Problem Set 2

Michael Fryer

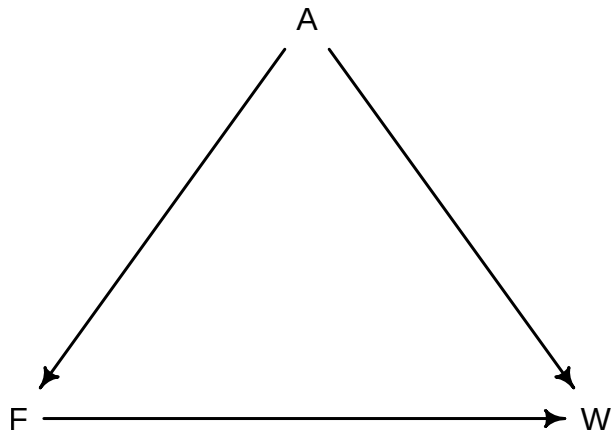
Collaborators: Florian Robrecht

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:

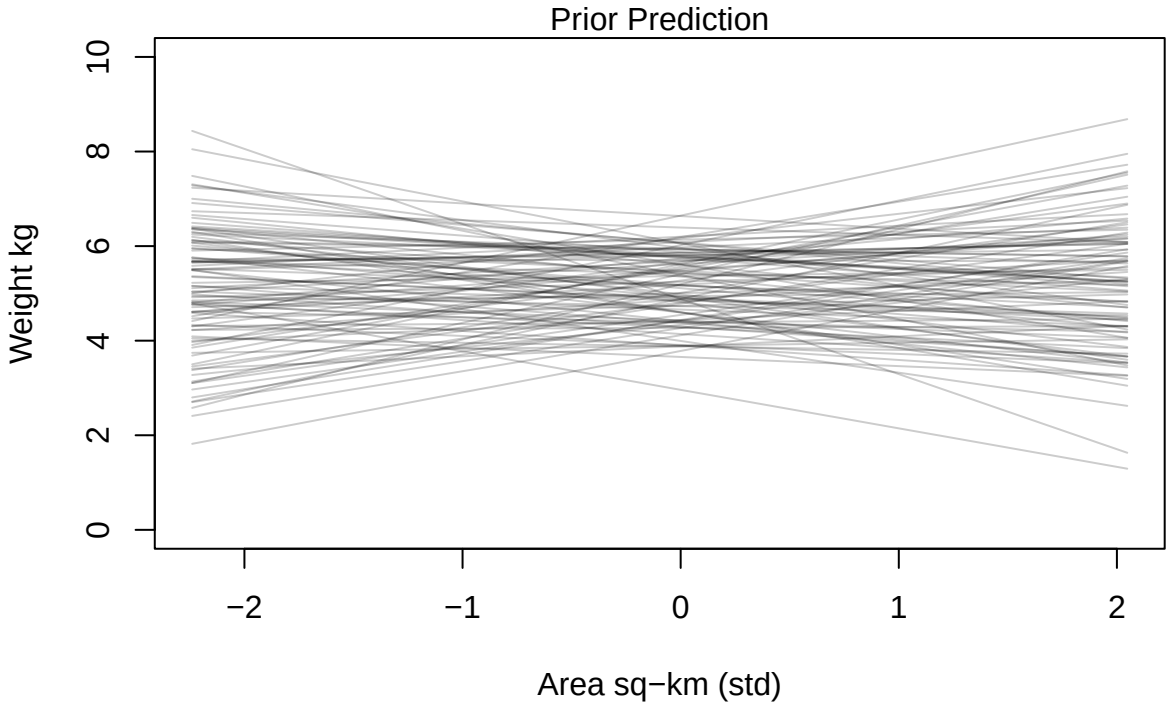
- Direct path: $A \rightarrow W$
- Indirect path: $A \rightarrow F \rightarrow W$

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.

