Problem Set 2

Michael Fryer

Collaborators: Florian Robrecht

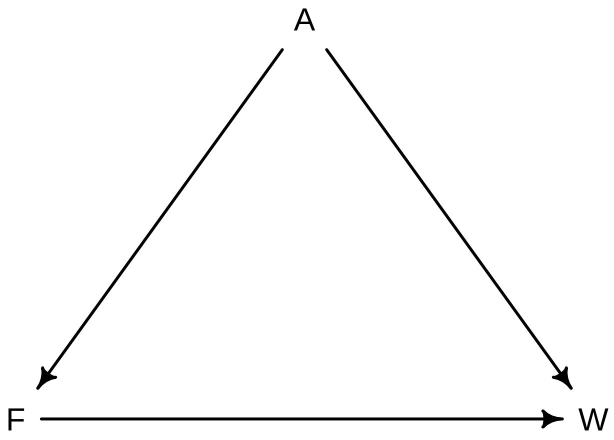
For this assignment, I collaborated with Florian and discussed our independent solutions.

1 Multiple Regression & Causal Models

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

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

1.1 Standardize the Values

```r
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: $A \to W$
- Indirect path: $A \to F \to W$

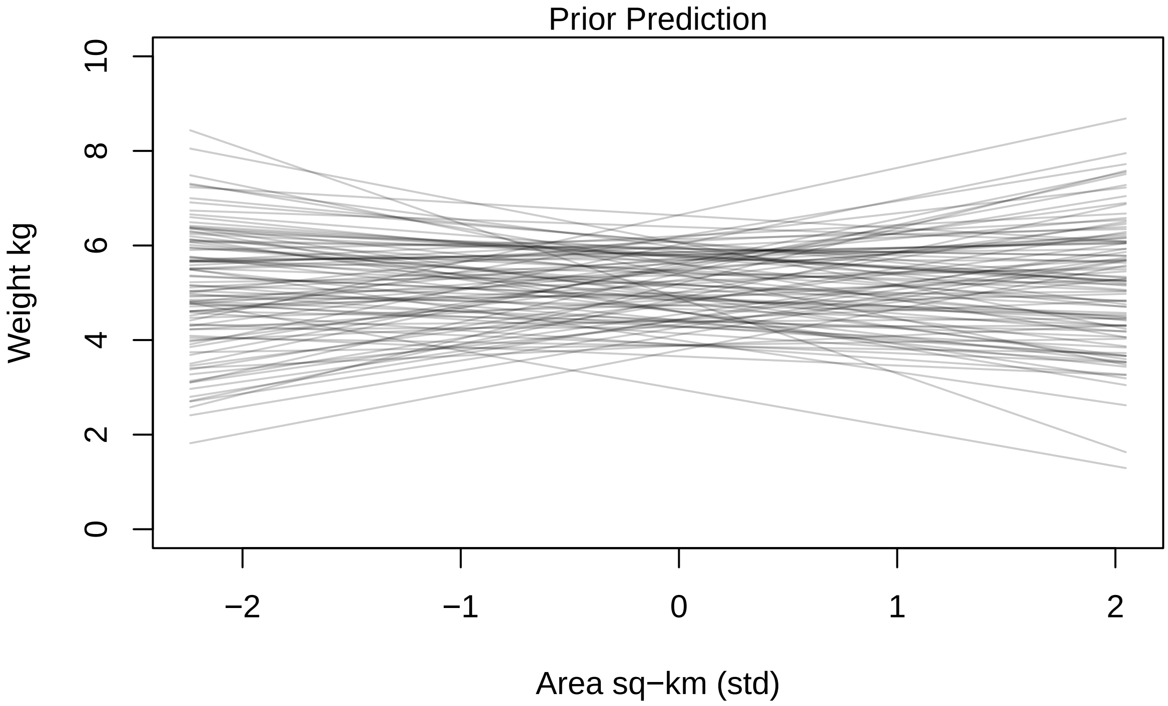
If we regress weight on territory size without including food, the coefficient should capture both pathways, the "total association" between A and W. Construct a linear regression (`m1a`) using `quap`. Urban foxes in this population have an average weight of 5kg. Use prior predictive simulation to assess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

Based on above, I am assuming that only the predictor variable, ***area****, should be standardized for this model.*

1.2.1 Prior Predictive Simulation

A note on priors: *For this model with a standardized predictor, the value of a represents our expected weight when the predictor* $A = 0$*. Since we are told foxes weigh 5kg on average in this population, we will create a normal distribution with* $\mu = 5$ *and* $\sigma = 0.75$ *which represents 15% of the mean. It is valid to use* $\mu = 5$ *here as 5kg is a population statistic. The value of b represents the rate of change between our predictor and outcome variables. A value of 1 implies that for every 1 standard deviation of change in our prediction variable, there is 1kg of change in our outcome variable. For this model, I will be using a normal distribution with* $\mu = 0$ *and* $\sigma = 0.5$*.*

