Statistical Reasoning 3 - Multiple Regression & DAGs

Ellie Kuskie & Rachel Parsons

Setup

```
# Run this to install some data packages
# devtools::install_github("rmcelreath/rethinking")
#library(rethinking)

# a function to scale and center. from rethinking package
standardize <- function(x) {
    x <- scale(x)
    z <- as.numeric(x)
    attr(z,"scaled:center") <- attr(x,"scaled:center")
    attr(z,"scaled:scale") <- attr(x,"scaled:scale")
    return(z)
}

  library(brms) # for statistics
```

```
Loading required package: Rcpp
```

```
Loading 'brms' package (version 2.23.0). Useful instructions
can be found by typing help('brms'). A more detailed introduction
to the package is available through vignette('brms_overview').
```

```
Attaching package: 'brms'
```

```
The following object is masked from 'package:stats':

    ar
```

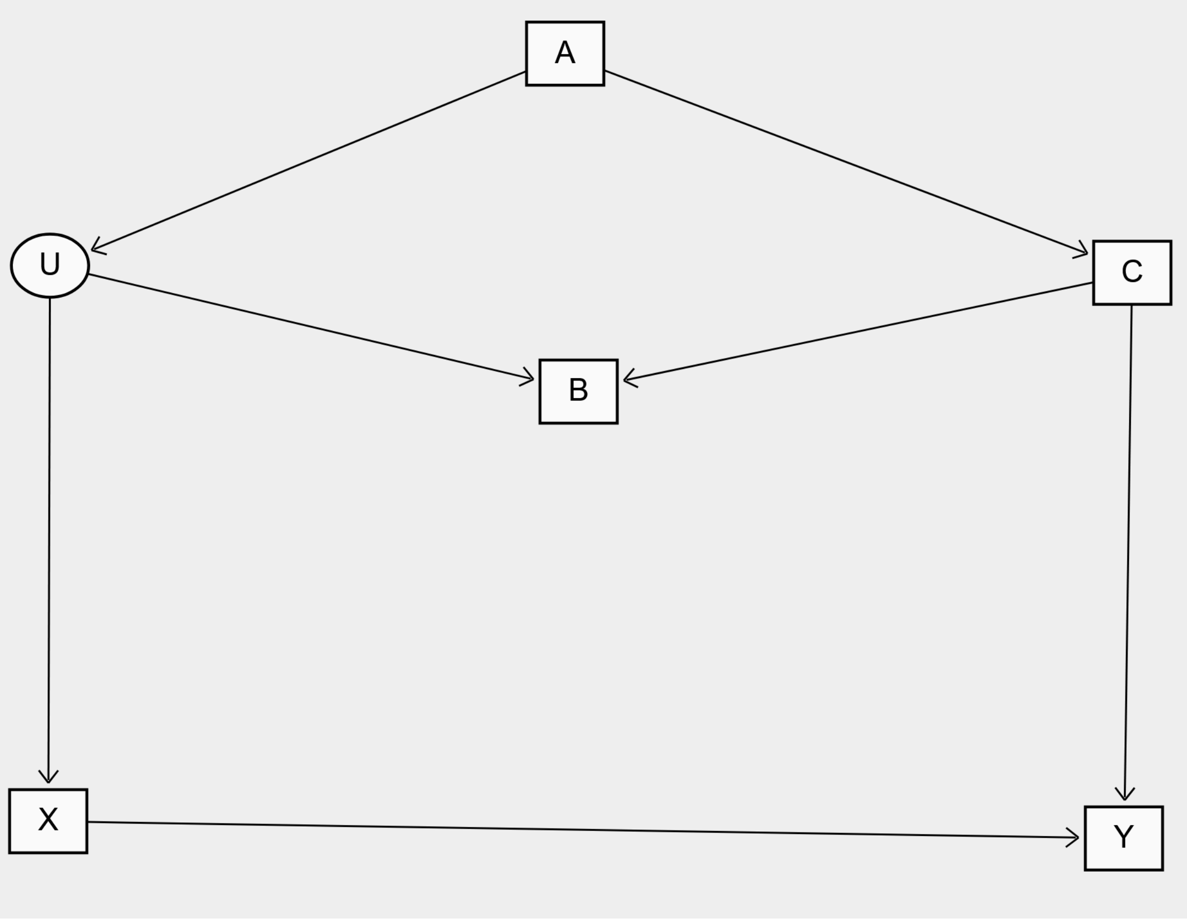
```r
library(tidyverse) # for data wrangling
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

1

Q1.1



dag { A [pos="-0.471,-1.755"] B [pos="-0.426,-0.301"] C [pos="1.428,-0.812"] U [latent,pos="-2.196,-0.843"] X [pos="-2.202,1.547"] Y [pos="1.400,1.621"] A -> C A -> U C -> B C -> Y U -> B U -> X X -> Y }

Q1.2

Forks

- U <- A -> C

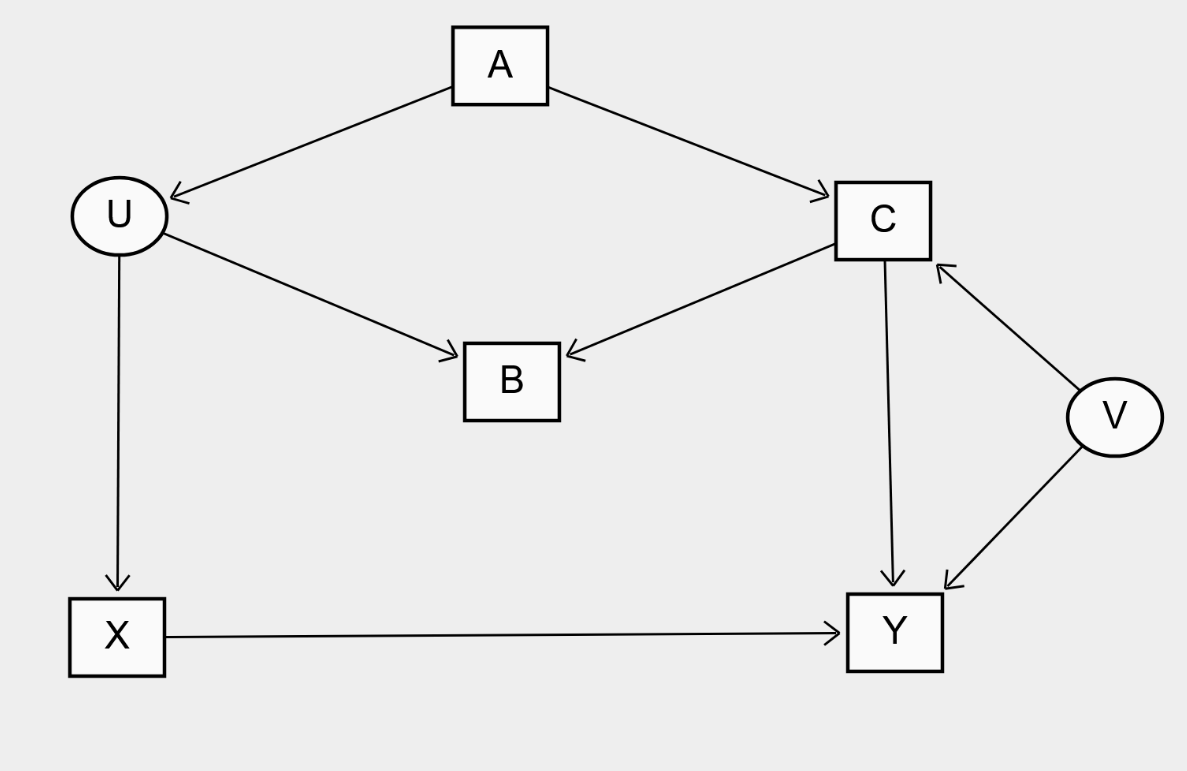
- B <- C -> Y
- B <- U -> X

Q1.3

Colliders

- X ->Y <- C
- U -> B <- C

Q1.4



Q1.5

Paths from X to Y

- X <- U -> B <- C -> Y
- X <- U <- A -> C -> Y
- X <- U -> B <- C <- V -> Y
- X <- U <- A -> C <- V -> Y

Q1.6

Open Paths

- X <- U <- A -> C -> Y

Q1.7

We should condition on A or C to estimate the direct effect of X on Y

2

