chapter 8 solutions

Easy

1. For each of the causal relationships below, name a hypothetical third variable that would lead to an interaction effect.

- Bread dough rises because of yeast -> temperature
- Education leads to higher income -> gender
- Gasoline makes a car go -> tire pressure

2. Which of the following explanations invokes an interaction?

(1) Caramelizing onions requires cooking over low heat and making sure the onions do not dry out.
(2) A car will go faster when it has more cylinders or when it has a better fuel injector.
(3) Most people acquire their political beliefs from their parents, unless they get them instead from their friends.
(4) Intelligent animal species tend to be either highly social or have manipulative appendages (hands, tentacles, etc.)

All of them?

3. For each of the explanations in 8E2, write a linear model that expresses the stated relationship.

```
m1 <- alist(carmelize ~ dnorm(mu, sigma),
            mu <- a + b*heat + c*dryness + d*heat*dryness)

m2 <- alist(car_speed ~ dnorm(mu, sigma),
            mu <- a + b*cylinders + c*fuel_injector + d*cylinder*fuel_injector)

m3 <- alist(belief ~ dnorm(mu, sigma),
            mu <- a + b*parents + c*friends + d*parents*friends)

m4 <- alist(intelligence ~ dnorm(mu, sigma),
            mu <- a + b*social + c*appendages + d*social*appendages)
```

Medium.

1. Recall the tulips example from the chapter. Suppose another set of treatments adjusted the temperature in the greenhouse over two levels: cold and hot. The data in the chapter were collected at the cold temperature. You find none of the plants grown under the hot temperature developed any blooms at all, regardless of the water and shade levels. Can you explain this result in terms of interactions between water, shade, and temperature?

Blooms are conditional on water and shade, which are conditional on temperature.

2. Can you invent a regression equation that would make the bloom size zero, whenever the temperature is hot?

```
# bloom = (a + b*water + c*shade + d*water*shade)*temp_cold
```

Where temp_cold = 0 if hot, 1 otherwise.

3. In parts of North America, ravens depend upon wolves for their food. This is because ravens are carnivorous but cannot usually kill or open carcasses of prey. Wolves however can and do kill and tear open animals, and they tolerate ravens co-feeding at their kills. This species relationship is generally described as a "species interaction." Can you invent a hypothetical set of data on raven population size in which this relationship would manifest as a statistical interaction? Do you think the biological interaction could be linear? Why or why not?

```
wolf_pop <- rnorm(1000, 500, 100)
raven_pop <- rnorm(1000, wolf_pop, 30)

d <- data.frame(wolf_pop = wolf_pop, raven_pop = raven_pop)
head(d)
```

```
##   wolf_pop raven_pop
## 1 538.2936  525.1487
## 2 599.2484  614.4311
## 3 410.8140  402.2192
## 4 416.6360  389.6639
## 5 414.6542  410.9839
## 6 461.6666  450.0765
```

I think it's possible the biological interaction could be linear but there is probably some carrying capacity where more wolves wouldn't necessarily mean more ravens, if for example the wolves were all starving, or something.