```
N <- 100
a <- rnorm(N, 5, 0.75)
b <- rnorm(N, 0, 0.5)
```

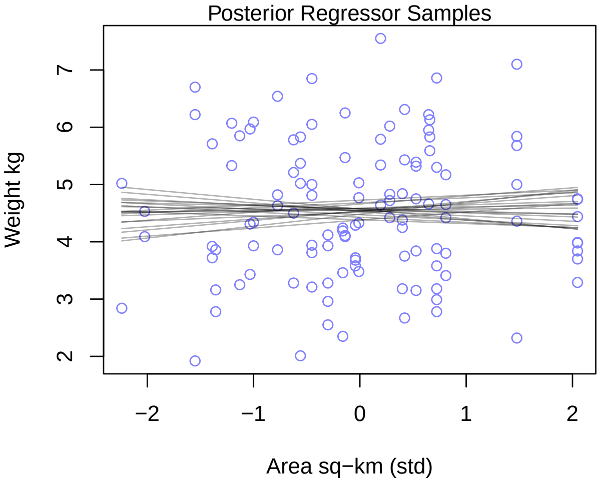


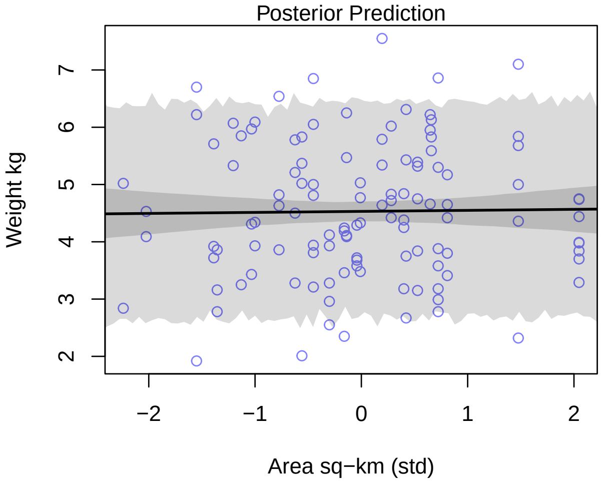
1.2.2 Linear Regression

```r
set.seed(rseed)
m <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    # No need to subtract mean as our predictor is standardized
    mu <- a + b*A,
    # Priors from earlier
    a ~ dnorm(5, 0.75),
    b ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d
)
```