```
# Load in the fox data
foxes <- read.csv('https://raw.githubusercontent.com/rmcelreath/rethinking/refs/heads/master/
```

```
head(foxes)
```

```
  group avgfood groupsize area weight
1     1    0.37         2 1.09   5.02
2     1    0.37         2 1.09   2.84
3     2    0.53         2 2.05   5.33
4     2    0.53         2 2.05   6.07
5     3    0.49         2 2.12   5.85
6     3    0.49         2 2.12   3.25
```

Q2.1

Fork

- groupsize <- avgfood -> weight

Pipe

- avgfood -> groupsize -> weight
- area -> avgfood -> weight
- area -> avgfood -> groupsize

Collider

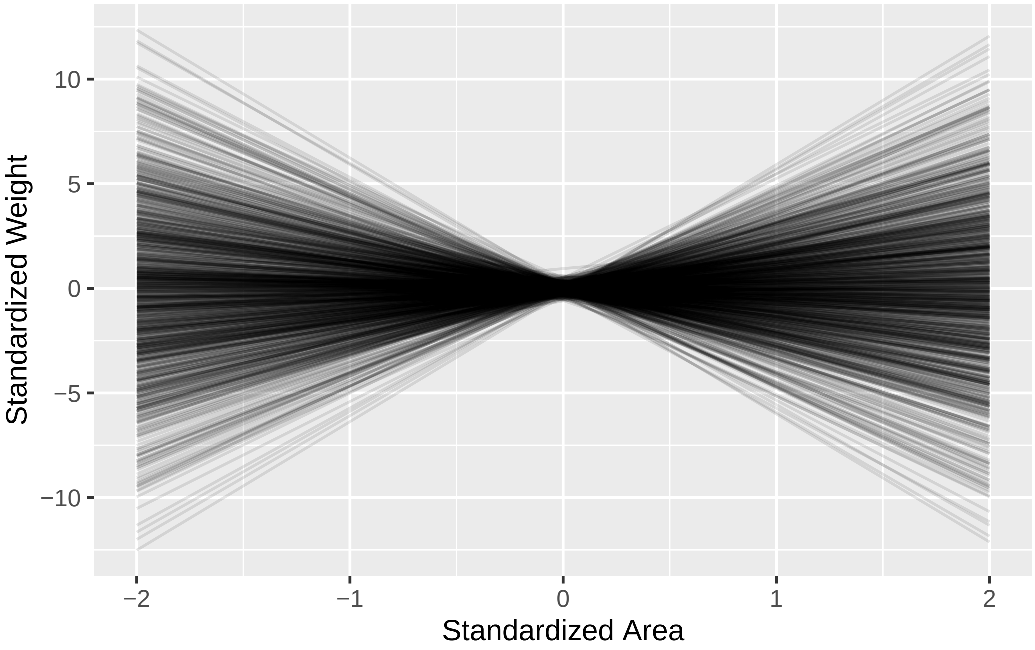
- avgfood -> weight <- groupsize

Total causal influence of area on weight

```r
fox_dat <- foxes %>%
  as_tibble() %>%
  select(area, avgfood, weight, groupsize) %>%
  mutate(across(everything(), standardize))
```