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

1.2.1 Prior Predictive Simulation

```
N <- 100
# If our prediction variable is not a factor, mean 0, we would expect the
#   predicted weight to be 5kg. A std of 0.75 is chosen as that represents
#   15% of the mean
a <- rnorm(N, 5, 0.75)
# b represents the rate of change between our predictor and prediction variables.
#   A value of 1 implies that for every one sd of change in our predictor
#   variable there is 1kg of change in our prediction variable
b <- rnorm(N, 0, 0.5)
```



1.2.2 Linear Regression

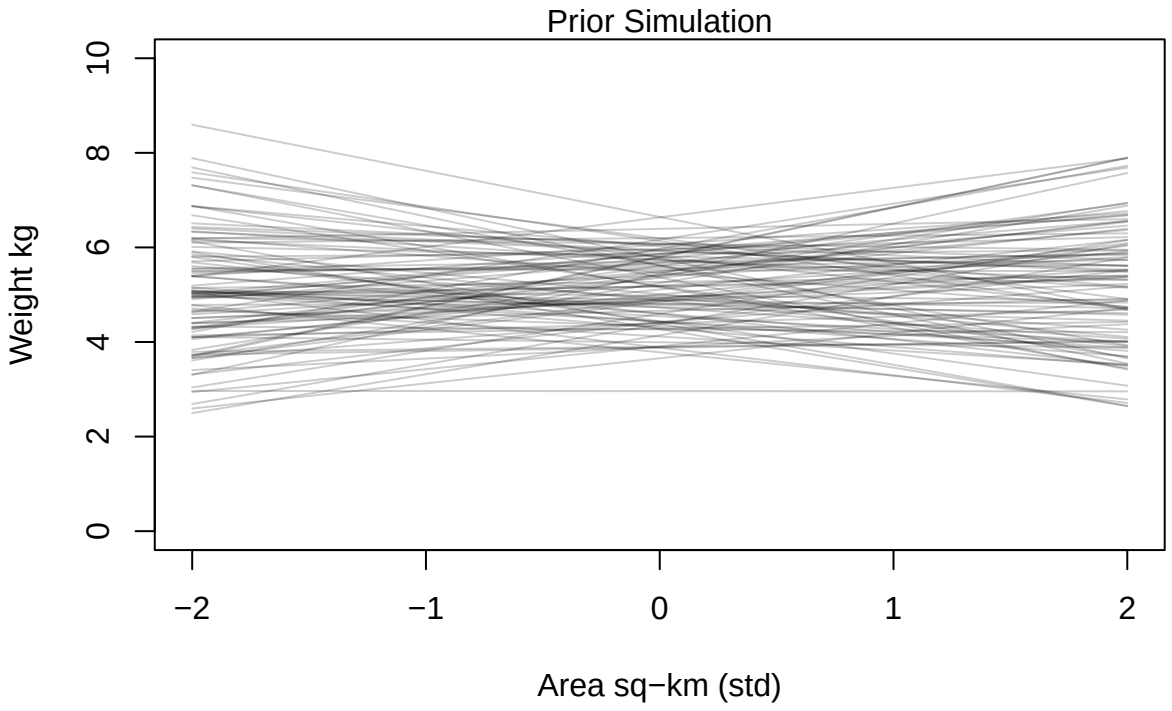
```
m <- quap(
  alist(
    weight ~ dnorm(mu, sigma),
    # No need to subtract mean as our predictor is standardized
    mu <- a + b*A,
```

```
    # 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     4.53936466 0.10776998  4.367127 4.7116019
## b     0.02200797 0.10683937 -0.148742 0.1927579
## sigma 1.17281090 0.07642636  1.050667 1.2949550
```

