

# Homework Assignment 10: Estimating with DAGs

36-402, Advanced Data Analysis, Spring 2011

## Solutions

1. (a) ANSWER:

Variable	Parents
cancer	cellular damage
cellular damage	tar, asbestos
tar	smoking
teeth	smoking, dental care
dental care	occupation
smoking	occupation
asbestos	occupation
occupation	None

(b) ANSWER:

Variable	Parents
cancer	None
cellular	cancer
tar	cellular
teeth	None
dental	teeth
smoking	tar, teeth
asbestos	cellular
occupation	asbestos, smoking, dental

2. ANSWER: In any graphical model, the joint distribution “factors according to the graph”:

$$p(X_1, X_2, \dots, X_p) = \prod_{i=1}^p p(X_i | X_{\text{parents}(i)})$$

where if  $X_i$  has no parents, we read  $p(X_i | X_{\text{parents}(i)})$  as just the marginal distribution  $p(X_i)$ . Here, there is only one variable with no parents, “occupation”, so we start with its marginal distribution:  $p(\text{occupation})$ . Then we need the conditional distributions of its children,  $p(\text{dental} | \text{occupation})$ ,  $p(\text{smoking} | \text{occupation})$ , and  $p(\text{asbestos} | \text{occupation})$ . Next we move on to the children of these variables:  $p(\text{tar} | \text{smoking})$  and  $p(\text{teeth} | \text{smoking, dental})$ .

(Notice that since “teeth” has two parents, we need to condition on both of them.) Then  $p(\text{cellular}|\text{asbestos}, \text{tar})$ , and finally  $p(\text{cancer}|\text{cellular})$ .

3.
  - (a) ANSWER: Teeth and Cancer share Occupation and Smoking as ancestors, so they are dependent. All paths from teeth to cancer have a positive product of signs, so these two are positively associated.
  - (b) ANSWER: Still dependent (unblocked path through dental to occupation to asbestos, positive product of signs).
  - (c) ANSWER: Dependent, unblocked path through smoking, positive association.
  - (d) ANSWER: All paths are blocked, therefore they are independent and there is no association.
  - (e) ANSWER: Dependent, path from smoking to occupation to asbestos to cancer, positive sign.
  - (f) ANSWER: Independent (therefore no association), as conditioning on cellular damage blocks the asbestos  $\leftarrow$  occupation  $\rightarrow$  smoking  $\rightarrow$  tar  $\rightarrow$  damage path.
  - (g) ANSWER: The path through tar and cellular damage is open, so dependent, with positive association.
  - (h) ANSWER: Conditioning on a common effect makes them dependent, and would produce a negative association (see the first full paragraph on page 8, lecture 21), but they are already dependent with a positive association from occupational prestige, so while they are dependent the sign of the association is indeterminate.
  - (i) ANSWER: Dependent, positive (only the straight path from tar to cancer is unblocked; teeth is a collider, and conditioning on it activates it, but all paths which involve it are blocked by conditioning on the non-colliders at smoking and asbestos).
  - (j) ANSWER: Dependent, positive. Conditioning on occupation blocks all paths from smoking to cancer except smoking  $\implies$  tar  $\implies$  damage  $\implies$  cancer.
4.
  - (a) ANSWER: To estimate the conditional risk of cancer given smoking, we merely regress cancer on smoking using our favorite regression technique (logistic regression might work, or a generalized additive model).

A more interesting question is to ask whether smoking *causes* cancer. of cancer on smoking. To answer this, we should control for any other factor that could potentially allow us to predict cancer from smoking, via an indirect chain of dependence. We should *not* control for any variables on the directed path from smoking to cancer. Examining the figure (and our previous answers), every indirect path from smoking to cancer goes through asbestos, so conditioning on that would

control for the indirect dependence, not create any dependence (it is a non-collider), and not block the direct path. We could also control for occupation. If we “control for” yellow teeth, we activate a collider and would have to block it by conditioning on occupation or asbestos or dental care. This would still be an unbiased estimate of the causal effect, but a needlessly noisy one, since controlling for things we don’t need to uses up degrees of freedom. The two minimal sets of controls are asbestos, by itself, and occupation, by itself.

