

Problem Set 2

Michael Fryer

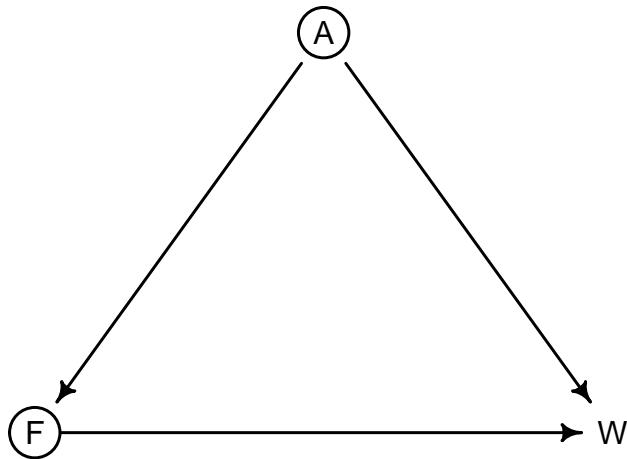
2025-11-16

1 Multiple Regression & Causal Models

The `foxes` dataset contains data on urban fox populations.

```
# First, load the foxes dataset
data(foxes)
d <- foxes
# You must set random seed to 390
rseed <- 390
set.seed(rseed)
```

Consider the following hypothesized causal relationship between **territory size** and **body weight** in foxes.



where A , F and W represent random variables `area` (territory size), `avgfood`, and `weight`, respectively.

If this DAG correctly describes the causal relationships, it makes specific predictions about what we should observe in the data. Your task is to test whether the observed patterns match these predictions.

- Territory size (A) has a **direct** effect on weight (W) : $A \rightarrow W$
- Food availability (F) has a **direct** effect on weight (W) : $F \rightarrow W$
- Territory size (A) has an **indirect** effect on weight (W) through food (F) : $A \rightarrow F \rightarrow W$

1.1 Standardize the Values

```
d$A <- standardize(d$area)
d$F <- standardize(d$avgfood)
d$W <- standardize(d$weight)
```

1.2 Part a

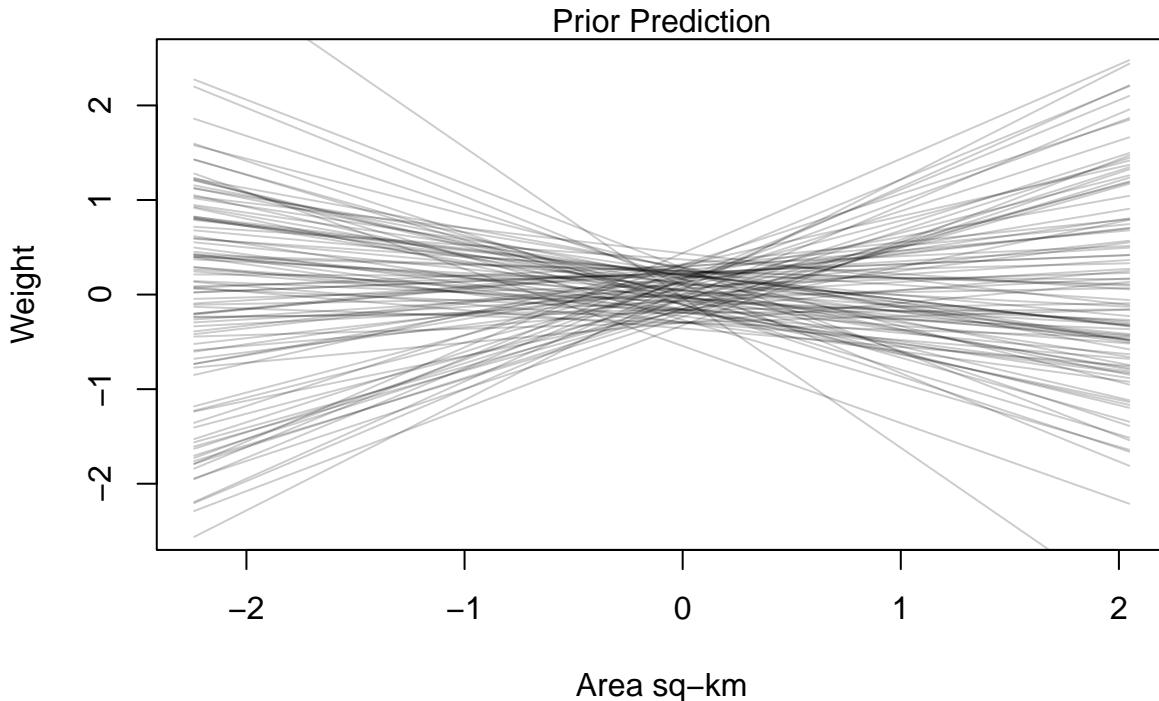
a) According to the DAG, territory size effects weight through two paths:

