Short case study 2:
Identifying predictors of tree leafout

Steps 1-2 using Lasso regression for leafout
From: A four-step simulation-based workflow for ecological analysis and science

The four-step workflow we outline is designed for bespoke model building, but we argue it can be helpful for any model building, especially by helping researchers to think through how model structures connect to their underlying concepts and theory. In bespoke model building we often can match the model that we fit exactly to our generative model—that is, we can simulate data from the model we then fit to the simulated data. In other approaches, we may use more of a heuristic model to simulate data. This means we simulate data using a model that matches a basic understanding of our ecological system or process, but that may not match the statistical model we will then fit to simulated data.

For example, consider a case where we're interested in an ecological process driven by climate, but we are not fully sure which climate variables drive the process, though we have some basic ideas. The timing of tree leafout is a good example where many climate variables may matter; experiments in small controlled chambers show that winter cool (also called chilling), daylength and spring warm temperatures all matter for some tree species, but other work suggests these do not matter in natural conditions (that is, trees outside in nature) and that precipitation, clouds and other factors may matter instead or in addition.

Using this example, we might consider lasso regression—a method that fits many possible predictors, but penalizes ones with low predictive power based on fitting a parameter (λ) using cross-validation. We can approach lasso regression for this problem by simulating data starting with a simple model. The simplest model of leafout assumes leafout happens after a certain thermal sum is reached, which is commonly called a 'growing degree days' (GDD) model in agriculture. Assuming plants accumulate temperatures above some baseline (say, 0 C), these temperatures are added up each day until the threshold is crossed, at which point the tree leafs out.

We can simulate this using simulated climate data or actual climate data. We think testing out such models, at some point in the workflow, on actual climate data can be helpful as climate variables can have natural trends and correlations that may show up via this workflow, thus we use empirical climate data here.

We read in some climate data summaries including a few variables of interest (related to winter temperatures, spring temperatures and precipitation):

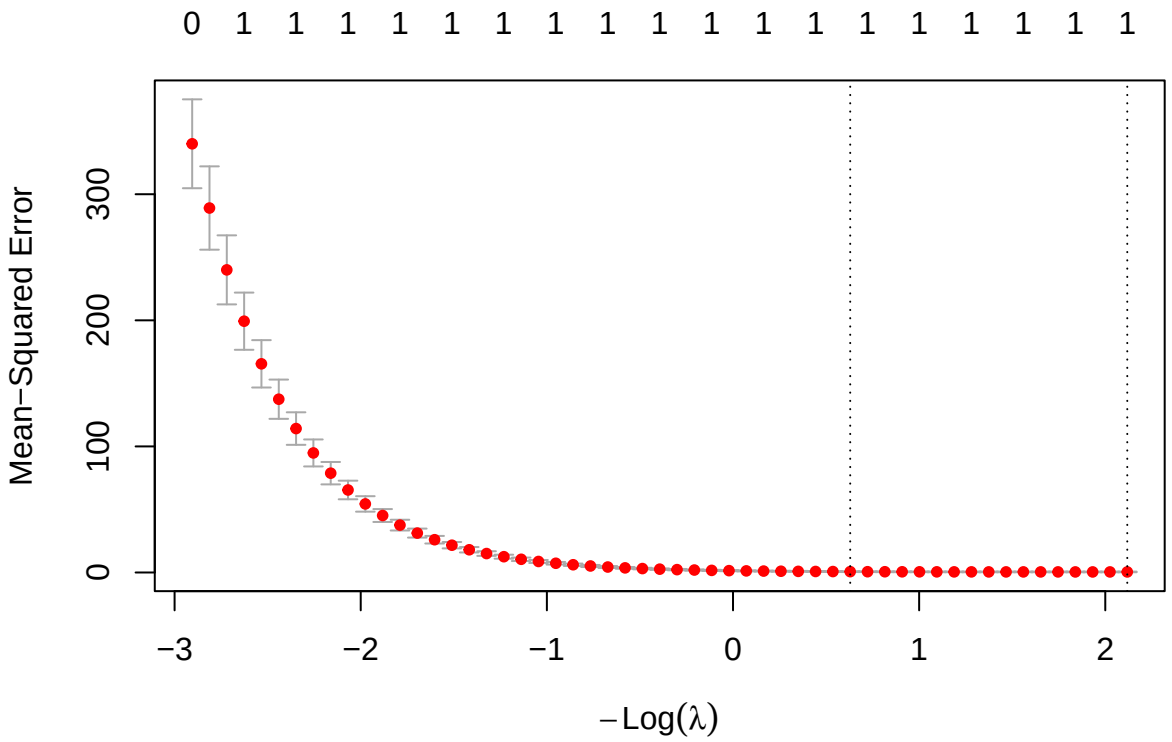
```
# housekeeping
rm(list=ls())
options(stringsAsFactors = FALSE)

wd <- '/Users/lizzie/Documents/git/projects/misc/miscmisc/bayesianflows/examples/lasso'

# libraries
library(geosphere) # get daylength
library(glmnet) # for lasso

# get the climate data
climdat <- read.csv(file.path(wd, "output/climdatwide.csv"))
tmeandaily <- read.csv(file.path(wd, "output/climdattmean.csv"))
```