```r
n <- 1000
priorsims <- tibble(group = seq_len(n),
       alpha = rnorm(n, 0, 0.2), # prior for alpha
       beta = rnorm(n, 0, 2)) %>% # prior for beta
  expand(nesting(group, alpha, beta), # the expand function gives us all possible combination
         area = seq(from = -2, to = 2, length.out = 100)) %>% # set up a range of areas
  mutate(weight = alpha + beta * area) # calculate weight from the parameters and area
```

```r
ggplot(priorsims, aes(x = area, y = weight, group = group)) +
  geom_line(alpha = 1 / 10) +
  labs(x = "Standardized Area", y = "Standardized Weight")
```



Q2.2

Reasonable minimum is 2 kg

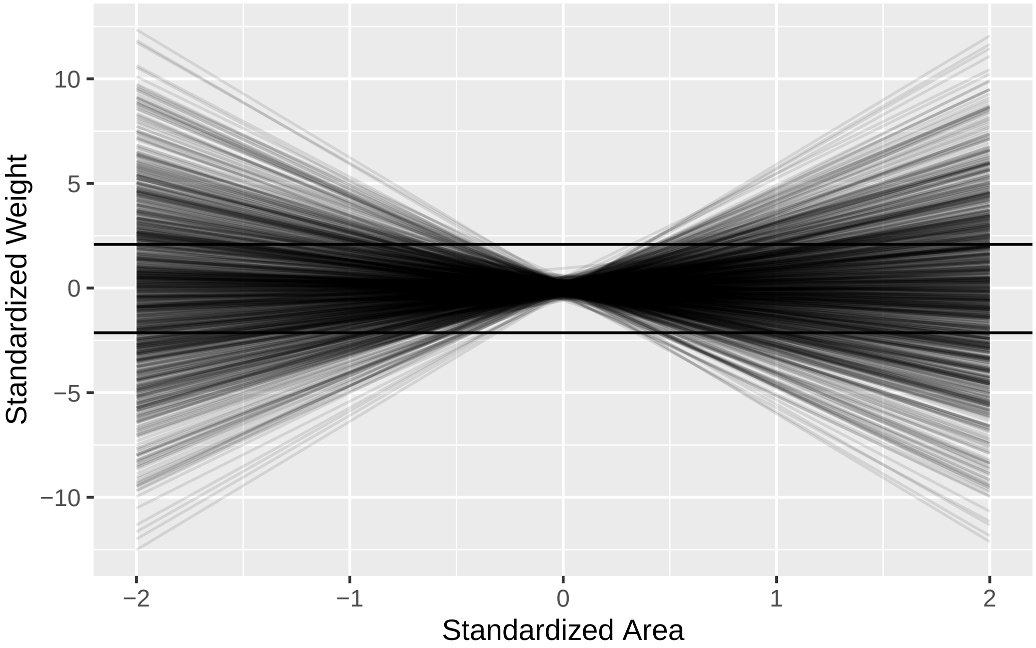
Q2.3

Reasonable maximum is 7 kg

Q2.4

```r
# get mean of fox weight
fox_weight_mean <- mean(foxes$weight)
# calculate fox weight standard deviation
fox_weight_sd <- sd(foxes$weight)
# calculate the standardized value of the "reasonable" minimum weight
standardizedmin <- (2 - fox_weight_mean)/fox_weight_sd
# calculate the standardized value of the "reasonable" maximum weight
standardizedmax <- (7 - fox_weight_mean)/fox_weight_sd
```

