Homework 10: Estimating with DAGs

36-402, Advanced Data Analysis

Due at the start of class, Tuesday, 19 April 2011

This homework will illustrate some of the advantages of using a known DAG structure. You will need to read the lectures on graphical models carefully in order to do it.

Figure 1 is an elaboration of the graph used in lectures. All problems refer to it, unless otherwise specified.

The file `fake-smoke.csv` contains some (synthetic) data, for use in problem 5.

1. *Parents and children* (10 points)
   (a) (5 points) For each variable in the model, list its parents; or, if it has no parents, say so.
   (b) (5 points) For each variable in the model, list its children. (Some variables have no children.)
2. *Joint distributions and factorization* (10 points) 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. *Associations* (20 points) Should there be a positive association, a negative association, or no association between the following variables? Explain with reference to the graph. (2 points each)
   (a) Yellowing of teeth and cancer?
   (b) Yellowing of teeth and cancer, controlling for smoking?
   (c) Yellowing of teeth and cancer, controlling for occupational prestige?
   (d) Yellowing of teeth and cancer, controlling for smoking and exposure to asbestos?
   (e) Smoking and cancer, controlling for the amount of tar in the lungs?
   (f) Asbestos and cancer, controlling for cellular damage?
   (g) Smoking and cancer, controlling for asbestos?
   (h) Smoking and asbestos, controlling for cellular damage?
   (i) Tar in lungs and cancer, controlling for asbestos, smoking, and yellowing of teeth?

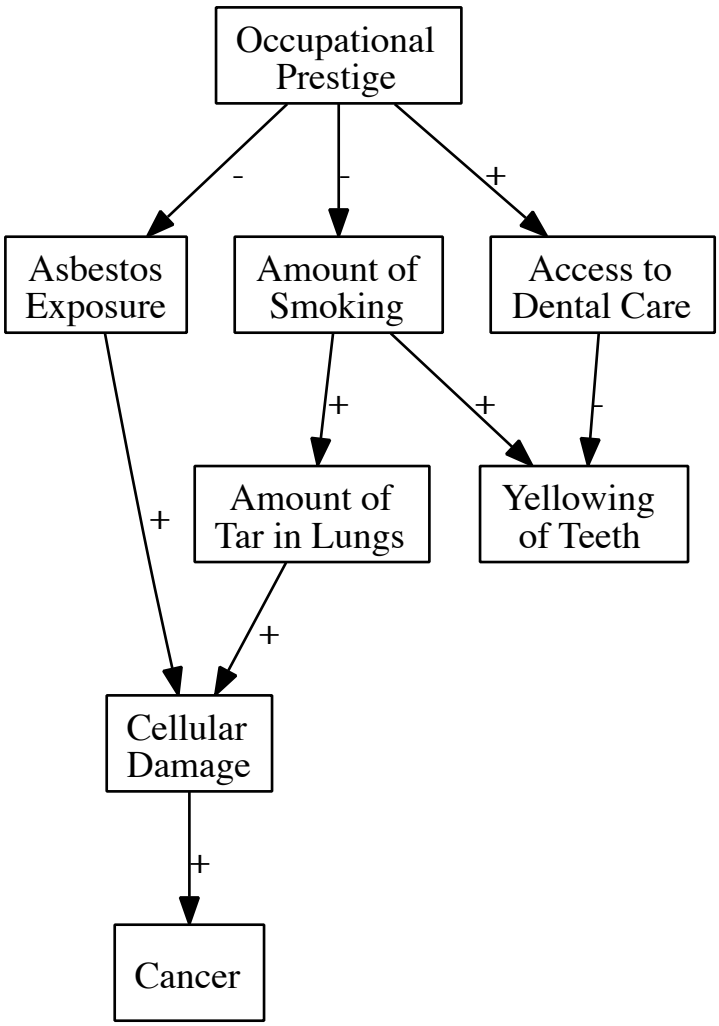

Figure 1: Graphical model for use in all problems, except part of the last. Signs on arrows indicate the sign of the associations (not necessarily linear) between parents and children.

