Lab 6: My Dataset

Chickadee Nestling Diet Composition

Cori Carver

4 March 2021

General Information

This lab is due Wednesday, March 10th by 11:59 pm and is worth 10 points. You must upload your .rmd file and your knitted PDF to the assignment folder on Canvas.

Overview of the Independent Project

Each student will undertake an independent project during the semester. The primary objective is to have students statistically analyze a dataset of their own choosing. Students are encouraged to identify a real dataset that is interesting or meaningful to them, but the dataset must be approved by the instructor to ensure that it will meet the learning objectives of the course. Students, who are unable to find an appropriate dataset should consult with the instructor who will provide them with an appropriate dataset.

The objective is to conduct the analyses necessary to address the outlined hypothesis. The focus will be on understanding the data, developing a model appropriate to answer the question of interest, testing and assessing the model, and interpreting and displaying the results. This will include two assignments. Partway through the semester, students will be asked to turn in a document identifying their chosen dataset, briefly summarizing at least three scientific articles relevant to their topic, and preliminary plots (in R) of their response variable(s) and relationships with selected predictors. The final project will constitute a full report including Statistical Methods, Results, and Graphics in the style of a manuscript being prepared for submission. This means students will have to critically address which results to report, how to report them, and which figures and tables to present. This report will be graded on the basis of statistical practice, quality of reporting, support for interpretation, and quality of figures.

What is an Appropriate Dataset?

Philosophy

The most important thing is that the dataset addresses a biological question in which you are interested. Second, the point is NOT to recreate an existing analysis, but to instead discover something new entirely on your own. So if you choose to use a dataset from a previously published paper, then you will need to use these data in a fresh new way. The whole point of this exercise is to move beyond 'canned labs' where the results are already known (by the professor at least) and to use the skills that we have learned this semester to discover something entirely new! In previous courses, I have had some students who have published the results of their independent project in the peer reviewed literature. This is not my expectation for all projects, but I hope that you will strive to complete novel and impactful work!

Some specific guidelines

- At least 50 records (or observations or rows)

- A continuous response variable or a discrete response variable (0 or 1; counts; etc)
- At least 5 covariates/predictors (continuous and categorical)

Where should I look for a dataset?

If you do not already have data from an undergraduate or graduate thesis project then talk to your supervisor or one of your instructors who does research that interests you and ask whether they have data that you might use. Remember that the goal is to do something new and not to repeat a previous analysis. You can also look for datasets online (e.g. Dryad, FigShare, Ecological Archives, etc). If you still cannot find a dataset then come talk to me. I can provide you with one, but it will be easiest for you if the data already mean something to you and contains variables that you are interested in. I will also post some possible datasets from the course in previous years as well as some explanations of their variables on Canvas.

This Assignment

The goal of this assignment is to make sure that everyone has both a dataset to use for their final project.

Part A (6 points)

Provide a brief description of the biology behind your dataset. Who collected the data and for what purpose? What level of biological organization was sampled as the unit of replication (e.g. individual, population, community, ecosystem, etc.)? Were these units sub-sampled? Were the data collected as part of an experiment or were they observational? How were replicate samples collected (i.e. randomly, systematically, opportunistically)? You do not need to have a research question or hypothesis at this stage.

I have collected the data in collaboration with the Boulder Chickadee Study for my Honors Thesis. Individual chickadees were sampled as the unit of replication; this unit was not further sub-sampled. The data is observational, and replicate samples were collected opportunistically.

I would like to further discuss replicates with you next lab. In my study, I opportunistically collected nestling fecal samples, sometimes multiple from the same nest. I have a total of 69 fecal samples, but I am not sure if fecal samples collected from the same nest are replicates or not. I have included the total number of fecal samples for this lab, but all of the breeding phenology data for the samples collected from the same nests are the same.

Part B (4 points)

In the code box below, import your dataset into R. First make sure that your dataset is flat and 2-dimensional. Each row should correspond to a replicate or a sub-replicate and each column should correspond to a variable. There should be only one data file. You might find it easiest to make some changes to your variable names at this point so that they are relatively simple but informative for import into R.

```
fecal<-read.csv(file="fecal.csv")
#View(fecal) This call was not allowing my rmd to be knitted
```

Export or save this dataset from its current format to a comma-separated format (filename.csv). Here is some example code to export the object 'filename' as a csv file called "output_filename.csv":

```
# File is already a CSV
```

The csv file will be output to your working directory.