```r
ggplot(priorsims, aes(x = area, y = weight, group = group)) +
  geom_line(alpha = 1 / 10) +
  labs(x = "Standardized Area", y = "Standardized Weight") +
  # added both horizontal lines
  geom_hline(yintercept = standardizedmax) +
  geom_hline(yintercept = standardizedmin)
```



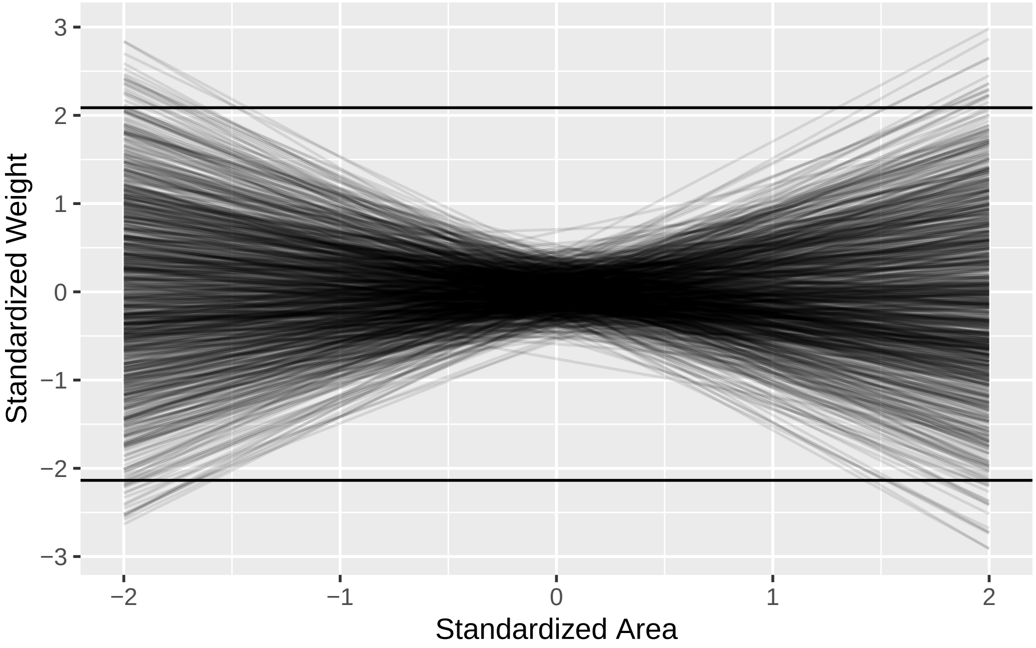
Q2.5

My priors fit the data reasonably well as my horizontal lines are both around 2 standard deviations away from the mean. A good amount of the priors do exceed these values.

Q2.6

```
n <- 1000
priorsims <- tibble(group = seq_len(n),
       alpha = rnorm(n, 0, 0.2), # prior for alpha
       beta = rnorm(n, 0, 0.5)) %>% # prior for beta
  expand(nesting(group, alpha, beta), # the expand function gives us all possible combinatio
         area = seq(from = -2, to = 2, length.out = 100)) %>% # set up a range of areas
  mutate(weight = alpha + beta * area) # calculate weight from the parameters and area
```

```
# remake of previous grpah from above using new priors simulation
ggplot(priorsims, aes(x = area, y = weight, group = group)) +
  geom_line(alpha = 1 / 10) +
  labs(x = "Standardized Area", y = "Standardized Weight") +
  geom_hline(yintercept = standardizedmax) +
  geom_hline(yintercept = standardizedmin)
```



Run Models

```
food_on_area <- brm(avgfood ~ 1 + area,
                    data = fox_dat,
                    family = gaussian,
                    # Here we set the priors that we investigated earlier
                    prior = c(prior(normal(0, 0.2), class = Intercept),
                              prior(normal(0, 2), class = b,),
                              prior(exponential(1), class = sigma)),
                    iter = 4000, warmup = 2000, chains = 4, cores = 4, seed = 1234,
                    file = "output/food_on_area")
```

```r
summary(food_on_area)
```

```
 Family: gaussian
  Links: mu = identity
Formula: avgfood ~ 1 + area
   Data: fox_dat (Number of observations: 116)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.00      0.04    -0.08     0.08 1.00     8484     6273
area          0.88      0.04     0.79     0.97 1.00     7603     5628

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.48      0.03     0.42     0.54 1.00     8026     5911