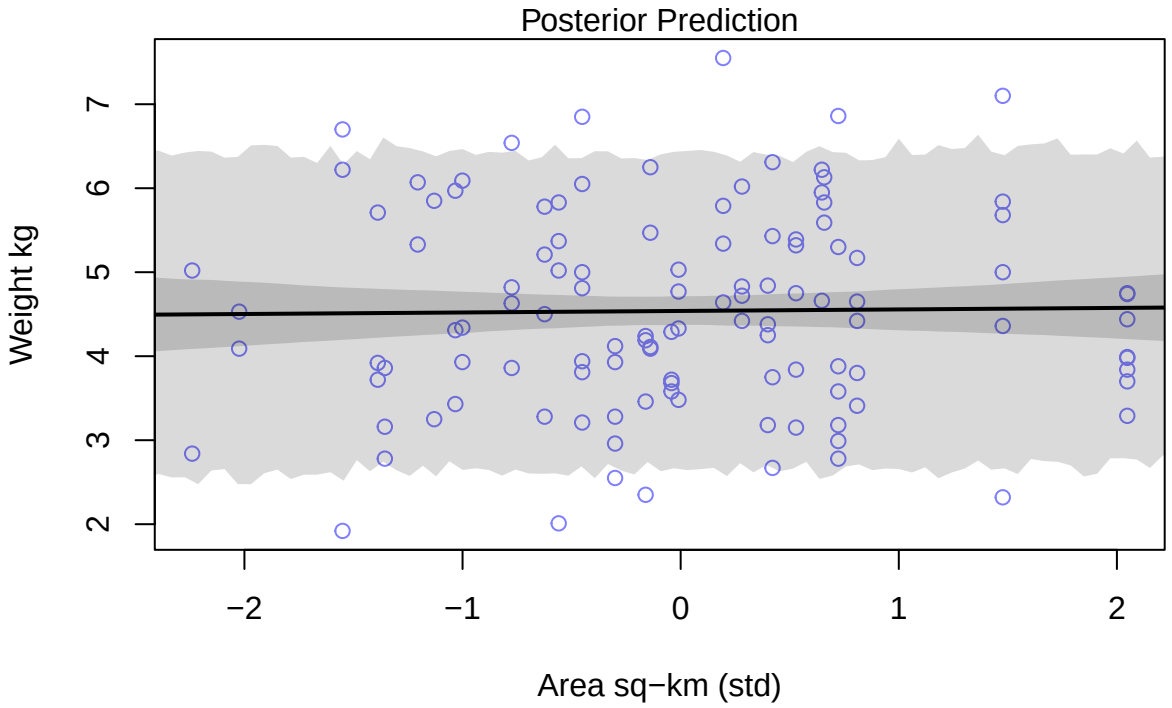
1.2.3 Simulate the Priors

```
set.seed(rseed)
prior <- extract.prior(m)
mu <- link(m, post=prior, data=list(A=c(-2, 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)
sim.weight <- sim(m, data=list(A=A.seq))
weight.PI <- apply(sim.weight, 2, PI, prob=0.89)
```



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

With a mean of 0.02 and a standard deviation of 0.11 we observe neither a strong nor precise relationship, illustrated in the figure above. 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.*

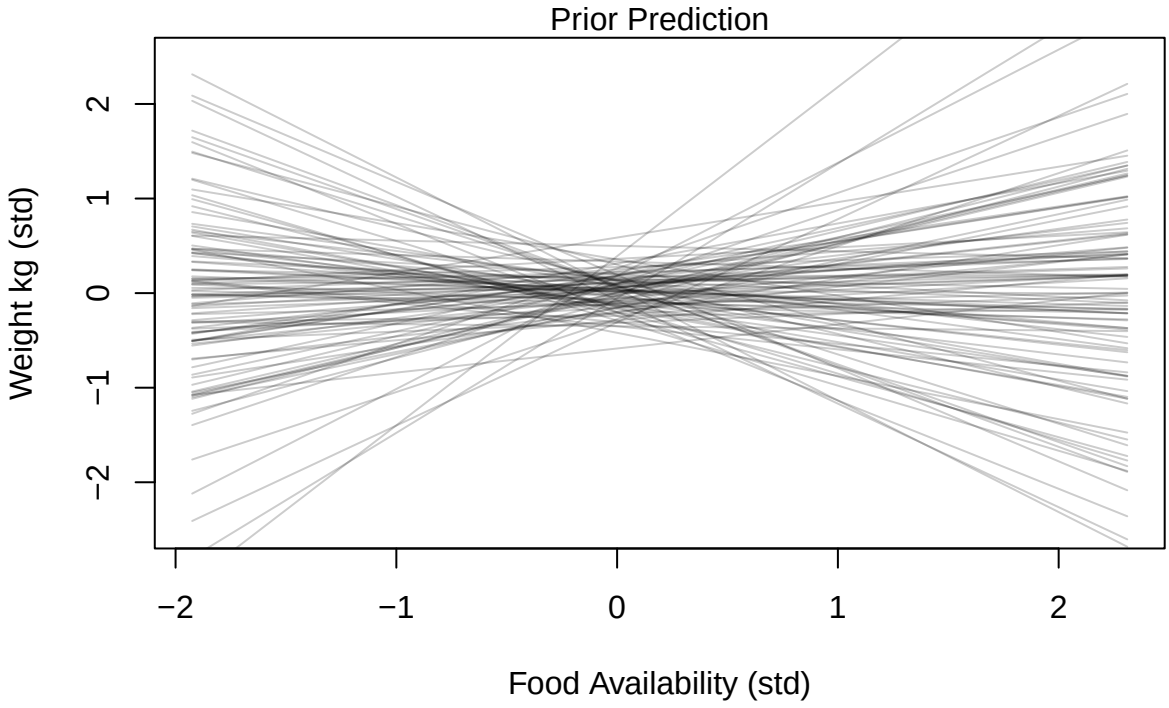
Use prior predictive simulation to assess the implication of your priors. Write 1-2 sentences to justify your priors.

