Bayesian Networks and Causal Inference

Assignment 1: Building a Bayesian Network

Guillem Soler Sanz Diego Bartoli Geijo Marta Español López

8th of December 2022

Problem domain

Life expectancy could be affected by many different factors, strongly interrelated between them. Most of the studies consider primarily demographic variables, government expenditure on health and mortality rates. However, health related factors and in particular immunization from diseases like Hepatitis B, Polio and Diphtheria could also play a central role in affecting the life expectancy of a country. In this project we use a dataset that considers together demographic, economic and health factors to build a Bayesian Network to analyse the dependencies between the variables that could influence life expectancy.

Data description

Dataset We obtain the data from the 'Life Expectancy (WHO)' [1] dataset, avalaible on Kaggle. This dataset presents data for 183 countries from 2000 to 2016. The data was obtained from the *Global Health Observatory* and the *United Nations* websites. The final merged file consists of 3111 rows and 32 variables, of which 29 are numerics and 3 are non ordinal categorical ones.

Variable selection Based on personal evaluation of which variables would be more useful and interesting, we choose 7 variables to start working with:

1. Life expectancy: life expectancy value in years at birth;
2. Adult mortality: probability of dying between 15 and 60 years per 1000 persons;
3. Infant mortality: probability of dying before the age of 1, expressed to 1;
4. Alcohol: litres of pure alcohol consumed per capita among people older than 15 years old;
5. BMI: age standardized estimate of the Body Mass Index[2] of the population older than 18 in kg/m^2;
6. Polio: Polio(Pol3) immunization coverage percentage among 1-year-olds;
7. Domestic general government health expenditure: government health expenditure as a percentage of gross domestic product [3];

The variables selected are reported in the table below with the type and the value range.

[1] https://www.kaggle.com/datasets/mmattson/who-national-life-expectancy
[2] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
[3] https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4953

Name	Type	% Range
Life expectancy	continous	36.30 - 84.17
Adult Mortality	continous	49.2 - 696.6
Infant Mortality	continous	0.00 - 0.16
Alcohol	continous	0.00 - 20.18
BMI	continous	19.80 - 32.20
Polio	continous	8.00 - 99.00
Domestic GGHE	continous	0.06 - 12.06

Exploratory analysis The figure 1 reports a summary analysis of the variables. We check the values searching for information on the measures considered and they all make sense. We analyse the number of missing values for each variable, figure 7. The ratio of rows without missing values in the whole dataset is 0.96, considering it an acceptable value we decide to just remove those rows. Regarding outliers, figure 9, we don't consider necessary to intervene. The dataset has now 2980 rows and 7 variables. The variables distribution is reported in figure 2. No variable has a distribution similar to life expectancy. Adult mortality and infant mortality have similar distributions between them, both right-skewed. Polio has a strongly left-skewed distribution, most of the values are very close to 100. BMI has a bimodal distribution. You can find the code used for the exploratory analysis, as well as other pre-processing outcomes in Appendix A.

```
 life_expect    adult_mortality  infant_mort         alcohol          bmi          polio
Min.   :36.23   Min.   : 49.2   Min.   :0.001470   Min.   : 0.000   Min.   :19.80   Min.   : 8.00
1st Qu.:63.20   1st Qu.:108.3   1st Qu.:0.008255   1st Qu.: 1.198   1st Qu.:23.30   1st Qu.:81.00
Median :71.60   Median :164.8   Median :0.019995   Median : 3.994   Median :25.50   Median :93.00
Mean   :69.15   Mean   :193.5   Mean   :0.032496   Mean   : 4.835   Mean   :25.05   Mean   :86.61
3rd Qu.:75.54   3rd Qu.:250.8   3rd Qu.:0.051720   3rd Qu.: 7.723   3rd Qu.:26.50   3rd Qu.:97.00
Max.   :84.17   Max.   :696.9   Max.   :0.164515   Max.   :20.182   Max.   :32.20   Max.   :99.00
                                                   NA's   :50       NA's   :34      NA's   :19

    gghe.d
Min.   : 0.06236
1st Qu.: 1.53344
Median : 2.60130
Mean   : 3.12293
3rd Qu.: 4.27811
Max.   :12.06273
NA's   :100
```

Figure 1: **Summary analysis of the variables.** For each variable it is displayed a general analysis with: minimum and maximum values, 1st and 3rd quartiles, median, mean and number of missing values.