Next we want to simulate a leafout date to occur after a thermal sum of 150 C (starting 1 January and accumulating values above 0 C). We also calculate the daylength on the date of leafout. Researchers commonly used this to estimate the effect of daylength in models of leafout (see https://doi.org/10.1073/pnas.2019411117), so—for this example—we figure we should too.

```
# add daylength
tmeandaily$daylength <- NA
for(i in 1:nrow(tmeandaily)){
  tmeandaily$daylength[i] <- daylength(50, tmeandaily$doy[i])
}

# simulate leafout after 150
fstar <- 150
lodf <- data.frame(year=unique(tmeandaily$year),
  loday=rep(NA, length(unique(tmeandaily$year))),
  gdd=rep(NA, length(unique(tmeandaily$year))),
  daylength=rep(NA, length(unique(tmeandaily$year))))

for(yearhere in unique(tmeandaily$year)) {
  thisyear <- tmeandaily[which(tmeandaily$year==yearhere),]
  leafoutdate <- min(which(cumsum(thisyear[["gddtemp"]]) > fstar))
  lodf$loday[which(lodf$year==yearhere)] <- leafoutdate
  lodf$gdd[which(lodf$year==yearhere)] <- cumsum(thisyear[["gddtemp"]])[leafoutdate]
  lodf$daylength[which(lodf$year==yearhere)] <- thisyear$daylength[leafoutdate]
}

# merge summaries and simulated leafout and daylength
simdat <- merge(climdat, lodf, by="year", all.x=TRUE)
```

Now we fit the lasso regression and look at what parameters it finds are most important. We simulated leafout based on GDD from temperature so we may expect metrics related to temperatures will appear important.

```
## Now fit lasso regression with all potential variables
# Create matrix of predictors (X) and response (y)
X <- as.matrix(simdat[, c("tminwinter",
    "tmeanspring",
    "precspring",
    "totalprec",
    "chillwinter",
    "daylength")])
y <- simdat$loday

# Run cross-validated to get lambda and plot results
# (the number of predictor variables is shown on the top)
cv_lasso <- cv.glmnet(X, y, alpha = 1, standardize = TRUE, nfolds = 10)
plot(cv_lasso)
```