- Direct path: $A \rightarrow W$
- Indirect path: $A \rightarrow F \rightarrow W$

If we regress weight on territory size without including food, the coefficient should capture both pathways, the “total association” between A and W . Construct a linear regression (`m1a`) using `quap`. Urban foxes in this population have an average weight of 5kg. Use prior predictive simulation to assess the implications of your priors. Standardize the prediction variable.

1.2.1 Prior Predictive Simulation

```
N <- 100
# As our data is standardized, we would expect our intercept, a, to be very
# close to 0
a <- rnorm(N, 0, 0.2)
# b represents the rate of change between our observed and unobserved variables.
# A value of 1 implies that for every one sd of change in our observed var.
# there is 1 sd of change in our unobserved variables. We will choose a sd of
# 0.5 since that means we expect most of these slopes to be in -1 < b < 1
b <- rnorm(N, 0, 0.5)
```



1.2.2 Linear Regression

```
m <- quap(
  alist(
    W ~ dnorm(mu, sigma),
    # data is standardized so no need to subtract xbar
    mu <- a + b * A,
    # Priors from earlier
```

```

    a ~ dnorm(0, 0.2),
    b ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
),
data=d
)

##           mean        sd      5.5%     94.5%
## a     -3.353656e-07 0.08360863 -0.1336231 0.1336224
## b      1.883325e-02 0.09089577 -0.1264357 0.1641022
## sigma  9.912655e-01 0.06466638  0.8879161 1.0946148

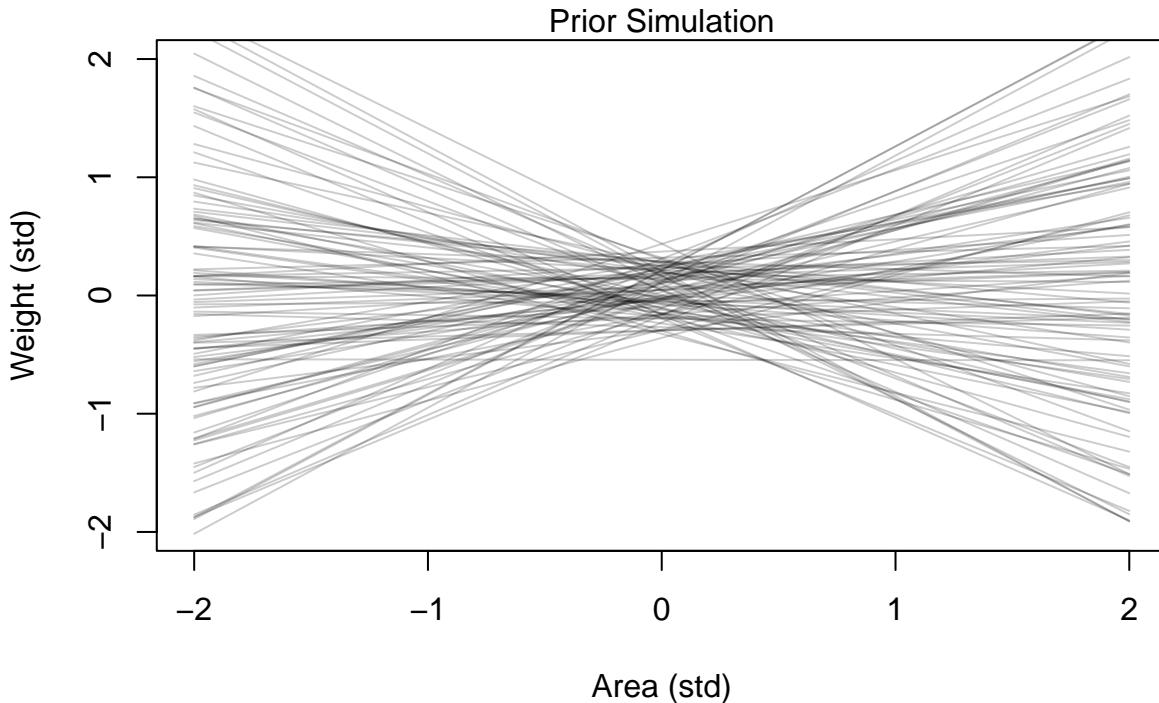
```