In these instructions I will provide examples such as "filename" above. Please name your files and variables in a way that makes most sense to you. My names are just examples.

The point of exporting the csv file is because I will need a csv copy of your dataset in order to check and trouble-shoot your analyses. Please make sure I have a copy of your up-to-date datafile and do not make any changes to it without either including these in your code (here and in your final project submission) or sending me an updated (modified) file.

Use the summary command to provide descriptive statistics for the variables in your data file.

```
summary(fecal)
```

```
##      band               site       nest.id           utm.easting
##  Length:68          Min.   :1.00   Length:68          Min.   :451435
##  Class :character   1st Qu.:1.00   Class :character   1st Qu.:469861
##  Mode  :character   Median :1.00   Mode  :character   Median :476586
##                     Mean   :1.75                      Mean   :472631
##                     3rd Qu.:2.25                      3rd Qu.:477181
##                     Max.   :4.00                      Max.   :479002
##                                                       NA's   :4
##   utm.northing       elevation    tree.species           height
##  Min.   :4333410   Min.   :1613   Length:68          Min.   : 97.0
##  1st Qu.:4428243   1st Qu.:1649   Class :character   1st Qu.:121.5
##  Median :4429229   Median :1653   Mode  :character   Median :132.0
##  Mean   :4428538   Mean   :1879                      Mean   :154.5
##  3rd Qu.:4432016   3rd Qu.:1947                      3rd Qu.:157.0
##  Max.   :4433469   Max.   :3254                      Max.   :298.0
##  NA's   :4         NA's   :4                         NA's   :4
##  orientation         canopy.cover   immediate.habitat    shavings
##  Length:68          Min.   :15.00   Length:68          Length:68
##  Class :character   1st Qu.:30.00   Class :character   Class :character
##  Mode  :character   Median :40.00   Mode  :character   Mode  :character
##                     Mean   :45.63
##                     3rd Qu.:60.00
##                     Max.   :80.00
##                     NA's   :5
##   year.placed     species          clutch.initiation   hatch.date
##  Min.   :2018   Length:68          Min.   :115.0     Min.   :134
##  1st Qu.:2019   Class :character   1st Qu.:119.0     1st Qu.:138
##  Median :2019   Mode  :character   Median :122.0     Median :141
##  Mean   :2019                      Mean   :125.8     Mean   :146
##  3rd Qu.:2019                      3rd Qu.:125.0     3rd Qu.:147
##  Max.   :2020                      Max.   :169.0     Max.   :192
##  NA's   :1                         NA's   :16        NA's   :5
##      day12      clutch.size      nestlings        fledged
##  Min.   :141   Min.   :5.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:151   1st Qu.:5.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :153   Median :6.000   Median :5.000   Median :4.000
##  Mean   :158   Mean   :6.045   Mean   :4.642   Mean   :4.339
##  3rd Qu.:162   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :201   Max.   :8.000   Max.   :8.000   Max.   :8.000
##  NA's   :6     NA's   :2       NA's   :1       NA's   :6
##    male.id           female.id
##  Length:68          Length:68
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Indicate which variables are dependent (i.e. response) variables (hopefully only one or a few) and which variables are independent (i.e. predictor) variables (hopefully > 5 as indicated in the first lecture). Indicate which variables, if any, were experimentally applied. For variables that are coded (e.g. Treatment = 1, 2, 3) be sure to indicate what each of these codes represents. Provide units for variables as necessary.

Here is an example from a recent file that I put together:

Variables and Their Explanations

band - A unique identifier for each nestling a fecal sample was collected from.

site - Field site where fecal sample was collected from: 1->Boulder, 2->CU Campus, 3->Sugar Loaf, 4->MRS.

nest.id - A unique identifier for each nest.

utm.easting - Location of nest.

utm.northing - Location of nest.

elevation - Elevation of nest in meters.

tree.species - Species of tree that nesting box is attached to.

height - Height of nesting box from ground in tree in cm.

orientation - Direction the opening of the nesting box is facing.

canopy.cover - Percent canopy cover above the nesting box.

immediate.habitat - Notes concerning habitat around nesting box.

shavings - Notes the addition of shavings (Y=shavings included, N=shavings not included).

year.placed - The year in which the box was deployed.

species - The species of chickadee that nested in the box and the fecal sample was collected from. MOCH = mountain chickadee and BCCH = black-capped chickadee

clutch.initiation - Julian date on which the first egg was laid.

hatch.date - Julian date on which the eggs hatched.

day.12 - Julian date on which the nestlings reached 12 days old and fecal sample was collected.

clutch.size - The number of eggs laid.

nestlings - The number of eggs that hatched.

fledged - The number of chickadees banded on day 12.

male.id - The band colors for the male associated with the nest.

female.id - The band colors for the female associated with the nest.

Summarizing the data

Indicate the number of rows in your data file by using:

