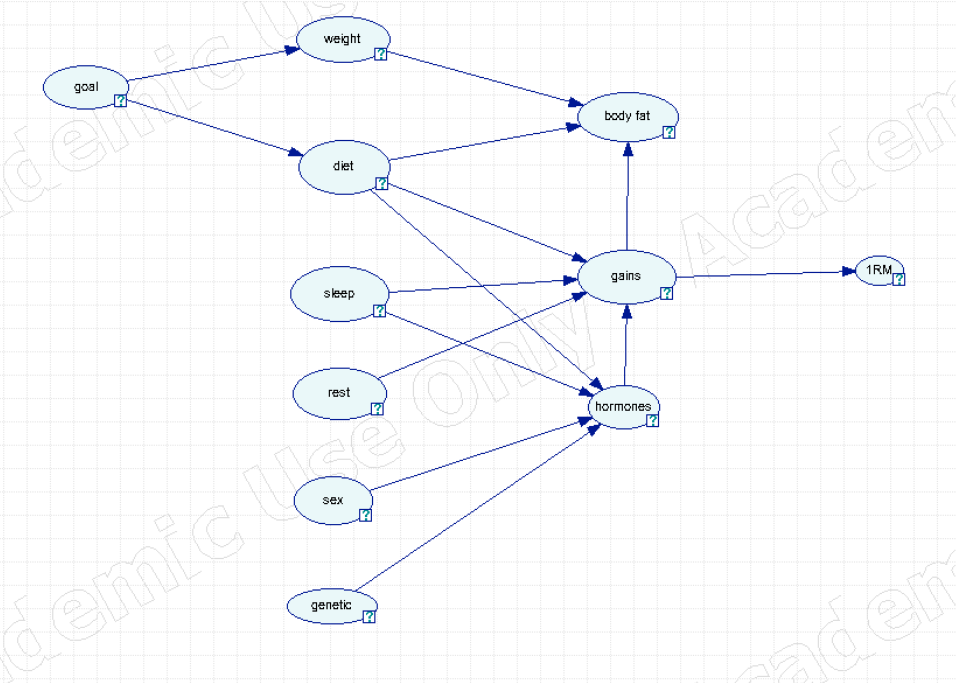
Second Homework: Causal Inference

Francesco Saverio Zuppichini

January 20, 2019

1 Structure of the network

My model represents weightlifting.



The variables are eleven in total. In the program, I renamed the states for better readability. It follows a description of each variable in order of appearance from left to right.

- goal. The final goal of the athlete. It can be bulking, gain muscle, or cut, loose body fat. States are bulk and cut

- weight. If true, the athlete gains weight from last month, if false otherwise. The states are called increased and not_increased

- diet. Represent what the athlete eats. If it's outcome is true, athlete eats around 200˜400 kcal more that his baseline, false otherwise. These states are called surplus and defict

- sleep. How much the athlete sleeps. We assume that sleep time begin around 23:00 ˜ 00:00. The outcome is 1 if the athlete sleeps more that 7.5 hours, 0 otherwise.

- rest. The hours between two workouts. enough if more than 36 hours, low otherwise.

- sex. male and female.

- genetic. Some people are more lucky than others. A good genetic yields to better hormones production thus better gains.

- body fat. increased and not_increased from last month.

- genetic. good or average. The genetic of the athlete.

- hormones. low if hormones production is under average, high if more. Hormones such as gh and testosterone are essential to gains.

- gains. Measured in grams of muscle per month. increase if more than 300g, same otherwise.

- 1RM_increase. In % how much we increase the one range of motion from last month on the base exercises: deadlift, squat and bench press. increase if 5% more than last month, not increase otherwise.

It follows an explanation of the arcs in the graph. body fat depends by the weight, the diet, booth depend of the goal, and the gains. The more you gain the more you weight the more body fat you gain.

gains is the most important node. It depends by weight, diet, sleep, rest and hormone production. The late, directly depend on sex, sleep and diet.

1RM increase depends on the gains.

1.1 State which is the objective of the network: for instance, highlight a couple of situations in which decision making could be difficult and in which the graph could provide valuable indications

Probably a mix of situations for **gains**. For instance, if the athlete sleeps more than 7.5 hours, eat enough but does not rest more than 36 hours. An other could be where the diet prevents a correct hormones production but the athlete trains, sleep and rest well.