```
##             mean         sd       5.5%     94.5%
## a     4.53936547 0.10776999  4.3671282 4.7116027
## b     0.02200862 0.10683938 -0.1487413 0.1927586
## sigma 1.17281103 0.07642638  1.0506669 1.2949551
```

1.2.3 Posterior Predictions





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

We did a good job with prior prediction of a as we observe $\mu = 4.54$. With b we observe $\mu = 0.02$ and $\sigma = 0.11$ we observe neither a strong nor precise relationship between these two variables, illustrated in the figures above. While the slope is positive, this analysis tells us that territory size gives very little information about 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.

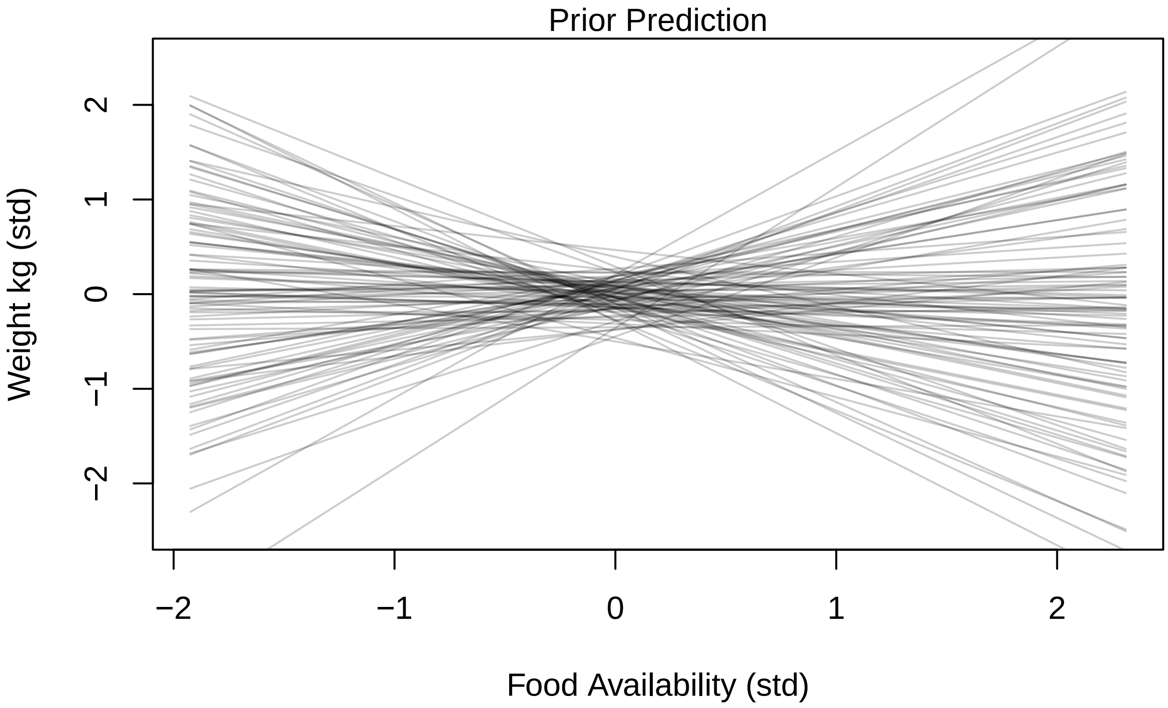
Hint: *With standardized variables, regression slopes represent standardized effect sizes. A slope of 1.0 would indicate a perfect positive relationship, while slopes >2 would be implausibly large for most ecological relationships.*

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

A note on priors: *Since both our predictor and outcome variables are standardized, we will simulate priors with $\mu = 0$ and a small standard deviation. For our intercept prior, a, we will use $\sigma = 0.2$ as we expect our data to be relatively tight around the mean. For our slope prior, b, we will use $\sigma = 0.5$ as we expect most slopes to be in $-1 < b < 1$.*

1.3.1 Prior Predictive Simulation