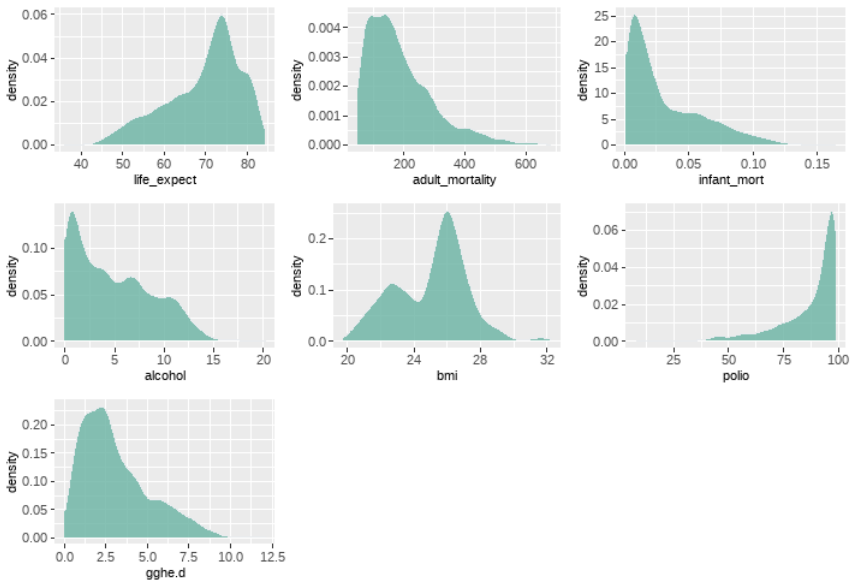

Figure 2: **Variables distribution.** Plot of each variable's distribution after pre-processing.

Network implementation

The programming language used all along the project for both network building and testing is R (ver. 4.0.3). Before building the network, an exploratory analysis has been performed on the data in order to make a better understanding of it, and to be able to pre-process it more accurately, when needed. For the exploratory analysis ggplot2 (ver. 3.3.2) [1] was used to visualize the variable distributions, and ggcorrplot (ver. 0.1.3) [2] to have a first glance at the correlation between all the selected variables from the data set. Regarding the pre-processing of the data, visdat (ver. 0.5.3) [3] was used to visualize the amount of missing values for each of the variables and dlookr (ver. 0.6.0) [4] to diagnose the presence of outliers in our data.

The network has been first built using DAGitty (ver. 0.3-1) [5] online in the browser, and the generated R code from the causal diagram has been then used for further testing and re-structuring in RStudio (ver. 1.2.5042) [6]. To assess d-separation and correlation, some functions from the DAGitty package were used. The function 'impliedConditionalIndependencies()' initially checked for existing conditional independencies among the variables and further on, the 'localTests()' and 'plotLocalTestResults()' functions helped to determine which edges to add or to remove from the network. In order to make the model fitting and explore the effect of each edge to each other, bnlearn (ver. 4.8.1) [7] was used.

Network creation and testing

Introduction. For the network creation we follow the general instructions reported in the R Companion of the course and in one of the given papers [8]. Therefore, we first propose an almost fully-connected graph and then remove the connections that do not seem reasonable to us. We reserve the right to also add connections if the indipendencies obtained are not significant. We don't test more than three different networks to not cause overfitting. You can find the code used for the network building and testing in Appendix B. The code is the same for all the networks, only the description of the network structure changes.