1.2 Explaining how you decide the arcs orientation, in case they are not self- explaining

They are all straightforward.

1.3 Which arrows can be reversed without being detectable by a statistical test? Explain why

The following set of edges can be reversed without being detectable by a statistical test. $(e(goal, weight), e(goal, diet), e(gains, 1RMincrease))$

1.4 Identify at least 4 couple of nodes (the node of each couple should be not directly linked to each other) and analyze their d-separation properties possibly conditioning on others

The nodes of the net are denominated by the first two consonant in the name for simplicity.

1. (**goal**, **body_fat**) These are the paths: $\{GL, WT, BF\}, \{GL, DT, BF\}, \{GL, DT, GN, BF\}$, $\{GL, DT, HR, GN, BF\}$
2. (**goal**, **gains**) These are the paths: $\{GL, DT, GN\}, \{GL, WT, BF, GN\}, \{GL, DT, BF, GN\}$, $\{GL, DT, HR, GN\}, \{GL, WT, BF, DT, GN\}, \{GL, WT, BF, DT, HR, GN\}$

   BF is a collider for WT, DT, GN so it already blocks some paths. We can condition on DT to block all paths
3. (**goal**, **hormones**) These are the paths : $\{GL, WT, BF, GN, HR\}, \{GL, WT, BF, DT, HR\}$, $\{GL, WT, BF, DT, GN, HR\}, \{GL, DT, BT, GN, HR\}, \{GL, DT, GN, HR\}\{GL, DT, HR\}$.

   As before, we can block on DT to block all paths between GL and HR.

4. (**diet**, **sleep**) These are the paths: $\{DT, GN, SL\}, \{DT, BF, GN, SL\}, \{DT, HR, GN, SL\}, \{DT, GL, WT, BF, GN, SL\}, \{DT, GL, WT, BF, GN, HR, SL\}, \{DT, HR, SL\}$

   All the paths are already blocked by GN and HR.

1.5 Discuss how d-connected variables are in fact dependent in the real problem, while d-separated variables are instead independent in the real problem.

- `goal` and `weight` are dependet since the goal that we choose determinate the weight we want to have

- `goal` and `diet` are dependent since the goal we picked also must be followed by a correct diet. If we want to loose weight, we must eat less

- `diet` and `sleep`. They independent, since what a person eat does not have repercussion on the sleep time an quality.

- `hormones` and `body fat` are likely dependent since if an athlete produces more hormones it can gains more muscle and increase his body fat. Unfortunally, it is scientific prooved that is impossibile to gain muscle and loose fat at the same time. For the interested reader it follows a very simple explanation. To syntetize new muscle tissue the body needs to have a surplos of energy. Thus we must eat more than our base metabolism needs, this is called 'bulking'. Having more energy leads to gain more weight and some body fat. The amount of body fat is directly proportional at the amount of kcal in surplus. Hoewer, they are several factors that also can influence the amount of body fat in a person, such as a history of bad diet and poor training.

- `gains` and `body fat` are depending. This can be seen very intuitivelly by following the explanation in the last paragraph. Again, if we gain muscle then we must had a surplus of kcal in the dies, so we have gain also some body fat

- `gains` and `1RM increase`. Surely, in we increase the amount of muscle we also will lift more.

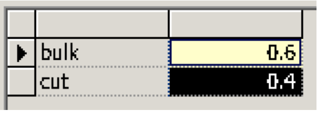
- `rest` and `1RM increase`. Resting between workouts avoid over training and leave to our body the time to build new muscles to be able to adapt and lift more.

  Ab interesting consideration is that if we set `gains` to true and I know the current `diet` is in surplus then I can correctly guess the outcome of `sleep` since if a person eat enough and had gains then it must have slept for a correct amount of time.

2 Conditional probability tables (CPTs)

For the data I have relied on your personal experience/common sense.

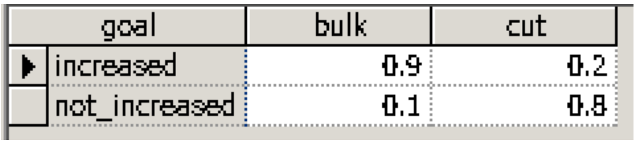
- goal



bulk	0.6
cut	0.4

Usually people that do weightlifting want to build muscle.