```
N <- 100
a <- rnorm(N, 0, 0.2)
b <- rnorm(N, 0, 0.5)
```

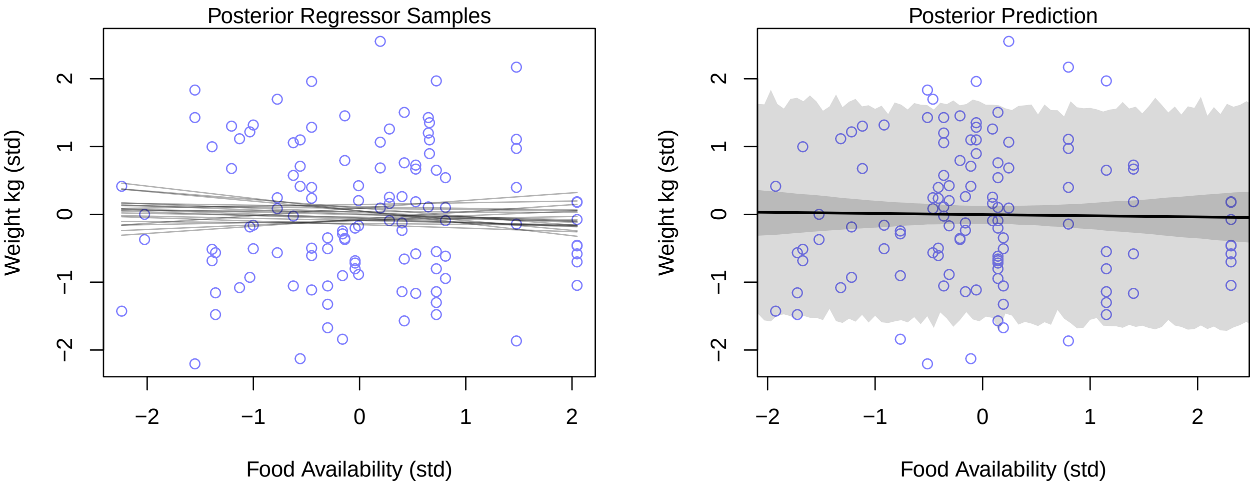


1.3.2 Linear Regression

```
set.seed(rseed)
mF <- quap(
  alist(
    W ~ dnorm(mu, sigma),
    # No need to subtract mean as our predictor is standardized
    mu <- a + bF*F,
    # Priors from earlier
    a ~ dnorm(0, 0.2),
    bF ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d
)
```

```
##                mean         sd       5.5%     94.5%
## a     -2.225552e-07 0.08360017 -0.1336094 0.1336090
## bF    -2.421130e-02 0.09088501 -0.1694631 0.1210405
## sigma  9.911439e-01 0.06465857  0.8878070 1.0944808
```

1.3.3 Posterior Predictions



As with ***a)****, we observe neither a strong nor precise association with* $\mu = -0.02$ *and* $\sigma = 0.09$*. This implies food availability gives little information about weight. It is notable that food availability appears to have a negative impact on weight whereas territory size has a positive impact.*

1.4 Part C

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.

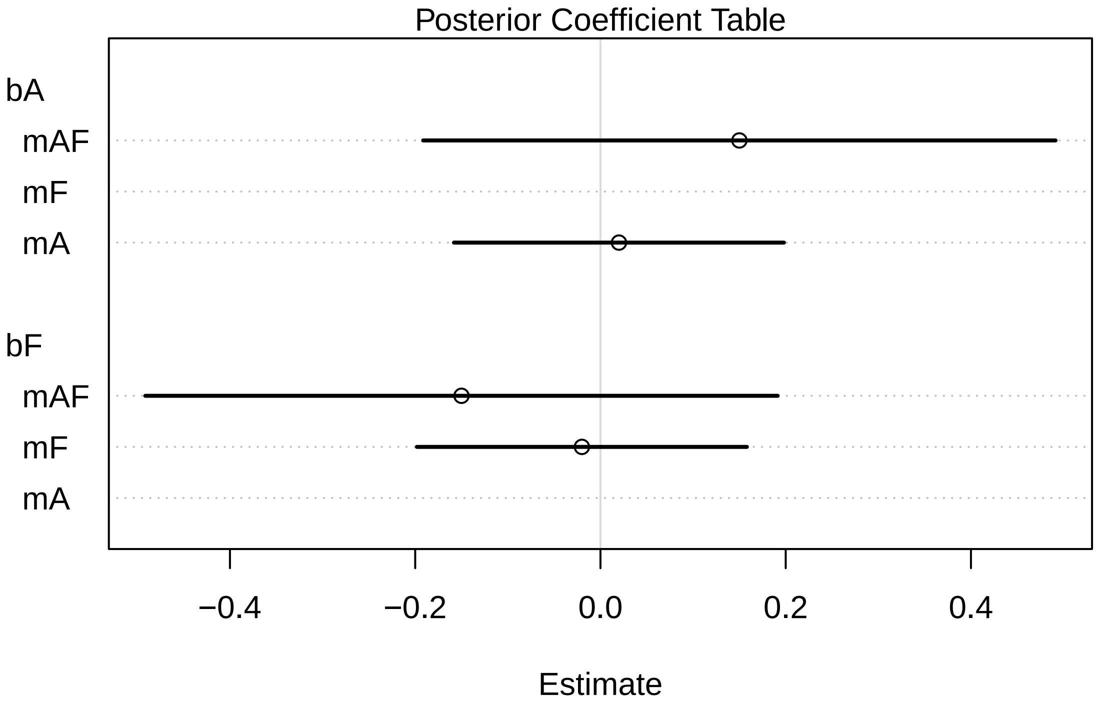
In the below analysis, I will standardize both the predictor variables and the outcome variable. This is done so we can accurately compare the models to one another later on.

1.4.1 Linear Regression

