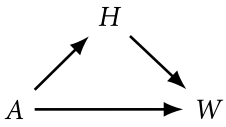
STATISTICAL RETHINKING 2025
WEEK 2 SOLUTIONS

1. The DAG you need is:



To turn this DAG into a generative model, we simulate *H* and *W* from values of *A*. We also need to make assumptions about how *A* influences *H* and *W*, as well as how *H* influences *W*.

The most important thing is to get the order right for simulating each variable. We start with variables that have no parents—this means there are no arrows into them. Then we can simulate variables in turn for which we have already simulated their parents. And so on, until we have simulated all of the variables. THis might sound confusing, because it is, but in the context of an example, it makes more sense.

In this case, the only variable with no parents is *A*. But we will provide *A* as an input to the simulation. That way we can easily have a uniform grid of ages. Next we find variables that are descendants of *A* only. Both *H* and *W* are descendants of *A*. But *W* is also a descendant of *H*. So we need to simulate *H* first, then *W*.

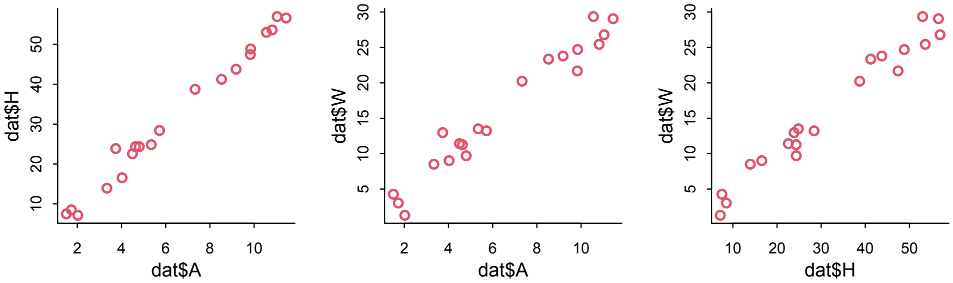
Here is a template function, based on the example from the book/lecture.

```
sim_HW <- function(A,bAH=5,bHW=0.5,bAW=0.1) {
    N <- length(A) # number of individuals
    H <- rnorm(N,bAH*A,2)
    W <- rnorm(N,bHW*H+bAW*A,2)
    data.frame(A,H,W)
}
```

The important thing is not the values I've chosen for the causal effects. You should play around with those values—they are not hypotheses or anything of the kind. They are parameters to explore and understand.

Let's make a example synthetic sample and plot it, to see how these simulated people look. Remember, we are considering only ages under 13.

```
dat <- sim_HW( runif(20,1,12) )
plot( dat$A , dat$H , lwd=2 , col=2 )
plot( dat$A , dat$W , lwd=2 , col=2 )
plot( dat$H , dat$W , lwd=2 , col=2 )
```



These relationships are not very realistic. The simulated children are too short, and young kids also do not grow in a linear fashion. But the simulation is structured like the DAG at least.

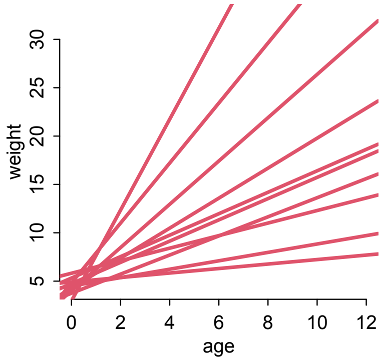
2. Since we want the total effect of age, we just need a linear regression of weight on age. Why? Because if we include height in the regression, then the coefficient on age will not be what we want to know. Age acts through height, but there is no need to measure height to get the total effect of age (according to the DAG). One way to better understand this is that every causal relationship has mediators, variables between the cause and the outcome of interest. Those mediators are usually unmeasured. Ignoring them is fine (usually).

For example a light switch is a cause of a lightbulb turning on. But there are many mediators in this system—the wire, the electricity, and many others. We don't have to measure those to estimate the total association between the switch and the bulb.

Let's set up the data and then simulate some priors.

```
library(rethinking)
data(Howell1)
d <- Howell1
d <- d[ d$age < 13 , ]

# sim from priors
n <- 10
a <- rnorm(n,5,1)
b <- runif(n,0,10)
plot( NULL , xlim=range(d$age) , ylim=range(d$weight) ,
    xlab="age" , ylab="weight" )
for ( i in 1:n ) abline( a[i] , b[i] , lwd=3 , col=2 )
```



These were my first guess, given that the relationship must be positive and that weight at age zero is birth weight, and average birth weight is around 5 kilograms (but varies a lot).

Here's the model.

```
m2 <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + b*A,
        a ~ dnorm(5,1),
        b ~ dunif(0,10),
        sigma ~ dexp(1)
    ), data=list(W=d$weight,A=d$age) )
precis(m2)
```

```
      mean   sd 5.5% 94.5%
a     7.17 0.34 6.62  7.71
b     1.38 0.05 1.29  1.46
sigma 2.51 0.15 2.28  2.74
```

The total causal effect of each year of growth is given (in this case) by the parameter b. So its 89% interval is 1.29 to 1.46 kilograms per year. There is nothing to marginalize in this case, because there are no covariates.