8H1. Return to the data(tulips) example in the chapter. Now include the bed variable as a predictor in the interaction model. Don't interact bed with the other predictors; just include it as a main effect. Note that bed is categorical. So to use it properly, you will need to either construct dummy variables or rather an index variable, as explained in Chapter 6.

```
library(rethinking)
```

```
## Loading required package: rstan

## Loading required package: StanHeaders

## Loading required package: ggplot2

## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

## Loading required package: parallel

## rethinking (Version 2.13)

##
## Attaching package: 'rethinking'
## The following object is masked from 'package:stats':
##
##     rstudent
```

```
data(tulips)
d <- tulips

d$blooms_std <- d$blooms / max(d$blooms)
d$water_cent <- d$water - mean(d$water)
d$shade_cent <- d$shade - mean(d$shade)

mbed <- quap(
alist(
blooms_std ~ dnorm( mu , sigma ) ,
mu <- a[bed] + bw*water_cent + bs*shade_cent + bws*water_cent*shade_cent ,
a[bed] ~ dnorm( 0.5 , 0.25 ) ,
bw ~ dnorm( 0 , 0.25 ) ,
bs ~ dnorm( 0 , 0.25 ) ,
bws ~ dnorm( 0 , 0.25 ) ,
sigma ~ dexp( 1 )
) , data=d )


m_nobed <- quap(
alist(
blooms_std ~ dnorm( mu , sigma ) ,
mu <- a + bw*water_cent + bs*shade_cent + bws*water_cent*shade_cent ,
a ~ dnorm( 0.5 , 0.25 ) ,
bw ~ dnorm( 0 , 0.25 ) ,
bs ~ dnorm( 0 , 0.25 ) ,
bws ~ dnorm( 0 , 0.25 ) ,
sigma ~ dexp( 1 )
) , data=d )
```

```
precis(mbed, depth = 3)
```

```
##             mean         sd        5.5%       94.5%
## a[1]   0.2732671 0.03571442  0.21618852  0.33034560
## a[2]   0.3964002 0.03569695  0.33934962  0.45345085
## a[3]   0.4091123 0.03569588  0.35206341  0.46616123
## bw     0.2074365 0.02537452  0.16688313  0.24798990
## bs    -0.1138485 0.02536989 -0.15439451 -0.07330255
## bws   -0.1438906 0.03099535 -0.19342710 -0.09435400
## sigma  0.1081852 0.01469373  0.08470183  0.13166867
```

2. Use WAIC to compare the model from 8H1 to a model that omits bed. What do you infer from this comparison? Can you reconcile the WAIC results with the posterior distribution of the bed coefficients?

```
compare(mbed, m_nobed, func = WAIC)
```

```
##               WAIC        SE     dWAIC       dSE     pWAIC     weight
## mbed    -23.27772  9.993045 0.000000       NA 9.792544 0.6539642
## m_nobed -22.00470 10.458586 1.273021 7.680831 6.575237 0.3460358
```

What do you infer: The model that includes bed is slightly better than the model without bed, but not by a whole lot (they are given comparable weights)

Can you reconcile this with the posterior distribution of bed coefficients: All of the bed coefficients are pretty similar, (b and c are almost identical), and bed a is a little different than b and c. But since this indicator is kind of like a "single level indicator" knowing the bed isn't a huge information gain. Especially, knowing b or c there's basically no difference, but there's some small difference between bed a and beds b/c.

```
precis(mbed, depth = 3)
```

```
##             mean         sd        5.5%       94.5%
## a[1]   0.2732671 0.03571442  0.21618852  0.33034560
## a[2]   0.3964002 0.03569695  0.33934962  0.45345085
## a[3]   0.4091123 0.03569588  0.35206341  0.46616123
## bw     0.2074365 0.02537452  0.16688313  0.24798990
## bs    -0.1138485 0.02536989 -0.15439451 -0.07330255
## bws   -0.1438906 0.03099535 -0.19342710 -0.09435400
## sigma  0.1081852 0.01469373  0.08470183  0.13166867
```

3. Pretty sure they mean return to model m8.3.

```
data(rugged)
d <- rugged
# make log version of outcome
d$log_gdp <- log( d$rgdppc_2000 )
# extract countries with GDP data
dd <- d[ complete.cases(d$rgdppc_2000) , ]
# rescale variables
dd$log_gdp_std <- dd$log_gdp / mean(dd$log_gdp)
dd$rugged_std <- dd$rugged / max(dd$rugged)


dd$cid <- ifelse( dd$cont_africa==1 , 1 , 2 )
```