Since both our predictor and prediction variable are standardized, we will simulate priors with mean 0 and a small standard deviation. For our slope pior, b, a value of 1 represents a change of 1 standard deviation in our predicted variable for 1 standard deviation of change in our predictor variable.

1.3.1 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 predictor and prediction variables.
#   A value of 1 implies that for every one sd of change in our predictor var.
#   there is 1 sd of change in our prediction 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)
```



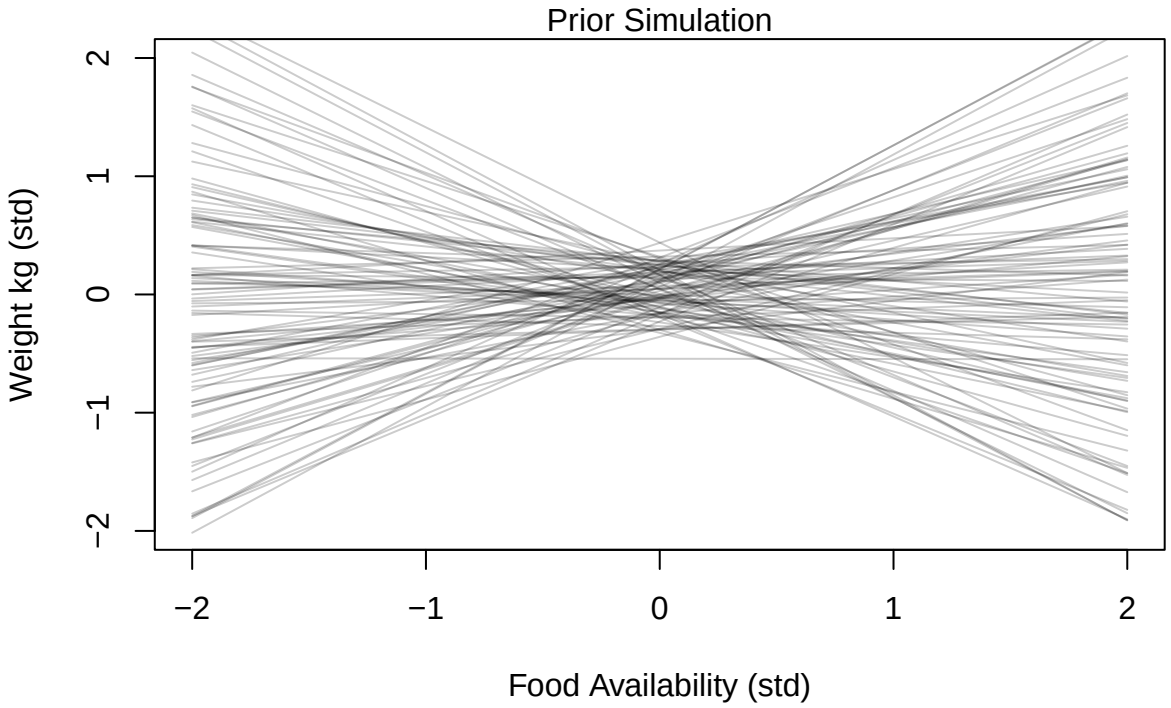
1.3.2 Linear Regression

