Midterm Exercise

Stephen R. Proulx

2/03/2023

For this midterm exercise we will use data from this paper: https://onlinelibrary.wiley.com/doi/10.1111/oik.07674 . We will use only a portion of their data, but our analysis will involve similar models to the ones discussed in the paper. You are free to read the paper or look to it for modeling inspiration, but you can complete this entire exercise without looking at the paper.

In this study, clown fish were observed in breeding groups associated with sea anemones. In the dataset, each row is an observation.

The dataframe has 236 observations of 61 unique fish groups (labeled by Anemone_ID). Anemone size is taken to represent the food-richness of the area the fish live in, so anemone size might influence fish health and therefore fish reproductive output.

The dataset includes a treatment, which is that some of the groups of fish were fed additional food. The column "FedIndex" is 1 if the fish were not fed, and 2 if they were fed. The treatment was performed in the middle of the season, so that some clutches of eggs were produced before the treatment, and others were produced after the treatment. The column "PostTreatment" is 1 if the clutch was laid before the treatment and 2 if it was after. Note that fish in the FedIndex=1 category were never fed additional food, even if PostTreatment=2.

First clear your working environment and load the data.

```
load("ClownFishData.RData")
```

You now have an object called "data" in your environment. Take a few minutes to inspect it.

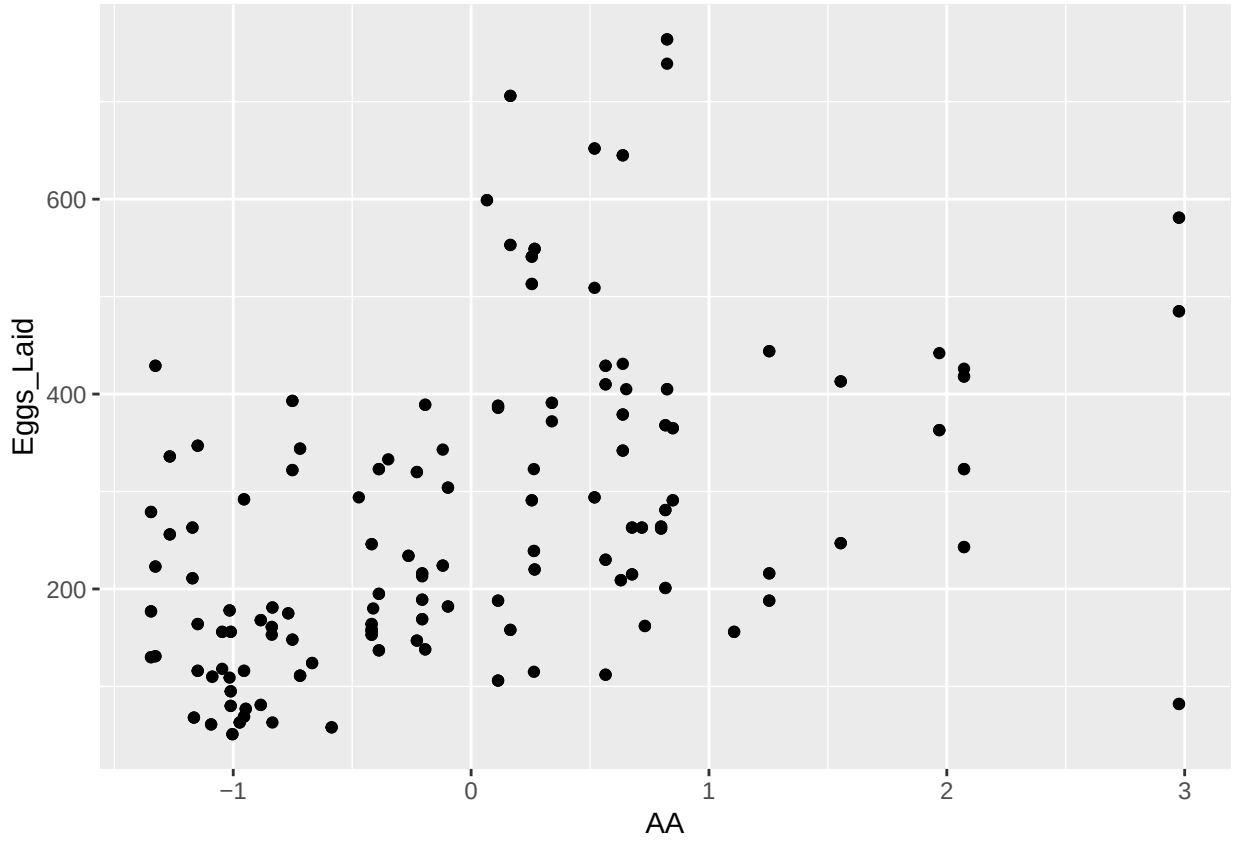
(1) Standardize and plot