```
# Area predictor with standardized weight
set.seed(rseed)
mA <- quap(
  alist(
    W ~ dnorm(mu, sigma),
    mu <- a + bA*A,
    a ~ dnorm(0, 0.2),
    bA ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d
)
# Food availability predictor reused from above
```

```r
# Both predictors with standardized weight
set.seed(rseed)
mAF <- quap(
  alist(
    W ~ dnorm(mu, sigma),
    mu <- a + bA*A + bF*F,
    a ~ dnorm(0, 0.2),
    bA ~ dnorm(0, 0.5),
    bF ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d
)
```

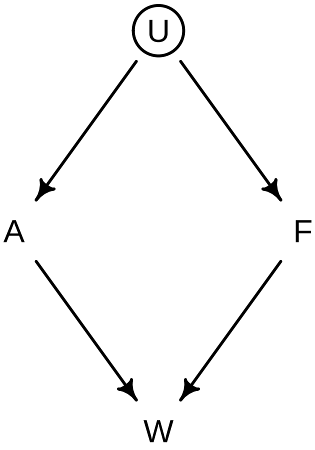
```
## Standardized Area Predictors:

##                mean         sd       5.5%     94.5%
## a     -1.309274e-07 0.08360865 -0.1336229 0.1336226
## bA     1.883348e-02 0.09089579 -0.1264356 0.1641025
## sigma  9.912658e-01 0.06466643  0.8879163 1.0946152