- weight

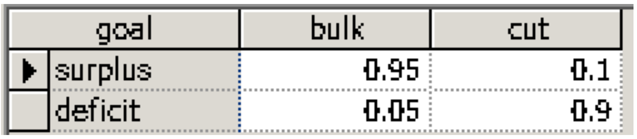


goal	bulk	cut
increased	0.9	0.2
not_increased	0.1	0.8

Of course, if we bulk we expect in almost all cases to increase our weight. Sometimes, due to some other factors, like stress or bad habits, this may influence our way to eat, train etx and thus we are not gain new weight.

If the cut period is not properly planned, some people can eat a lot after a too strong diet and thus exponentially increase their weight instead of reduce it.

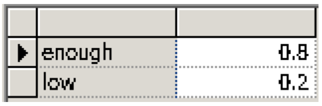
- diet



goal	bulk	cut
surplus	0.95	0.1
deficit	0.05	0.9

If we bulk we have no eat more, since this is very easy we have a very low probably to still be in deficit. This reflect the case where we think we are eating enough but we are still in deficit. Cutting is harder since an athlete needs to correctly tracks is kcal so we have a 10% probability of still beeing in surplus.

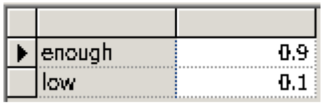
- sleep



▶	enough	0.8
	low	0.2

Some people does not sleep enough. Be aware that sleep does not only take in account the number of hours slept but also the time we go to bed. There is a big difference between an athlete that goes to bed at 23 and wakes up at 7 than one that goes to bed ad 01 and wakes up at 9.

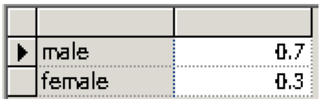
- rest



▶	enough	0.9
	low	0.1

Even if is well know that our body need some rest time between workouts, some people still over train.

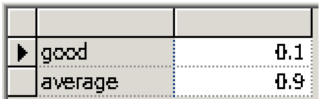
- sex



▶	male	0.7
	female	0.3

Most of the weightlifters are male.

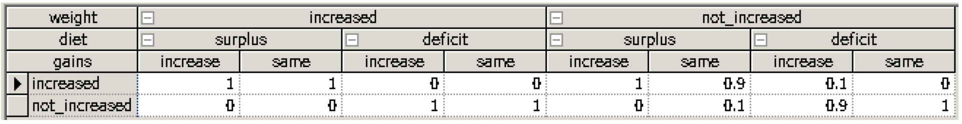
- genetic



▶	good	0.1
	average	0.9

Only a small part of the population has a good genetic

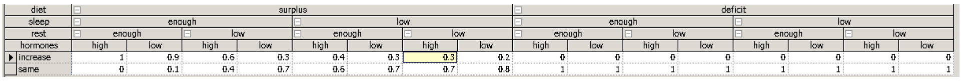
- body fat



weight	increased				not_increased			
diet	surplus		deficit		surplus		deficit	
gains	increase	same	increase	same	increase	same	increase	same
increased	1	1	0	0	1	0.9	0.1	0
not_increased	0	0	1	1	0	0.1	0.9	1

Body fat depends on weight, diet and gains. If we are in surplus then our body fat will increase no matters what.

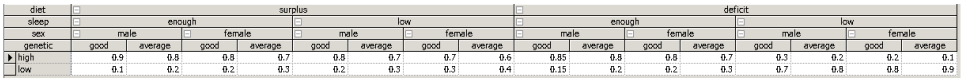
- gains



diet	surplus								deficit							
sleep	enough				low				enough				low			
rest	enough		low		enough		low		enough		low		enough		low	
hormones	high	low	high	low	high	low	high	low	high	low	high	low	high	low	high	low
increase	1	0.9	0.6	0.3	0.4	0.3	0.3	0.2	0	0	0	0	0	0	0	0
same	0	0.1	0.4	0.7	0.6	0.7	0.7	0.8	1	1	1	1	1	1	1	1

Gains are the core node in the graph. They depends on a lot of factor. Intuitivelly, in everything goes well, first column, we will gain muscle. If some of the variables such as rest, sleep and diet are not positive we will probably not gain a lot