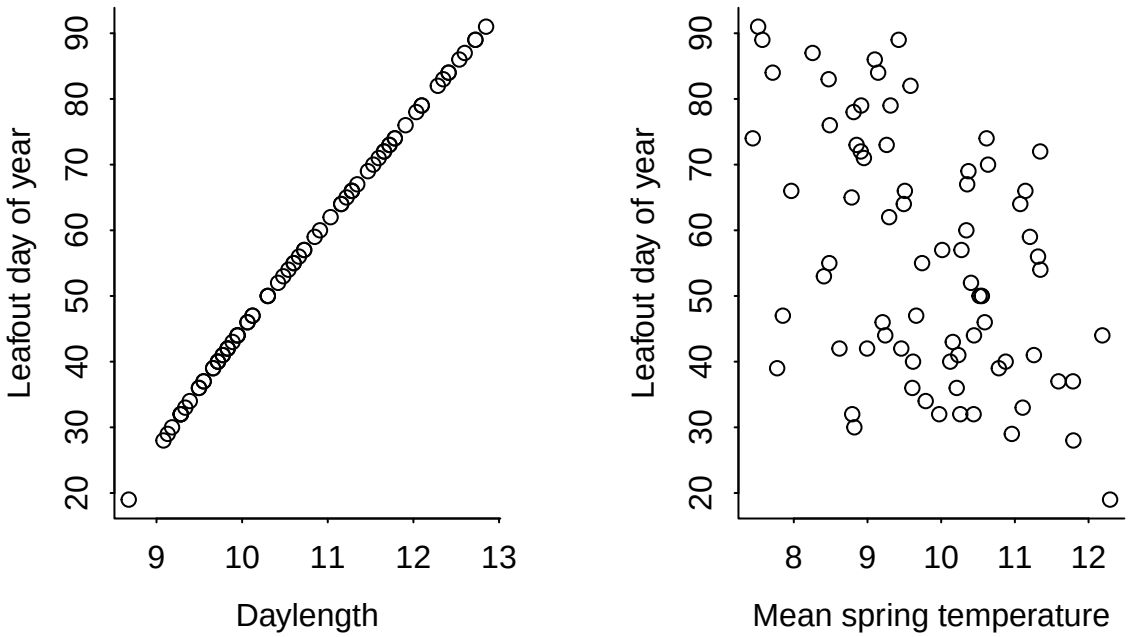

```
# Now check out the coefficients
coef(cv_lasso, s = "lambda.min")
```

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                lambda.min
## (Intercept) -120.51045
## tminwinter     .
## tmeanspring    .
## precspring     .
## totalprec      .
## chillwinter    .
## daylength     16.50365
```

But we find that none of the temperature metrics matter, even though we only used mean temperature above 0, which mostly happens in the spring, to simulate our leafout, and we included mean spring temperature in our lasso regression predictor set.

What's going on here? Hopefully at this point in the workflow we would start to dig into our data more. In visualizing our simulated leadout data versus our predictors, we realize that daylength almost perfectly correlates with day of year in the spring. Even though this is a common practice in the field (see https://doi.org/10.1073/pnas.2019411117) we may wonder if this is perhaps a bad idea. Using a heuristic model based on only temperatures to simulate leadout data, we have found daylength appears to be all that matters in our lasso regression, and begun to question our statistical approach.

```
par(mfrow=c(1,2), mgp=c(2, 0.5, 0), tck=-0.01)
plot(loday~daylength, simdat, bty="l", ylab="Leafout day of year",
     xlab="Daylength")
plot(loday~tmeanspring, simdat, bty="l", ylab="Leafout day of year",
     xlab="Mean spring temperature")
```



In visualizing the data more, we also see that a variable we expected to matter, mean spring temperature, looks pretty weakly related to leafout. We could continue on and try dropping daylength from our model (we encourage you to try it yourself adapting the above code), but that would show us that many variables now matter, including a number of ones that we did not use to simulate our data either.

These two first steps in the workflow have highlighted a disconnect between the statistical model we want to use and the underlying biological model. First, we did not include the actual predictor we used in our heuristic (biological) model—GDD—in our statistical model. While leaving out this metric (GDD) is a surprisingly common approach in research on tree leafout, the issues with it become more clear by using this workflow. Even if we had included GDD, we have used a linear regression model when our thermal sum model (GDD) is not actually linear. Addressing this likely requires thinking through a better statistical approach that more closely aligns to the biology of a thermal sum model to start (see https://doi.org/10.1111/gcb.15746).