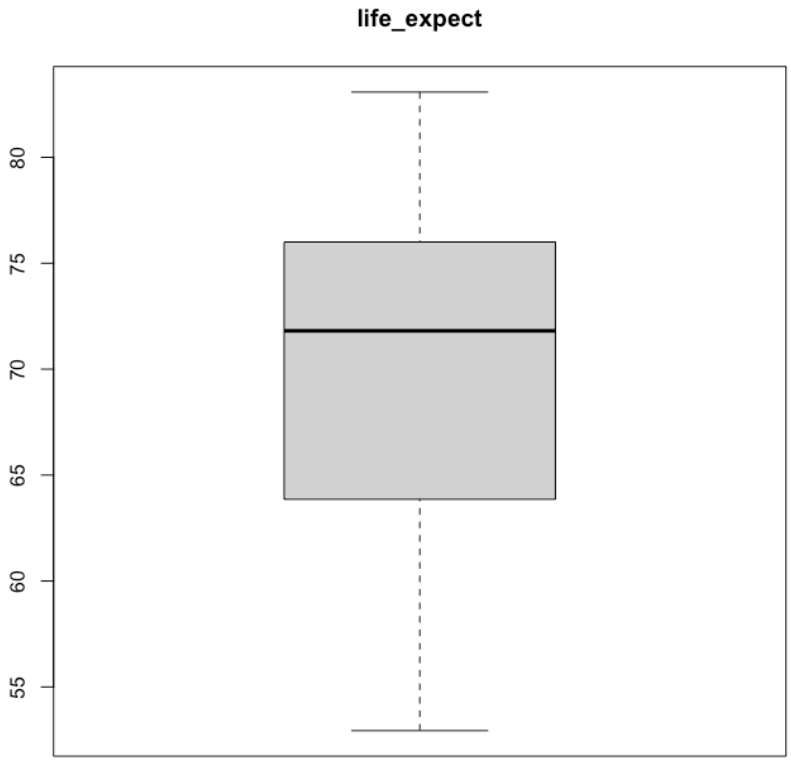
- Højsgaard, Søren, et al. *Graphical Models with R*. Springer, 2012.

- James, Gareth, et al. An Introduction to Statistical Learning with Applications in R. Springer, 2021.

- In-class presentations slides by Prof. Nicolussi

5. Appendix

Appendix A



Box plot of life expectancy

Appendix B

```
Model: A cModel with 12 variables
 -2logL    :        7831.42 mdim :   27 aic :      7885.42
 ideviance :        1018.89 idf  :   15 bic :      7960.91
 deviance  :         200.74 df   :   51
```

Model for simultaneous p-values

Appendix C

```
Model: A cModel with 13 variables
 -2logL    :       10743.98 mdim :   46 aic :     10835.98
 ideviance :        1077.82 idf  :   33 bic :     10964.58
 deviance  :         172.53 df   :   45
```

Model for convex optimisation

Appendix D

The respective R code can be found from my GitHub, from:

https://github.com/maresiaerik/probablilistic_modeling_exam