First network. We build the first network, figure 3, based on personal knowledge of the variables with the help of online research. All the variables except Polio and Bmi are connected to the outcome, this is because we think that these two variables don't have a direct influence on life expectancy. We briefly explain the reasoning behind all the connections added:

- Alcohol $\rightarrow$ Bmi: we think that alcohol is one of the main reasons for weight problems.
- Alcohol $\rightarrow$ Adult mortality: we believe that alcohol has a negative influence on people's health.
- Alcohol $\rightarrow$ Health expenditure(gghe.d): as we explained our opinion is that alcohol has a negative influence on people's health so we think that a huge alcohol consumption could increase the health expenditure.
- Alcohol $\rightarrow$ Life expectancy: the same arguments as before, alcohol has a strong influence on people's health such as to influence also life expectancy.
- Bmi $\rightarrow$ Adult mortality: some mortal diseases are caused by weight problems, so we consider that there is a relation between both variables.
- Bmi $\longleftrightarrow$ Infant mortality: there is a bidirected edge because we think that there are latent variables that could influence both variables, for example the percentage of mothers that have obesity problems.
- Bmi $\longleftrightarrow$ Polio: there is a bidirected edge because we think that there are latent variables that could influence both variables.
- Health expenditure(gghe.d) $\rightarrow$ Bmi: some diseases are caused by weight problems, so we believe that there is a relation between both variables.

- Health expenditure(gghe.d) → Life expectancy: we think that the amount of resources that a country invests in health has a direct relation with the life expectancy of this country.
- Health expenditure(gghe.d) → Polio: we suppose that if the percentage of people who are vaccinated against polio is high it is because this country invests money in health.
- Health expenditure(gghe.d) ⟷ Infant mortality: there is a bidirected edge because we think that there are latent variables that could influence both variables, for example the quality of health care services in that country.
- Infant mortality → Life expectancy: we think that infant mortality of a country is directly related to the life expectancy of that country.
- Polio → Adult mortality: we think that a high number of people vaccinated against Polio reduces the risk of getting the disease and so the deaths caused by it.
- Polio → Infant mortality: we think that a high number of people vaccinated against Polio reduces the risk of getting the disease and so the deaths caused by it.
- Adult mortality → Life expectancy: we think that the adult mortality of a country is directly related to the life expectancy of that country.
- Adult mortality ⟷ Health expenditure(gghe.d): there is a bidirected edge because we think that there are latent variables that could influence both variables, for example the quality of health care services in that country.

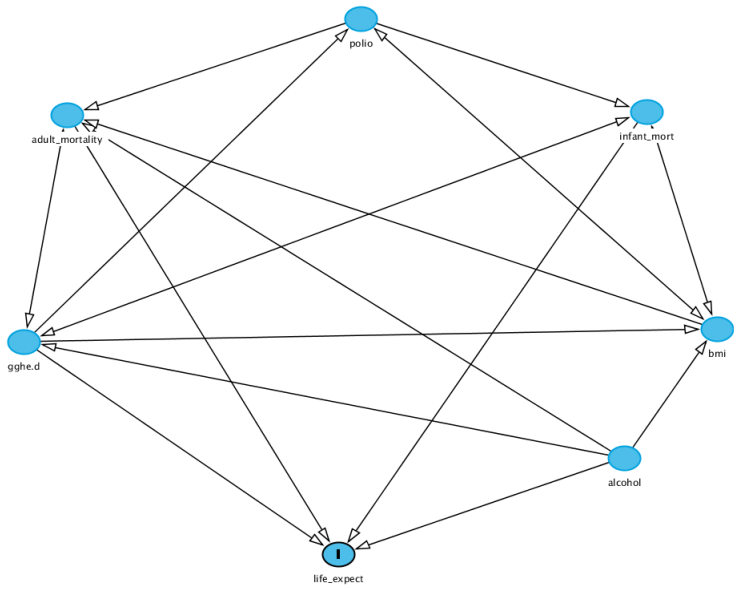

Figure 3: **First network.** First version of the DAG based on theoretical knowledge of the subject, maximizing the number of edges added.

We then proceed to perform the first local conditional independence test on the DAG and we obtain the following conditional independencies (full results of the local test can be found at Appendix B):

- alcohol ⫫ polio | ggh. (p-value=2.5053e-5)
- bmi ⫫ life_expectancy | adult_mort, alcohol, ggh. infant_mort (p-value=0.0322)
- life expectancy ⫫ polio | adult_mort, alcohol, ggh. infant_mort (p-value=0.1314)

Looking at the p-values, the 2 bottom independencies show to be fairly significant, while the first one remains a bit less likely to happen. Also, taking a look at the estimates obtained for the last 2 independencies, they are very close to zero values, meaning that the non-existence of an edge between those variables is appropriate, as they don't influence directly each other, thus, they shall remain unconnected.

