Assignment 2 - Methods 4 - Group 03 (Lydia Pauli Rambrand, Frederik E. Solberg, Alexandra Ciulisová, Valeria Alladio)

Laurits Lyngbaek

2025-03-18

Second assignment

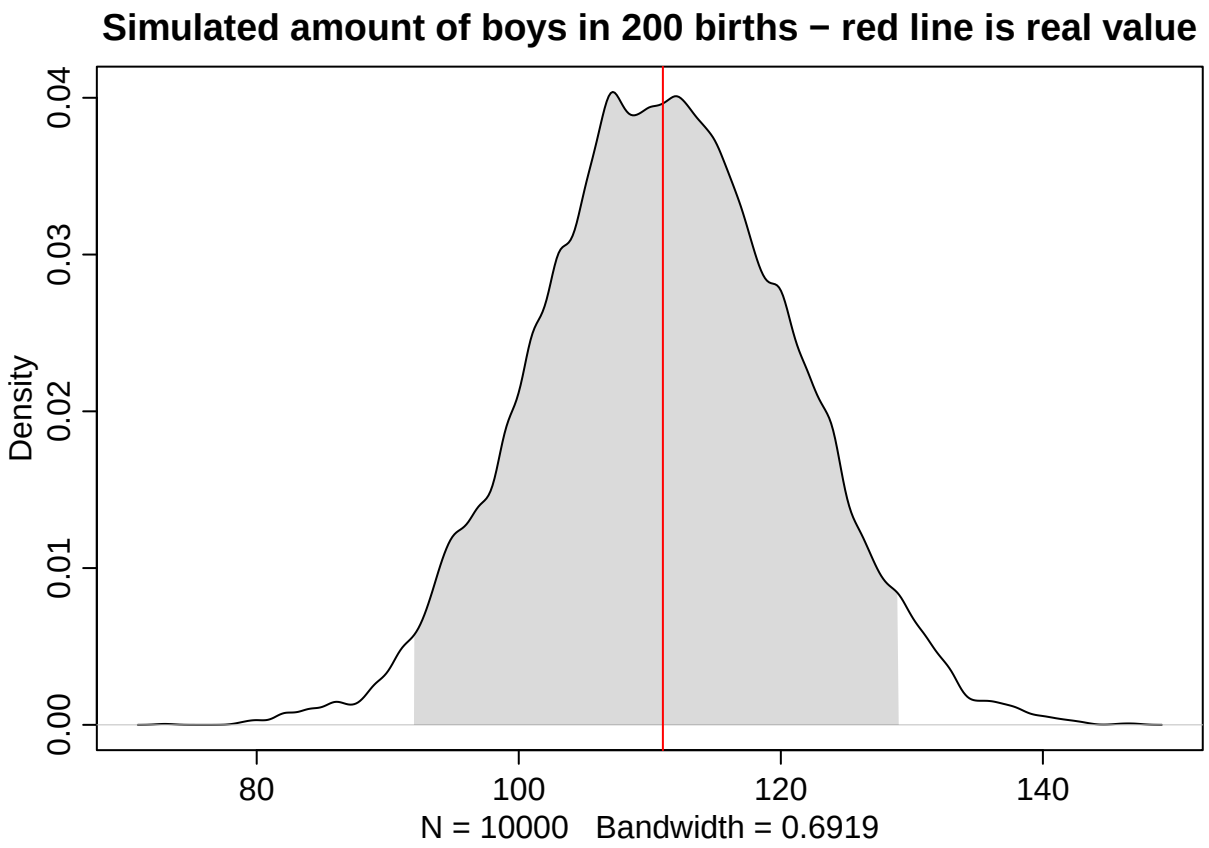
The second assignment uses chapter 3, 5 and 6. The focus of the assignment is getting an understanding of causality.

Chapter 3: Causal Confussion

Reminder: We are tying to estimate the probability of giving birth to a boy I have pasted a working solution to questions 6.1-6.3 so you can continue from here:)

3H3 Use rbinom to simulate 10,000 replicates of 200 births. You should end up with 10,000 numbers, each one a count of boys out of 200 births. Compare the distribution of predicted numbers of boys to the actual count in the data (111 boys out of 200 births).

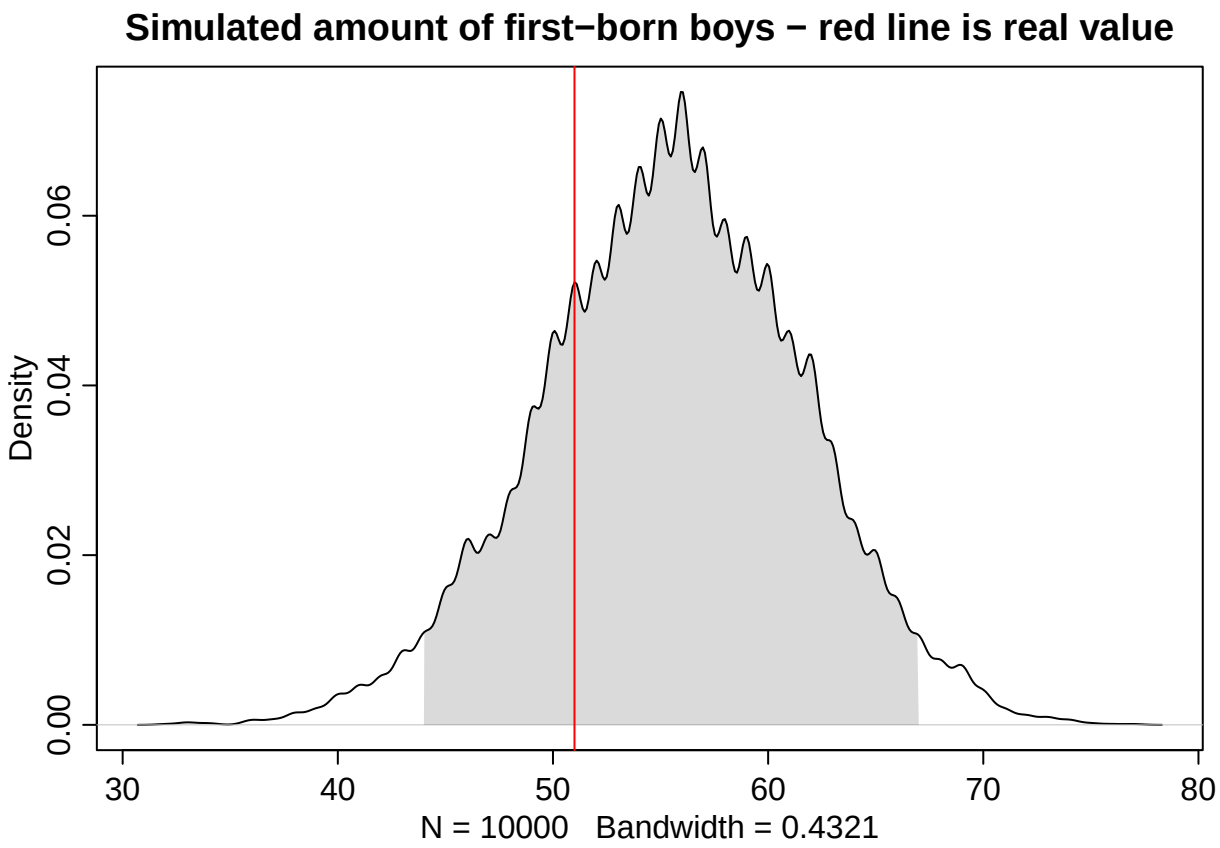
```
# 3H1
# Find the posterior probability of giving birth to a boy:
pacman::p_load(rethinking)
data(homeworkch3)
set.seed(1)
W <- sum(birth1) + sum(birth2)
N <- length(birth1) + length(birth2)
p_grid <-seq(from =0, to = 1, len =1000)
prob_p <- rep(1,1000)
prob_data <- dbinom(W,N,prob=p_grid)
posterior <-prob_data * prob_p
posterior <- posterior / sum(posterior)

# 3H2
# Sample probabilities from posterior distribution:
samples <- sample (p_grid, prob = posterior, size =1e4, replace =TRUE)


# 3H3
# Simulate births using sampled probabilities as simulation input, and check if they
 ↪  allign with real value.
simulated_births <- rbinom(n = 1e4, size = N, prob = samples)
rethinking::dens(simulated_births,show.HPDI = 0.95)
abline(v=W, col="red")
title("Simulated amount of boys in 200 births - red line is real value")
```



3H4. Now compare 10,000 counts of boys from 100 simulated first borns only to the number of boys in the first births, birth1. How does the model look in this light?

A: (Frederik)

```
# Simulate first-born births using sampled probabilities as simulation input, and check
 ↪  if they allign with real value.
C_first <- length(birth1)
simulated_first_births <- rbinom(n = 1e4, size = C_first, prob = samples)
rethinking::dens(simulated_first_births, show.HPDI = 0.95)
abline(v = sum(birth1), col = "red")
title("Simulated amount of first-born boys - red line is real value")
```

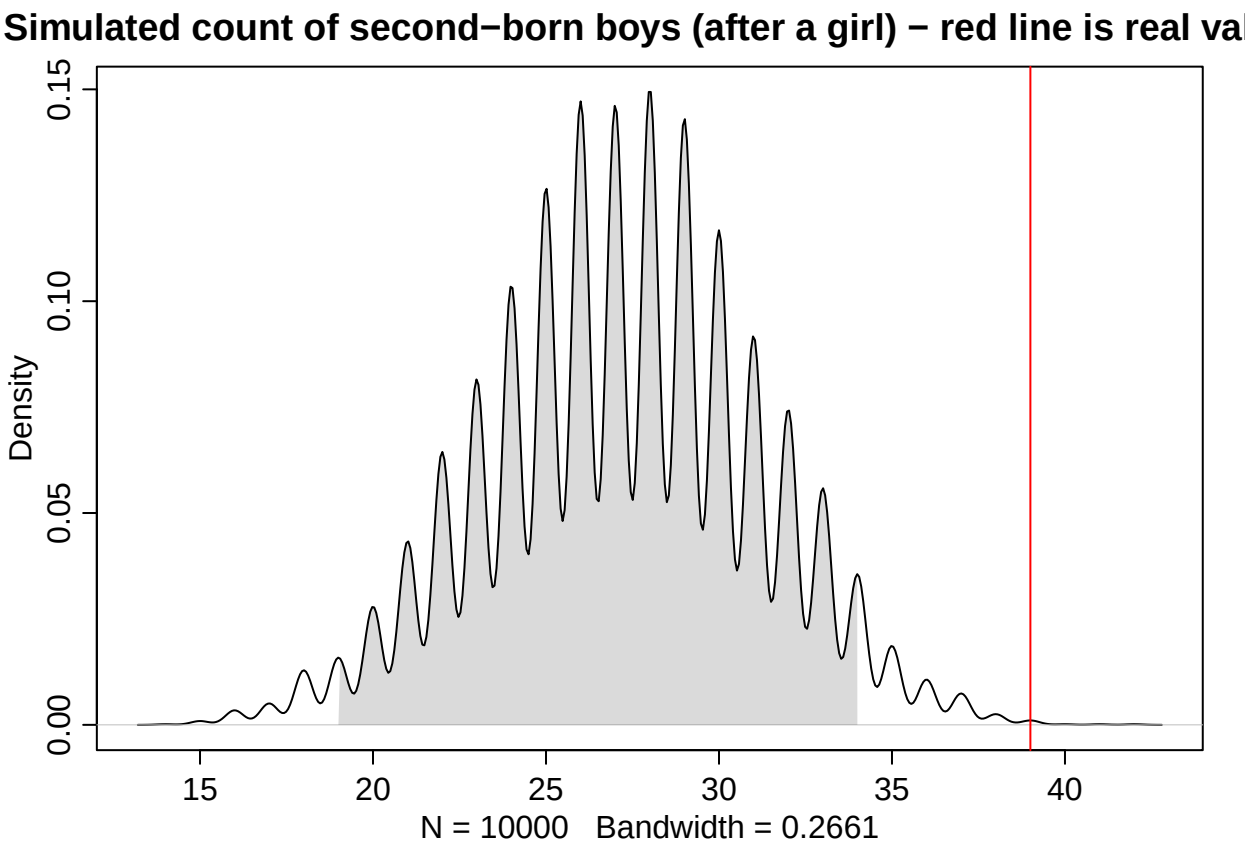


This time we see that the red line indicating the actual count is left off center, of our distribution of simulations indicating that our simulations over predict the number of first born being boys.