## Standardized Area and Food Availability Predictors:

##                mean         sd       5.5%     94.5%
## a     -7.930838e-06 0.08334335 -0.1332067 0.1331908
## bA     1.461574e-01 0.17418687 -0.1322269 0.4245416
## bF    -1.490664e-01 0.17418701 -0.4274509 0.1293181
## sigma  9.874582e-01 0.06444010  0.8844704 1.0904459
```



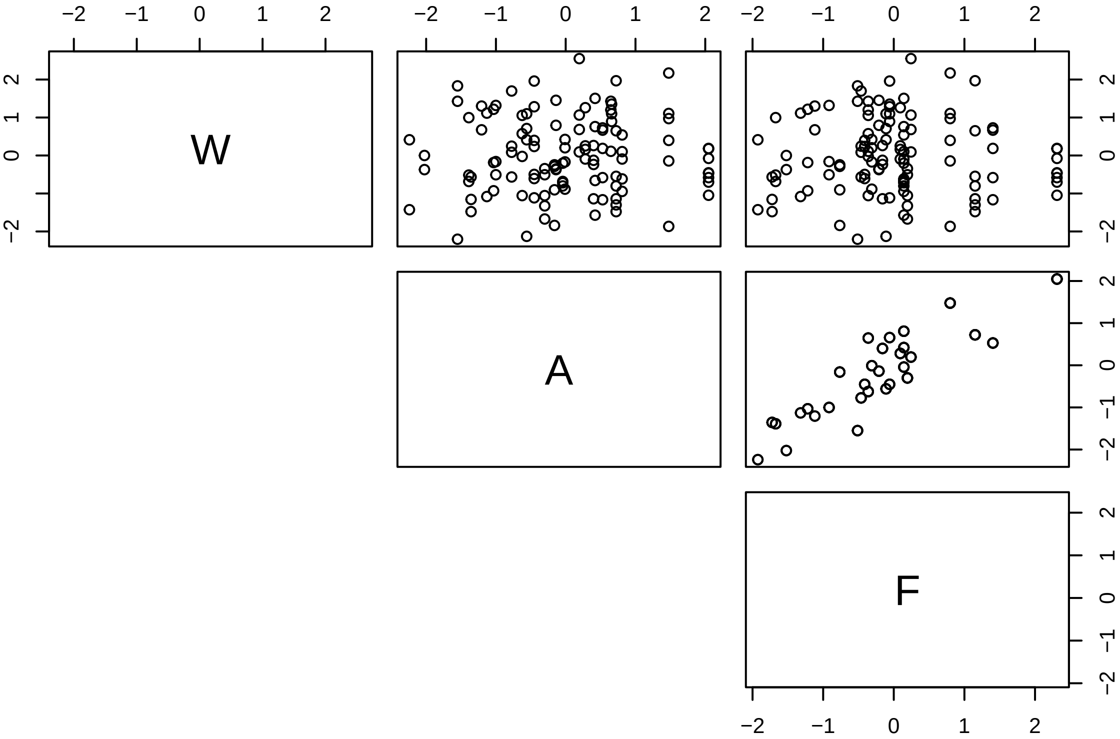
In the above figure, we can see that A and F are each more correlated with W in the combined model, mAF, than in either of the bivariate models mA and mF. However, we also note that the uncertainty, standard deviation, has increased drastically. Observing results like this is a strong indicator of a potential masked relationship.

1.5 Masked Relationship

Although this is not part of the assignment, I will consider the following relationship for my own edification.

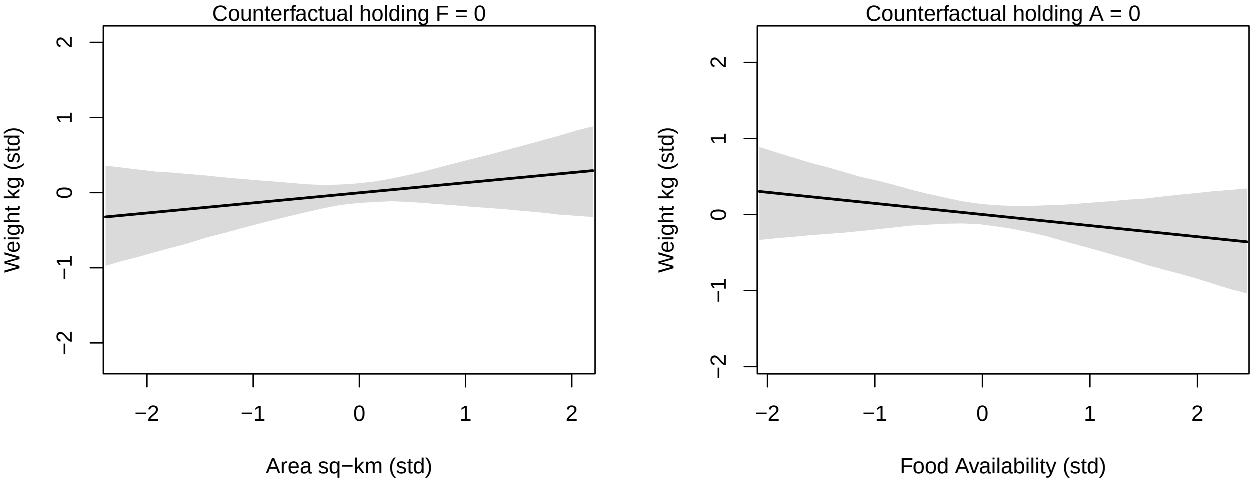


From a simple pairs plot, we can see that A and F are positively correlated with one another. The result of this pattern is that the two predictors tend to cancel eachother out.



1.5.1 Counterfactuals

Going of the assumption that the above DAG is accurate, we will simulate breaking the links from $U \rightarrow A$ *and* $U \rightarrow F$ *by producting counterfactual plots wherin one predictor is held at 0.*



Although we can't be sure that this DAG wholly represents this data, in fact there are many DAGs that fit this data, we can see that the DAG(U, A, F, W) fits this observed data. That is, there is likely some unobserved U *that is a common cause of* A *and* F*.*

2 AI Declaration

AI was not used for this assignment.

3 Appendix