1.2.3 Simulate the Priors

```

set.seed(rseed)
prior <- extract.prior(m)
mu <- link(m, post=prior, data=list(A=c(-2, 2)))
plot(NULL,
      xlim=c(-2, 2), ylim=c(-2, 2),
      xlab="Area (std)", ylab="Weight (std)"
)
mtext("Prior Simulation")
for (i in 1:N) lines(c(-2, 2), mu[i,], col=col.alpha("black", 0.2))

```



1.2.4 Posterior Predictions

```

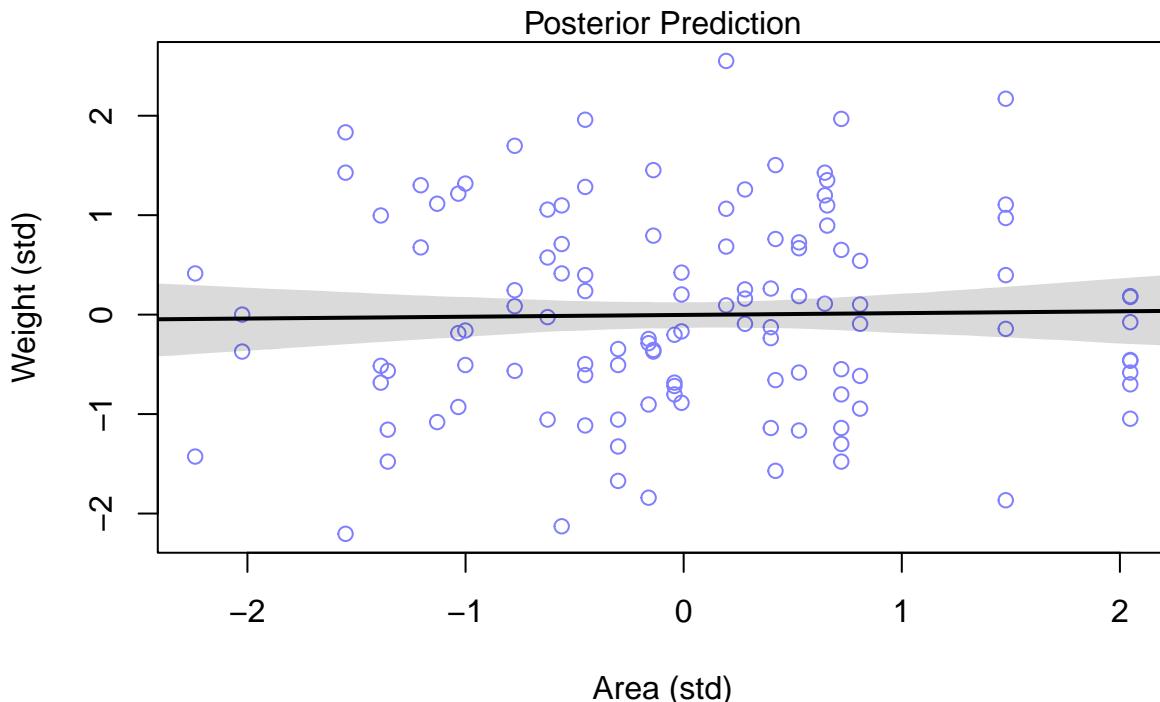
A.seq <- seq(from=-3, to=3, length.out=N)
mu <- link(m, data=list(A=A.seq))
mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI)
# Plot raw data

```

```

plot(W ~ A, data=d, col=rangi2, xlab="Area (std)", ylab="Weight (std)")
mtext("Posterior Prediction")
# Plot MAP line
lines(A.seq, mu.mean, lwd=2)
# Plot HDPI region for line
shade(mu.PI, A.seq)

```



Question: What association do you observe? What does your analysis suggest about how territory size relates to weight?