Second Network. Considering the output of the previous analysis, we decide to just add a latent variable edge from Alcohol⟷Polio. We keep bmi and life expectancy unconnected, even if the p-value is a bit lower than the usual 0.5 treshold, because the estimate value is close to zero and because from a theoretical point of view we believe that the variables should be independent. Regarding life expectancy and polio, both from a theoretical point of view and looking at the statistics obtained, it makes sense to also keep them unconnected.
With regard to the newly added edge, we believe that the obtained independency is adequate from a theoretical point of view, but looking at the result from the test we can see that is the least significant of all the results with a very low p-value, from which we can deduce that they both may be related in some way. We then decide to connect them but as a latent connection, so their relationship depends on some other external variable, as we believe that alcohol does not have a direct effect on the proportion of vaccinated babies, nor the other way around.

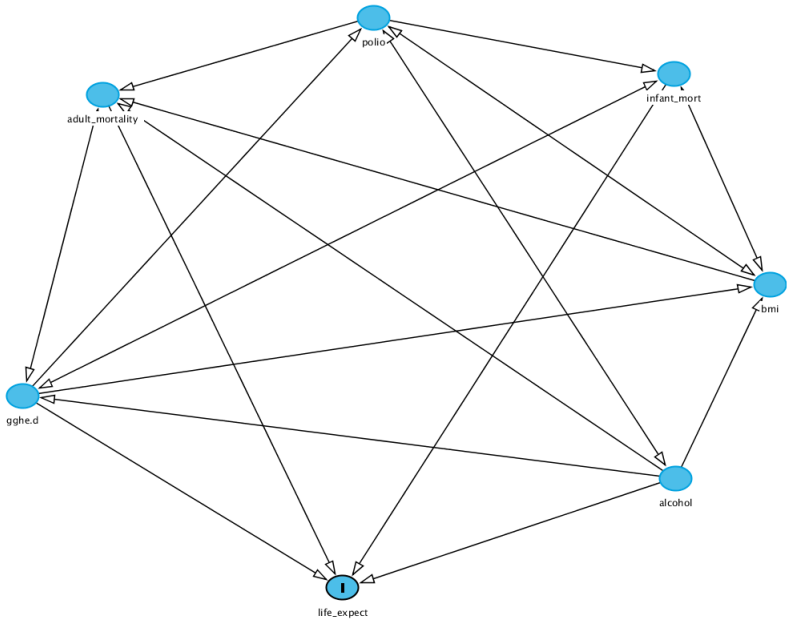

Figure 4: **Second network.** Second version of the DAG after adding the latent variable connection between Alcohol⟷Polio.

After adding the new connection to the DAG we re-test the network to see how this change influences the causal relationships between the variables, and we obtain the following outcome:

- bmi $\perp\!\!\!\perp$ life_expectancy | adult_mort, alcohol, ggh. infant_mort (p-value=0.0322)
- life expectancy $\perp\!\!\!\perp$ polio | adult_mort, alcohol, ggh. infant_mort (p-value=0.1314)

Third Network. After re-testing the network, we can see that as a result of the previously added latent variable connection between alcohol and polio we now have left only two of the earlier independencies.
We take another look at the DAG and we consider that bmi and alcohol may not be directly related, based on our own knowledge of the matter, so we decide to remove the existing connection and check whether alcohol's influence on bmi is significant enough or not.