```
############
# Plot a DAG
############
fox_dag <- dagitty("dag {
  A -> F
  A -> W
  F -> W
}")
coordinates(fox_dag) <- list(x=c(A=0, F=-0.5, W=0.5), y=c(A=0, F=1, W=1))
drawdag(fox_dag, xlim=c(-1, 1), ylim=c(-1, 1))

#####################
# Plot an updated DAG
#####################
fox_dag <- dagitty("dag {
  U -> A
  U -> F
  A -> W
  F -> W
}")
coordinates(fox_dag) <- list(x=c(U=0, A=-0.25, F=0.25, W=0), y=c(U=0, A=0.5, F=0.5, W=1))
drawdag(fox_dag, xlim=c(-1, 1), ylim=c(-1, 1), shapes=c(U="c"))

##############
# Plot Helpers
##############
# Plots a standardized predictor x
plot_prior_prediction <- function(x, ylim, xlab, ylab) {
  plot(NULL, xlim=range(x), ylim=ylim, xlab=xlab, ylab=ylab)
  mtext("Prior Prediction")
  # Graph all the lines
  for (i in 1:N) {
    curve(a[i] + b[i]*x,
      from=min(x), to=max(x),
      col=col.alpha("black", 0.2), add=TRUE
    )
  }
}
# Plots a subset of the predictions from the posterior
plot_posterior_samples <- function(m, x, y, xlab, ylab) {
  post <- extract.samples(m, n=20)
  plot <- plot(x, y,
    xlim=range(x), ylim=range(y),
    xlab=xlab, ylab=ylab,
    col=rangi2
  )
  mtext("Posterior Regressor Samples")
  for (i in 1:nrow(post)) {
    curve(post$a[i] + post$b[i]*x, col=col.alpha("black", 0.3), add=TRUE)
  }
}
# Plots the posterior prediction for a standardized predictor
```

```
plot_posterior_prediction <- function(d, m, dist, predictor, xlab, ylab) {
  # Compute all the posterior prediction information
  data <- list()
  x.seq <- seq(from=-3, to=3, length.out=N)
  data[[predictor]] <- x.seq
  mu <- link(m, data=data)
  mu.mean <- apply(mu, 2, mean)
  mu.PI <- apply(mu, 2, PI)
  sim.val <- sim(m, data=data)
  val.PI <- apply(sim.val, 2, PI, prob=0.89)
  # Plot raw data
  plot(dist, data=d, col=rangi2, xlab=xlab, ylab=ylab)
  mtext("Posterior Prediction")
  # Plot MAP line
  lines(x.seq, mu.mean, lwd=2)
  # Plot HDPI region for line
  shade(mu.PI, x.seq)
  # Plot PI region for simulated weights
  shade(val.PI, x.seq)
}

######################
# Plot counterfactuals
######################
par(mfrow = c(1, 2))
# Counterfactual holding F=0
x.seq <- seq(from=min(d$A) - 0.15, to=max(d$A) + 0.15, length.out=30)
mu <- link(mAF, data=data.frame(A=x.seq, F=0))
mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI)
plot(NULL,
     xlim=range(d$A), ylim=range(d$A),
     xlab="Area sq-km (std)", ylab="Weight kg (std)"
)
mtext("Counterfactual holding F = 0")
lines(x.seq, mu.mean, lwd=2)
shade(mu.PI, x.seq)
# Counterfactual holding A=0
x.seq <- seq(from=min(d$F) - 0.15, to=max(d$F) + 0.15, length.out=30)
mu <- link(mAF, data=data.frame(F=x.seq, A=0))
mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI)
plot(NULL,
     xlim=range(d$F), ylim=range(d$F),
     xlab="Food Availability (std)", ylab="Weight kg (std)"
)
mtext("Counterfactual holding A = 0")
lines(x.seq, mu.mean, lwd=2)
shade(mu.PI, x.seq)
par(mfrow = c(1, 1))
```