3H5. The model assumes that sex of first and second births are independent. To check this assumption, focus now on second births that followed female first borns. Compare 10,000 simulated counts of boys to only those second births that followed girls. To do this correctly, you need to cound the number of first borns who were girls and simulate that many births, 10,000 times. Compare the counts of boys in your simulations to the actual observed count of boys following girls. How does the model look in this light? Any guesses what is going on in these data?

A: (Frederik)

```
# Count of number of first-born girls
C_fb_girls <- sum(birth1 == 0)  # Number of second births after a girl

# number of boys among these second births
C_boy_after_fb_girl <- sum(birth2[birth1 == 0])

# Simulate births of second-born given that first-born was a girl,
# using sampled probabilities as simulation input, and check if they allign with real
 ↪  value.
simulated_boys_after_girl <- rbinom(n = 1e4, size = C_fb_girls, prob = samples)
rethinking::dens(simulated_boys_after_girl, show.HPDI = 0.95)
abline(v = C_boy_after_fb_girl, col = "red")
title("Simulated count of second-born boys (after a girl) - red line is real value")
```



We can see from the plot that simulations highly under predict number of the second born after a girl which are boys, which might suggest that sex of first and second born might not be independent.

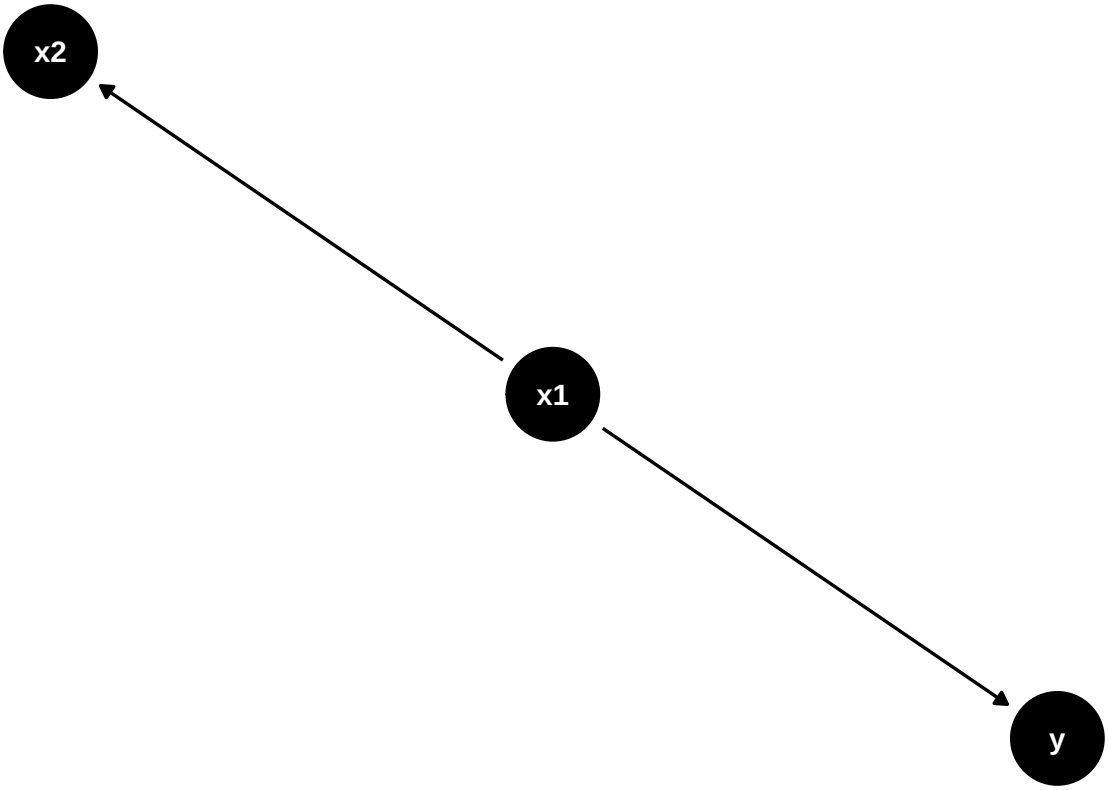
Chapter 5: Spurrious Correlations

Start of by checking out all the spurrious correlations that exists in the world. Some of these can be seen on this wonderfull website: https://www.tylervigen.com/spurious/random All the medium questions are only asking you to explain a solution with words, but feel free to simulate the data and prove the concepts.

5M1. Invent your own example of a spurious correlation. An outcome variable should be correlated with both predictor variables. But when both predictors are entered in the same model, the correlation between the outcome and one of the predictors should mostly vanish (or at least be greatly reduced).

A: (Alexandra) This could be seen with predicting sunflower growth (y) with the amount of sun (x1) and how much sunscreen is sold on the day (x2). The fork correlates x2 and y, unless their relationship is conditioned on x1

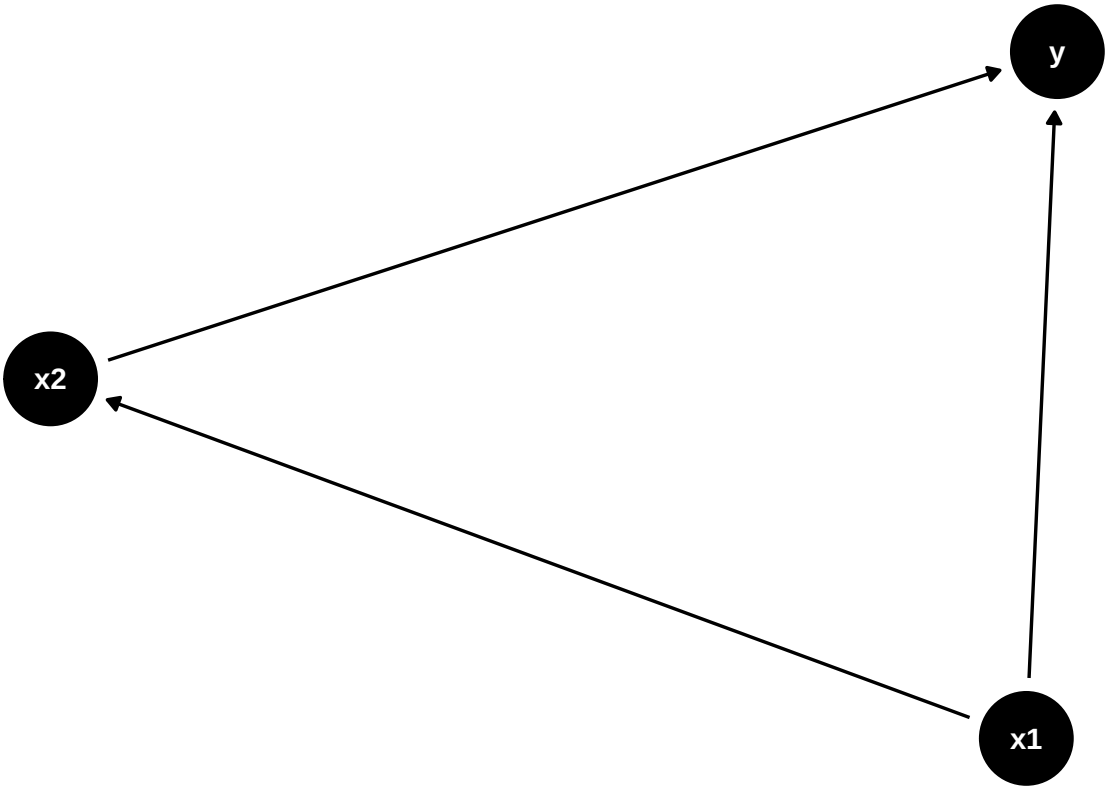
```
theme_set(theme_dag())
dagify(
  x2 ~ x1,
  y ~ x1
) %>%
  ggdag()
```



5M2. Invent your own example of a masked relationship. An outcome variable should be correlated with both predictor variables, but in opposite directions. And the two predictor variables should be correlated with one another.

A: (Alexandra) This could be seen with predicting a grade on a test (y) based on amount of time studying the night before (x1) and the tiredness level on the day of testing (x2).