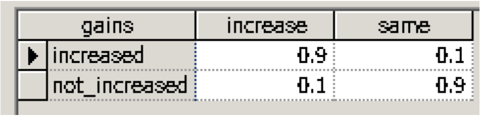
- hormones



diet	surplus								deficit							
sleep	enough				low				enough				low			
sex	male		female		male		female		male		female		male		female	
genetic	good	average	good	average	good	average	good	average	good	average	good	average	good	average	good	average
high	0.9	0.8	0.8	0.7	0.8	0.7	0.7	0.6	0.85	0.8	0.8	0.7	0.3	0.2	0.2	0.1
low	0.1	0.2	0.2	0.3	0.2	0.3	0.3	0.4	0.15	0.2	0.2	0.3	0.7	0.8	0.8	0.9

In our graph, hormones depends on diet, sleep, sex and genetic. With a diet in surplus, enough sleep we see a higher hormones in the population. If we also have a good genetic this value increases. Females have a slightly lower hormone productions than mans. On the other hand, if we don't sleep enough our body won't produce as much hormones as if we slept enough, this is reflected on the lower probabilities on having an hormone's high production showed in the table.

- 1RM



gains	increase	same
increased	0.9	0.1
not_increased	0.1	0.9

3 Causal Inference

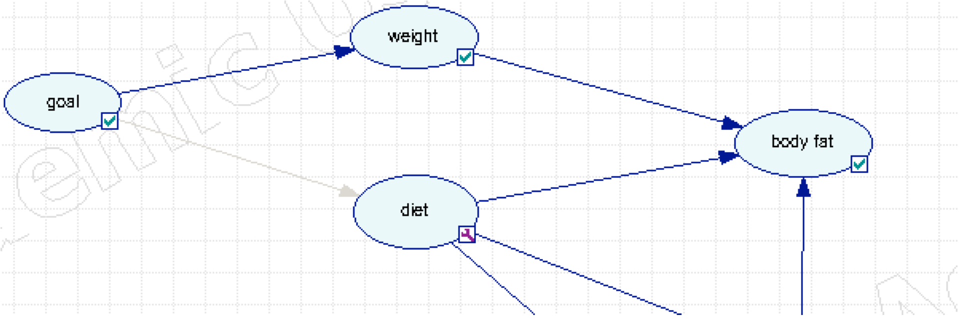
I used the control value feature from genie (`https://support.bayesfusion.com/docs/GeNIe/bn_cont`

3.1 Calculate the causal effect of X on Y.

I pick X equal to *diet* and Y equal to *body fat*. X can assume two values, $deficit$ and $surplus$, while Y can be *increased* and *not_increased*. We first want to find out

$$P(Y = y \mid do(X = x)) \tag{1}$$

We want to observe Y, so by plugging the values into 1, we want to calculate $P(Y = increased \mid do(X = surplus))$ and $P(Y = increased \mid do(X = deficit))$. We can easily do so with genie by controlling the value of X.



We obtain:

$$P(Y = increased \mid do(X = surplus)) = 0.993 \tag{2}$$

$$P(Y = not\_increased \mid do(X = surplus)) = 0.007 \tag{3}$$

The results make sense, since if we eat more than our baseline, $X = surplus$, we have an almost certain probability to weight more than last month. On the other hand, if we eat less than our baseline we are almost sure we can't increase our weight.

3.2 Identify possible confounders between X and Y.

There is no confounder between X and Y

3.3 Would it be practically possible in your specific problem to perform also a randomized controlled study to disentangle the causal effect between the variables from their correlation?

Yes, sure. In my problem perform a randomized controlled study is possible. The diet can be imposed to two groups of the population and then we can observe the effect on body fat.