(b) ANSWER: We implement the model discussed in part (a):

```
logist.b = glm(cancer~smoking+asbestos, data = smoke, family = binomial)
summary(logist.b)
```

```
# model coefficients:
```

```
Coefficients:
```

```

              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.494553   0.176858 -14.105 < 2e-16 ***
smoking      0.365560   0.168330   2.172  0.0299 *
asbestos     0.052367   0.009247   5.663 1.49e-08 ***
Null deviance: 714.09 on 600 degrees of freedom
Residual deviance: 338.13 on 598 degrees of freedom
```

The model finds, with marginal statistical significance, that smoking causes cancer. Before taking this conclusion very seriously we should pay attention to the goodness-of-fit of the model (see lecture 14). Note that the residual deviance (338) is vastly lower than the null deviance (714) with about the same degrees of freedom. This says that the model fits much better than if we had only fitted an intercept, which is a good sign. As a sanity check we can also compare the predictions to the truth:

```
# model fit:
```

```
table(smoke$cancer, round(logist.b$fitted.values))
  0   1
0 427   5
1  48 121
```

The table, with truth tabulated on the rows and the estimates in the columns, shows that the model correctly fits 427 of the 432 no-cancer cases, and and correctly fits 121 of the 169 cancer cases.

(c) ANSWER: We fit a logistic regression model using all covariates:

```
logist.c = glm(cancer~., data = smoke, family = binomial)
summary(logist.c)
```

```
# model coefficients:
```

```
Coefficients:
```

```

              Estimate Std. Error z value Pr(>|z|)
```

(Intercept)	-7.99767	4.66365	-1.715	0.0864	.
cellular	-2.28934	3.03429	-0.754	0.4506	
tar	24.01178	23.38684	1.027	0.3046	
teeth	9.87549	7.41977	1.331	0.1832	
dental	2.35654	2.24426	1.050	0.2937	
smoking	-0.04837	1.23852	-0.039	0.9688	
asbestos	0.16040	0.15257	1.051	0.2931	
occupation	-0.11653	0.28330	-0.411	0.6808	
Null deviance: 714.09 on 600 degrees of freedom					
Residual deviance: 334.28 on 593 degrees of freedom					

The model fit as assessed by deviance is approximately the same as before. This time, however, the estimated effect for smoking is that it slightly reduces the risk of cancer. However, this effect is not even close to being statistically significant. In fact, nothing is, not even cellular damage, which (in the model) is the true direct cause of cancer.

- (d) ANSWER: The insurance company should use the model from part (c). Assuming that the insurance company will not involve itself with the personal health decisions of its clients (telling them whether or not to smoke), the company is interested only in estimating the cancer risk for each client given all factors. What *causes* the risk is not important to them.
- (e) ANSWER: The doctor should use the model from part (b), which attempts to reveal the causative affect of smoking on cancer. Normally the patient is interested in knowing how to change their lifestyle to minimize cancer risk, and less interested in indirect statistical clues about their risk factors.
5. (a) ANSWER:
- Teeth and cancer are still positively associated through occupation. **Difference:** the path through smoking is now blocked.
  - Same as part (3b).
  - Difference:** all paths blocked, no association.
  - Same as part (3d).
  - Same as part (3e).
  - Same as part (3f), minus the explanation for the path through tar and damage, which is now missing.
  - Difference:** now independent, without the tar  $\rightarrow$  damage path.
  - Difference:** still dependent, but they are dependent and positively associated regardless of whether we control for damage.
  - Difference:** independent, since we are controlling for asbestos and since the tar  $\rightarrow$  damage path is gone.
  - Difference:** independent, all paths are blocked.

- (b) ANSWER: Any of the relations which switched from being dependent in problem 3 to being independent in problem 5 could be used to distinguish between the two DAGs. In particular, in the first DAG, cancer is dependent on tar after controlling for smoking, but in the second DAG cancer and tar are independent given smoking. This thought is continued in the extra credit.
6. EXTRA CREDIT To distinguish between the two DAGs, we look at whether  $\text{cancer} \perp\!\!\!\perp \text{tar} \mid \text{smoking}$  (DAG 1), or whether  $\text{cancer} \perp\!\!\!\perp \text{tar} \mid \text{smoking}$  (DAG 2). Specifically, let's look at the conditional distribution  $p(\text{cancer} \mid \text{tar}, \text{smoking})$ : in DAG 2, tar should drop out of this as irrelevant, but not in DAG 1. To avoid parametric specification issues, I'll use a non-parametric conditional density estimator, as in Lecture 6:

```
library(np)
tar.npc <- npcdens(factor(cancer)~smoking+tar,data=smoke)
```