```
length(fecal$band)
```

```
## [1] 68
```

If you have factors (i.e. categorical variables) then make sure that they are included as factors. If not, use the factor command to change them.

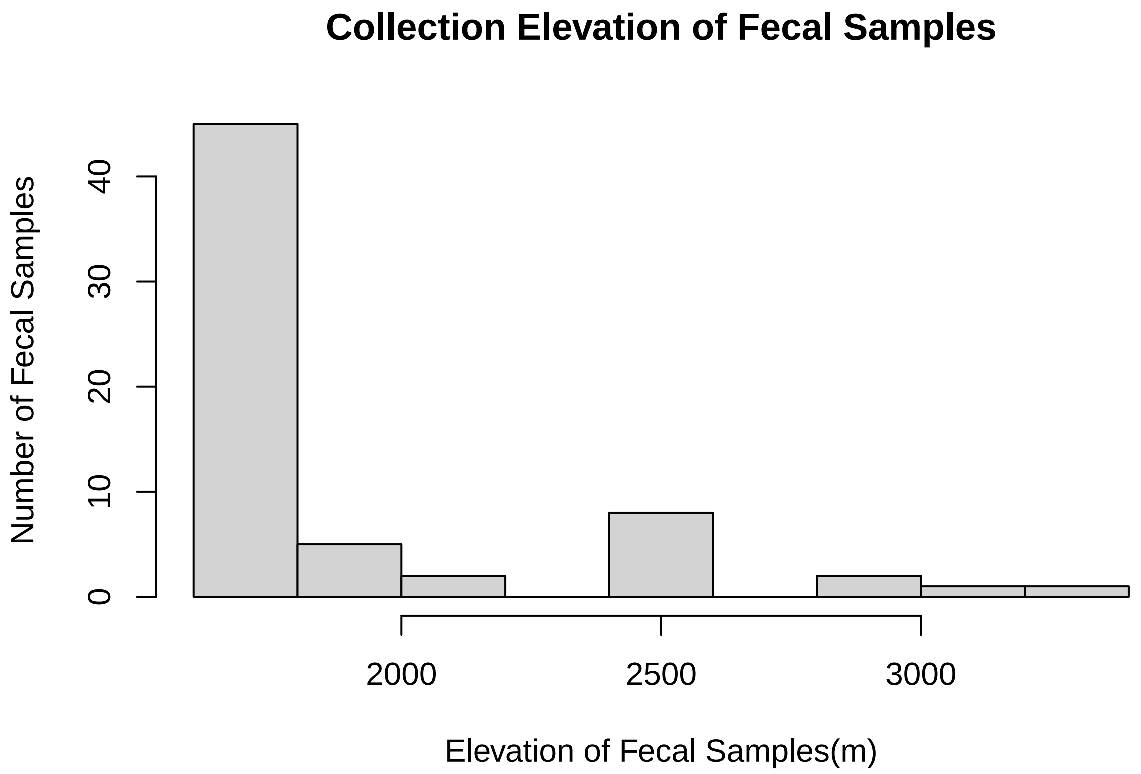
```
fecal$site<-factor(fecal$site) #1=Boulder, 2=CU Campus, 3=Sugar Loaf, 4=MRS
fecal$tree.species<-factor(fecal$tree.species) #Separated by species
fecal$orientation<-factor(fecal$orientation) #Seperated by cardinal directions
fecal$shavings<-factor(fecal$shavings) #Y=presence of shavings, N=no shavings
fecal$year.placed<-factor(fecal$year.placed) #Year deployed: 2018, 2019, 2020
fecal$species<-factor(fecal$species) #Species of bird: MOCH = mountain chickadee,
#BCCH = black-capped chickadee
```

Any individual, site or plot identifiers should be factors. Recall the summary command. Indicate what the variables and their levels mean as above.