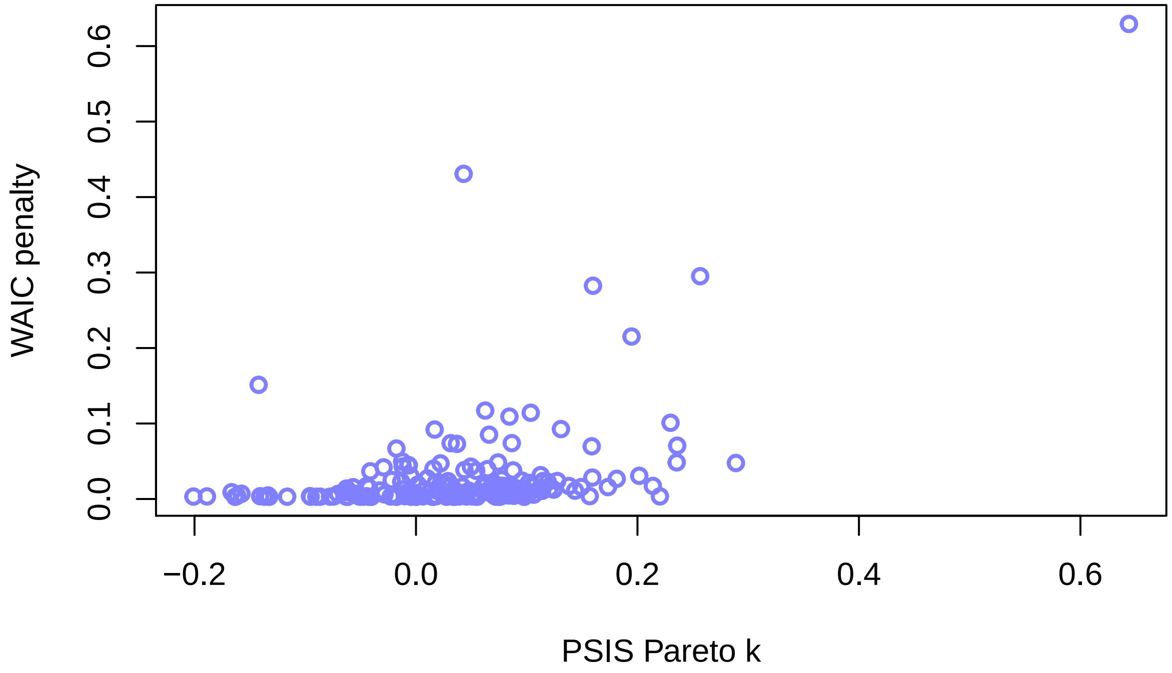
```
m8.3 <- quap(
alist(
log_gdp_std ~ dnorm( mu , sigma ) ,
mu <- a[cid] + b[cid]*( rugged_std - 0.215 ) ,
a[cid] ~ dnorm( 1 , 0.1 ) ,
b[cid] ~ dnorm( 0 , 0.3 ) ,
sigma ~ dexp( 1 )
) , data=dd )
```

3. Use WAIC pointwise penalties and PSIS Pareto k values to measure relative influence of each country.

```
PSIS_model <- PSIS(m8.3,pointwise=TRUE)
```

```
## Some Pareto k values are high (>0.5). Set pointwise=TRUE to inspect individual points.
```

```
WAIC_model <- WAIC(m8.3,pointwise=TRUE)
plot( PSIS_model$k , WAIC_model$penalty , xlab="PSIS Pareto k" ,
ylab="WAIC penalty" , col=rangi2 , lwd=2 )
```



3. By these criteria, is Seychelles influencing the results?

```
# get row of seycells
seychelles_index = which(dd$country == 'Seychelles')

# get k value for seycells
PSIS_model[145,]
```

```
##         PSIS       lppd   penalty  std_err         k
## 145 1.276007 -0.6380034 0.6187556 15.34518 0.6438172
```

```
dd[seychelles_index, ]
```

```
##     isocode isonum    country rugged rugged_popw rugged_slope rugged_lsd
## 199     SYC    690 Seychelles  4.885       1.802       11.129      1.278
##     rugged_pc land_area    lat    lon   soil desert tropical dist_coast
## 199    54.101        46 -6.723 51.924 13.043      0      100          0
##     near_coast gemstones rgdppc_2000 rgdppc_1950_m rgdppc_1975_m rgdppc_2000_m
## 199        100         0    17957.47      1912.258      3250.874      6353.528
##     rgdppc_1950_2000_m q_rule_law cont_africa cont_asia cont_europe
## 199           3625.007       0.58           1         0           0
##     cont_oceania cont_north_america cont_south_america legor_gbr legor_fra
## 199            0                  0                  0         0         1
##     legor_soc legor_deu legor_sca colony_esp colony_gbr colony_fra colony_prt
## 199         0         0         0          0          1          0          0
##     colony_oeu africa_region_n africa_region_s africa_region_w africa_region_e
## 199          0               0               0               0               1
##     africa_region_c slave_exports dist_slavemkt_atlantic dist_slavemkt_indian
## 199               0             0                 11.457                1.742
##     dist_slavemkt_saharan dist_slavemkt_redsea pop_1400 european_descent
## 199                 4.635                2.253        0               NA
##      log_gdp log_gdp_std rugged_std cid
## 199 9.795761    1.150126  0.7876491   1
```

According to this, Seychelles has a k value of .44 (less than .5). We can see from the plot it is the most influential point in this data. So yes, it is influencing the results.

