Causal Data Science for Business Decision Making

Problem Set

Fall 2021

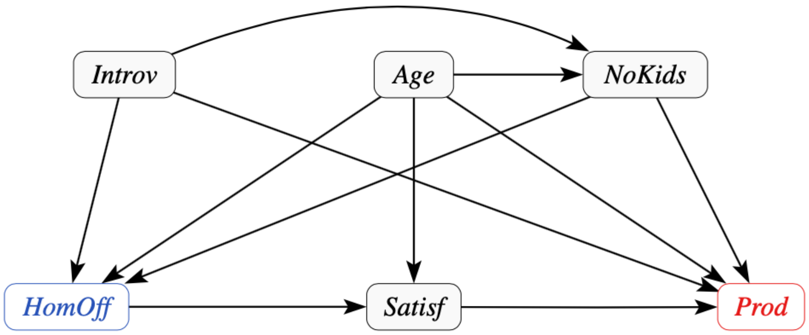
Suppose the company KSI Dronovon wants to save on costs for renting office spaces and therefore thinks about implementing a home office policy for its employees. Currently, already one third of staff regularly works from home (i.e., more than 50% of their time) and people generally seem to be quite happy with the situation. However, top management is worried that working from home might have a negative effect on productivity. To investigate this question, HR conducts a large study based on an extensive employee survey as well as internal records, and collects the following data for 5,364 employees:

$HomOff$:	An indicator whether the employee has worked regularly from home in the past
Age:	Age of the employee
$NoKids$:	Number of kids
$Introv$:	Introversion (based on a personality test that includes several Likert-scale items)
$Satisf$:	Employee satisfaction (on a scale of 1–10)
$Prod$:	Employee productivity

You have been recently hired as a data analyst at KSI Dronovon and because your boss knows about your great analytics skills that you picked up at university, she asked you to help out the HR team with the project. Upon reflecting on the problem, you decide to proceed in the following steps.

1.) Open the data file (`problem_set_data.csv`) in Excel and run a univariate regression of $Prod$ on $HomOff$. What can you conclude about the productivity of employees who are regularly working from home?

2.) After some research and talks with relevant stakeholders in the organization, you come up with the model below. Input the causal diagram into Fusion. What are the testable implications (conditional independencies) of this DAG?



3.) Load the data file into Fusion. Is the data compatible with the testable implications of the DAG?

4.) What happens if you try to learn the model from data alone? Use the PC algorithm and Fisher's z-Transformation as conditional independence test. How does the resulting output differ from the diagram in 2.)?

5.) Go back to the diagram in 2.) What happens if you introduce an unobservable factor that exerts a causal influence on number of kids and productivity: $NoKids \leftarrow - - - - \rightarrow Prod$? Do the testable implications of the DAG change? What kind of variables could this unobserved factor represent?

6.) How many causal paths are there in the causal diagram? And how many biasing paths are there? List all admissible backdoor adjustment sets.

7.) Estimate $P(HomOff|do(Prod))$ using generalized additive models and three-fold cross validation. How does the causal effect of $HomOff$ on $Prod$ differ from the correlation in 1.)?

8.) Now assume that data on employee introversion is not available. Set $Introv$ to a latent node in the causal diagram. Would you still be able to identify $P(HomOff|do(Prod))$?

9.) Now additionally assume that there is a direct effect of working from home on productivity. Add the edge $HomOff \rightarrow Prod$ to the diagram. Would your proposed solution from 8.) still work?

10.) Suppose you could run an experiment in which you randomly pay out commuting allowances to a group of employees. These allowances reduce the cost of commuting for employees and are thus an incentive to come to the office more often. Add the edge $ComAll \rightarrow HomOff$ to the diagram. Does this help to identify the effect of working from home on productivity? Demonstrate how to do this in Fusion. (Bonus question: is z-identification possible in this scenario?)

11.) What could be possible obstacles for implementing an experiment like in 10.)?