```
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     -1.073596e-06 0.08360234 -0.1336138 0.1336116
## bF    -2.421160e-02 0.09088778 -0.1694678 0.1210446
## sigma  9.911751e-01 0.06466365  0.8878301 1.0945201
```

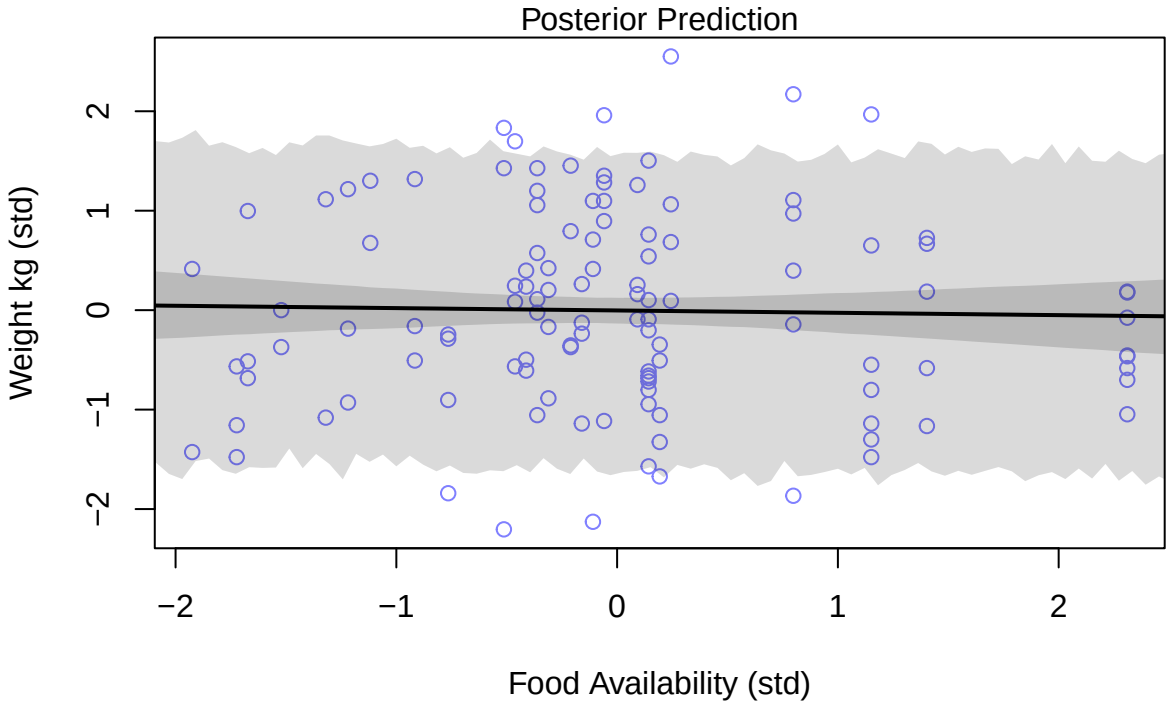
1.3.3 Simulate the Priors

```
set.seed(rseed)
prior <- extract.prior(mF)
mu <- link(mF, post=prior, data=list(F=c(-2, 2)))
```



1.3.4 Posterior Predictions

```
F.seq <- seq(from=-3, to=3, length.out=N)
mu <- link(mF, data=list(F=F.seq))
mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI)
sim.W <- sim(mF, data=list(F=F.seq))
W.PI <- apply(sim.W, 2, PI, prob=0.89)
```



As with part a, we observe neither a strong nor precise association with a predicted mean of -0.02 and a standard deviation of 0.09. As with above, thi 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 prediction variables and the predictor variable. This is done so we can accurately compare the models to one another.

1.4.1 Linear Regression