(We need to let `np` know that `cancer` is a categorical variable, hence the `factor()` wrapper.) After a little thought, we get a fitted conditional density function. Examining the bandwidths shows that smoking, rather than tar, has been smoothed away almost entirely:

```
>summary(tar.npc)
```

```
Conditional Density Data: 601 training points, in 3 variable(s)
(1 dependent variable(s), and 2 explanatory variable(s))
```

```

                factor(cancer)
Dep. Var. Bandwidth(s):  3.200268e-14
                smoking      tar
Exp. Var. Bandwidth(s): 1149858 0.02514876
```

```
Bandwidth Type: Fixed
Log Likelihood: -169.9323
```

```
Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 2
```

```
Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Dependent Vars.: 1
```

The bandwidth of `smoking` is over a million, while its standard deviation is 1.4. The bandwidth of `tar`, on the other hand, is quite small compared to its own standard deviation.

Plotting like so (after Lecture 6)

```
tar.seq <- seq(from=min(smoke$tar),to=max(smoke$tar),length.out=100)
smoke.seq <- seq(from=min(smoke$smoking),to=max(smoke$smoking),length.out=100)
tar.grid <- expand.grid(tar=tar.seq,smoking=smoke.seq,cancer=1)
tar.npc.predict <- predict(tar.npc,newdata=tar.grid)
library(lattice)
levelplot(tar.npc.predict~tar.grid$tar*tar.grid$smoking)
```

produces Figure 1. The contours of equal probability are almost (but not quite) lines of equal levels of tar. So, very far from being independent of tar given smoking, cancer is almost independent of smoking given tar. This is not at all what we would expect under DAG 2, but quite in line with DAG 1. We conclude that the data came from DAG 1. (Which they did.)

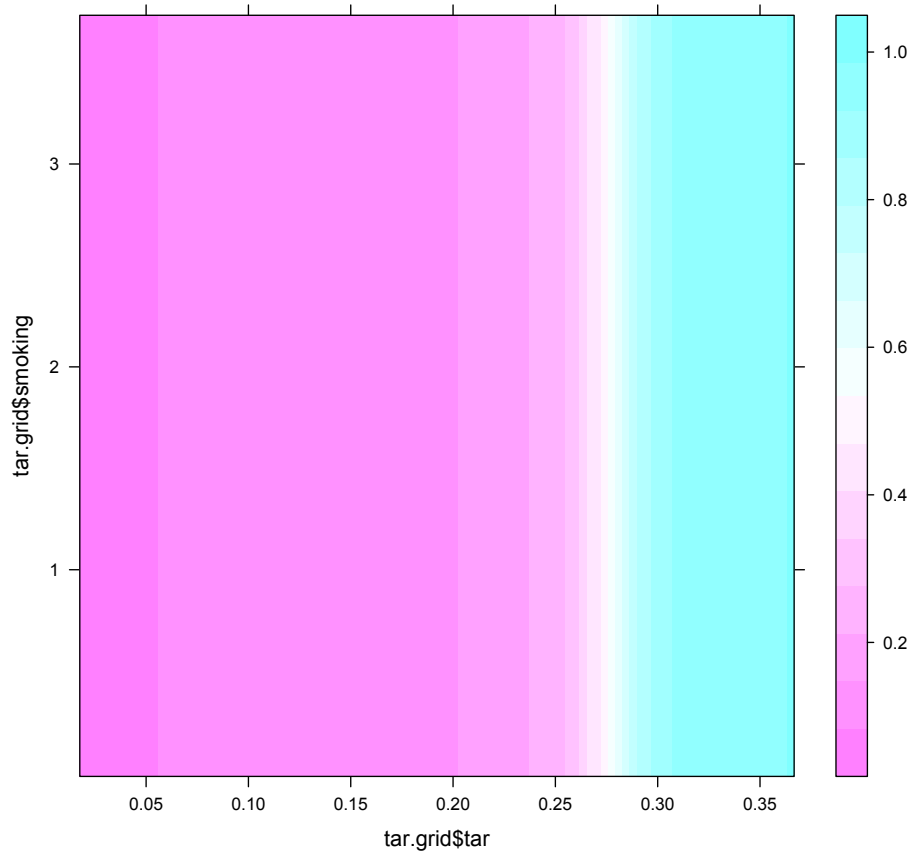


Figure 1: Conditional probability of cancer (indicated by color, see bar at right) as a function of tar levels (horizontal axis) and smoking (vertical axis).