3.4 Compute the ACE of X on Y

Average Causal Effect (ACE) is defined as follow:

$$P(Y = 1 \mid do(X = 1)) - P(Y = 1 \mid do(X = 0)) \tag{4}$$

Using Equation 4 in our network

$$P(Y = increased \mid do(X = surplus)) = 0.993 \tag{5}$$
$$P(Y = increased \mid do(X = deficit)) = 0 \tag{6}$$

Thus

$$P(Y = increased \mid do(X = surplus)) - P(Y = increased \mid do(X = deficit)) = \\ = 0.993 - 0 = 0.993$$

The ACE is 0.993

3.5 Choose another variable C and calculate the c-specific effect of X on Y.

I choose $C = bulk$ equal to `goal`. In order calculate the c-specific effect, we need to estimate $P(Y = increase \mid do(X = surplus), C = bulk)$. So the probability to gain fat given we eat enough and we are in a bulking phase. Following rule 2 from the book, we also need the set of variables that blocks all the backdoor paths from X to Y, called S, in our case C already blocks all the such paths. Thus

$$\begin{aligned} P(Y = increased \mid do(X = surplus), C = bulk) &= \\ P(Y = increased \mid X = surplus, C = bulk) &= 0.998 \end{aligned}$$

The results makes perfect sense since we want to find out the probability of gain weight if we eat more than the baseline and we are in a bulking phase.

3.6 Identify a minimal set of variables that must be measured in order to estimate the c-specific effect of X on Y.

In this case, C, X and Y since we have not any set S of variable that we need to ensure the paths from X to Y are closed.

3.7 Choose a function g and compute the effect of the conditional intervention of X=g(C) on Y

Let

$$g(x) = \begin{cases} 1 & if\ x = 1 \\ 0 & if\ x = 0 \end{cases} \tag{7}$$

So, by using Equation 3.17 from the book we get:

$$\begin{aligned} P(Y = increase \mid do(X = g(C)) = P(Y = increase \mid do(X = surplus), C = bulk)P(C = bulk) \\ + P(Y = increase \mid do(X = deficit), C = cut)P(C = cut) \end{aligned}$$

From the 2.6 we know

$$P(Y = increase \mid do(X = surplus), C = bulk) = 0.998$$

Thus:

$$\begin{aligned} &P(Y = increase \mid do(X = g(C)) = \\ &0.998 \cdot 0.6 + P(Y = increase \mid do(X = deficit), C = cut)P(C = cut) = \\ &0.59 + 0 = 0.59 \end{aligned}$$

3.8 Identify possible mediating variables between X and Y and calculate the CDE of Y changing the value of X.

There are not mediating variables between X and Y. In the case we have a mediating variable, we need to calculate

$$P(Y = y \mid do(X = x), do(Z = z)) - P(Y = y \mid do(X = x'), do(Z = z)) \tag{8}$$

4 Simulation

The only parent of X is goal, we assume that this variable is not measurable.

4.1 Calculate the causal effect of X on Y.

To calculate the *casual effect* of X on Y we can use again equation 1. We know the casual effect rule (Rule 1 on the book) that we need the parents of X, unfortunately we can not rely on that since goal cannot be measured by assumption. The backdoor criterion (Definition 3.3.1 form the book) cannot help us either since there is not a set Z that satisfies it. The front-door criterion (Theorem 3.4.1) cannot be used since the there is no variable that blocks all the directed path from X to Y.

4.2 Identify possible confounders between X and Y.

There is no confounder between X and Y

4.3 Would it be practically possible in your specific problem to perform also a randomized controlled study to disentangle the causal effect between the variables from their correlation?

Same as 3.3.

4.4 Compute the ACE of X on Y

We cannot computed since `goal` is not measurable by assumption.

4.5 Choose another variable C and calculate the c-specific effect of X on Y.

4.6 Identify a minimal set of variables that must be measured in order to estimate the c-specific effect of X on Y.

We cannot computed since we cannot measure any set of variables S such that $S \cup C$ satisfies the backdoor criterion. See Rule 2 on the book.

4.7 Choose a function g and compute the effect of the conditional intervention of X=g(C) on Y

The c-specific effect cannot be computed, thus we are not able to answer this question.

4.8 Identify possible mediating variables between X and Y and calculate the CDE of Y changing the value of X.

There are not mediating variables between X and Y.

5 Comment on the results

5.1 What kind of experience have you got with this model? E.g., is the causal model responding in a sensible way to your queries? What should be changed/modified to make it more realistic?

Working with this model give me the possibility to use all the material we studied in class. Moreover, it helps me a lot to understand how intervention works. The casual model responded correctly to my queries with any surprises probably due to the fact that the outcome of the variable is already easy to guess by just using logic. To make it more realistic lots of factors should be included, for example the exercises done by the athlete (compound vs isolation), the volume, the intensity and others factors that could in affect the hormones production.