Homework 9: Circles and Arrows and a Paragraph on the Back Explaining Each One

36-402, Spring 2024

Due at 6 pm on Thursday, 4 April 2024

AGENDA: Practice at using graphical-models rules; thinking about when regressions do and do not give us causal information; distinguishing between different causal structures.

Refer to the DAG from Figure 1, which elaborates on the example from the text, except in Q5.

1. *Parents and children*
   (a) (5) For each variable in the model, list its parents; or, if it has no parents, say so.
   (b) (5) For each variable in the model, list its children. (Some variables have no children.)
2. *Joint distributions and factorization* (10) Using the graph, list the smallest collection of marginal and conditional distributions which must be estimated in order to get the joint distribution of all variables.
3. *Making and breaking dependence*
   (a) (5) Explain, from the graph, why smoking and cancer are dependent.
   (b) (5) Explain, from the graph, why access to dental carw and cancer are dependent.
   (c) (5) List all the sets of variables which we could condition on, in order to make cancer and smoking statistically independent.
   (d) (5) If we have a set of variables which make smoking and cancer statistically independent, can we restore the dependence by adding a conditioning variable? Either give an example, or explain why none could exist.
   (e) (5) What, if anything, do we have to condition on to make yellowing of teeth independent of asbestos exposure? What additional conditioning variable would make them dependent again? Could we make that dependency go away by adding yet another conditioning variable?

4. *Regressions and causal effects* The file `sim-smoke.csv` contains data for the variables in Figure 1.

   (a) (3) Run a logistic regression of cancer on smoking. Report the coefficient on smoking and explain its interpretation.

   (b) (3) Run a logistic regression of cancer on smoking, controlling for yellowing of teeth. Report the coefficient on smoking and explain its interpretation.

   (c) (3) Run a logistic regression of cancer on smoking, controlling for asbestos exposure. Report the coefficient on smoking and explain its interpretation.

   (d) (3) Run a logistic regression of cancer on all the covariates. Report the coefficient on smoking and explain its interpretation.

   (e) (7) What covariates *should* be controlled for, in order to estimate the effect of smoking on cancer? Explain your reasoning for including or excluding each variable as a control.

   (f) (5) Assume Figure 1 gets the causal structure right. Which of these regressions, if any, would be most suitable to a doctor advising patients about whether they need to quit smoking? Carefully explain your reasoning. *Hint:* You can answer this question, and the next, without having actually run the regressions.

   (g) (5) Similarly, which of these models would be most useful to an insurance company deciding how much to charge customers for life or medical insurance? Again, carefully explain your reasoning.

5. *Which DAG?* Consider the DAG from Figure 2.

   (a) (5) Find a conditional independence relation which holds in Figure 2 but not in Figure 1.

   (b) (5) Is there a conditional independence which holds in Figure 1 but not in Figure 2? If so, what is it? If not, explain why not.

   (c) (5) Can you tell whether the data came from Figure 1 or Figure 2? If you can, explain how, and your guess. If you do not think you can, explain why not.

6. (1) *Timing* How long, roughly, did you spend on this assignment? How much of that time was spent on math, on coding/debugging, and on writing?

Presentation Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All plots and tables are generated by code included in the R

Markdown file. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for. Text from the homework assignment, including this rubric, is included only when relevant, not blindly copied.