We observe a b value of 0.06 on average. This suggests that territory size has a very slightly positive effect on weight.

1.3 Part b

b) Regress weight on food availability. That is, construct a quap linear regression (`m1b`) to estimate the association of food availability and fox weight. Before fitting the model, standardize both `avgfood` and `weight` to have mean 0 and standard deviation 1.

```

centered.d <- d
centered.d$avgfood <- scale(centered.d$avgfood)
centered.d$weight <- scale(centered.d$weight)

```

Hint: With standardized variables, regression slopes represent standardized effect sizes. A slope of 1.0 would indicate a perfect positive relationship, while slopes >2 would be implausibly large for most ecological relationships.

1.3.1 Prior Prediction

```

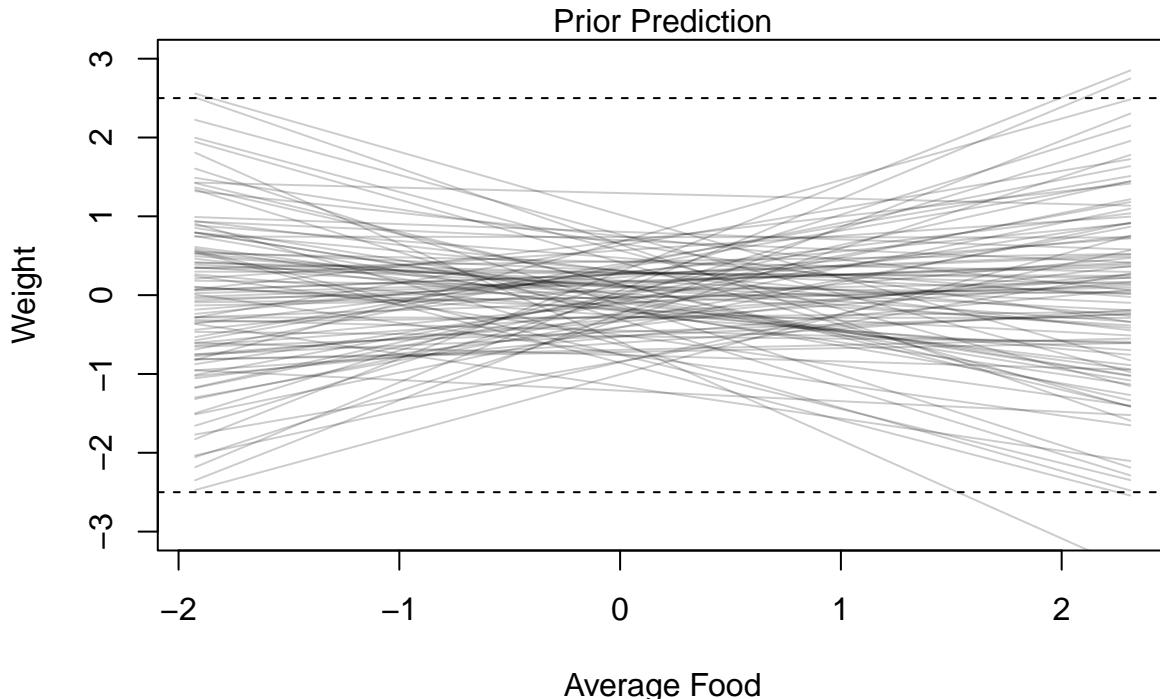
N <- 100
# a is the intercept. This means a fox with avg food of 0 has this weight.
# Normally, it is not possible to have a negative weight but since our data
# is standardized we will assume a mean 0 and but use a smaller std of 0.5

```

```

a <- rnorm(N, 0, 0.5)
# b represents the rate of change of a fox's weight with respect to food access.
# Similarly to above, the rates being compared are standardized meaning, a
# rate of 1 or -1 means a 1:1 relationship. To not discount the possibility
# of a negative correlation, we will center our predicted prior on 0 with std
# of 0.5
b <- rnorm(N, 0, 0.5)
xbar <- mean(centered.d$avgfood)

```



Use prior predictive simulation to assess the implication of your priors. Write 1-2 sentences to justify your priors.

In the above prior prediction, I used $a \sim \text{Normal}(0, 0.5)$ and $b \sim \text{Normal}(0, 0.5)$. I chose these values because our data is standardized. I used a std of 0.5 in the normal because I wanted to tighten the priors slightly from the standardized 1.