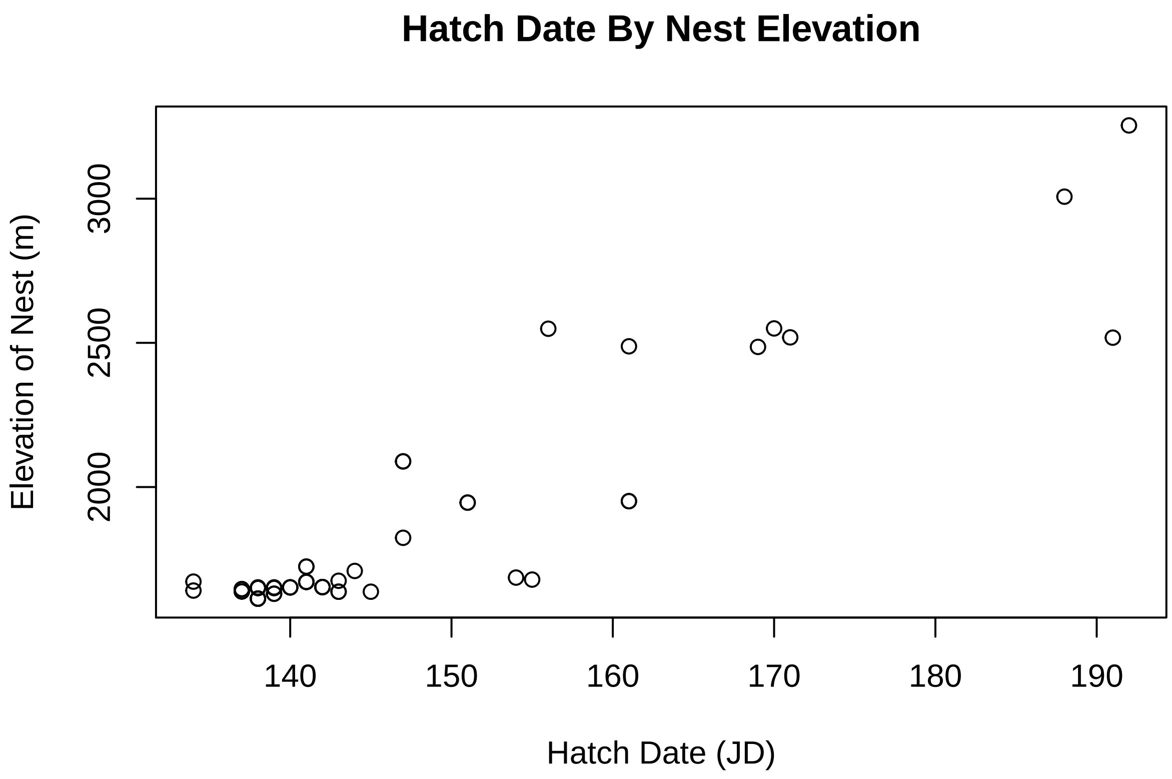
Some Basic Plots

Provide 2 preliminary plots (in R) of the response variable(s) and/or its relationship with selected predictors. What kind of plot will depend on the type of relationship. These could include a histogram of the distribution (raw or transformed), bivariate scatterplots with numerical predictors, or a stripchart or even mosaic plot for categorical predictors.

```
hist(fecal$elevation, ylab = "Number of Fecal Samples", xlab="Elevation of Fecal Samples(m)",
     main= "Collection Elevation of Fecal Samples")
```



```
plot(fecal$hatch.date, fecal$elevation, ylab = "Elevation of Nest (m)",
     xlab = "Hatch Date (JD)", main = "Hatch Date By Nest Elevation")
```



Save your R Markdown file and knit a PDF of the file and output.

What to submit

- A copy of your csv file
- A knitted PDF of your assignment
- A copy of your Rmd file