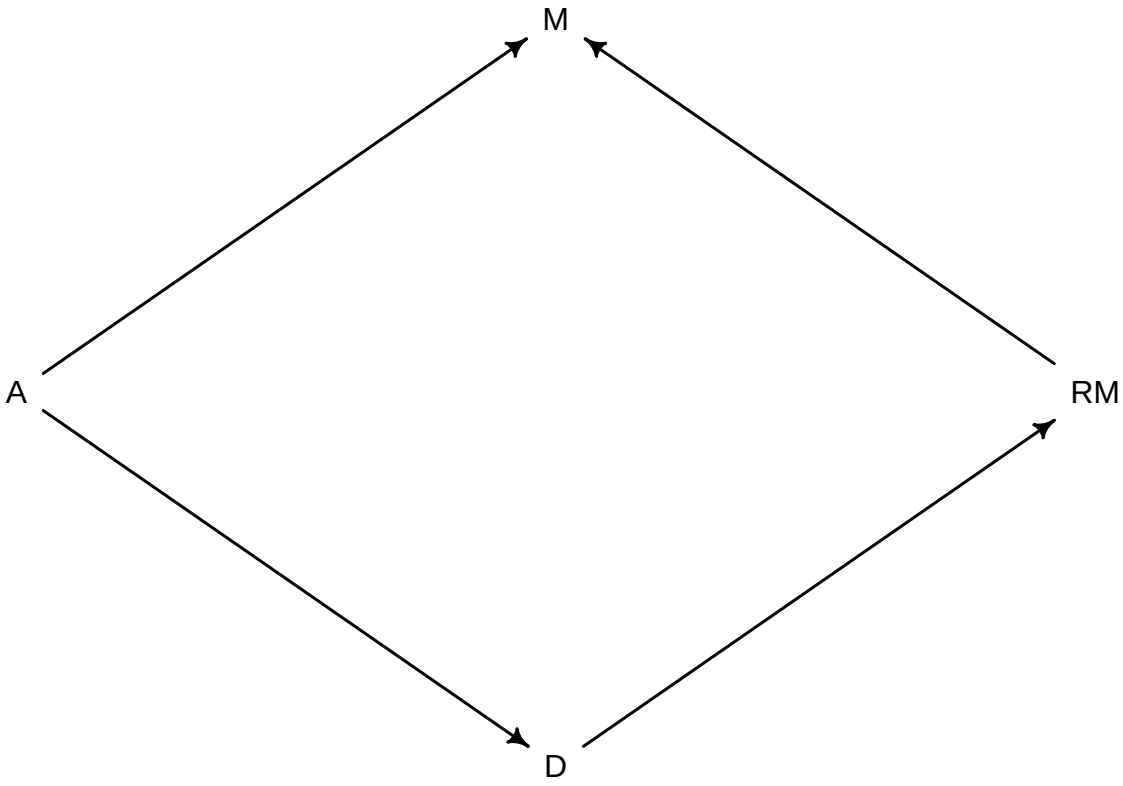
```
theme_set(theme_dag())
dagify(
  x2 ~ x1,
  y ~ x1,
  y ~ x2
) %>%
  ggdag()
```



5M3. It is sometimes observed that the best predictor of fire risk is the presence of firefighters— States and localities with many firefighters also have more fires. Presumably firefighters do not cause fires. Nevertheless, this is not a spurious correlation. Instead fires cause firefighters. Consider the same reversal of causal inference in the context of the divorce and marriage data. How might a high divorce rate cause a higher marriage rate? Can you think of a way to evaluate this relationship, using multiple regression

A: (Valeria) A higher divorce rate will increase the population of singles, which will, presumably, have the opportunity to marry again. Imagining a population that is not allowed to divorce (or die young), the maximum amount of marriages that could be contracted is equal to the integer part of half the number of adults. Divorced individuals, on the other hand, can marry multiple times, increasing the total number of marriages. Using DAGS:

```
dag.3 <-dagitty("dag{A->M;A->D;D->RM; RM->M}")
coordinates(dag.3) <-list(x=c(A=0.5,D=1,M=1,RM=1.5),y=c(A=0.5,D=1,M=0,RM=0.5))
drawdag( dag.3)
```

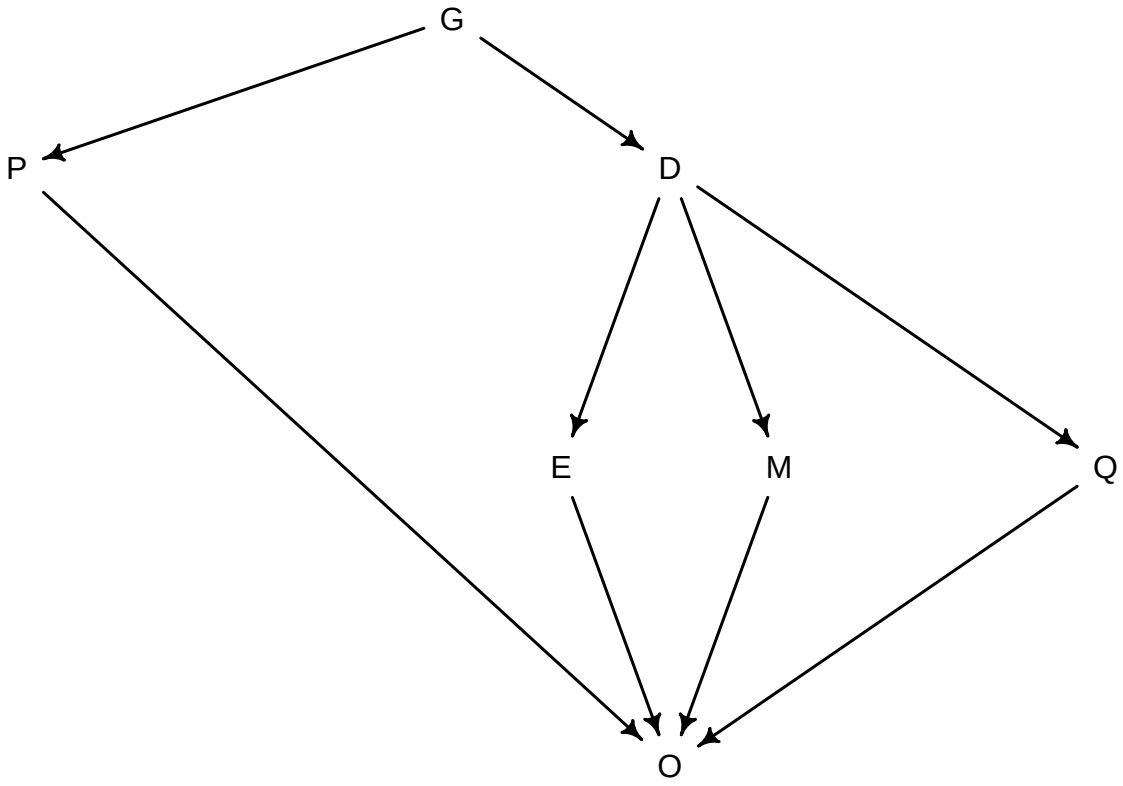


If I were to build a linear regression model with age predicting the divorce rate and condition such models on RM, the remarriage number, I would close the causal path from D to M and thus have: D _||_ M| A, RM Meaning that I would expect β_M to fall close to zero.

5M5. One way to reason through multiple causation hypotheses is to imagine detailed mechanisms through which predictor variables may influence outcomes. For example, it is sometimes argued that the price of gasoline (predictor variable) is positively associated with lower obesity rates (outcome variable). However, there are at least two important mechanisms by which the price of gas could reduce obesity. First, it could lead to less driving and therefore more exercise. Second, it could lead to less driving, which leads to less eating out, which leads to less consumption of huge restaurant meals. Can you outline one or more multiple regressions that address these two mechanisms? Assume you can have any predictor data you need.

A: (Lydia) I would investigate the mechanisms by multiple steps: First, I would draw a DAG, with causal relationships between gasoline price (G), kilometers traversed in car in a week (D), amount of large restaurant meals eaten in a week (M), hours spent exercising (E), and obesity rate (O). I would include in this scientific model two unobserved confounders, P and Q. The former is influenced by gasoline price, whilst the latter is influenced by driving, and both directly affect obesity rates. These confounders represent alternative mechanisms linking gas prices and driving to obesity rates.

```
gas_dag <-dagitty("dag{G->D;G->P;D->M;D->E;D->Q;E->O;M->O;Q->O;P->O}")
coordinates(gas_dag)
 ↪  <-list(x=c(G=0.4,D=0.6,M=0.7,E=0.5,P=0,Q=1,O=0.6),y=c(G=0,D=0.2,M=0.6,E=0.6,P=0.2,Q=0.6,O=1))
drawdag(gas_dag)
```



Assuming this DAG, I would make a linear regression of O predicted by G, D, E and M. The latter two are included to test each of the mechanisms involving them. It is important in this context to stratify by driving, as the number of large meals would otherwise be correlated with exercise and the unobserved variable(s) Q. The model's estimated effect of driving would represent effects of Q, being the only unblocked path from D to O. Conditioning further on G allows us to detect P and differentiate between P and Q, who would otherwise be correlated (as P and D are correlated through the fork with G, and all other paths from D to O are blocked). Thus, the estimates of the model would indicate answers to the questions: - Are each of the two proposed mechanisms supported by the data? - are other, unobserved mechanism related to either gasoline price or driving responsible (in part) for the correlation between obesity rate and gasoline price?

Based on a very simple simulation, the model seems able to provide reasonable answers to these questions:

```
gas_price <- rnorm(1000, 50, 3)
p_hidden <- rnorm(1000, 2*gas_price, 1)
driving <- rnorm(1000, 4 * gas_price, 4)
exercise <- rnorm(1000, 50 - driving/6, 2)
large_meals_out <- rnorm(1000, 60 - driving, 5)
q_hidden <- rnorm(1000, 2*driving - 70, 6)

obesity <- rnorm(1000, 100 - 4*exercise + 2*large_meals_out + p_hidden + q_hidden, 3)
gas_df <- data.frame(gas_price, driving, exercise, large_meals_out, p_hidden, q_hidden,
↪  obesity)
```

```
s <- lm(obesity ~ gas_price + driving + exercise + large_meals_out, gas_df) %>% summary()
s$coefficients[,1:2]
```

```
##                  Estimate Std. Error
## (Intercept)     30.146863 7.15622007
## gas_price        2.108626 0.24062914
## driving          2.039839 0.07431168
## exercise        -4.114961 0.10824062
## large_meals_out  2.085349 0.04521790
```

The model correctly identifies the effects of E and M (exercise and "large.meals.out"), while also indicating the presence of other relevant effect (P and Q, seen in the estimates for "gas.price" and "driving")

Chapter 5: Foxes and Pack Sizes

All five exercises below use the same data, data(foxes) (part of rethinking).84 The urban fox (Vulpes vulpes) is a successful exploiter of human habitat. Since urban foxes move in packs and defend territories, data on habitat quality and population density is also included. The data frame has five columns: (1) group: Number of the social group the individual fox belongs to (2) avgfood: The average amount of food available in the territory (3) groupsize: The number of foxes in the social group (4) area: Size of the territory (5) weight: Body weight of the individual fox

5H1. Fit two bivariate Gaussian regressions, using quap: (1) body weight as a linear function of territory size (area), and (2) body weight as a linear function of groupsize. Plot the results of these regressions, displaying the MAP regression line and the 95% interval of the mean. Is either variable important for predicting fox body weight?