```
# Area predictor with standardized weight
mA <- quap(
  alist(
    W ~ dnorm(mu, sigma),
    # No need to subtract mean as our predictor is standardized
    mu <- a + bA*A,
    # Priors from earlier
    a ~ dnorm(0, 0.2),
    bA ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d
)
# Food availability predictor reused from above
# Both predictors with standardized weight
mAF <- quap(
  alist(
    W ~ dnorm(mu, sigma),
    # No need to subtract mean as our predictor is standardized
    mu <- a + bA*A + bF*F,
    # Priors from earlier
    a ~ dnorm(0, 0.2),
    bA ~ dnorm(0, 0.5),
    bF ~ dnorm(0, 0.5),
    sigma ~ dexp(1)
  ),
  data=d
)
```

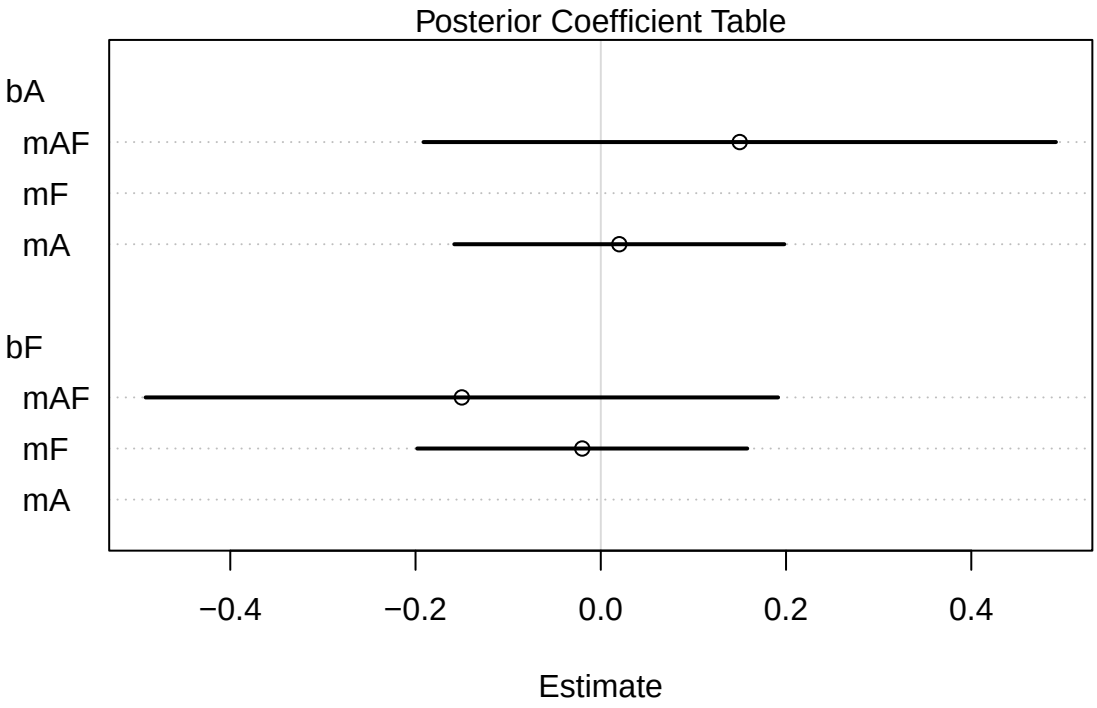
```
## Standardized Area Predictors:
```

```
##               mean         sd       5.5%     94.5%
## a     1.204066e-05 0.08360950 -0.1336121 0.1336362
## bA    1.882915e-02 0.09089688 -0.1264416 0.1640999
## sigma 9.912780e-01 0.06466843  0.8879254 1.0946307
```