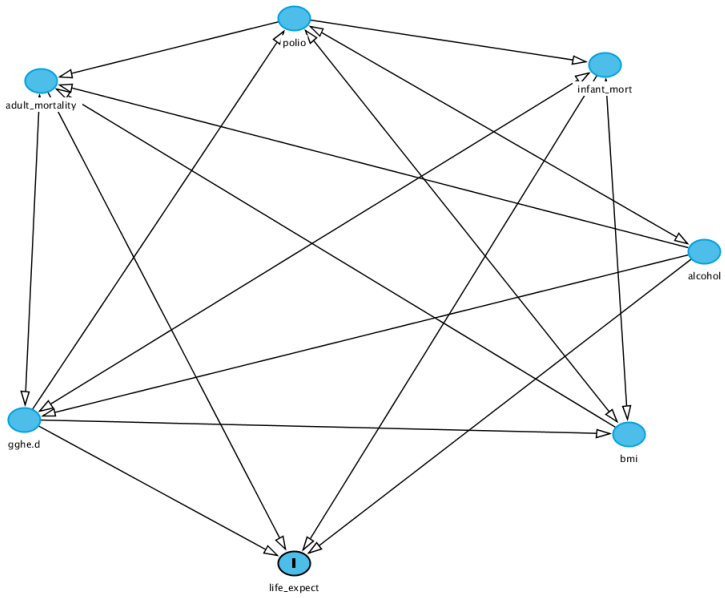

Figure 5: **Third Network.** Final version of the DAG after removing the connection Alcohol→Bmi.

We then proceed to perform the last conditional independence test on the network and the definitive independencies are the following:

- alcohol ⫫ bmi | ggh. (p-value=0.0575)
- bmi ⫫ life_expectancy | adult_mort, alcohol, ggh. infant_mort (p-value=0.0322)
- life expectancy ⫫ polio | adult_mort, alcohol, ggh. infant_mort (p-value=0.1314)

When removing the Alcohol→Bmi edge we successfully identify that alcohol and bmi actually have an independent relationship, which shows to be significant enough, so we decide to keep those two variables directly unrelated. With a very close to zero estimate we can say that the variables are not dependent, and with a high p-value indicating that this independency is very likely to happen, we can say that both our point of view and the test's outcome support our decision of removing the edge.
We then consider this to be the definitive DAG and that there is no further need for testing because, as mentioned before, the variables chosen are all highly correlated to each other and to the response variable, so we believe that the current network is the most appropriate to represent the relationships among the variables chosen.

Model fitting

We explore the effects of the edges to each other by fitting the model on data scaled to standard deviation 1. The results are reported in figure 6, the code used in Appendix C. Two interesting findings are that Bmi and Polio seem to be more important than alcohol in influencing adult mortality and that adult mortality is the most influencing variable for life expectancy. For life expectancy we obtain a low value of the standard deviation of the residuals, so we can suppose, even though we

have not actually implemented prediction, that the variables considered are enough to obtain good prediction results of life expectancy using linear regression.

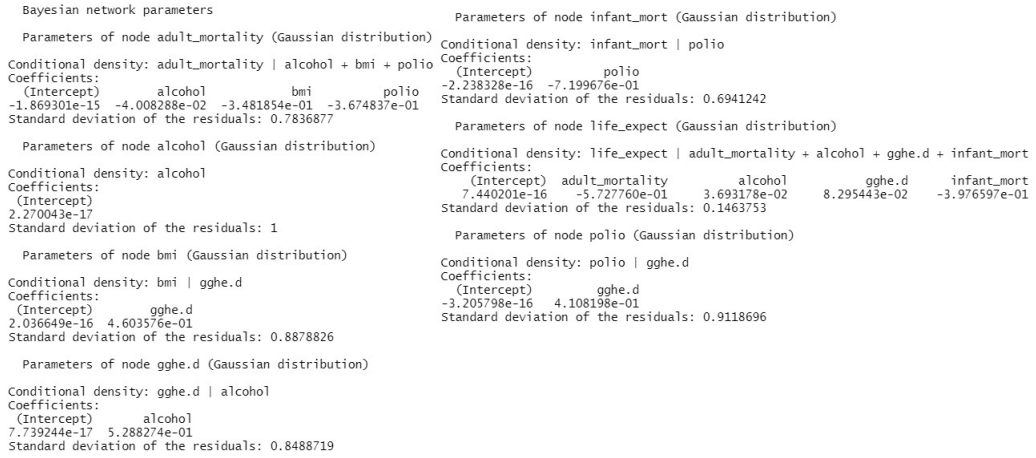

Figure 6: **Model fitting results.**

Discussion and Conclusions

Problems. Life expectancy is a complex measure that is influenced by a huge variety of factors and we found it difficult to logically identify the causal relationships between the variables. We probably used a too small sample of variables to provide a structured analysis of the dependencies existing between the factors influencing life expectancy, which led to the presence of a large number of latent variables in our networks. Furthermore, we believe that most of the variables that can be taken into account can influence each other, so it is difficult to derive a large number of independencies.