3. Are there other nations that are relatively influential?

We can see from the plot that there are a smattering of other points (at least two that jump out) that also seem to have above averagely high WAIC penalty scores, but none of them have k's over .5

3. If so, can you explain why?

I'd say the other countries shouldn't be overly influential as to hurt prediction, but they may influence the model more than some countries if that's fair?

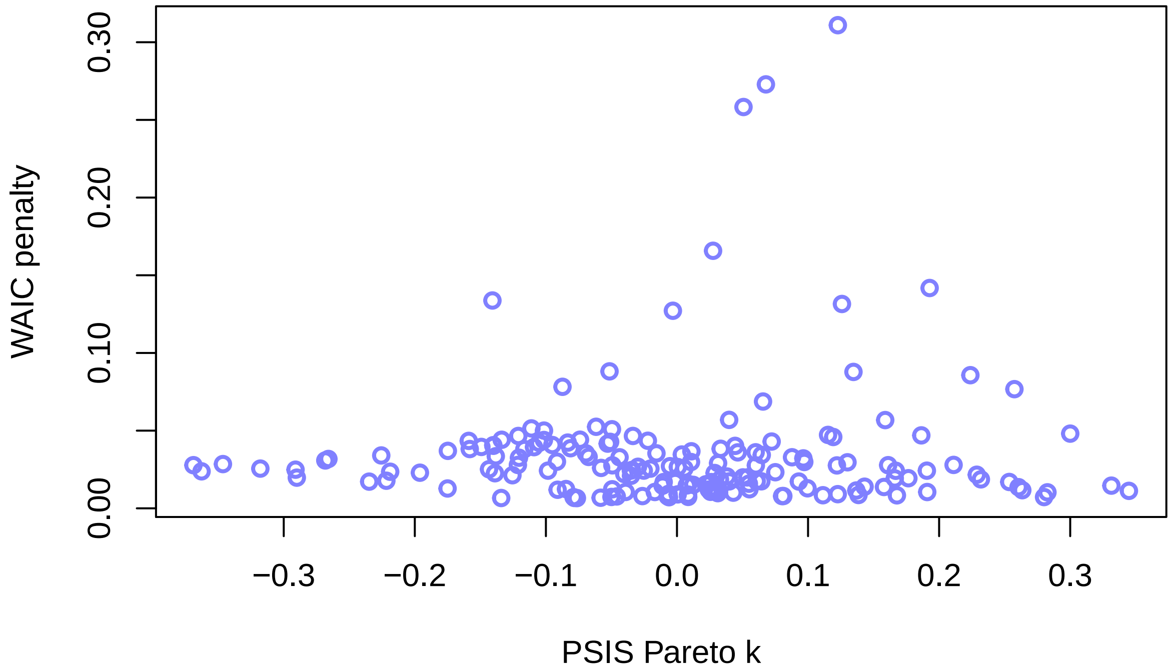
3 . Now use robust regression, as described in the previous chapter. Modify m8.5 to use a Student-t distribution with v = 2.

```
m8.3b <- quap(
alist(
log_gdp_std ~ dstudent(2, mu , sigma ) ,
mu <- a[cid] + b[cid]*( rugged_std - 0.215 ) ,
a[cid] ~ dnorm( 1 , 0.1 ) ,
b[cid] ~ dnorm( 0 , 0.3 ) ,
sigma ~ dexp( 1 )
) , data=dd )
```