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Q2.7

```r
# new model for the affect of avgfood on weight
food_on_weight <- brm(weight ~ 1 + avgfood,
                    data = fox_dat,
                    family = gaussian,
                    # Here we set the priors that we investigated earlier
                    prior = c(prior(normal(0, 0.2), class = Intercept),
                              prior(normal(0, 2), class = b,),
                              prior(exponential(1), class = sigma)),
                    iter = 4000, warmup = 2000, chains = 4, cores = 4, seed = 1234,
                    file = "output/food_on_weight")
```

```r
summary(food_on_weight)
```

```
 Family: gaussian
  Links: mu = identity
Formula: weight ~ 1 + avgfood
   Data: fox_dat (Number of observations: 116)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.00      0.09    -0.17     0.17 1.00     7611     5664
avgfood      -0.02      0.10    -0.21     0.17 1.00     8144     5754

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     1.01      0.07     0.89     1.16 1.00     7547     5472

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Q2.8

We need to condition on groupsize

Q2.9

I think that as the number of foxes in a group increases, the weight of foxes will decrease. This is because with larger groups, the food will have to be split between the individuals.

Run the Model

```r
group_on_weight <- brm(weight ~ 1 + groupsize,
                       data = fox_dat,
                       family = gaussian,
                       prior = c(prior(normal(0, 0.2), class = Intercept),
                                 prior(normal(0, 0.5), class = b,),
                                 prior(exponential(1), class = sigma)),
                       iter = 4000, warmup = 2000, chains = 4, cores = 4, seed = 1234,
                       file = "output/group_on_weight")
```

```r
summary(group_on_weight)
```

```
 Family: gaussian
  Links: mu = identity
Formula: weight ~ 1 + groupsize
   Data: fox_dat (Number of observations: 116)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.00      0.08    -0.16     0.16 1.00     8305     5924
groupsize    -0.15      0.09    -0.33     0.03 1.00     9398     5909

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     1.00      0.07     0.88     1.14 1.00     8612     5908

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```r
food_direct <- brm(weight ~ 1 + avgfood + groupsize,
                   data = fox_dat,
                   family = gaussian,
                   prior = c(prior(normal(0, 0.2), class = Intercept),
                             prior(normal(0, 0.5), class = b,),
                             prior(exponential(1), class = sigma)),
                   iter = 4000, warmup = 2000, chains = 4, cores = 4, seed = 1234,
                   file = "output/food_direct")

summary(food_direct)
```

```
 Family: gaussian
  Links: mu = identity
Formula: weight ~ 1 + avgfood + groupsize
   Data: fox_dat (Number of observations: 116)
  Draws: 4 chains, each with iter = 4000; warmup = 2000; thin = 1;
         total post-warmup draws = 8000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.00      0.08    -0.16     0.15 1.00     6399     5493
avgfood       0.47      0.18     0.10     0.84 1.00     3962     4290
groupsize    -0.57      0.19    -0.93    -0.20 1.00     4145     4524

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.96      0.06     0.84     1.10 1.00     5389     4998

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Q2.10a

Avgfood causes an increase in weight. For every 1 unit increase in average food, weight increases by 0.47 kg. The credible interval [0.10, 0.84] indicates that the relationship is causal.

Group size causes a decrease in weight. For every 1 increase in group size, weight decreases by 0.57 kg. The credible interval [-0.93, -0.20] indicates that the relationship is causal.

Q2.10b

This interpretation changes from the univariate regressions because now there is a causal relationship for both variables on weight. Their effects are opposite to each other so when only one was considered, it showed no relationship.

Q2.10c

A higher amount of food increases the weight of foxes because as foxes eat more, they gain more weight. A higher amount of food also affects the group size because more food can support more foxes. However, as group size increases, the same amount of food will have to support more individuals leading to a decrease in weight. I think the univariate regressions showed no effect because group size and avg food have opposite affects on weight. The multivariate analysis showed an effect because it considers the effect avgfood has on group size.