1.3.2 Linear Regression

```

m <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    mu <- a + b * (avgfood - xbar),
    # Priors from earlier
    a ~ dnorm(0, 0.5),
    b ~ dnorm(0, 0.5),
    sigma ~ dunif(0, 5)
  ),
  data=centered.d
)

##               mean        sd      5.5%     94.5%
## a      -6.127103e-07 0.09087832 -0.1452417 0.1452405

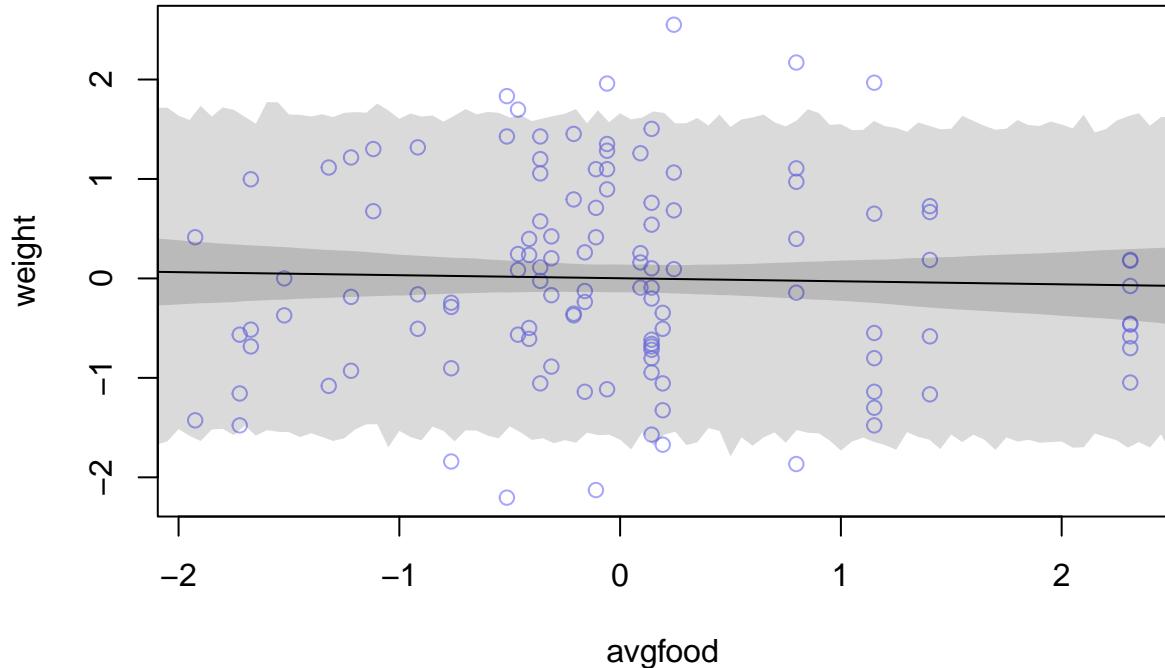
```

```
## b      -2.420430e-02 0.09125955 -0.1700547 0.1216461
## sigma 9.953688e-01 0.06534894  0.8909286 1.0998090
```

1.3.3 Plot the Predictions

```
# Adding mu uncertainty
avgfood.seq <- seq(from=-2.5, to=2.5, by=0.05)
mu <- link(m, data=data.frame(avgfood=avgfood.seq))
mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI, prob=0.89)
# Adding a prediction interval (prediction uncertainty due to std uncertainty)
sim.weight <- sim(m, data=list(avgfood=avgfood.seq))
weight.PI <- apply(sim.weight, 2, PI, prob=0.89)

# Plot raw data
plot(weight ~ avgfood, data=centered.d, col=col.alpha(rangi2, 0.7))
# Plot MAP line
lines(avgfood.seq, mu.mean)
# Plot HPDI region for line
shade(mu.PI, avgfood.seq)
# Plot PI region for simulated heights
shade(weight.PI, avgfood.seq)
```



- c) Now regress weight on *both* territory size and food availability Construct a `quap` model (`m1c`) that includes both predictors. Use the standardized variables. Explain your findings with 3-4 sentences and appropriate plots.

2 AI Declaration