We will be using anemone_area as a predictor. Since it is a continuous variable, it's a good idea to standardize it. Name the standardized version of this column "AA".

```
d2<-mutate(data,
           AA = standardize(anemone_area),
           FS = standardize(female_size),
           MS = standardize(male_size),
           EL = standardize(Eggs_Laid))
```

Make a figure showing the relationship between anemone area and number of eggs laid.

```
ggplot( d2 , aes(x=AA ,y=Eggs_Laid )) +geom_point()
```



(2) Plotting a prior

You will construct a linear regression model for the number of eggs laid with the anemone area (standardized) as the predictor. The model is

$$\text{Eggs\_Laid} \sim \text{Normal}(\mu, \sigma) \mu = a + b * \text{AA} a \sim \text{Normal}(400, 150) b \sim \text{Normal}(0, 200) \sigma \sim \text{Exponential}(0.01)$$

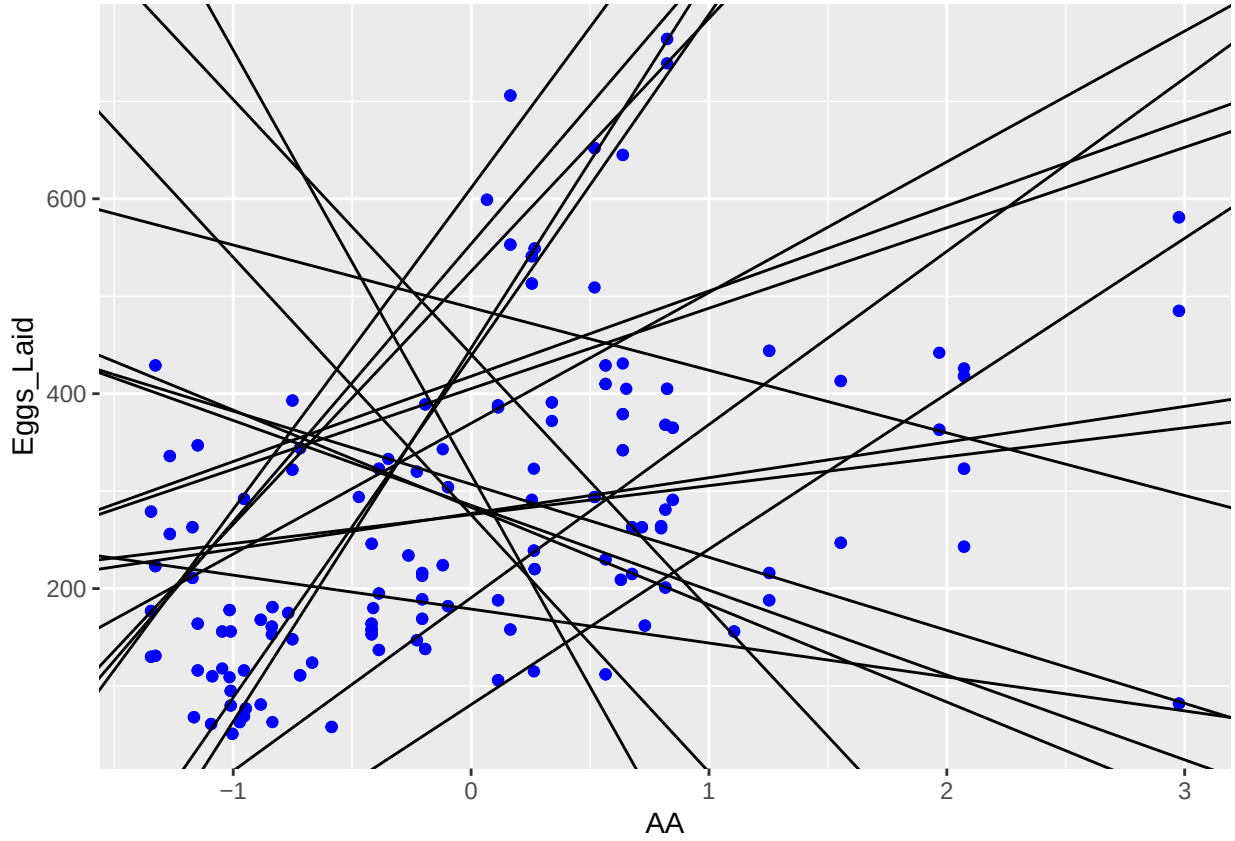
Plot the prior with the data. Explain what makes this a reasonable prior.

```
m.AA.eggs.laid <-
  quap( alist(
    Eggs_Laid ~ dnorm(mu,sigma),
    mu <- a + b * AA  ,
    a ~ dnorm(400,150),
    b ~ dnorm(0,200),
    sigma ~ dexp(0.01) ),
    data=d2,
    start= list(a=250,b=65,sigma=140))


precis(m.AA.eggs.laid)
```

```
##            mean       sd      5.5%     94.5%
## a     267.46918 8.976530 253.12296 281.81541
```

```
## b      68.15241 9.002507  53.76467  82.54016
## sigma 138.14620 6.330959 128.02810 148.26429
```

```
prior.samples<-extract.prior(m.AA.eggs.laid,prior=TRUE,n=20) %>% as_tibble()

ggplot(d2 , aes(x=AA,y=Eggs_Laid) )+
  geom_point(color="blue")+
  geom_abline(intercept=prior.samples$a, slope = prior.samples$b)
```



This prior produces possible relationships that span the range of the data. None of the lines are way below or way above the entire cloud of points. Some of the slopes are pretty extreme, and exceed the range of variation we see in the data. This is good, we want the priors to be broader than our data, but off in outer space.

(3) Linear regression model

Construct a *quap* model for the data and use `precis` to summarize the output.

The *quap* is a bit sensitive with these data, so it helps to specify the initial conditions. Use `start=list(a=250,b=65,sigma=140)` as an option for *quap*.

```
m.AA.eggs.laid <-
  quap( alist(
    Eggs_Laid ~ dnorm(mu,sigma),
    mu <- a + b * AA  ,
```

```
    a ~ dnorm(400,150),
    b ~ dnorm(0,200),
    sigma ~ dexp(0.01) ),
    data=d2,
    start= list(a=250,b=65,sigma=140))


precis(m.AA.eggs.laid)
```

```
##             mean       sd      5.5%     94.5%
## a     267.46918 8.976530 253.12296 281.81541
## b      68.15241 9.002507  53.76467  82.54016
## sigma 138.14620 6.330959 128.02810 148.26429
```

Note that if you get NaN for sigma values, the random part of the hill-climbing algorithm did not converge and you need to run it again. Quap is pretty fragile, if we start off in the wrong direction it won't converge.

(4) Explain the precis output

The precis output shows summary information about the parameters in our linear regression model of eggs laying based on anemone size. Because we standardized anemone area, we can interpret the intercept as the mean for average sized anemones. We find that this value is most probably between 253 and 281. Skipping to sigma, we find that the standard deviation is identified as being between 128 and 148, so most data should be within 300 units of the mean. Looking at the plot, the value of a and sigma are consistent with the range that we see. For the slope, b, we see a clearly positive relationship, with values between 53 and 82 being most probably.

(5) Plot the linear regression lines

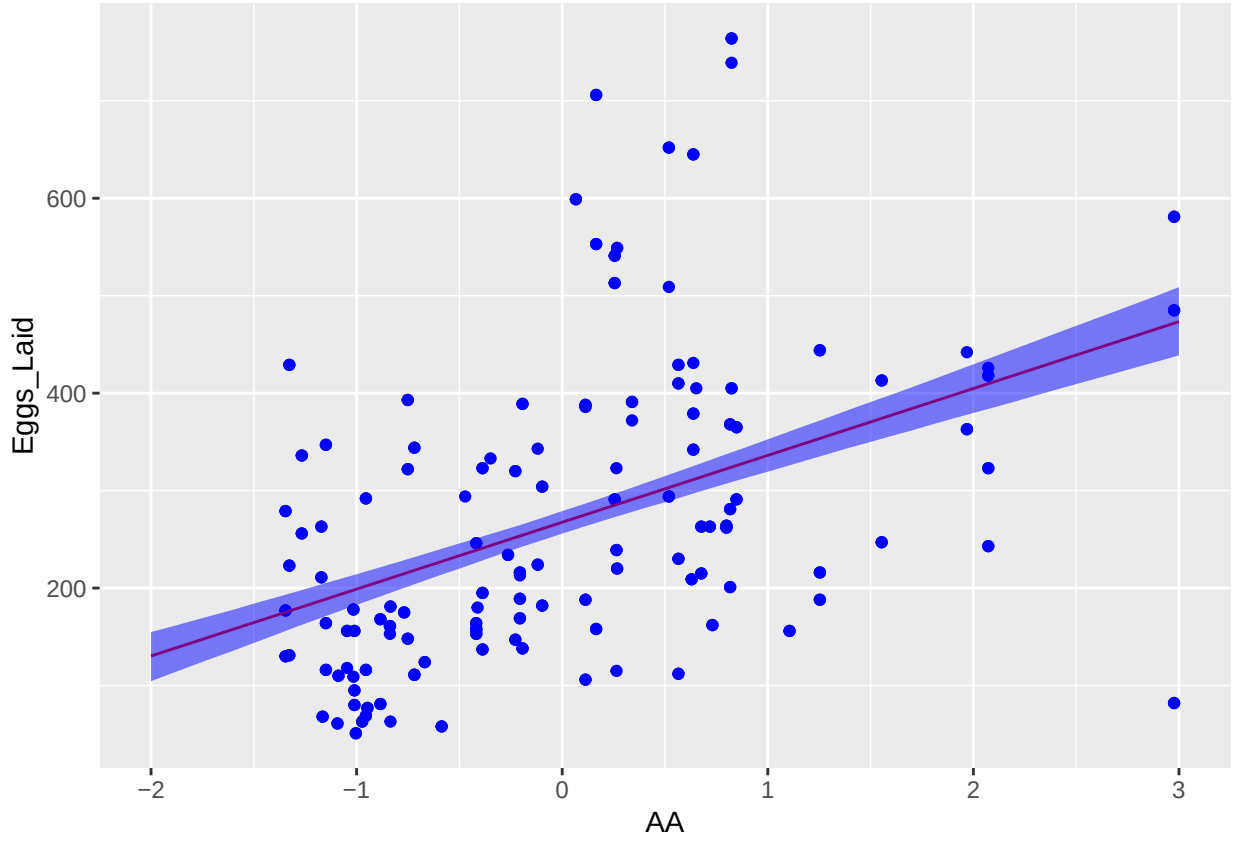
Use `link_df` to create samples of the *quap* fit. To do this, create a dataframe with evenly spaced out values of AA. Plot the mu values on a graph with the data.

```
sim_dat <- tibble(AA=seq(from=-2, to=3, by=0.1))
samples.AA.eggs.laid <- link_df(m.AA.eggs.laid,data=sim_dat)
```

```
## Warning: The ‘x‘ argument of ‘as_tibble.matrix()‘ must have unique column names if
## ‘.name_repair‘ is omitted as of tibble 2.0.0.
## i Using compatibility ‘.name_repair‘.
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use ‘all_of()‘ or ‘any_of()‘ instead.
##   # Was:
##   data %>% select(i)
##
##   # Now:
##   data %>% select(all_of(i))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

```r
samples.AA.eggs.laid.summarized <-samples.AA.eggs.laid %>%
  group_by(AA) %>%
  summarise(
    mean.mu = mean(mu),
    lower.mu = quantile(mu,0.1),
    upper.mu = quantile(mu,0.9)
  )%>%
  ungroup()


ggplot(d2 , aes(x=AA,y=Eggs_Laid) )+
  geom_point(color="blue")+
  geom_line(data=samples.AA.eggs.laid.summarized,aes(x=AA,y=mean.mu),color="red") +
  geom_ribbon(data=samples.AA.eggs.laid.summarized,inherit.aes = FALSE,aes(x=AA,ymin=lower.mu,ymax=upper
```



(6) List one "big world" explanations for why the data show more variability than the model fit does.

The model assumes that the number of eggs laid is normally distributed with a mean that depends on the anemone size. In reality, the fish are eating food and the food they get depends on conditions on the reef, their location on the reef, what predators of the fish are nearby, and a whole range of other conditions. This means that eggs laid will be determined by multiple random components, which may not be normally distributed, and which may be correlated. For example, anemones that are in a good spot might have fish that lay more eggs at both the pre- and post-treatment time point, and these values would be correlated, and cause the observed distribution to show more dispersion than a normal distribution.

(7) Including the treatment effect

Split the data into two datasets, one for fish who received the treatment, and the other for fish who did not receive the treatment.

You will analyze each of these datasets with a multivariate model that builds on your prior model. In addition to the effect of anemone size, include an effect based on whether or not the clutch of eggs was laid before or after the feeding occurred (remember PostTreatment=1 before feeding, and 2 after feeding).

For each dataset, perform the quap fit and use `precis` to summarize the results.

```
d2.notreat <- filter(d2, FedIndex==1)
d2.treat <- filter(d2, FedIndex==2)
```

```
m.nofed.time <-
  quap( alist(
    Eggs_Laid ~ dnorm(mu,sigma),
    mu <- a + b * AA + c* (PostTreatment-1)  ,
    a ~ dnorm(400,150),
    b ~ dnorm(0,200),
    c ~ dnorm(0,200),
    sigma ~ dexp(0.01) ),
    data=d2.notreat,
    start= list(a=250,b=65,sigma=140,c=100))


precis(m.nofed.time)
```

```
##            mean        sd      5.5%     94.5%
## a     201.52267 13.486023 179.96940 223.07594
## b      58.99430  8.867001  44.82312  73.16548
## sigma  99.68421  6.200513  89.77459 109.59382
## c      86.27706 17.744910  57.91727 114.63686
```

```
m.fed.time <-
  quap( alist(
    Eggs_Laid ~ dnorm(mu,sigma),
    mu <- a + b * AA + c* (PostTreatment-1)  ,
    a ~ dnorm(400,150),
    b ~ dnorm(0,200),
    c ~ dnorm(0,200),
    sigma ~ dexp(0.01) ),
    data=d2.treat,
    start= list(a=250,b=65,sigma=140,c=1))


precis(m.fed.time)
```

```
##            mean        sd      5.5%     94.5%
## a     179.13951 21.494374 144.78735 213.49168
## b      73.81605 13.231199  52.67004  94.96206
## sigma 137.85406  9.315611 122.96591 152.74220
## c     172.72002 27.135784 129.35179 216.08824
```

(8) Interpret the quap fits. What can you say about how the two datasets differ from each other in terms of their response to anemone size and to pre/post treatment?

Both models have similar values for a and b, although the mean of the psoterior for a is higher in the fed treatment, the range of both the fed and unfed overlap. This is also true for b. For c, the range of probably values is higher for the group of anemones/fish that were fed. This tells us that in the feeding treatment, after food was given, more eggs were laid. This supports the hypothesis that adding food increases egg laying.