Results. We have been able to identify 3 significant conditional independencies, though. We obtain that, given health expenditure (gghe.d), alcohol and bmi are not dependent, as we initially supposed. The influence of alcohol consumption on body weight is probably not strong enough to influence the bmi of an entire population. The other two independencies obtained confirm our intuition that bmi and polio don't have a direct influence on life expectancy, in particular the second one. We can also conclude that most of the selected variables greatly influence life expectancy. It can be a positive influence, there is higher life expectancy with an increase of the domestic general government health expenditure, or it can be a negative influence in the case of adult mortality, an increase of it directly involves a decrease in life expectancy.

Future work. For further implementations, we would try to take more variables from the original data set to improve the network. Increasing the number of variables might increase the difficulty of defining the network structure but the results obtained would be more complete. We would also try to make explicit some of the latent variables in the network, even though it could be hard to associate a bidirected edge to just a single variable. However we believe that, even with these improvements, it would still be difficult to obtain a large number of significant independencies because the variables in the field under consideration are all strongly intercorrelated.

References

[1] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.

[2] A. Kassambara and I. Patil. ggcorrplot: Visualization of a correlation matrix using ggplot2. *International Journal of Epidemiology*, 2022.

[3] N. Tierney. visdat: Visualising whole data frames. *JOSS*, 2(16):355, 2017.

[4] C. Ryu. *dlookr: Tools for Data Diagnosis, Exploration, Transformation*, 2022. R package version 0.6.1.9001.

[5] J. Textor, B. Van Der Zander, M. K. Gilthorpe, M. Liskiewicz, and Ellison G. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2016.

[6] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2020.

[7] M. Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22, 2010.

[8] Karl D Ferguson, Mark McCann, Srinivasa Vittal Katikireddi, Hilary Thomson, Michael J Green, Daniel J Smith, and James D Lewsey. Evidence synthesis for constructing directed acyclic graphs (ESC-DAGs): a novel and systematic method for building directed acyclic graphs. *International Journal of Epidemiology*, 49(1):322–329, 07 2019.

Appendix

A Code and tables for data exploration and pre-processing

```
  //required libraries
  library(dagitty)
  library(dlookr)
  library(ggplot2)
  library(ggcorrplot)
  library(visdat)
  library(bnlearn)


  //load the dataset
  setwd('./Desktop/bayesian/assignment1/')
  lexp <- read.csv("who_life_exp.csv", header = TRUE, stringsAsFactors = TRUE)
  str(lexp)

  //rename variables
  columns <- c("life_expect", "adult_mortality", "infant_mort", "alcohol", "bmi", "
    polio", "gghe.d")
  lexp2 <- lexp[columns]

  //variables selected, summary analysis
  str(lexp2)
  summary(lexp2)

  //missing value analysis
  vis_miss(lexp2)
  diagnose(lexp2)
  print(nrow(na.omit(lexp2)) / nrow(lexp2))
  lexp3 <- na.omit(lexp2)
  str(lexp3)

  //outlier analysis
  diagnose_outlier(lexp3)

  //plotting variable distribution
  My_theme_2 <- function() {
  theme(axis.title=element_text(size=8), axis.text=element_text(size=8))
}
p1 <- ggplot(lexp3, aes(x=life_expect)) + geom_density(fill="#69b3a2", color="#e9ecef
    ", alpha=0.8) + My_theme_2()
```

```
p2 <- ggplot(lexp3, aes(x=adult_mortality)) + geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) + My_theme_2()
p3 <- ggplot(lexp3, aes(x=infant_mort)) + geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) + My_theme_2()
p4 <- ggplot(lexp3, aes(x=alcohol)) + geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) + My_theme_2()
p5 <- ggplot(lexp3, aes(x=bmi)) + geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) + My_theme_2()
p6 <- ggplot(lexp3, aes(x=polio)) + geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) + My_theme_2()
p7 <- ggplot(lexp3, aes(x=gghe.d)) + geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8) + My_theme_2()
gridExtra::grid.arrange(p1, p2, p3, p4, p5, p6, p7, ncol=3)

//correlation matrix
  m1<-lavCor(lexp4)
  m1
```

variables	missing_count	missing_percent
<chr>	*<int>*	*<dbl>*
life_expect	0	0
adult_mortality	0	0
infant_mort	0	0
alcohol	50	1.61
bmi	34	1.09
polio	19	0.611
gghe.d	100	3.21

