MAC6916

Bayesian Networks and Causality

Outline

1. The data
2. Objective
3. Mean and Linear Regression
4. Bayesian Network
5. Structured Causal Model
6. Conclusion

The data

ENEM 2016: demographic + test results

- Source: https://www.kaggle.com/gbonesso/enem-2016
- Data Dictionary: https://github.com/lgmoneda/mac6916/blob/master/project/Dicionario_Microdados_Enem_2017.xlsx

The data

Students characteristics:

- Age (NU_IDADE)
- Region (transformed SG_UF_RESIDENCIA)
- Gender (TP_SEXO)
- School (TP_ESCOLA)
- Skin color (TP_COR_RACA)
- Monthly Income (Q006)

Variables to filter examples:

- Test attendance (TP_PRESENCA_CN, TP_PRESENCA_CH, TP_PRESENCA_LC, TP_PRESENCA_MT)

Variables of interest:

- Sciences grade (NU_NOTA_CN)
- Human Sciences grade (NU_NOTA_CH)
- Language grade (NU_NOTA_LC)
- Math grade (NU_NOTA_MT)
- Writing grade (NU_NOTA_REDACAO)

The data

Exclude students that...

- are not from public schools either from private;
- were eliminated in one of the tests;
- didn't attended one of the tests;

By doing so, we go from 8.6M examples to 512K.

Variable of interest: sum all grades.

The objective

Evaluate a public policy that plans to pay private schools to public school students aiming the impact of it on their ENEM grade.

Mean and Linear Regression

As a naive approach:

1) **Mean:** Do the difference between the mean grade from public schools and private schools students: **453**
2) **Linear Regression:** Learn a linear regression over the total grade and use the estimated beta for the private school variable: **260**

The results make sense so far. The mean is the naivest approach possible and it would only work if it's a randomized controlled trial, but it's clearly not the case. By assuming the private school is the only difference between the two groups, we hope it's overstimating its impact on the ENEM grade.

The linear regression controls for all the variables and by doing so the effect is splitted by all the variables used to explain the tota score. So we expect it to be lower than the mean, but also there's a change we're controlling unnecessarily for some of the variables, or even allowing confounding by opening a non-causal path between private schools and total grade.

Learning a bayesian network

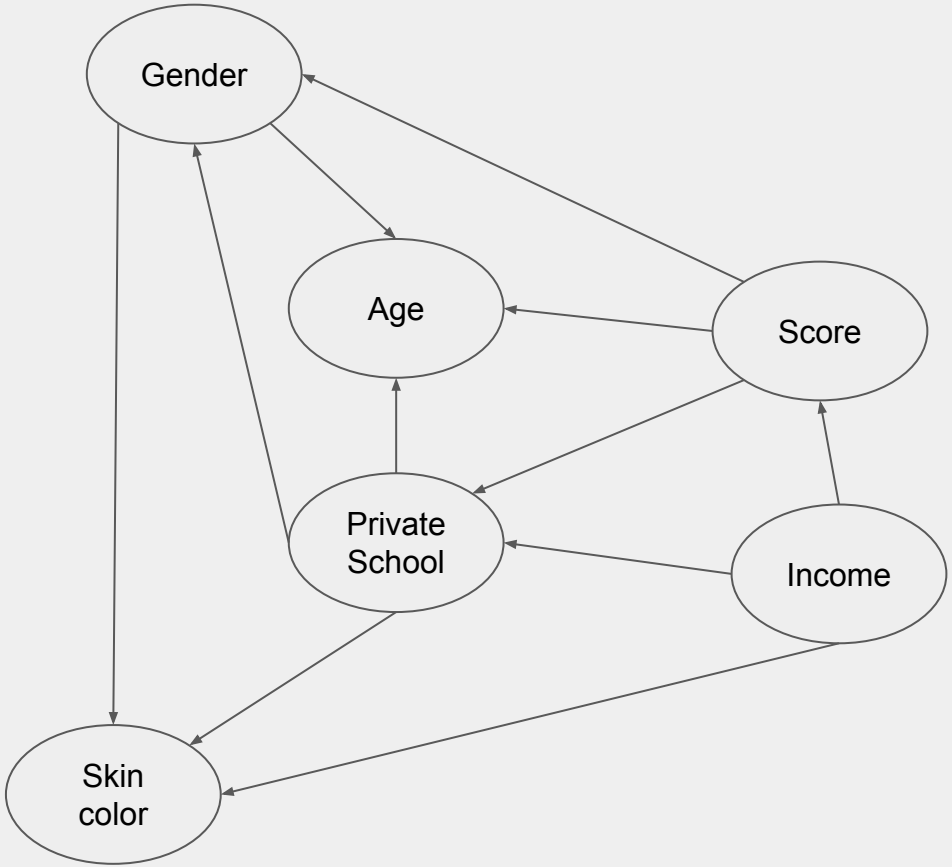
The Bayesian Network has two usages:

1) **Structure / Independences:** We want to learn how the variables involved in our model relate to each other;
2) **Inference:** We want to measure the impact

The Structured will be used for the SCM (Structured Causal Model) and also for the BN itself.

