Problem set 1

**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?**



*HomeOff ⫫ Prod | Age,Introv,NoKids,Satisf*

*Satisf ⫫ NoKids | Age,HomeOff*

*Introv⫫ Satisf | Age,HomeOff*

*Introv⫫Age*

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

**DAG?**

* Testable Implications is derived from the dag (d-seperation relationship)
* Then looking at the actual data, these can be tested, if validated, it implies a correct causal model
* Remember the choice of test have importance too, e.g., a fischer’s test assumes continuous variables. As such, choosing one over the other, may provide different results and needs to be grounded in some rules

*Model driven approach*

1. Specify model based on background knowledge
2. Check the d-seperatiuon implied by model and see if they hold in the model

*Other way around (causal discovery) 🡪 first exploratory step in a causal discovery, can’t eliminate background knowledge, always be skeptical*

1. Testing all the possible conditional independence relationship implied by the data
2. Construct the causal model compatible with those
   1. Learn model section in fusion

**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.)?**

* Changed from 4 to 5 implications, possibly due to small sample

Graphical user interface, text, application

Description automatically generated

Diagram

Description automatically generated

A picture containing text, indoor

Description automatically generated

**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 ← − − − → 𝑃𝑟𝑜𝑑?**

**Do the testable implications of the DAG change? What kind of variables could this**

**unobserved factor represent?**

* No change
* The variables could represent; available work-time, number of diaper changes, alcohol consumption

**6.) How many causal paths are there in the causal diagram? And how many biasing paths are**

**there? List all admissible backdoor adjustment sets.**

Causal paths: 1 (HomeOff 🡪 Satisf 🡪 Prod)

Biasing paths: 9 (using original dag)

Admissible backdoor adjustment sets: 1 {Introv,Age,NoKids}

* Controlling for these influence factors, we can get to the causal effect on homeOff on prod according to backdoor criterion

**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.)?**

E(X) Higher in the group working from home, contradicting the univariate 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))?**

Yes, but this time via frontdoor criterion

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

No, not identifiable from observational data alone. This is due to the frontdoor criterion requiring that the effect is FULLY mediated by the vatiable, here satis, not confounded by a latent note. Does not work if we have remaining effect, we need full mediation

**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 → 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 zidentification possible in this scenario?)**

Assumming linearity, some may be identified (IV estimation)

Z-identification is not possible if a direct unobserved confounder between the treatment and outcome

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