Figure 7: **Missing values distribution.** Missing value count and percentage for each variable.

	variables	outliers_cnt	outliers_ratio	outliers_mean	with_mean	without_mean
life_expect	life_expect	18	0.6040268	43.4210367	69.40210381	69.55999010
adult_mortality	adult_mortality	98	3.2885906	530.6870214	190.65933358	179.09697639
infant_mort	infant_mort	42	1.4093960	0.1258477	0.03171529	0.03036963
alcohol	alcohol	4	0.1342282	18.3546075	4.89226138	4.87416683
bmi	bmi	25	0.8389262	31.7440000	25.07030201	25.01384095
polio	polio	205	6.8791946	47.7317073	87.04261745	89.94666667
gghe.d	gghe.d	55	1.8456376	9.3378820	3.12471091	3.00788205

Figure 8: **Outliers analysis.** Shows the outlier count for each variable, and several outlier specific statistics.

B Code and tables for network building and testing

```
//BUILDING FINAL NETWORK
g <- dagitty('dag {
adult_mortality [pos="-1.555,-1.249"]
alcohol [pos="0.462,-0.383"]
bmi [pos="0.300,0.281"]
gghe.d [pos="-1.750,0.139"]
infant_mort [pos="0.202,-1.357"]
life_expect [outcome,pos="-0.786,1.193"]
polio [pos="-0.785,-1.446"]
adult_mortality -> life_expect
adult_mortality <-> gghe.d
alcohol -> adult_mortality
alcohol -> gghe.d
alcohol <-> polio
alcohol -> life_expect
bmi -> adult_mortality
bmi <-> infant_mort
bmi <-> polio
gghe.d -> bmi
gghe.d -> life_expect
gghe.d -> polio
gghe.d <-> infant_mort
infant_mort -> life_expect
polio -> adult_mortality
```

```
polio -> infant_mort
}')

//TESTING NETWORK
res <- localTests(g, lexp4, type='cis')
print(res3)
plot(g)
```

```
> print(res1)
                                          estimate      p.value         2.5%        97.5%
alch _||_ poli | ggh.                   0.07709882 2.505383e-05  0.041300746  0.112700424
bmi _||_ lf_x | adl_, alch, ggh., inf_  -0.03924450 3.228090e-02 -0.075069341 -0.003318651
lf_x _||_ poli | adl_, alch, ggh., inf_  0.02765999 1.314134e-01 -0.008278774  0.063527450
> print(res2)
                                          estimate   p.value         2.5%        97.5%
bmi _||_ lf_x | adl_, alch, ggh., inf_  -0.03924450 0.0322809 -0.075069341 -0.003318651
lf_x _||_ poli | adl_, alch, ggh., inf_  0.02765999 0.1314134 -0.008278774  0.063527450
> print(res3)
                                          estimate    p.value         2.5%        97.5%
alch _||_ bmi | ggh.                    0.03377041 0.06533356 -0.002144621  0.069598546
bmi _||_ lf_x | adl_, alch, ggh., inf_  -0.03924450 0.03228090 -0.075069341 -0.003318651
lf_x _||_ poli | adl_, alch, ggh., inf_  0.02765999 0.13141339 -0.008278774  0.063527450
```

Figure 9: **Local conditional independence tests.** Display of all three local conditional independence tests performed on the DAG. Res1 being the output from the initial unmodified network, Res2 the output after adding the latent variable connection Alcohol⟷Polio and Res3 being the output after removing the edge Alcohol→Bmi.

C Code for model fitting

```
m <- toString(g, 'bnlearn')
net <- model2network(m)
fit <- bn.fit(net, as.data.frame(scale(lexp)))
fit
```