(j) Smoking and cancer, controlling for asbestos and occupational prestige?

4. *Using conditional independence to specify regressions* (40 points)

   (a) (10 points) We wish to know the conditional risk of cancer given smoking. What other variables should be controlled for? Which other variables do not need to be controlled for?

   (b) (10 points) Using the `fake-smoke.csv` data from the class website, fit a logistic regression model for the risk of cancer given the level of smoking, controlling for any appropriate covariates.

   (c) (10 points) Using the same data set, fit another logistic regression for the risk of cancer using *all* the covariates. What does this say about the relationship between smoking and cancer? Why is this different than what is implied by the model in 4b?

   (d) (5 points) A medical insurance company needs to predict the risk of cancer among customers in order to set rates. Should it use the model from 4b or the one from 4c? Why? (Assume, for the sake of the problem, that the training data and the insurance customers are both representative samples of the general population.)

   (e) (5 points) A doctor wants to advise their patients about what actions to take to reduce their risk of cancer. Should they use the model from 4b or 4c? Why?

5. (20 points) Consider the alternative graph in Figure 2.

   (a) (10 points) Repeat problem 3 with the new graph. Clearly indicate in your response which associations differ for the two DAGs.

   (b) (10 points) Suggest an experiment, or an observational analysis, which could let us check which structure was right; explain, in terms of the graphs.

6. (10 points) Extra Credit: Which DAG did the example data come from? How can you tell?

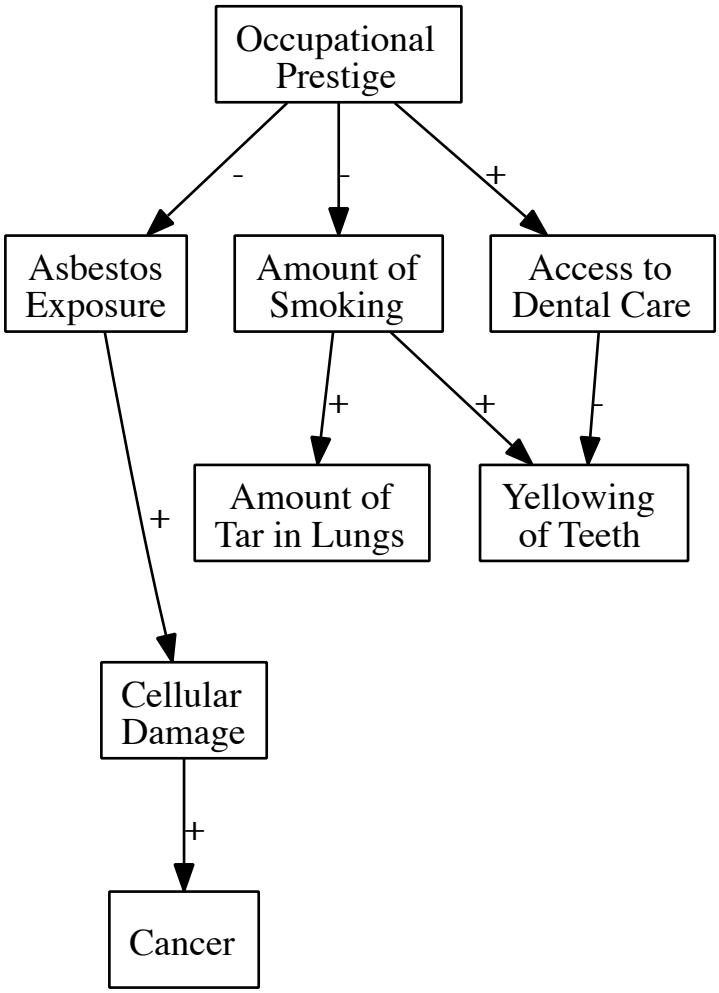

Figure 2: An alternative DAG for the same variables.