Learning the Structure, the parameters and doing an inference that asks: what would be the mean ENEM score for students from public schools if they were in a private one?

Learning a bayesian network



The question we care about is the impact of Private School on the Score. From a causal view, **we'd expect the arrow to be from Private School to Score**. Also, **Income as a confounder** was the expected thing: it impacts the treatment and the outcome.

Learning a bayesian network - Inference

P(TOTAL_NOTA | PRIVATE_SCHOOL = 0)

TOTAL_NOTA	phi(TOTAL_NOTA)
TOTAL_NOTA_0	0.0001
TOTAL_NOTA_1	0.0017
TOTAL_NOTA_2	0.0340
TOTAL_NOTA_3	0.2046
TOTAL_NOTA_4	0.3499
TOTAL_NOTA_5	0.2895
TOTAL_NOTA_6	0.0000
TOTAL_NOTA_7	0.1110
TOTAL_NOTA_8	0.0091
TOTAL_NOTA_9	0.0001

P(TOTAL_NOTA | PRIVATE_SCHOOL = 1)

TOTAL_NOTA	phi(TOTAL_NOTA)
TOTAL_NOTA_0	0.0004
TOTAL_NOTA_1	0.0174
TOTAL_NOTA_2	0.2105
TOTAL_NOTA_3	0.4790
TOTAL_NOTA_4	0.2328
TOTAL_NOTA_5	0.0533
TOTAL_NOTA_6	0.0000
TOTAL_NOTA_7	0.0063
TOTAL_NOTA_8	0.0002
TOTAL_NOTA_9	0.0001

If we do the difference between the dot product by the probability vector and the median grade value for each category, we find a difference of **500**

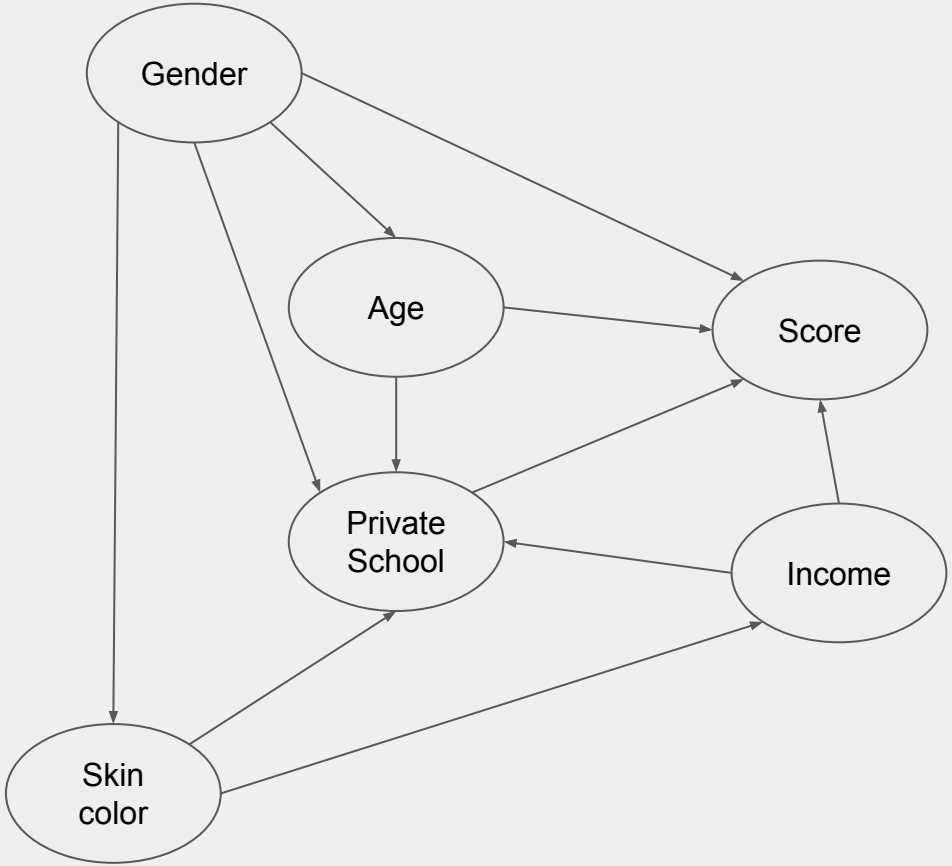
SCM

For SCM we're going to estimate two different models:

1) **From the structure learned by the BN:** Use it as the Causal Graph.
2) **"Fixing" causal directions from the learned BN:**

Using the exact same structure found by the BN the result is **273** (Linear Regression Estimator) and **283** (Propensity Score Stratification).

SCM - Changing directions



Changing the causal direction accordingly to my intuition, the effect is: **265**

Conclusion

The BN result shouldn't be taken seriously because it comes from a discrete result to continuous one due to make it able to compare. The differences in average is clearly the wrong way of doing and it's very different from the other approaches, but the LR and the SCMs have a very close result.

Approach	Effect
Differences in average	**453**
Linear Regression	**260**
BN	**500**
SCM using BN Structure	**273**
SCM using adjusted BN Structure	**265**