```
## Standardized Area and Food Availability Predictors:
```

```
##                mean         sd       5.5%     94.5%
## a     -1.614310e-07 0.08334412 -0.1332002 0.1331998
```

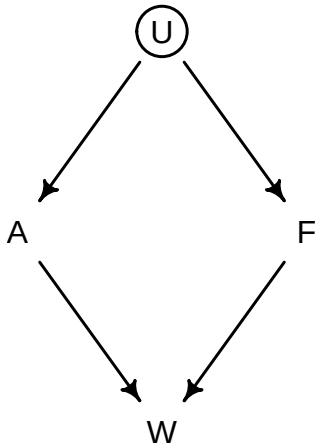
```
## bA     1.461363e-01 0.17418845 -0.1322505 0.4245231
## bF    -1.490368e-01 0.17418862 -0.4274238 0.1293503
## sigma  9.874691e-01 0.06444189  0.8844786 1.0904597
```



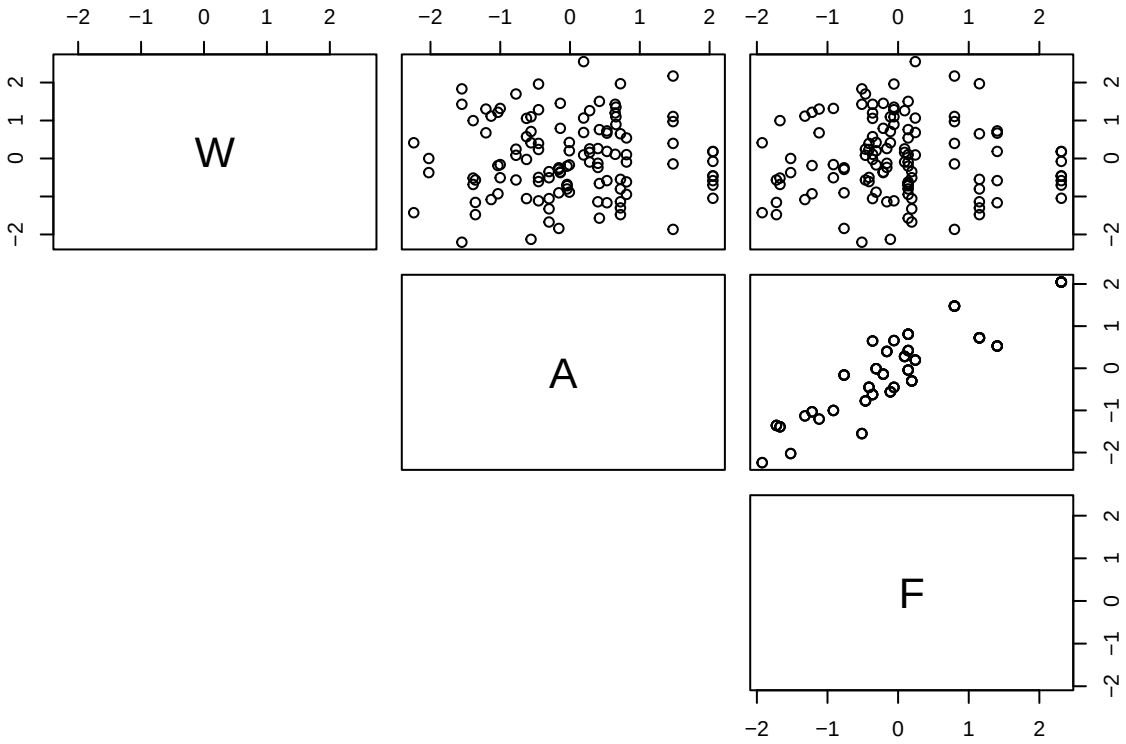
In the above figure, we can see that A and F are each more correlated with W in the combined model, mAF than in either of the bivariate models mA and mF. However, we also note that the uncertainty, standard deviation, has increased drastically. Observing results like this is a strong indicator of a potential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

1.5 Masked Relationship

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



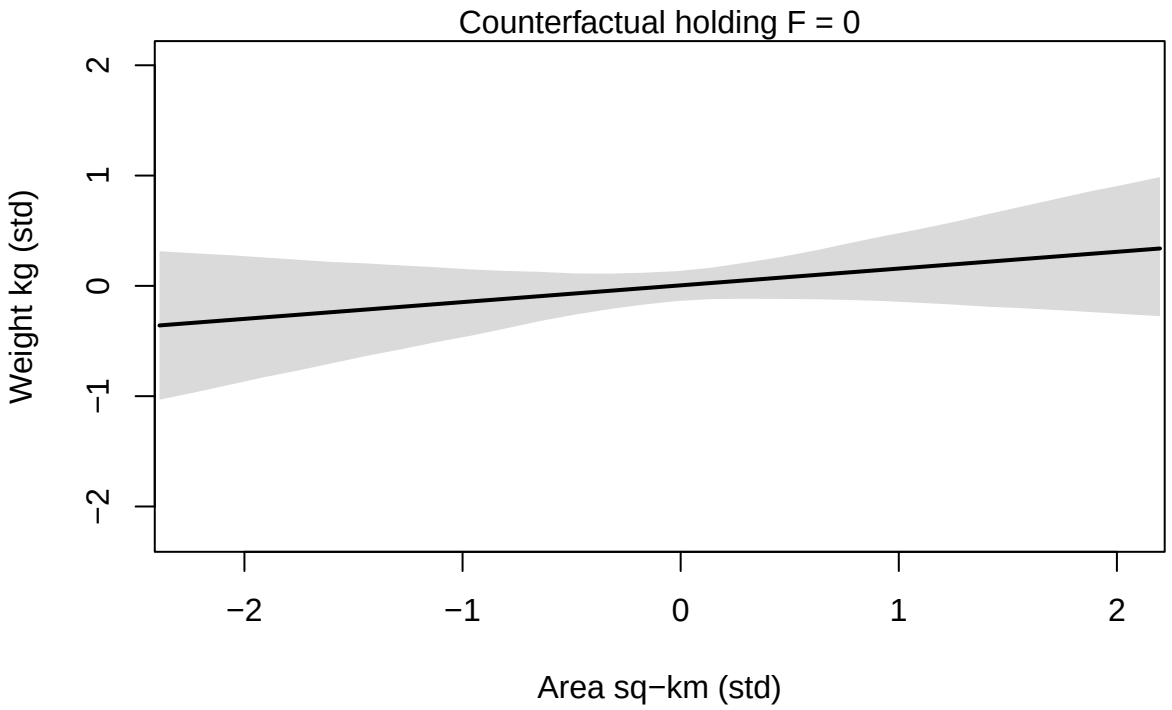
From a simple pairs plot, we can see that A and F are positively correlated with one another. The result of this pattern is that the two predictors tend to cancel eachother out.



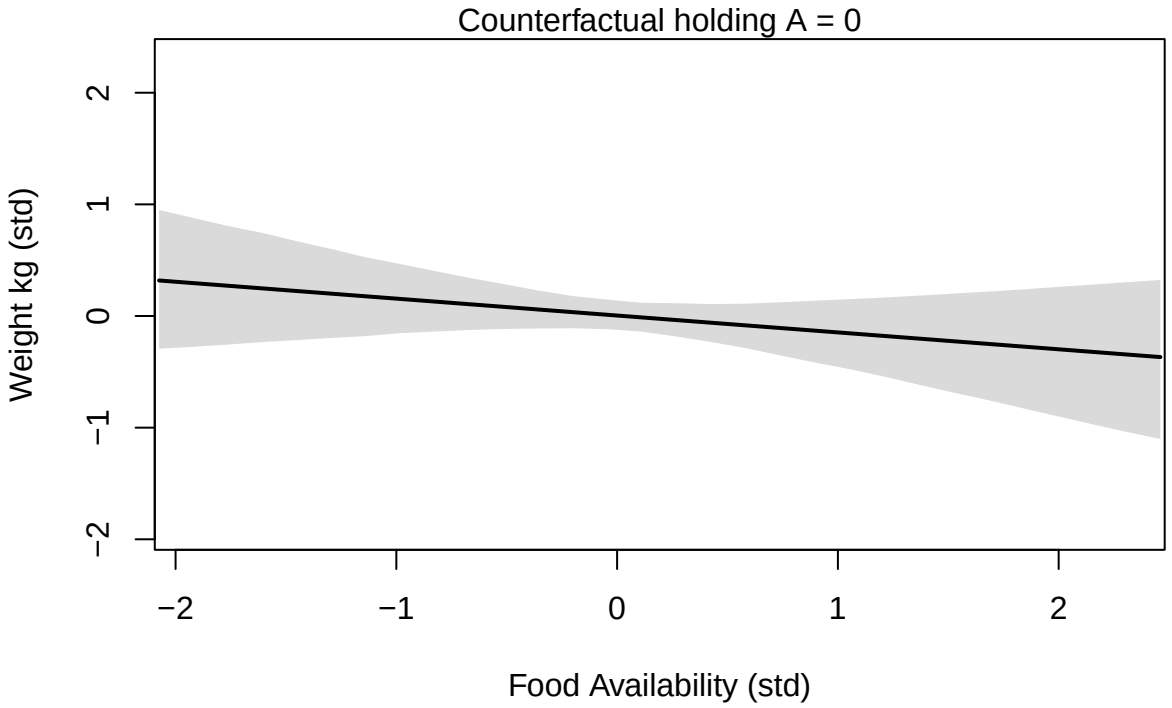
1.5.1 Counterfactuals

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.

```
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)
```




```
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)
```



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 M and N.

2 AI Declaration

AI was not used for this assignment.