A: (Lydia)

```
data(foxes)

# first weight ~ area

#standardising variables
foxes$weight <- scale(foxes$weight)
foxes$area <- scale(foxes$area)

area_model <- quap(flist = alist(
  weight ~ dnorm(mu,sigma),
  mu <- a + b_area*area,
  a ~ dnorm(0,1),
  b_area ~ dnorm(0,2),
  sigma ~ dunif(0,5)
),data=foxes)

# Now weight ~ size

foxes$groupsize <- scale(foxes$groupsize)

size_model <- quap(flist = alist(
  weight ~ dnorm(mu,sigma),
  mu <- a + b_size*groupsize,
  a ~ dnorm(0,1),
  b_size ~ dnorm(0,2),
  sigma ~ dunif(0,5)
),data=foxes)
```

```r
# look at the model outputs
precis(area_model)
```

```
##                mean         sd       5.5%     94.5%
## a      2.197304e-07 0.09203674 -0.1470923 0.1470927
## b_area 1.943545e-02 0.09273021 -0.1287653 0.1676362
## sigma  9.954913e-01 0.06535690  0.8910384 1.0999443
```

```r
precis(size_model)
```

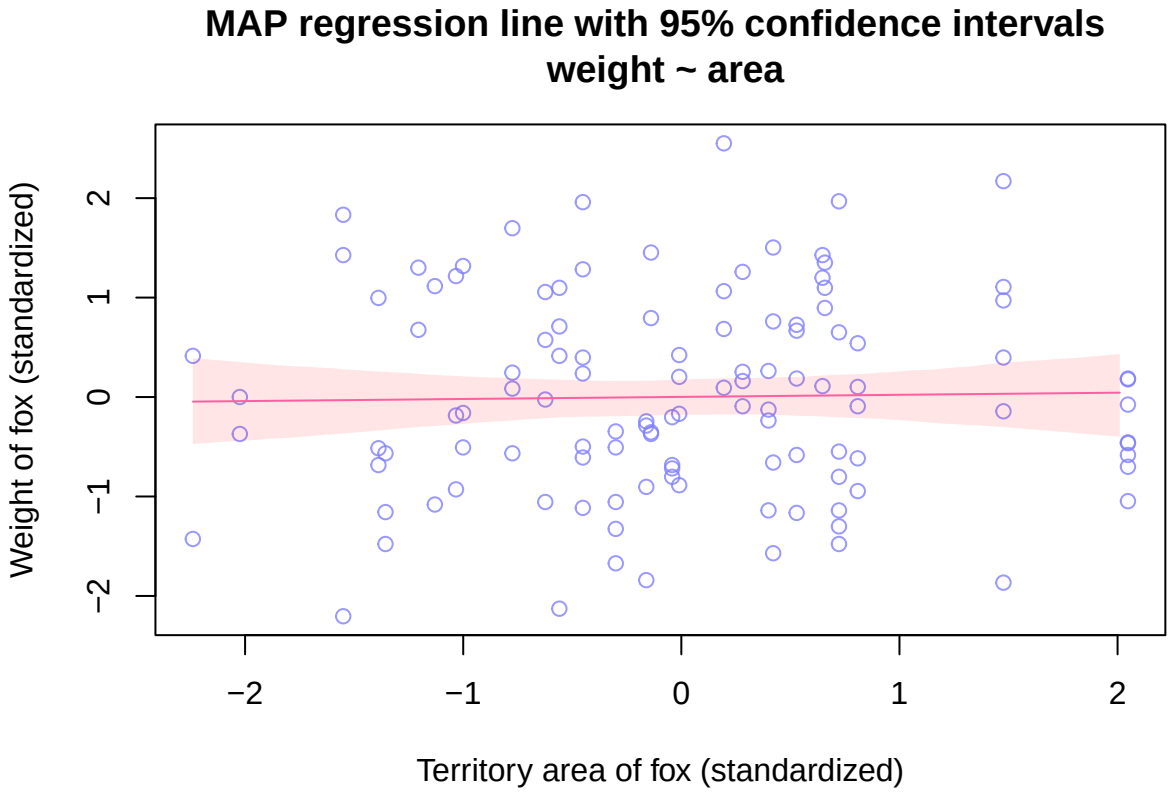
```
##                 mean         sd       5.5%       94.5%
## a       1.009776e-07 0.09086324 -0.1452169  0.14521711
## b_size -1.606565e-01 0.09154048 -0.3069559 -0.01435713
## sigma   9.826921e-01 0.06451659  0.8795821  1.08580205
```

groupsize is likely (but only weakly) related to lower body weight, and almost certainly not positively related to it, based on our regression results.

Territory area seems unrelated to weight, as the posterior distribution of the effect of area is centered around 0. However, the distribution is relatively wide, so we area might still have some positive or negative effect.

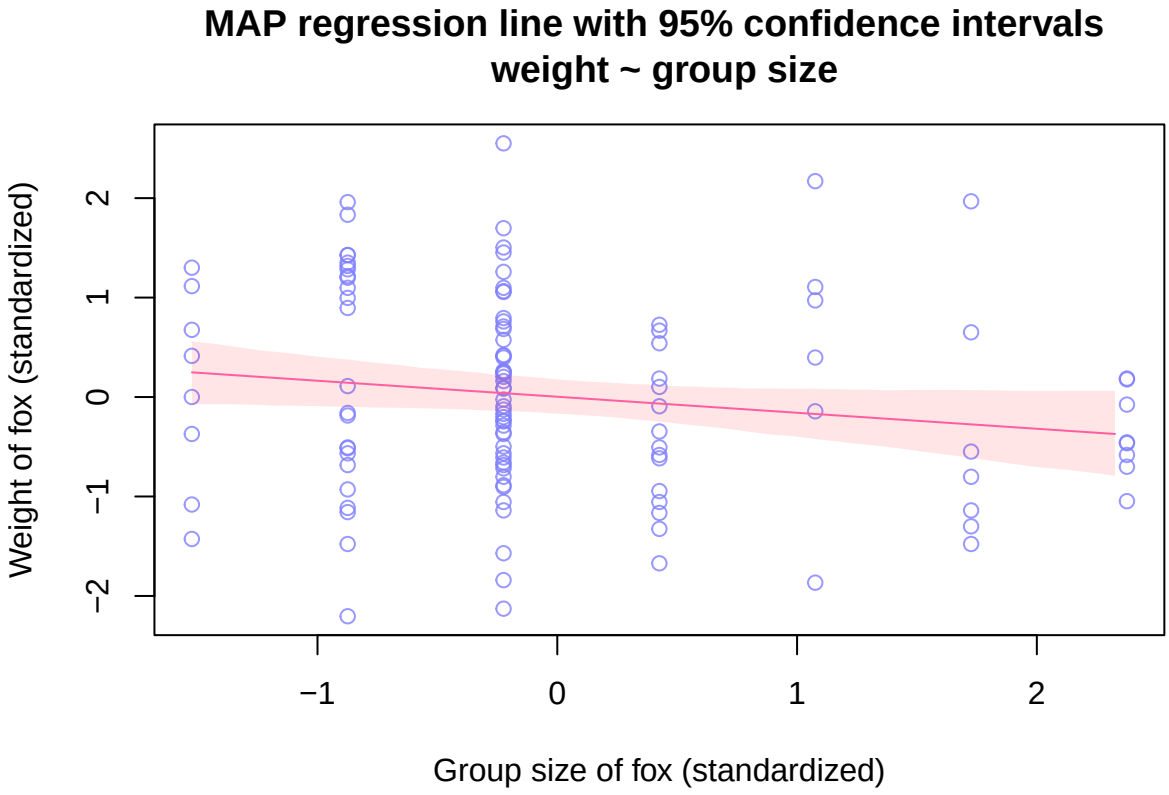
Making the plot for territory area

```r
x_values <- seq(min(foxes$area), max(foxes$area), by = 0.05)
mu_values <- link(area_model, data.frame(area = x_values))
mu.mean <-apply(mu_values,2,mean)
mu.PI <-apply(mu_values,2,PI,prob=0.95)

plot( weight~area,data=foxes,col=col.alpha(rangi2,0.8),
      main="MAP regression line with 95% confidence intervals \n weight ~ area",
      xlab="Territory area of fox (standardized)",
      ylab="Weight of fox (standardized)")
#MAP line
lines( x_values, mu.mean, col="hotpink")
#95% interval
shade( mu.PI, x_values, col=col.alpha("red",0.1))
```



And the plot for groupsize

```
x_values <- seq(min(foxes$groupsize), max(foxes$groupsize), by = 0.05)
mu_values <- link(size_model, data.frame(groupsize = x_values))
mu.mean <-apply(mu_values,2,mean)
mu.PI <-apply(mu_values,2,PI,prob=0.95)

plot( weight~groupsize,data=foxes,col=col.alpha(rangi2,0.8),
      main="MAP regression line with 95% confidence intervals \n weight ~ group size",
      xlab="Group size of fox (standardized)",
      ylab="Weight of fox (standardized)")
#MAP line
lines( x_values, mu.mean, col="hotpink")
#95% interval
shade( mu.PI, x_values, col=col.alpha("red",0.1))
```



5H2. Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean. What does this model say about the importance of each variable? Why do you get different results than you got in the exercise just above?

A: (Lydia) Now fit a multiple linear regression with weight as the outcome and both area and groupsize as predictor variables. Plot the predictions of the model for each predictor, holding the other predictor constant at its mean. What does this model say about the importance of each variable? Why do you get different results than you got in the exercise just above?

```
size_area_model <- quap(flist = alist(
  weight ~ dnorm(mu,sigma),
  mu <- a + b_size*groupsize + b_area*area,
  a ~ dnorm(0,1),
  b_size ~ dnorm(-0.1,2),
  b_area ~ dnorm(0,2),
  sigma ~ dunif(0,5)
),data=foxes)

precis(size_area_model)
```

```
##                 mean         sd       5.5%      94.5%
## a      -1.865978e-06 0.08738447 -0.1396591  0.1396554
## b_size -5.569709e-01 0.15614478 -0.8065204 -0.3074214
## b_area  4.793562e-01 0.15614488  0.2298066  0.7289059
## sigma   9.447736e-01 0.06204251  0.8456177  1.0439295
```

```
#let's also look at their correlation in the data
cor(foxes$groupsize, foxes$area)
```

```
##           [,1]
## [1,] 0.8275945
```

From this model, it looks like both group size and territory size are related to body weight; we are quite sure (within the small world context of our model) that the "effect" of group size is negative and that of area is positive. Both effects are larger and more certain in this model compared to the two previous ones, due to the two variables being positively correlated but with opposite effects on weight, partially masking each other when not accounted for.

The estimated effect of territory size was more strongly affected by the inclusion/exclusion of the other variable than vice-versa, due to groupsize having a larger effect on weight (after standardising).

5H3. Finally, consider the avgfood variable. Fit two more multiple regressions: (1) body weight as an additive function of avgfood and groupsize, and (2) body weight as an additive function of all three variables, avgfood and groupsize and area. Compare the results of these models to the previous models you've fit, in the first two exercises. (a) Is avgfood or area a better predictor of body weight? If you had to choose one or the other to include in a model, which would it be? Support your assessment with any tables or plots you choose. (b) When both avgfood or area are in the same model, their effects are reduced (closer to zero) and their standard errors are larger than when they are included in separate models. Can you explain this result?

A: (Valeria)

```
# Importing the data
data(foxes)
d <-foxes
#standardizing data
d$g_s <-standardize(d$groupsize)
d$w <-standardize(d$weight)
d$A <-standardize(d$area)
d$avf <- standardize(d$avgfood)
```

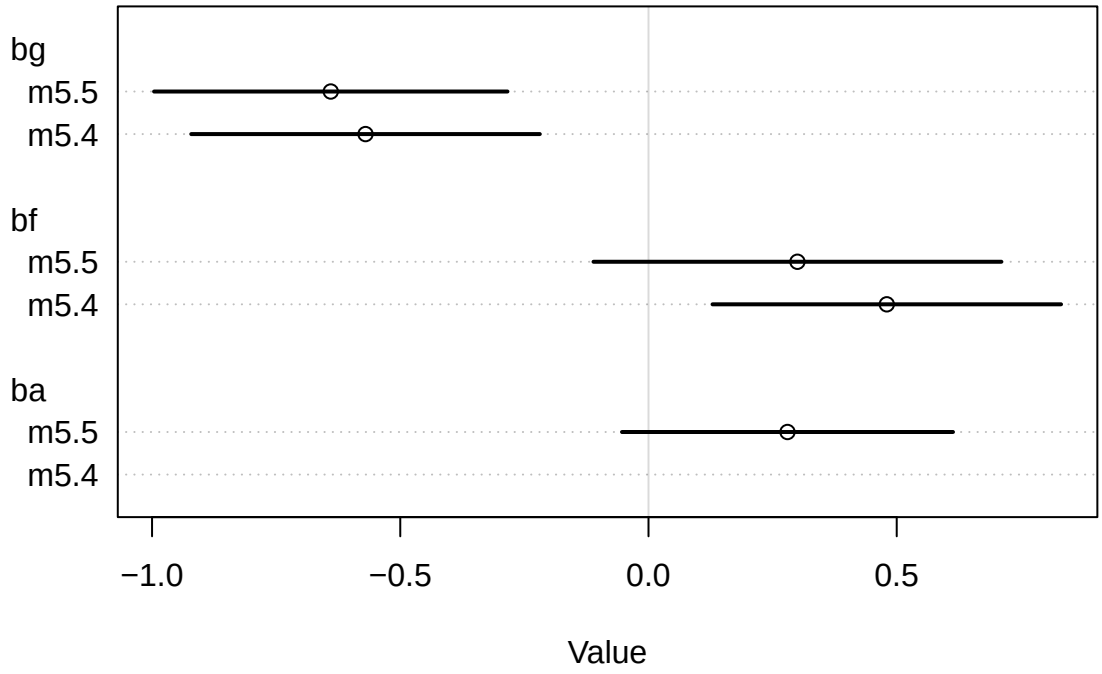
```
# Body weight as additive function of avgfood and groupsize
m5.4 <-quap(
  alist(
  w ~dnorm(mu,sigma),
  mu <-a+bg*g_s+bf*avf,
  a ~dnorm(0,0.2),
  bg ~dnorm(0,0.5),
  bf ~dnorm(0, 0.5),
  sigma ~dexp(1)
  ),data=d)

precis(m5.4)
```

```
##                mean         sd       5.5%      94.5%
## a      4.214989e-07 0.08013798 -0.1280756  0.1280764
## bg    -5.735245e-01 0.17914155 -0.8598273 -0.2872217
## bf     4.772525e-01 0.17912305  0.1909793  0.7635257
## sigma  9.420427e-01 0.06175236  0.8433505  1.0407349
```

```
# Body weight as additive function of all three
m5.5 <-quap(
  alist(
  w ~dnorm(mu,sigma),
  mu <-a+bg*g_s+bf*avf+ba*A,
  a ~dnorm(0,0.2),
  bg ~dnorm(0,0.5),
  bf ~dnorm(0, 0.5),
  ba ~dnorm(0, 0.5),
  sigma ~dexp(1)
  ),data=d)

precis(m5.5)
```

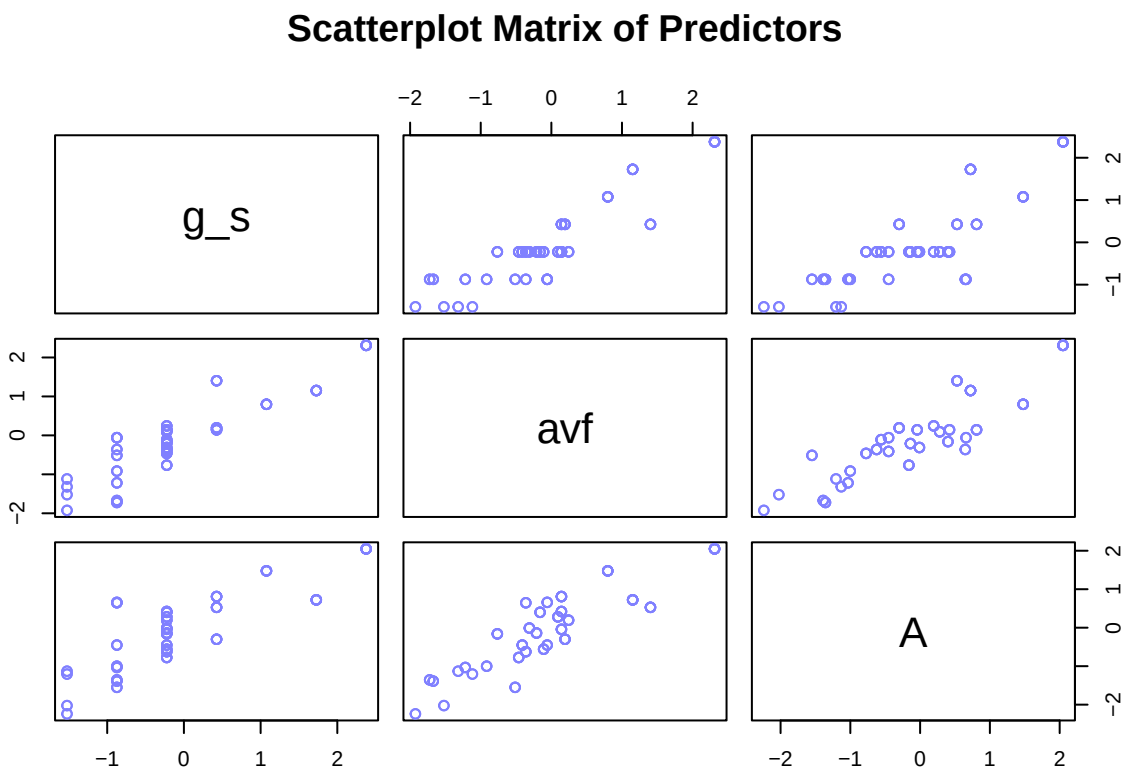
```
##                mean         sd         5.5%      94.5%
## a      3.576357e-08 0.07936202 -0.126835796  0.1268359
## bg    -6.396197e-01 0.18161485 -0.929875334 -0.3493641
## bf     2.968990e-01 0.20960025 -0.038082682  0.6318807
## ba     2.782384e-01 0.17011229  0.006366089  0.5501107
## sigma  9.312065e-01 0.06100011  0.833716506  1.0286964
```

```
plot( coeftab(m5.4,m5.5),pars=c("bg","bf","ba"))
```



I can see the coefficient β_{gs} remains fairly similar in both models, while in the model including all three predictors both β_{avgf} and β_{area} include zero in their interval, meaning that the effect I observed in previous models for the territory area has been erased by the inclusion of the average food.

```
# Checking correlation among predictors
pairs( ~ g_s+ avf+ A ,data=d,col=rangi2,
      main = "Scatterplot Matrix of Predictors", )
```



In fact all of these predictor appear to be strongly correlated.

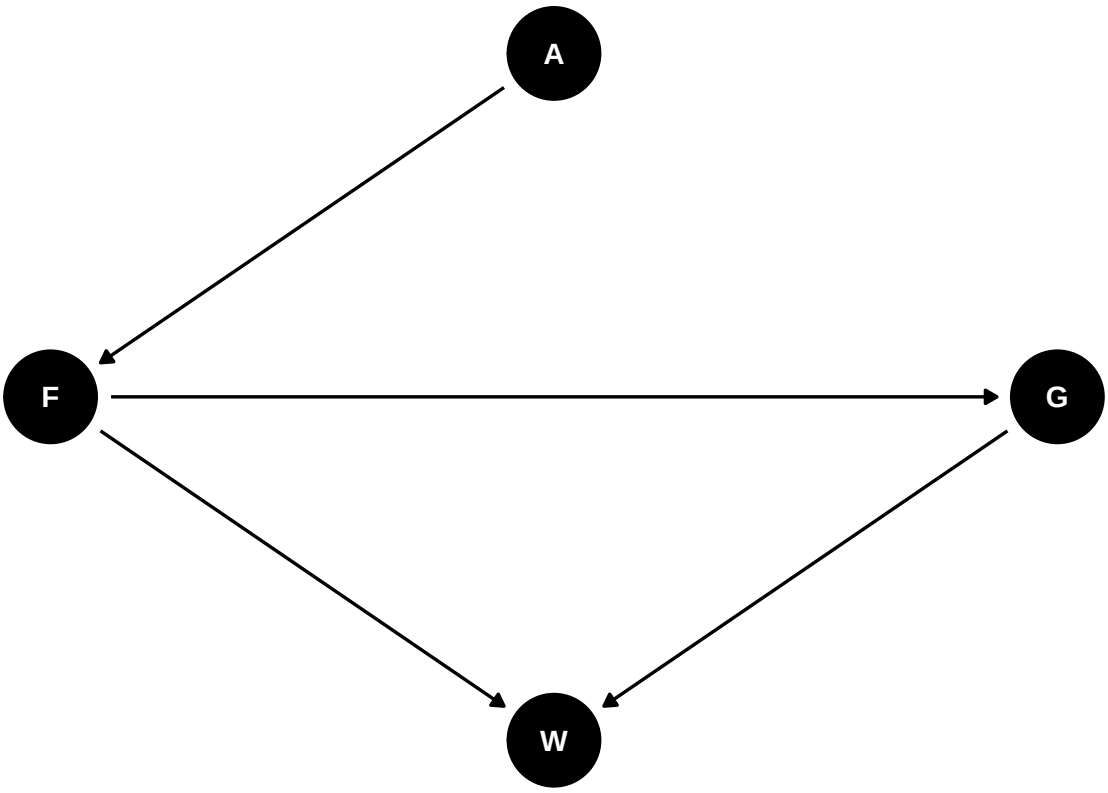
a) I would choose average area as a predictor because, despite the β coefficient associated with it in 'm5.3' being slightly lower then the one associated with average food in 'm5.5', the associated standard deviation is also lower. What is more, if average food is a consequence of territory size it would be better to use the more "fundamental" variable.

b) This effect could be due to the total area being directly related to the average food present, since larger areas naturally bring more food for the foxes. So "average food" becomes a redundant predictor, a post-treatment variable to territory size.

Defining our theory with explicit DAGs Assume this DAG as an causal explanation of fox weight:

```
#pacman::p_load(dagitty,
                #ggdag)
dag <- dagitty('dag {
A[pos="1.000,0.500"]
F[pos="0.000,0.000"]
G[pos="2.000,0.000"]
W[pos="1.000,-0.500"]
A -> F
```

```
F -> G
F -> W
G -> W
}')

# Plot the DAG
ggdag(dag, layout = "circle")+
  theme_dag()
```



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

Using what you know about DAGs from chapter 5 and 6, solve the following three questions:

1) 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 of it?

A: (Valeria) Given the DAG above there is only one direct path from A to F

```
# Confirming  the above
adjustmentSets(dag, exposure = "A", outcome = "F")
```

```
##  {}
```

```
# Using quap approximation to estimate causal influence of A on F
m5.6 <-quap(
  alist(
  avf ~dnorm(mu,sigma),
  mu <-a+ba*A,
  a ~dnorm(0,0.2),
  ba ~dnorm(0, 0.5),
  sigma ~dexp(1)
  ),data=d)

precis(m5.6)
```

```
##               mean         sd        5.5%      94.5%
## a     1.092202e-07 0.04231162 -0.06762202 0.06762224
## ba    8.764761e-01 0.04332441  0.80723528 0.94571685
## sigma 4.662637e-01 0.03052547  0.41747813 0.51504932
```

The Dag, and model it represents, imply a very strong positive effect of territory size on the average amount of food available for foxes. Thus increasing the territory size will increase the average availability of food.

2) Infer the **total** causal effect of adding food F to a territory on the weight W of foxes. Can you calculate the causal effect by simulating an intervention on food?

A: (Lydia) We can get the total causal effect of F on W by only including F as a predictor in our model, leaving the pipe through G open

```
foxes$avgfood <- scale(foxes$avgfood)
food_weight_model <- quap(flist = alist(
  weight ~ dnorm(mu,sigma),
  mu <- a + b_food*avgfood,
  a ~ dnorm(0,1),
  b_food ~ dnorm(0,2),
  sigma ~ dunif(0,5)
),data=foxes)

precis(food_weight_model)["b_food",]
```

```
##               mean        sd       5.5%     94.5%
## b_food -0.02955725 0.1098473 -0.2051144 0.1459999
```

The total effect of food on weight is likely close to zero, either positive or negative.

3) Infer the **direct** causal effect of adding food F to a territory on the weight W of foxes. In light of your estimates from this problem and the previous one, what do you think is going on with these foxes?

A: (Valeria) To only take into account the direct effect I have to close the pipe F $\rightarrow$ G $\rightarrow$ W. I can do so by conditioning my model on group size.

```
# Check
adjustmentSets(dag, exposure = "F", outcome = "W", effect = "direct")
```

{ G }

```
# Model
m5.8 <- quap(
  alist(
  w ~dnorm(mu,sigma),
  mu <-a+bf*avf+bg*g_s,
  a ~dnorm(0,0.2),
  bf ~dnorm(0, 0.5),
  bg ~dnorm(0, 0.5),
  sigma ~dexp(1)
  ),data=d)

precis(m5.8)
```

```
##                mean         sd       5.5%      94.5%
## a     -4.552763e-08 0.08013804 -0.1280761  0.1280760
## bf     4.772530e-01 0.17912317  0.1909796  0.7635264
## bg    -5.735260e-01 0.17914167 -0.8598290 -0.2872230
## sigma  9.420436e-01 0.06175250  0.8433512  1.0407360
```

Considering only the direct effect of increasing food availability (F), meaning: we are keeping all other predictors (in this specific case only group size) constant, then there is a strong positive causal effect of F on the weight of foxes. At the same time increasing group size has a definite negative effect on weight, quite reasonably, since resources would have to be shared between more individuals.

The difference between these two estimates could be ascribed to a masking effect, where the group size increases with the more resources available, as individuals migrate to such advantageous areas. Still this has a negative effect on the outcome variable W. Since in bigger groups less resources can go to each individual. The two predictors have opposite effects on the outcome variable, so including both in the regression separates their contributions.

Chapter 6: Investigating the Waffles and Divorces

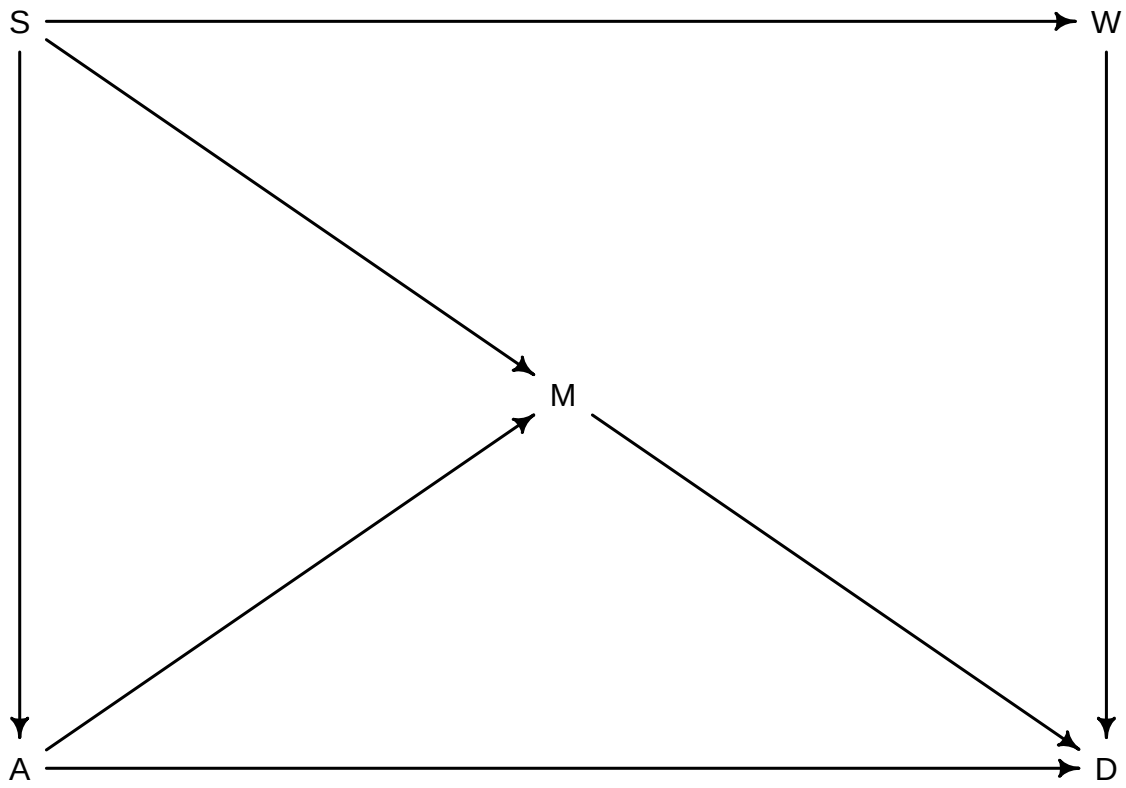
6H1. Use the Waffle House data, data(WaffleDivorce), to find the total causal influence of number of Waffle Houses on divorce rate. Justify your model or models with a causal graph.

A: (Valeria)

```
# Data loading and standardizing
data(WaffleDivorce)
d6 <-WaffleDivorce

#Standardising
d6$D <-standardize(d6$Divorce)
d6$M <-standardize(d6$Marriage)
d6$A <-standardize(d6$MedianAgeMarriage)

# Building an index variable to indicate a state's southern-ness or not so that
# priors'uncertainty can be defined in both cases (whether state is in the South
# or not)
d6$S <-ifelse(d6$South==1,2,1)
```

Based on previous experience with this dataset I define a, rather naive, DAG assuming a direct causal effect of W on D.

```
#Defining a DAG
dag.6 <-dagitty("dag{
  A ->D
  A ->M->D
  A <-S->M
  S ->W->D
  }")
coordinates(dag.6) <-list(x=c(A=0,M=0.5,S=0,D=1,W=1),
                          y=c(A=1,M=0.5,S=0,D=1,W=0))

drawdag(dag.6)
```




```
# I can check whether my reasoning about the pathways for this DAG is correct
adjustmentSets( dag.6,exposure="W",outcome="D", effect = "total")
```

```
## { A, M }
## { S }
```

```
# And check the conditional independencies of this DAG
impliedConditionalIndependencies( dag.6)
```

```
## A _||_ W | S
## D _||_ S | A, M, W
## M _||_ W | S
```

So, if this DAG corresponds to an accurate model for these data: 1) Age of marriage is independent from the number of waffle houses once I condition the model on the state being southern or not 2) Divorce rate is independent on the state being southern once I condition the model on age at marriage, marriage rate, and number of waffle houses. This since these three variable contain, according to this simplified DAG all the information necessary to determine whether the state is southern. 3) Marriage rate is independent from the number of waffle houses once I condition the model on a state being southern or not.

To find the total causal effect of the number of waffle houses on divorce rate I have to block the back pathways from W to D. In the DAG those starting at W and ending at D. There are three:

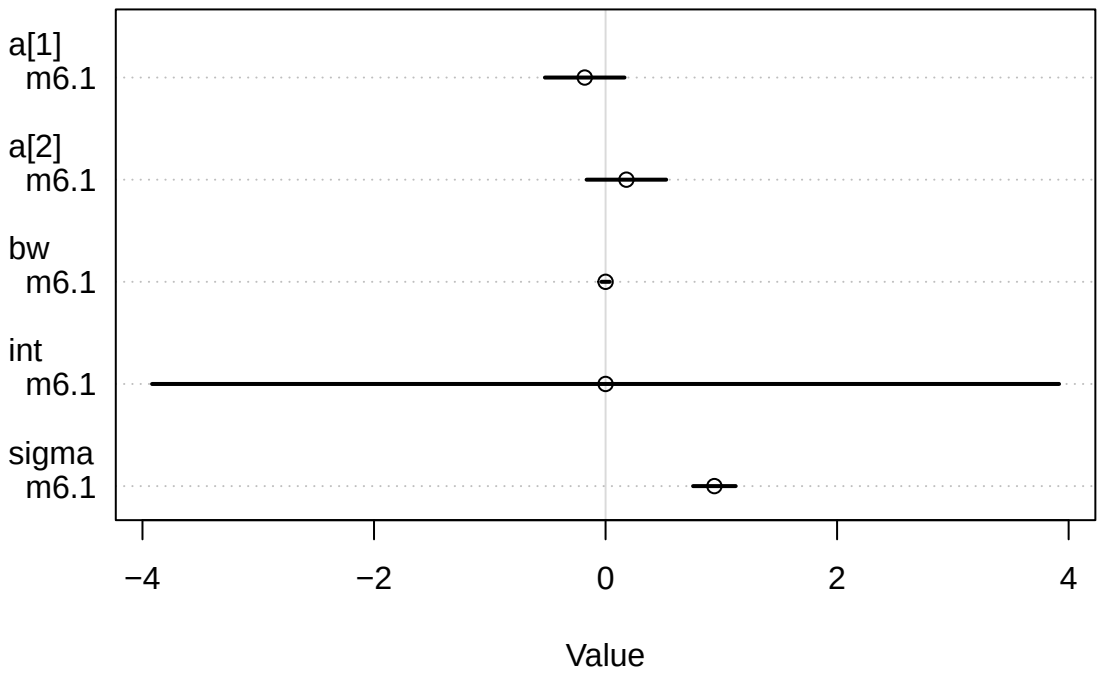
a) $W \leftarrow S \rightarrow M \rightarrow D$
b) $W \leftarrow S \rightarrow A \rightarrow D$
c) $W \leftarrow S \rightarrow A \rightarrow M \rightarrow D$

As confirmed above they can all be blocked simply conditioning the model on S, since they all pass through this parameter node.

```
# Testing
# Model, using quadratic approximation
m6.1 <-quap(
  alist(
  D ~dnorm(mu,sigma),
  mu <-int + a[S]+bw*W, # model conditioned on index variable S
  a[S] ~dnorm(0,0.2),
  bw ~dnorm(0,0.5),
  int ~ dnorm(0,2),
  sigma ~dexp(1)
  ) ,data=d6)

precis(m6.1,depth=2)
```

```
##                mean         sd        5.5%      94.5%
## a[1]  -1.821177e-01 0.17547315 -0.46255769 0.09832228
## a[2]   1.821186e-01 0.17547350 -0.09832198 0.46255911
## bw     7.210072e-04 0.01809526 -0.02819871 0.02964073
## int    9.999351e-05 1.99870267 -3.19421290 3.19441289
## sigma  9.372924e-01 0.09403381  0.78700824 1.08757663
```

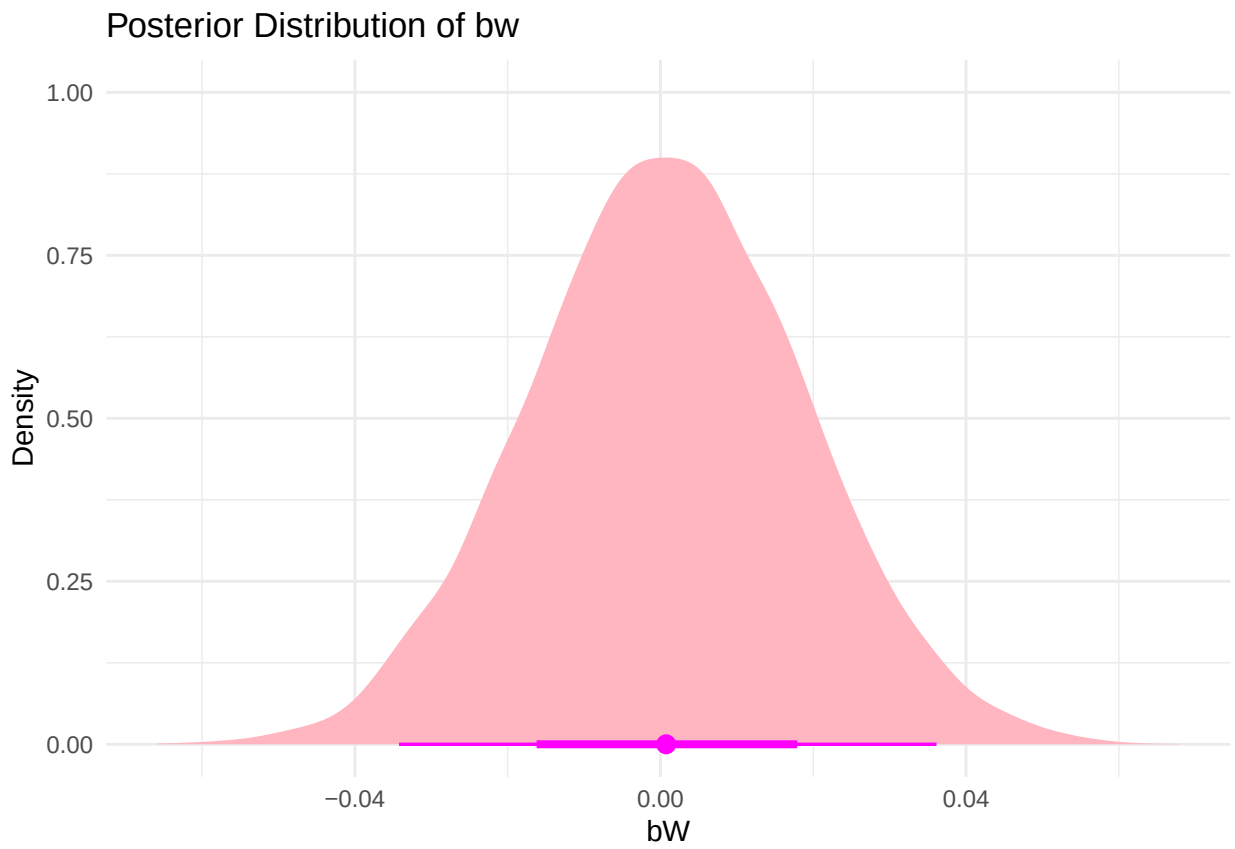
```
plot( coeftab(m6.1))
```



As expected I see no causal effect of number of waffle houses on divorce rate (β_w=0). What is more both intercepts, while having their means on either side of zero, include zero in their percentage intervals

Looking at the posterior distribution of β_W

```
# Posterior samples of bw
post_samp <- extract.samples(m6.1)
# Converting to a tibble for ggplot
bw_samp <- data.frame(bw = post_samp$bw)

# Plotting
ggplot(bw_samp, aes(x = bw)) +
  stat_halfeye(.width = c(0.66, 0.95), fill = "lightpink", color = "magenta") +
  labs(x = "bW",
       y = "Density",
       title = "Posterior Distribution of bw") +
  theme_minimal()
```



I can also compute the contrast between divorce rate on northern and southern states

```
# Contrast computation
# Sampling from posteriors to have an expected difference between northern
# and southern states
post <-extract.samples(m6.1)
post$diff_fm <-post$a[,1]-post$a[,2]
precis( post,depth=2)
```

```
##                  mean         sd        5.5%       94.5%       histogram
## bw       0.0008668464 0.01776662 -0.02739123  0.02921642
## int     -0.0165357365 1.96661959 -3.18099752  3.14384979
## sigma    0.9375927078 0.09489317  0.78793661  1.08906089
## a[1]    -0.1815667490 0.17293592 -0.45870397  0.09614535
## a[2]     0.1815953531 0.17455773 -0.09761357  0.46170732
## diff_fm -0.3631621021 0.20433344 -0.68524159 -0.03140412
```

The contrast is negative, indicating a higher mean divorce rate in the southern states than in the northern ones.

6H2. *A*: *(Valeria)* Build a series of models to test the implied conditional independencies of the causal graph you used in the previous problem. If any of the tests fail, how do you think the graph needs to be amended? Does the graph need more or fewer arrows? Feel free to nominate variables that aren't in the data.

From the exercise above I know the conditional independences implied by the chosen DAG to be:

A _||_ W| S

D _||_ S| A, M, W

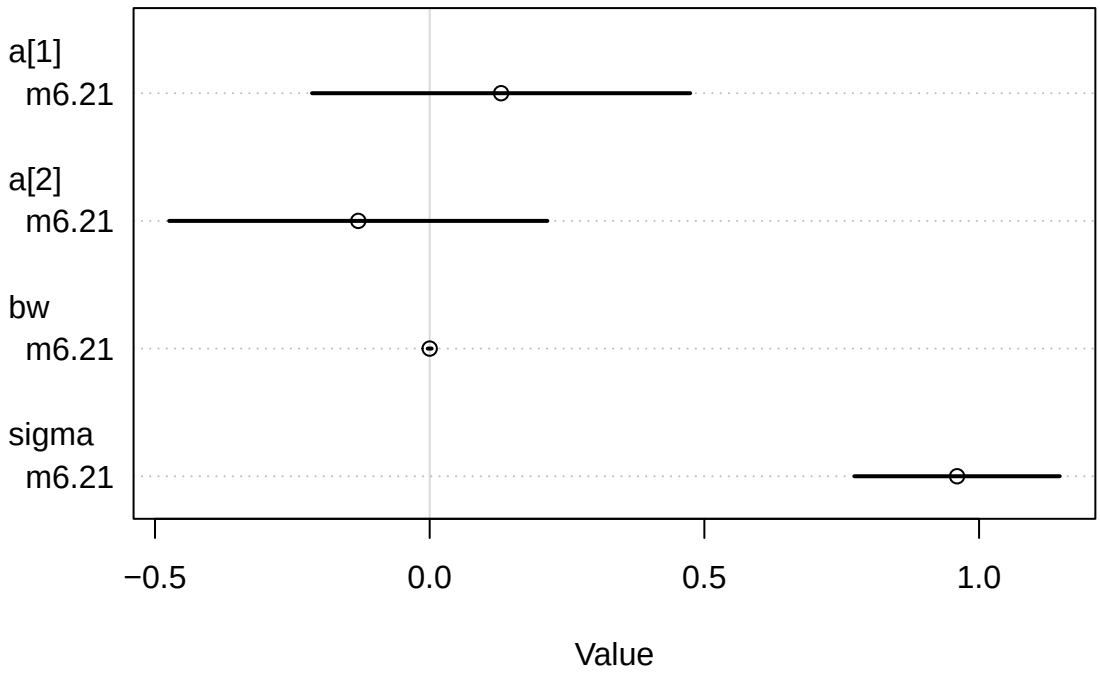
M _||_ W| S

To test my DAG I have to test them.

```
# Testing first independency
m6.21 <-quap(
  alist(
  A ~dnorm(mu,sigma),
  mu <-a[S]+bw*W, # model including S and W as predictors
  a[S] ~dnorm(0,0.2),
  bw ~dnorm(0,0.5),
  sigma ~dexp(1)
  ) ,data=d6)

precis(m6.21,depth=2)
```

```
##                mean           sd         5.5%       94.5%
## a[1]   0.1277891262 0.175604098 -0.152860139 0.408438391
## a[2]  -0.1277909909 0.175604451 -0.408440820 0.152858838
## bw    -0.0005065113 0.001812617 -0.003403424 0.002390401
## sigma  0.9587527406 0.095336406  0.806386750 1.111118731
```

```
plot( coeftab(m6.21))
```

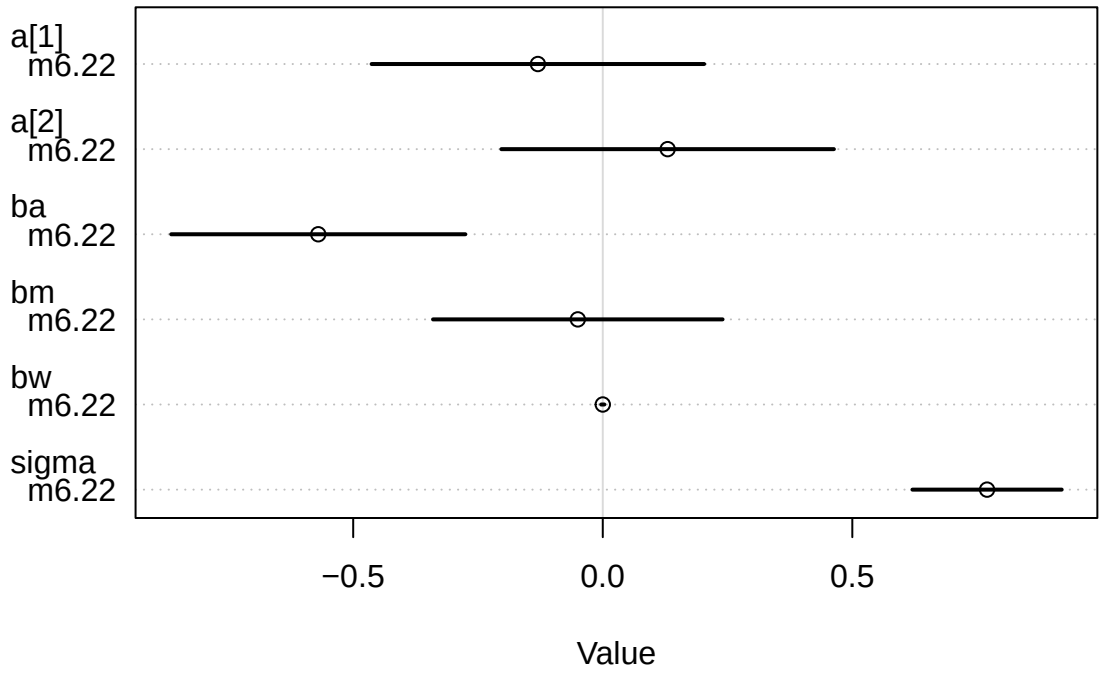


The estimated coefficient β_W being zero confirms this independency

```
# Testing second independency
m6.22 <-quap(
  alist(
  D ~dnorm(mu,sigma),
  mu <-a[S]+ba*A+bm*M+bw*W, # model including S and conditioned on A, M, W
  a[S] ~dnorm(0,0.2),
  ba ~dnorm(0,0.5),
  bm ~dnorm(0,0.5),
  bw ~dnorm(0,0.5),
  sigma ~dexp(1)
  ) ,data=d6)

precis(m6.22,depth=2)
```

```
##                mean          sd         5.5%       94.5%
## a[1]  -0.1318523473 0.169980619 -0.403514206  0.13980951
## a[2]   0.1318530869 0.169980918 -0.139809250  0.40351542
## ba    -0.5748964420 0.150359093 -0.815199313 -0.33459357
## bm    -0.0482968152 0.148009174 -0.284844062  0.18825043
## bw     0.0005226538 0.001647471 -0.002110322  0.00315563
## sigma  0.7654972184 0.076327147  0.643511695  0.88748274
```

```
plot( coeftab(m6.22))
```



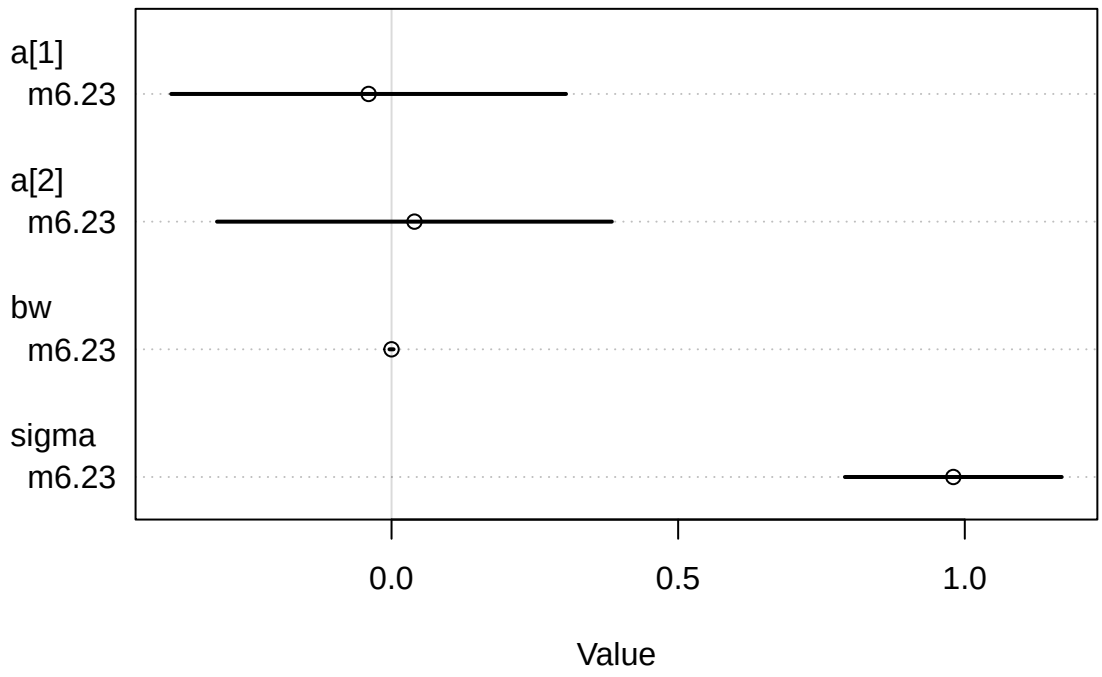
This second test is more dubious. While the SD associated with the intercepts is incredibly large, including zero, there is still a difference in means between southern and norther states. Rather than a situation where

the predictor is independent here there seems to be a masking effect from some outside variable acting in opposite ways in southern and northern states.

```
# Testing third independency
m6.23 <-quap(
  alist(
  M ~dnorm(mu,sigma),
  mu <-a[S]+bw*W, # model including W and conditioned on S
  a[S] ~dnorm(0,0.2),
  bw ~dnorm(0,0.5),
  sigma ~dexp(1)
  ) ,data=d6)

precis(m6.23,depth=2)
```

```
##                mean         sd         5.5%       94.5%
## a[1]  -0.0412427561 0.17570308 -0.322050217 0.239564705
## a[2]   0.0412429828 0.17570344 -0.239565043 0.322051009
## bw     0.0001634833 0.00182946 -0.002760347 0.003087314
## sigma  0.9780907337 0.09649114  0.823879253 1.132302214
```

```
plot( coeftab(m6.23))
```

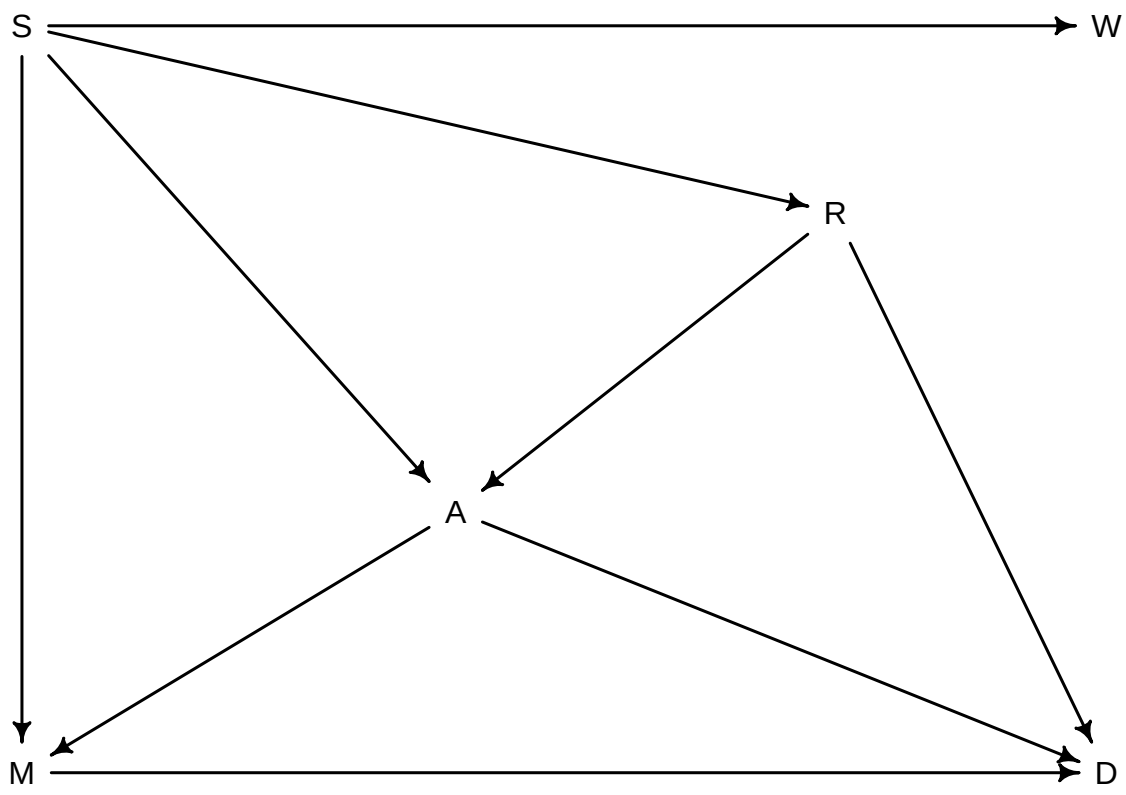


Here too the result of the test is quite positive: marriage rate is independent from the number of waffle houses once the southerness of a state has been included in the model.

Ultimately there might be a masking variable acting on states. It could be a religiosity index ('R') being higher in the southern states and lower in the northern ones. This R parameter would, in turn affects age of marriage, and divorce rates, creating a DAG with many more causal arrows.

What is more it makes little sense to have a causal arrow between W and D, since the previous tests have proven the effect of W is mediated entirely by S. So I can semove the W -> D direct path.

```
#Defining a new DAG including a religiosity predictor
dag.62 <-dagitty("dag{
  A ->D
  A ->M->D
  A <-S->M
  S ->W
  R <-S
  R ->A
  R ->D

  }")
coordinates(dag.62) <-list(x=c(A=0.4,M=0,S=0,D=1,W=1, R=0.75),
                           y=c(A=0.65,M=1,S=0,D=1,W=0, R=0.25))

drawdag(dag.62)
```




```
# I can check whether my reasoning about the pathways for this DAG is correct
adjustmentSets( dag.62,exposure="W",outcome="D", effect = "total")
```

```
## { A, M, R }
## { S }
```

```
# And check the conditional independencies of this DAG
impliedConditionalIndependencies( dag.62)
```

```
## A _||_ W | S
## D _||_ S | A, M, R
## D _||_ W | S
## D _||_ W | A, M, R
## M _||_ R | A, S
## M _||_ W | S
## R _||_ W | S
```

This revised DAG is still good for the previously found independencies A _||_ W| S and M _||_ W| S. Still doesn't imply that divorce rate is independent from the southerness of the state, once A, W and M have been included in the model, which was the condition giving us trouble, since the new variable R mediates the effect of the S predictor on D (divorce rate). Furthermore this DAG carries the implication D _||_ W| S, proven in the previous exercise.

This DAG would, still, have to be tested for four additional independencies:

1) D _||_ S| A,M,R
2) D _||_ W| A,M,R
3) M _||_ R| A,S
4) R _||_ W| S

This without accounting for other unobserved variables.