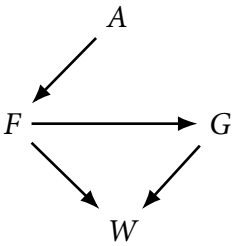
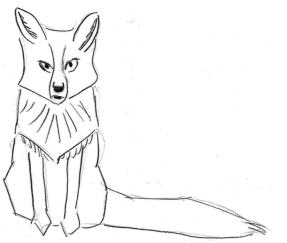
STATISTICAL RETHINKING 2022
WEEK 3

As usual, there are three normal problems and one optional challenge problem. Please submit your homework through the URL provided via email by Friday 28 January 2022.

1. The first two problems are based on the same data. The data in `data(foxes)` are 116 foxes from 30 different urban groups in England. These fox groups are like street gangs. Group size (`groupsize`) varies from 2 to 8 individuals. Each group maintains its own (almost exclusive) urban territory. Some territories are larger than others. The `area` variable encodes this information. Some territories also have more `avgfood` than others. And food influences the `weight` of each fox. Assume this DAG:



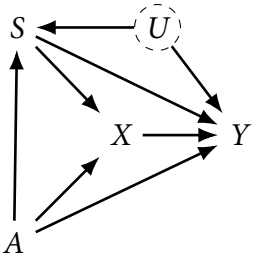


where F is `avgfood`, G is `groupsize`, A is `area`, and W is `weight`.

Use the backdoor criterion and estimate the total causal influence of A on F. What effect would increasing the area of a territory have on the amount of food inside it?

2. Now infer both the **total** and **direct** causal effects of adding food F to a territory on the weight W of foxes. Which covariates do you need to adjust for in each case? In light of your estimates from this problem and the previous one, what do you think is going on with these foxes? Feel free to speculate—all that matters is that you justify your speculation.

3. Reconsider the Table 2 Fallacy example (from Lecture 6), this time with an unobserved confound U that influences both smoking S and stroke Y. Here's the modified DAG:



First use the backdoor criterion to determine an adjustment set that allows you to estimate the causal effect of X on Y, i.e. $P(Y|\text{do}(X))$. Second explain the proper interpretation of each coefficient implied by the regression model that corresponds to the adjustment set. Which coefficients (slopes) are causal and which are not? There is no need to fit any models. Just think through the implications.

4-OPTIONAL CHALLENGE. Write a synthetic data simulation for the causal model shown in **Problem 3**. Be sure to include the unobserved confound in the simulation. Choose any functional relationships that you like—you don't have to get the epidemiology correct. You just need to honor the causal structure. Then design a regression model to estimate the influence of X on Y and use it on your synthetic data. How large of a sample do you need to reliably estimate $P(Y|\text{do}(X))$? Define "reliably" as you like, but justify your definition.