3. Does this change the results in a substantial way?

Yes, it does change the result in a substaintial way. Seychelle's k value dropped by an order of magnitude and switched direction(.44 to -.04) !

```
PSIS_model <- PSIS(m8.3b,pointwise=TRUE)
WAIC_model <- WAIC(m8.3b,pointwise=TRUE)
plot( PSIS_model$k , WAIC_model$penalty , xlab="PSIS Pareto k" ,
ylab="WAIC penalty" , col=rangi2 , lwd=2 )
```



```
PSIS_model[145,]
```

```
##         PSIS       lppd  penalty  std_err         k
## 145 1.900719 -0.9503596 0.308239 18.12506 0.1227343
```

4. Try to honestly evaluate the main effects of both mean.growing.season and sd.growing.season on the outcome

```
data(nettle)
d <- nettle

d$lang.per.cap <- d$num.lang / d$k.pop
d$outcome = log(d$lang.per.cap)
```

4a. Evaluate the hypothesis that language diversity, as measured by log(lang.per.cap), is positively associated with the average length of the growing season, mean.growing.season. Consider log(area) in your regression(s) as a covariate (not an interaction).

```
d$l_area = log(d$area)

flista <- alist(outcome ~ dnorm(mu, sigma),
                mu <- a + b*mean.growing.season + c*l_area,
                a ~ dnorm(0,1),
                b ~ dnorm(0,1),
                c ~ dnorm(0,1),
                sigma ~ dexp(1))
```

```
mod_a <- quap(flista, d)
precis(mod_a)
```

```
##              mean         sd        5.5%      94.5%
## a     -0.80725214 0.89632706 -2.23975589  0.6252516
## b      0.09692787 0.04922838  0.01825141  0.1756043
## c     -0.41055670 0.06750475 -0.51844232 -0.3026711
## sigma  1.39865166 0.11410088  1.21629641  1.5810069
```

4a. Interpret your results.

These results suggest that the growing season is positively associated with languages per capita...

4b. Now evaluate the hypothesis that language diversity is negatively associated with the standard deviation of length of growing season, sd.growing.season. This hypothesis follows from uncertainty in harvest favoring social insurance through larger social networks and therefore fewer languages. Again, consider log(area) as a covariate (not an interaction).

```
flistb <- alist(outcome ~ dnorm(mu, sigma),
                mu <- a + b*sd.growing.season + c*l_area,
                a ~ dnorm(0,1),
                b ~ dnorm(0,1),
                c ~ dnorm(0,1),
                sigma ~ dexp(1))

mod_b <- quap(flistb, d)
precis(mod_b)
```

```
##             mean         sd       5.5%      94.5%
## a     -0.4636220 0.88071426 -1.8711735  0.9439295
## b     -0.1449119 0.17077606 -0.4178450  0.1280212
## c     -0.3659783 0.07650968 -0.4882556 -0.2437011
## sigma  1.4311867 0.11620293  1.2454720  1.6169014
```

4b. Interpret your results.

These results do not necessarily suggest that the standard deviation of growing season is negatively associated with languages.

4c. Finally, evaluate the hypothesis that mean.growing.season and sd.growing.season interact to synergistically reduce language diversity

```
flistc <- alist(outcome ~ dnorm(mu, sigma),
                mu <- a + b*sd.growing.season + c*l_area + e*mean.growing.season + f*mean.growing.seaso
                a ~ dnorm(0,1),
                b ~ dnorm(0,1),
                c ~ dnorm(0,1),
                e ~ dnorm(0,1),
                f ~ dnorm(0,1),
                sigma ~ dexp(1))

mod_c <- quap(flistc, d)
precis(mod_c)
```

```
##              mean         sd        5.5%        94.5%
## a     -1.18527359 0.92348624 -2.66118297  0.290635788
## b      0.39405903 0.37021032 -0.19760856  0.985726617
## c     -0.40612640 0.07862570 -0.53178545 -0.280467350
## e      0.17888722 0.06678560  0.07215093  0.285623503
## f     -0.07626488 0.04517991 -0.14847110 -0.004058647
## sigma  1.35607082 0.11165261  1.17762838  1.534513253
```

In this model, f is the coefficient of the interaction between mean.growing.season and sd.growing.season. The 89% CI suggests that this term is in fact negatively associated with languages. So, it seems to be the case that this hypothesis is true.