homework5

Christopher Huong

2024-02-26

5E1. Which of the linear models below are multiple linear regressions?

2 and 4.

5E2. Write down a multiple regression to evaluate the claim: Animal diversity is linearly related to latitude, but only after controlling for plant diversity. You just need to write down the model definition.

L ~ N(mu, sigma)
mu = a + bA * A[i] + bP * P[i]

5E3. Write down a multiple regression to evaluate the claim: Neither amount of funding nor size of laboratory is by itself a good predictor of time to PhD degree; but together these variables are both positively associated with time to degree. Write down the model definition and indicate which side of zero each slope parameter should be on.

T ~ N(mu, sigma)
mu = a + bF * F[i] + bL * L[i]

Slope should be greater than zero since both variables are positively associated with T. The predictors should be negatively associated with each other.

5E4. Suppose you have a single categorical predictor with 4 levels (unique values), labeled A, B, C and D. Let Ai be an indicator variable that is 1 where case i is in category A. Also suppose Bi , Ci , and Di for the other categories. Now which of the following linear models are inferentially equivalent ways to include the categorical variable in a regression? Models are inferentially equivalent when it’s possible to compute one posterior distribution from the posterior distribution of another model.

1, 3, 4 are equivalent. Maybe 5 too.

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).

The share of state GDP by agriculture is positively associated with illegal abortions. The percentage of republican voters by state are positively associated with illegal abortions.
Farmers tend to vote republican (who tend to vote against legal abortion), thus the share of state GDP by agriculture influences the percentage of republican voters.

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.

Depression is positively influenced by chronic illness and negatively influenced by exercise. Exercise and chronic illness are strongly negatively associated.
Thus, in a zero-order correlation matrix, Exercise and chronic illness will show weak associations with depression.

A multiple regression predicting depression with both illness and exercise as a predictor will show strong effects of each. Thus they were masked in the correlation matrix

```
n <- 100

Ex <- rnorm(n)
Ill <- rnorm(n, mean=Ex)
Depr <- rnorm(n, mean = Ex-Ill)
d <- data.frame(Ex=Ex, Ill=Ill, Depr=Depr)

cor(d)
```

```
##              Ex        Ill        Depr
## Ex   1.00000000  0.7019339  0.04282736
## Ill  0.70193385  1.0000000 -0.47039350
## Depr 0.04282736 -0.4703935  1.00000000
```

```
lm(Depr ~ Ex + Ill, data=d) |> summary()
```

```
##
## Call:
## lm(formula = Depr ~ Ex + Ill, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85325 -0.81893  0.01795  0.70357  2.26979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0516     0.1057   0.488    0.626
## Ex            1.1530     0.1588   7.262 9.57e-11 ***
## Ill          -1.1085     0.1138  -9.743 4.80e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.037 on 97 degrees of freedom
## Multiple R-squared:  0.4955, Adjusted R-squared:  0.4851
## F-statistic: 47.64 on 2 and 97 DF,  p-value: 3.861e-15
```

5M3. 5H1. In the divorce example, suppose the DAG is: M → A → D. What are the implied conditional independencies of the graph? Are the data consistent with it?

M is independent of D given A.

```
library(dagitty)
library(rethinking)

data("WaffleDivorce");d<-WaffleDivorce
d$D <- standardize(d$Divorce)
d$M <- standardize(d$Marriage)
d$A <- standardize(d$MedianAgeMarriage)
```

Check linear effect of M on D

```
M_on_D <- quap(
  alist(
    D ~ dnorm(mu,sigma),
    mu <- a + bM*M,
    a ~ dnorm(0, 0.2),
    bM ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),data=d
)

precis(M_on_D)
```

```
##                mean         sd       5.5%     94.5%
## a     -7.874230e-11 0.10824650 -0.1729988 0.1729988
## bM     3.500539e-01 0.12592757  0.1487974 0.5513105
## sigma  9.102663e-01 0.08986264  0.7666485 1.0538842
```

Positive slope of M on D
Now add A to the model

```
MA_on_D <- quap(
  alist(
    D ~ dnorm(mu,sigma),
    mu <- a + bM*M + bA*A,
    a ~ dnorm(0, 0.2),
    bM ~ dnorm(0, 0.5),
    bA ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),data=d
)

precis(MA_on_D)
```

```
##                mean         sd       5.5%      94.5%
## a     -5.669547e-08 0.09707612 -0.1551465  0.1551463
## bM    -6.538082e-02 0.15077322 -0.3063456  0.1755839
## bA    -6.135133e-01 0.15098376 -0.8548145 -0.3722120
## sigma  7.851190e-01 0.07784364  0.6607098  0.9095282
```

The slope of M on D disappears when A is included in the model. The data are consistent with the DAG.

5H2. Assuming that the DAG for the divorce example is indeed M → A → D, fit a new model and use it to estimate the counterfactual effect of halving a State's marriage rate M. Use the

counterfactual example from the chapter (starting on page 140) as a template.

```
m1 <- quap(
  alist(
    # M -> A
    A ~ dnorm(mu_MA, sigma_MA),
    mu_MA <- a_MA + bM*M,
    a_MA ~ dnorm(0, 0.2),
    bM ~ dnorm(0, 0.5),
    sigma_MA ~ dexp(1),

    # A -> D
    D ~ dnorm(mu_AD, sigma_AD),
    mu_AD <- a_AD + bA*A,
    a_AD ~ dnorm(0, 0.2),
    bA ~ dnorm(0, 0.5),
    sigma_AD ~ dexp(1)
  )
  ,data=d
)


precis(m1)
```

```
##                    mean         sd       5.5%      94.5%
## a_MA     -4.328736e-06 0.08683947 -0.1387906  0.1387819
## bM       -6.947070e-01 0.09571628 -0.8476801 -0.5417339
## sigma_MA  6.816560e-01 0.06756012  0.5736818  0.7896301
## a_AD      4.395913e-06 0.09737386 -0.1556178  0.1556266
## bA       -5.684247e-01 0.10999268 -0.7442143 -0.3926352
## sigma_AD  7.882736e-01 0.07799848  0.6636170  0.9129302
```

Simulating manipulation of marriage rate on divorce rate

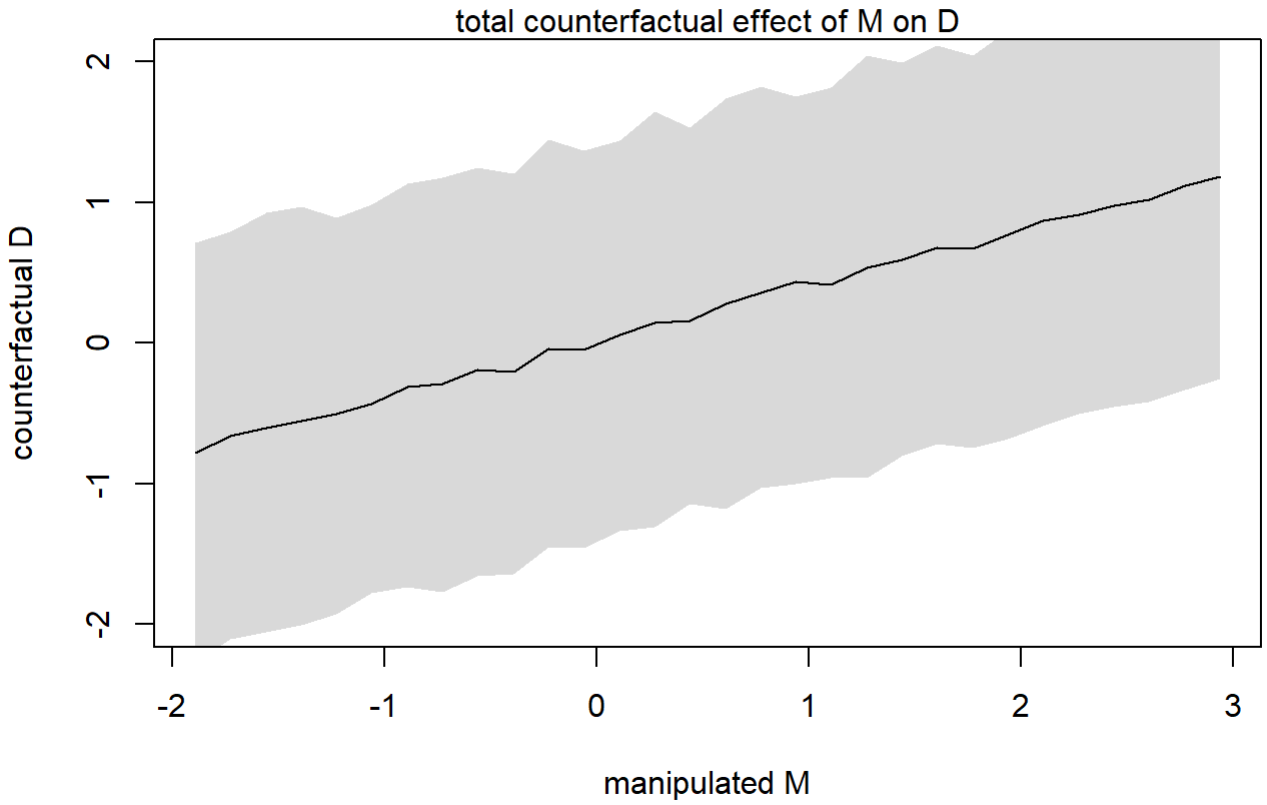
```
M_seq <- seq(from=min(d$M)-0.15, to=max(d$M)+0.15, length.out=30)

sim_dat <- data.frame(M=M_seq)

s <- sim(m1, data=sim_dat, vars=c("A", "D"))
```

Plot

```
plot(sim_dat$M, colMeans(s$D),
     ylim=c(-2,2), type="l",
     xlab="manipulated M", ylab="counterfactual D")
shade(apply(s$D,2,PI), sim_dat$M)
mtext("total counterfactual effect of M on D")
```



Effect of halving marriage rate

```
#half of mean marriage rate
(mean(d$Marriage) / 2)
```

```
## [1] 10.057
```

```
#convert to standardized units
((mean(d$Marriage) / 2) - mean(d$Marriage))/sd(d$Marriage)
```

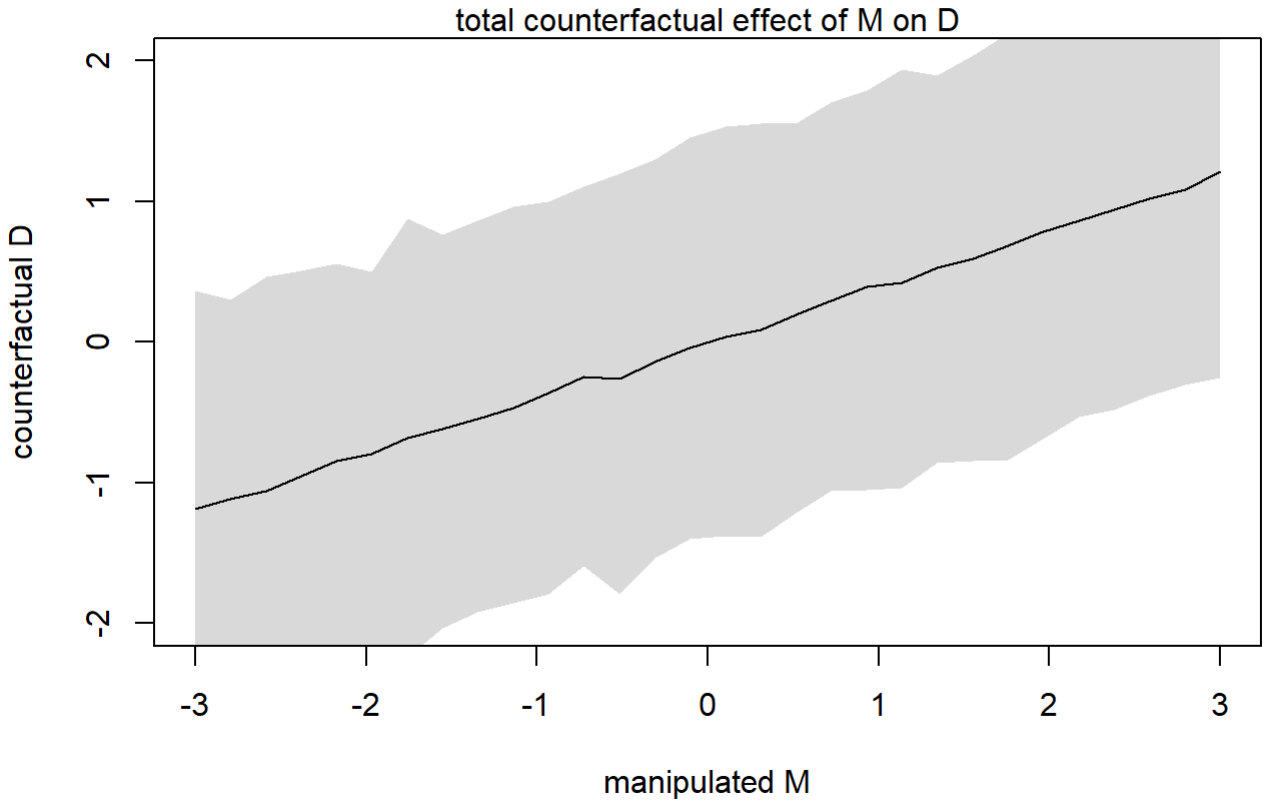
```
## [1] -2.648039
```

Increase range of x-axis to see effect

```
M_seq <- seq(from=-3, to=3, length.out=30)

sim_dat <- data.frame(M=M_seq)

s <- sim(m1, data=sim_dat, vars=c("A", "D"))
plot(sim_dat$M, colMeans(s$D),
     ylim=c(-2,2), type="l",
     xlab="manipulated M", ylab="counterfactual D")
shade(apply(s$D,2,PI), sim_dat$M)
mtext("total counterfactual effect of M on D")
```



A -2.6 decrease in standardized marriage rate (which corresponds to halving the marriage rate) would have a counterfactual effect of decrease of approximately 1 standardized divorce rate.

```
M_seq <- c(0, -2.648039)
sim_dat <- data.frame(M=M_seq)
s <- sim(m1, data=sim_dat, vars=c("A", "D"))

mean(s$D[,2]) - mean(s$D[,1])
```

```
## [1] -0.9924691
```

5H3. Return to the milk energy model, m5.7. Suppose that the true causal relationship among the variables is:

M -> N -> K
M -> K

Now compute the counterfactual effect on K of doubling M. You will need to account for both the direct and indirect paths of causation. Use the counterfactual example from the chapter (starting on page 140) as a template.

```
rm(list=ls())
data("milk"); d <- milk
d$K <- standardize(d$kcal.per.g)
d$M <- standardize(d$mass)
d$N <- standardize(d$neocortex.perc)
d <- d[complete.cases(d$K, d$M, d$N),]
```

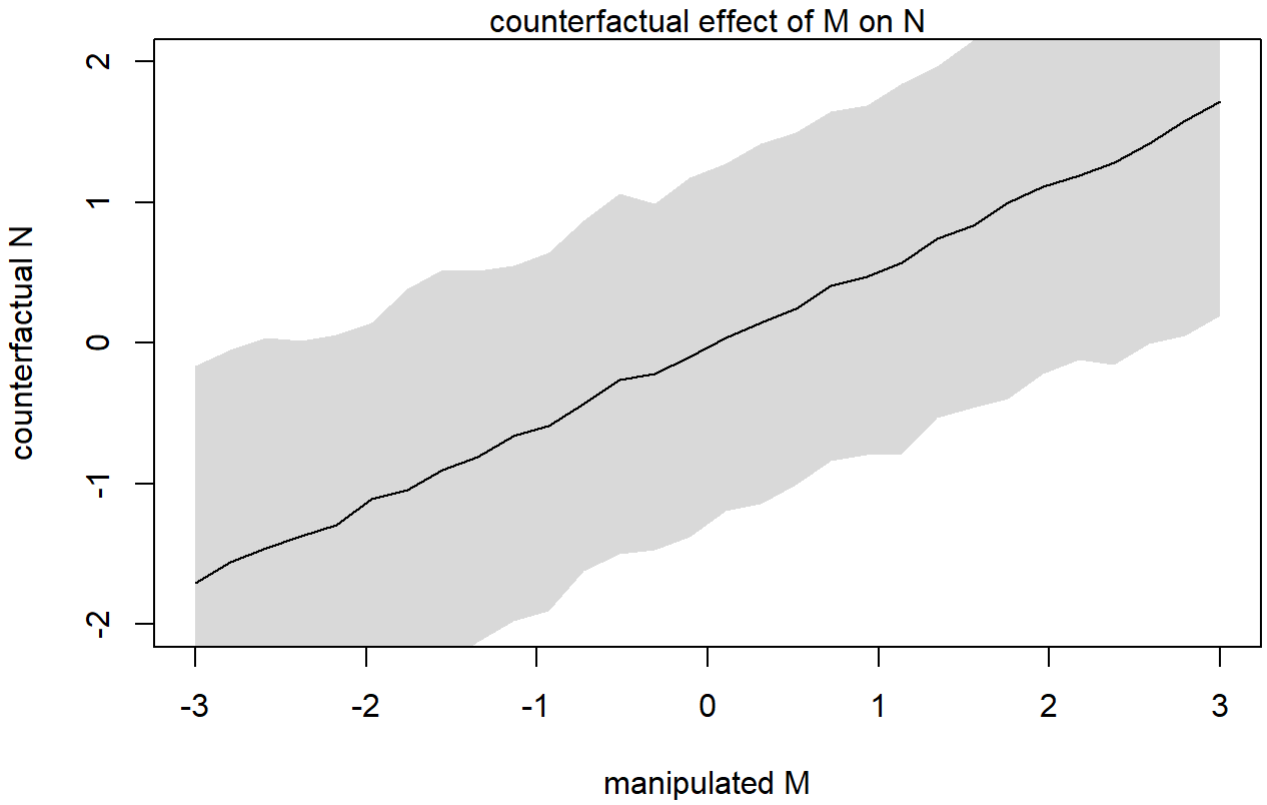
```
m2 <- quap(
  alist(
    # M -> K <- N
    K ~ dnorm(mu, sigma),
    mu <- a + bM*M + bN*N,
    a ~ dnorm(0, 0.2),
    bM ~ dnorm(0, 0.5),
    bN ~ dnorm(0, 0.5),
    sigma ~ dexp(1),
    # M -> N
    N ~ dnorm(mu_MN, sigma_MN),
    mu_MN <- a_MN + bMN * M,
    a_MN ~ dnorm(0, 0.2),
    bMN ~ dnorm(0, 0.5),
    sigma_MN ~ dexp(1)
  ),
  data=d
)

precis(m2)
```

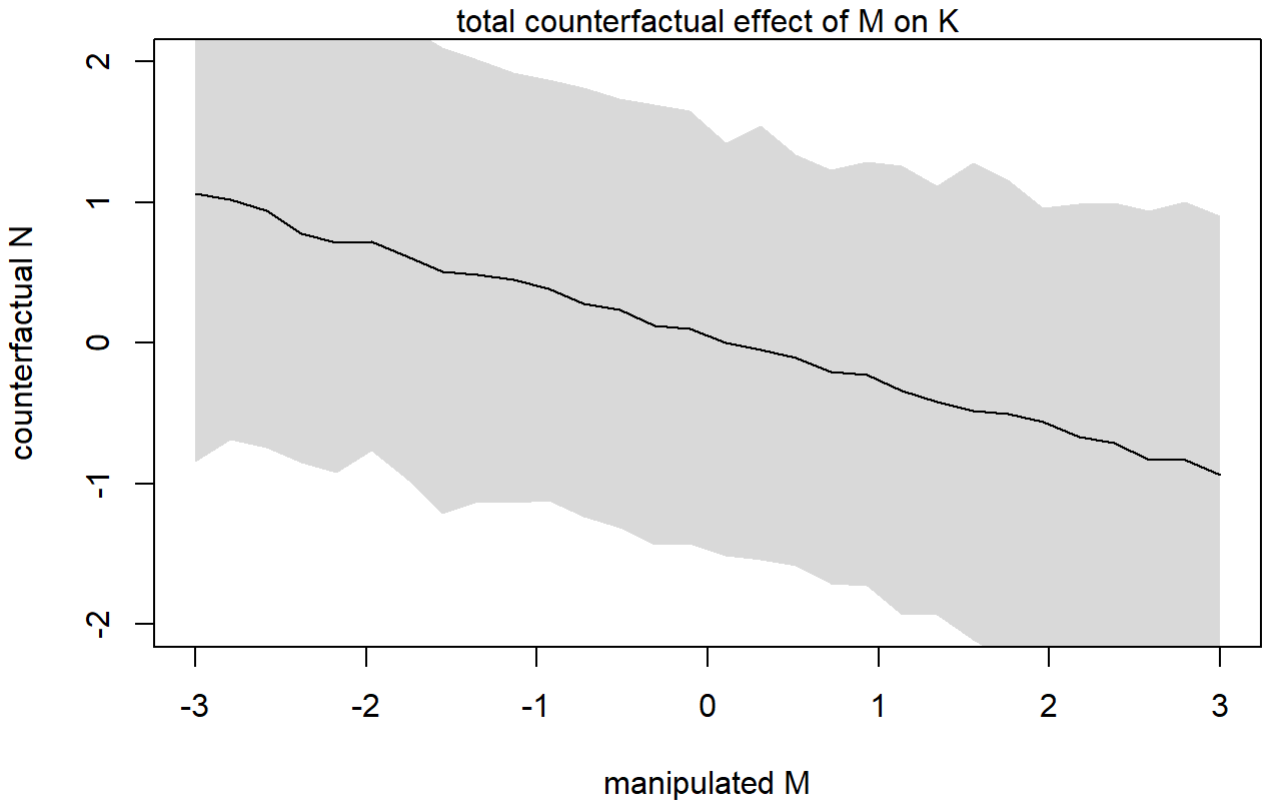
```
##                mean        sd        5.5%      94.5%
## a          0.06963521 0.1437690 -0.16013535  0.2994058
## bM        -0.56042876 0.2432292 -0.94915594 -0.1717016
## bN         0.42041149 0.2322509  0.04922973  0.7915932
## sigma      0.84043514 0.1448770  0.60889364  1.0719766
## a_MN      -0.02344729 0.1351873 -0.23950273  0.1926082
## bMN        0.55688853 0.1854419  0.26051659  0.8532605
## sigma_MN   0.75325094 0.1263448  0.55132753  0.9551744
```

```
M_seq <- seq(from=-3,to=3,length.out=30)
sim_dat <- data.frame(M=M_seq)
s <- sim(m2, data=sim_dat, vars=c("N", "K"))
```

```
plot(sim_dat$M, colMeans(s$N),
     ylim=c(-2,2), type="l",
     xlab="manipulated M", ylab="counterfactual N")
shade(apply(s$N,2,PI), sim_dat$M)
mtext("counterfactual effect of M on N")
```




```
plot(sim_dat$M, colMeans(s$K),
     ylim=c(-2,2), type="l",
     xlab="manipulated M", ylab="counterfactual N")
shade(apply(s$K,2,PI), sim_dat$M)
mtext("total counterfactual effect of M on K")
```



Effect of doubling M on K

```
mean(d$mass) * 2
```

```
## [1] 33.27529
```

```
(33.27529 - mean(d$mass)) / sd(d$mass)
```

```
## [1] 0.7055134
```

Doubling M corresponds to an increase in 0.706 standardized units

```
M_seq <- c(0, 0.7055134)
sim_dat <- data.frame(M=M_seq)
s <- sim(m2, data=sim_dat, vars=c("N", "K"))

mean(s$K[,2]) - mean(s$K[,1])
```

```
## [1] -0.2494746
```

Doubling M has a counterfactual effect of -0.24 standardized units on K