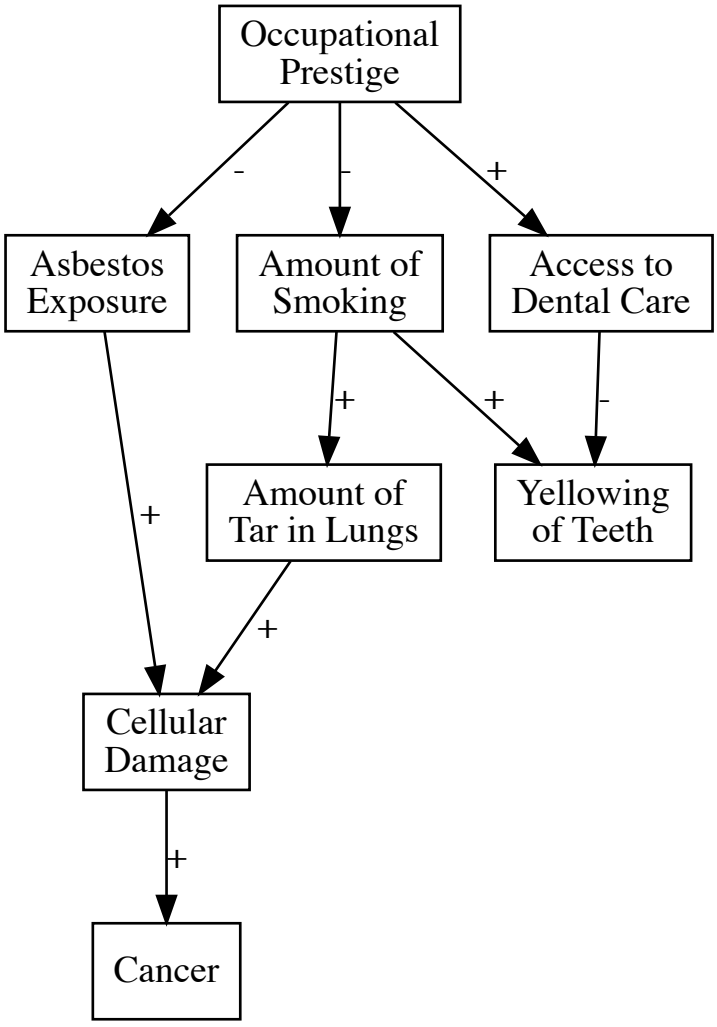
Occupational Prestige
-
-
+
Asbestos Exposure
Amount of Smoking
Access to Dental Care
+
+
-
+
Amount of Tar in Lungs
Yellowing of Teeth
+
Cellular Damage
+
Cancer

Figure 1:

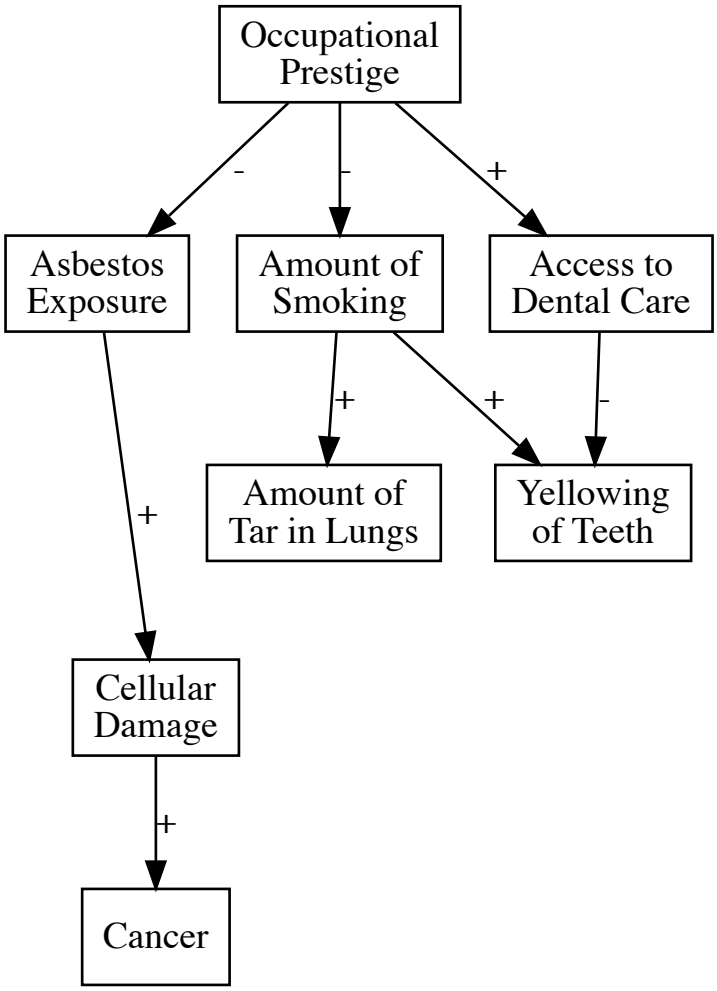
Occupational Prestige
-
-
+
Asbestos Exposure
Amount of Smoking
Access to Dental Care
+
+
-
+
Amount of Tar in Lungs
Yellowing of Teeth
Cellular Damage
+
Cancer

Figure 2: