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

36-402, Spring 2017

Due at 11:59 pm on Wednesday, 12 April 2017

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

TIMING: Problems 1, 2 and 3 are all theory/math; you will need to do some algebra in problems 1 and 2 especially. Problems 4 and 5 combine some theory with fitting some (given) models, and interpreting the estimates, but no novel programming or extensive bootstrapping.

The extra credit requires a *lot* of algebra.

In problems 1 and 2, when questions ask for variances, covariances, or regression coefficients, give your answer as an algebraic expression involving the parameters (α, σ^2, etc.), not particular numbers. As always, show your work for full credit.

General hint for problems 1 and 2: If we are linearly regressing Y on $X_1, X_2, \ldots X_p$, the population coefficients are given by $\beta = \mathbf{v}^{-1}\mathbf{c}$, where $\mathbf{v}$ is the $p \times p$ matrix of covariances among the predictors, $v_{ij} = \mathrm{Cov}[X_i, X_j]$, and $\mathbf{c}$ is the $p \times 1$ matrix of covariances with Y, $c_{i1} = \mathrm{Cov}[X_i, Y]$. (See chapter 2 of the notes.)

1. Consider the DAG from Figure 1. Suppose that $X \sim \mathcal{N}(0,1)$, $Y = \alpha X + \epsilon$, and $Z = \beta_1 X + \beta_2 Y + \eta$, where ϵ and η are mean-zero Gaussian noise with common variance σ^2, uncorrelated with X and with Y.

   (a) (4) Find $\mathrm{Var}[Y]$ and $\mathrm{Var}[Z]$. *Hint:* express them as functions of X, ϵ and η.

   (b) (2) List all the paths linking X and Y which are open when nothing is conditioned on; when conditioning on Z.

   (c) (2) List all the paths linking X and Z which are open when conditioning on nothing; when conditioning on Y.

   (d) (4) Find the covariances between X and Y, between X and Z, and between Y and Z. *Hint:* Use the path-tracing rules from §20.3.2.

(e) (3) Find the population (i.e., true) coefficient for a linear regression of Y on X.

(f) (3) Find the population coefficient for a linear regression of Z on X.

(g) (3) Find the population coefficients for a linear regression of Z on X and Y.

(h) (3) Consider linearly regressing Y on X and Z. Find the population coefficient for X.

2. Consider the DAG from Figure 2. Suppose that $U \sim \mathcal{N}(0,1)$, $X = \alpha U + \epsilon$, $R = \beta X + \eta$, $Q = \gamma_1 X + \gamma_2 R + \zeta$, and $Y = \delta_1 R + \delta_2 Q + \delta_3 U + \xi$, where $\epsilon, \eta, \zeta, \xi$ are independent Gaussian noises with mean zero and variance σ^2.

   (a) (4) List all the paths linking X to Y which are open when conditioning on nothing; when conditioning on U; when conditioning on R; when conditioning on Q; when conditioning on R and Q; when conditioning on U, R and Q.

   (b) (4) List all the paths linking R to Y which are open when conditioning on nothing; when conditioning on X; when conditioning on X and U; when conditioning on X, U and Q.

   (c) (4) List all the paths linking X to Q which are open when conditioning on nothing; when conditioning on Y; when conditioning on U; when conditioning on U and Y; when conditioning on R; when conditioning on R and Y.

   (d) (5) Find the 5×5 matrix of variances and covariances among U, X, R, Q and Y.

   (e) (2) Find the population coefficient on X in a linear regression of Y on X (alone).

   (f) (2) Find the population coefficient on X in a linear regression of R on X.

   (g) (3) Find the population coefficients in a linear regression of Q on X and R.

3. Consider the DAG from Figure 3, which elaborates on the example from class.

   (a) (1) Explain, from the graph, why smoking and cancer are dependent.

   (b) (3) List all the sets of variables which we could condition on, in order to make cancer and smoking statistically independent.

   (c) (4) 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.

(d) (4) 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. The file `smoke.csv` contains data for the variables in Figure 3.

   (a) (2) 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) (5) Assume Figure 3 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.

   (f) (5) Similarly, which of these models would be most useful to an insurance company setting a premium? Again, carefully explain your reasoning.

5. Consider the DAG from Figure 4.

   (a) (3) Find a conditional independence relation which holds in Figure 4 but not in Figure 3.

   (b) (2) Is there a conditional independence which holds in Figure 3 but not in Figure 4? If so, what is it? If not, explain why not.

   (c) (4) Can you tell whether the data came from Figure 3 or Figure 4? If you can, explain how, and your guess. If you do not think you can, explain why not.

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. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. 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 quantitative and mathematical claims are supported by appropriate derivations, included in the text,

or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.

Extra credit (5 points): In Figure 2, find the population coefficients for a linear regression of Y on R and Q.

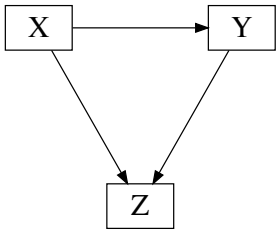

Figure 1:

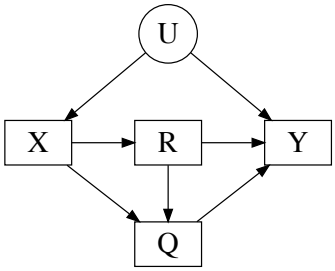

Figure 2:

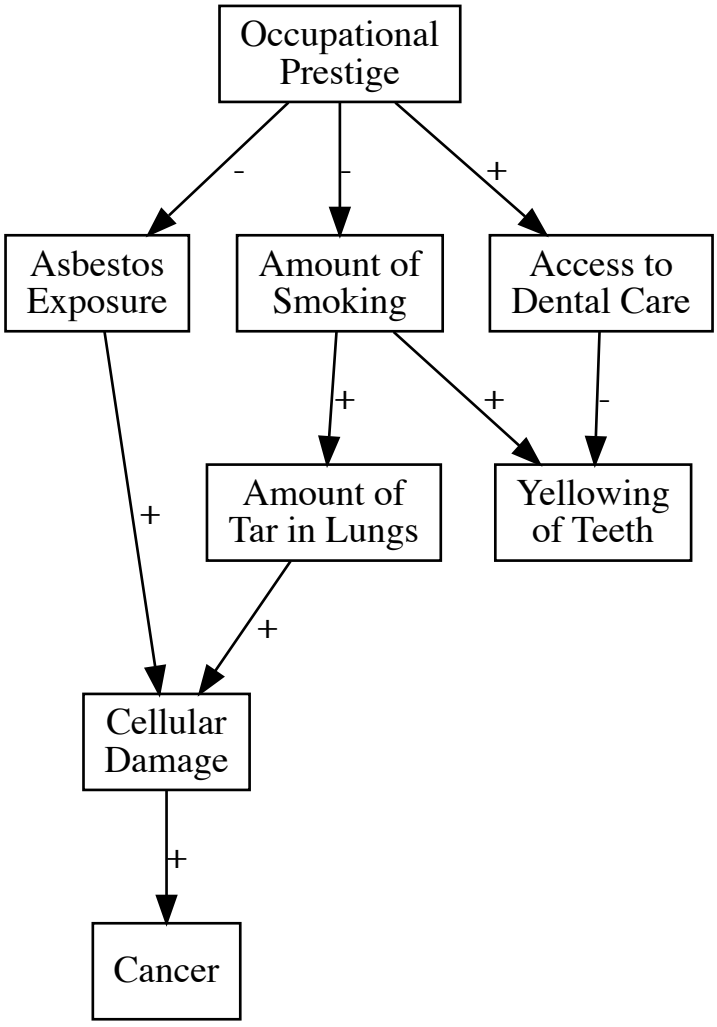
Occupational Prestige
-
-
+
Asbestos Exposure
Amount of Smoking
Access to Dental Care
+
+
-
+
Amount of Tar in Lungs
Yellowing of Teeth
+
Cellular Damage
+
Cancer

Figure 3:

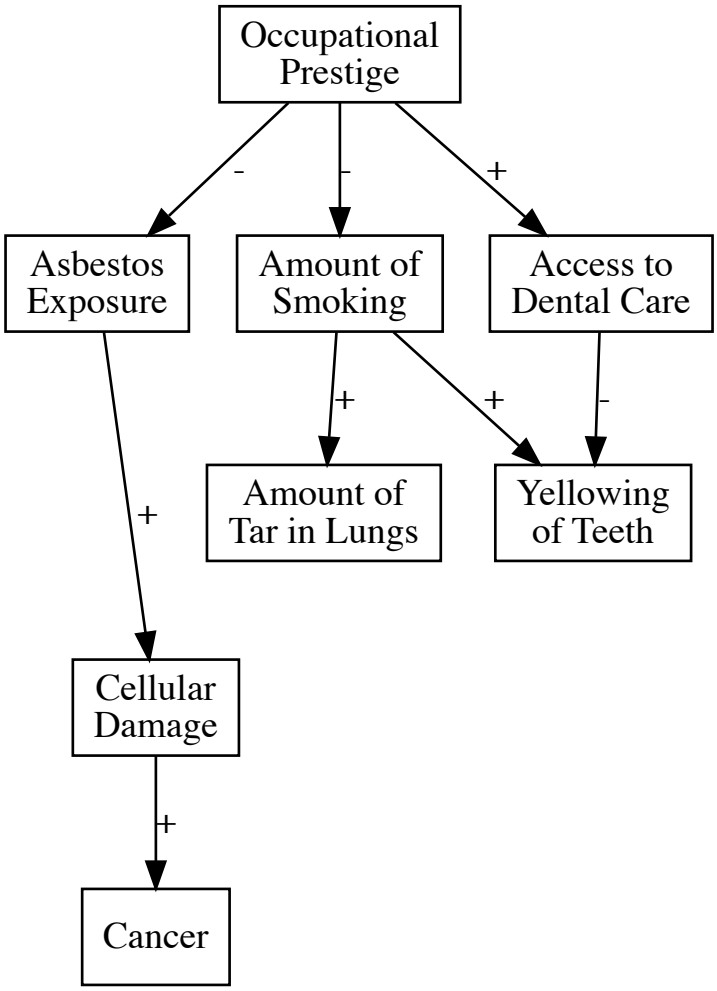
Occupational Prestige
-
-
+
Asbestos Exposure
Amount of Smoking
Access to Dental Care
+
+
-
+
Amount of Tar in Lungs
Yellowing of Teeth
Cellular Damage
+
Cancer

Figure 4: