BIO 655 Assignment 1

Maggie Swift

1/15/2020

Exercise 1

Write out the multiplication for $\mathbf{x}'\boldsymbol{\beta}$.

Given the following:

$$\mathbf{x}' = \begin{bmatrix} 1, x_{1,2}, ..., x_{1,p} \\ 1, x_{2,2}, ..., x_{2,p} \\ \vdots \\ 1, x_{n,2}, ..., x_{n,p} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad i \in \{1, 2, ..., n\}$$

We can multiply like so:

$$\mathbf{x}'\boldsymbol{\beta} = \begin{bmatrix} \beta_1 + \beta_2 x_{1,2} + ... + \beta_p x_{1,p} \\ \beta_1 + \beta_2 x_{2,2} + ... + \beta_p x_{2,p} \\ \vdots \\ \beta_1 + \beta_2 x_{n,2} + ... + \beta_p x_{n,p} \end{bmatrix} \implies \mathbf{x}_i'\boldsymbol{\beta} = \beta_1 + \beta_2 x_{i,2} + ... + \beta_p x_{i,p}$$

Exercise 2

Repeat the analysis using function `lm()` *with a different species as a response. Interpret the analysis.*

For this analysis, I use data from *Carya tomentosa* (mockernut hickory). From Fig.1, I can see that the relationship between (a) moisture deficit and winter temperature, and (b) site moisture and winter temperature is about the same as it was with *Quercus alba*.

Following the same steps as the example, I fit a linear regression for `biomass` response to a series of variables:

```
fit1 <- lm(biomass ~ moisture, data)
```

```
fit2 <- lm(biomass ~ moisture + I(moisture^2), data)
```

```
fit3 <- lm(biomass ~ soil + moisture*stdage + temp + deficit, data)
```

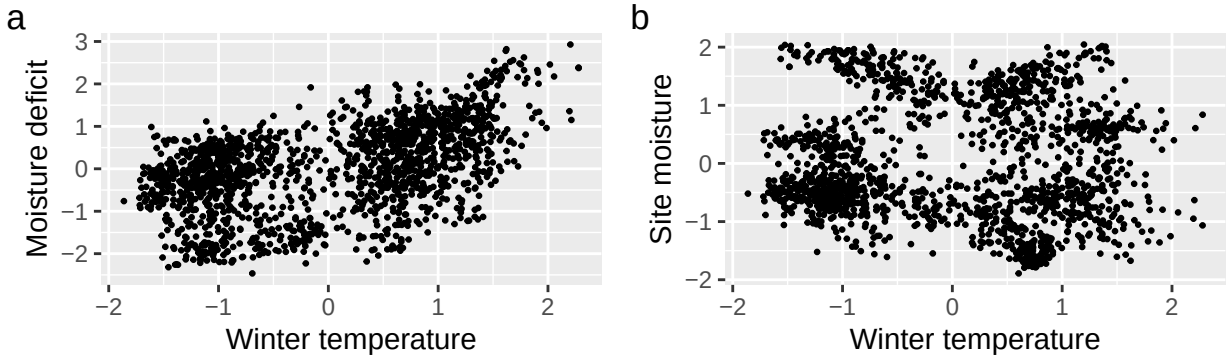

Figure 1: Some variables in the data.frame data

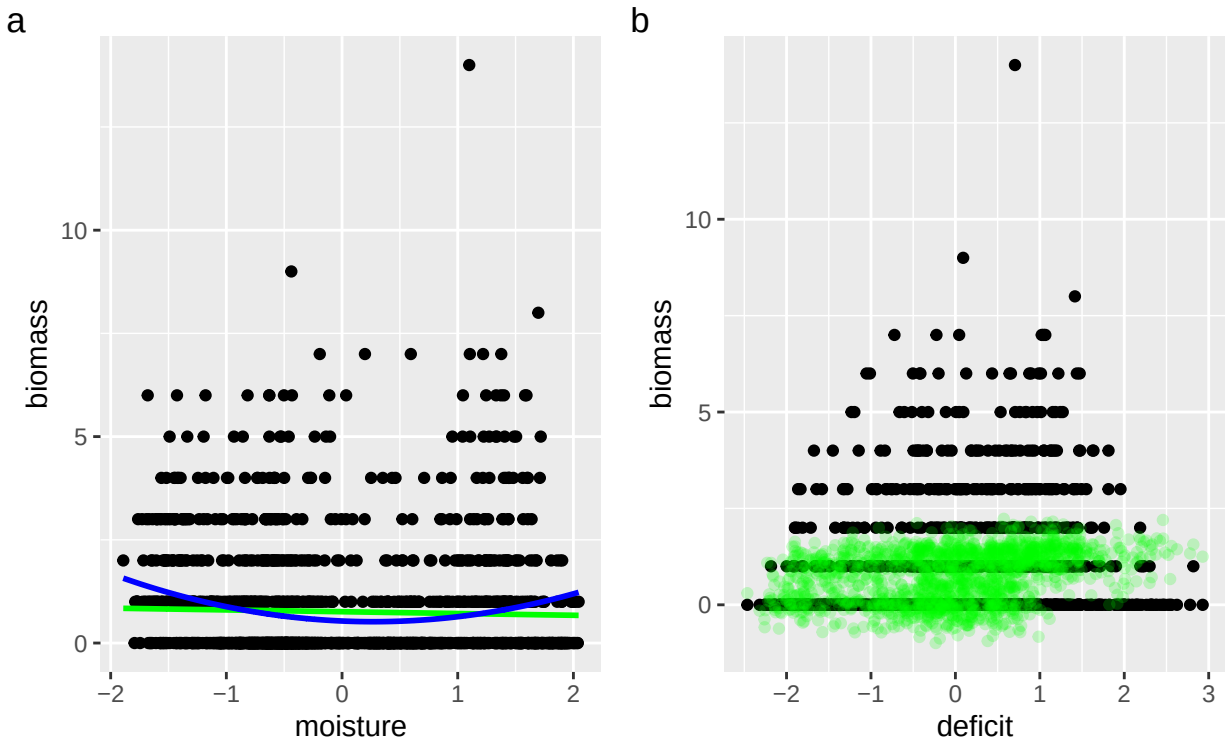

Figure 2: Biomass of *C. tomentosa* fitted to three models. (a) Biomass is predicted by moisture (green) and moisture^2 (blue). (b) Biomass is predicted by soil, winter temperature, moisture deficit, and the interactions between moisture and stand age; predictions for the model are shown in green.

From the summary for `fit3` (see Appendix), I can see that `stdage`, `temp`, and a few of the `soil` levels are significant; all others are decidedly not (p-value > 0.05). All of the significant parameters had positive values for the betas, indicating a positive relationship between `stdage` and `temp` on biomass.

The levels for `soil` are "EntVert", "Mol", "reference", "SpodHist", and "UltKan". Looking at the p- and t-values for each, a switch from the baseline "EntVert" soil type to either "reference" or "UltKan" both will cause some increase in biomass. Any of the other soil types ("Mol" & "SpodHist") would not produce significant increases or decreases in biomass when compared to "EntVert"

For the entire model, both multiple and adjusted R-squared are less than 0.25, indicating that the model explains less than 25% of the variation in this dataset (0.2327 and 0.2284, respectively). In addition, the F-statistic is quite high (54.16), confirming my conclusion that this model leaves much to be desired.

Exercise 3

Using a different species and predictors, interpret differences between estimates for the Tobit and standard regression model.

For this analysis, I chose to focus on sassafras (*Sassafras albidum*) and the predictors `moisture`, `soil`, and `deficit`. As I read on its Wikipedia article that it "prefers rich, well-drained sandy loam", I thought these three would be interesting to investigate.

The summaries (given in the Appendix) are actually relatively close to one another; both find that `moisture`, along with the `soil` levels "reference" and "SpodHist" have 95% credible intervals that do not include zero. The Tobit regression flagged `deficit` as well. A linear regression model with the same predictors (also in Appendix) gives a significance answer identical to the standard Bayesian; that is, `moisture` is significant, as well as `soil` levels "SpodHist" and "reference".

As for the beta coefficients for `deficit`, `moisture`, and `soil` levels "SpodHist" and "reference", the two standard models give similar estimates, while Tobit was radically different:

	deficit	moisture	soil.SpodHist	soil.reference
Linear Regression	-0.034	0.078	-0.206	0.152
Standard Bayes	-0.034	0.078	-0.204	0.152
Tobit	-0.326	0.433	-4.758	0.881

The Tobit regression gave beta-coefficients of much higher magnitude (although all of the same sign) to each predictor. This is to be expected, as Tobit regression's extension of zeroes into the negative parameter space often causes the line of fit's slope to be much steeper than that of a traditional regression.

Exercise 4

In a Bayesian model, identify the marginal, joint, and conditional distributions. What is random in the posterior distribution? What is random in the likelihood?

In a Bayesian model we have:

[D, P] = [D|P][P] = [P|D][D]

The *joint* distribution is [D,P]. That is, we are searching for the distribution of both D and P together. The *marginal* distribution is given by [P] in the middle term, and [D] in the last. The *conditional* distribution is either [D|P] or [P|D], read as 'D given P' or 'P given D', respectively.

In the posterior distribution ([P|D]), the process is random, as we already know the data. In the likelihood ([D|P]), the entire term is random, as it is itself a probability distribution realizing a random process (we don't know the exact process).

Exercise 5

What are the three elements of a design? What does each contribute?

The three elements of a design are replication, stratification, and randomization (RSR). Replication ensures that conclusions aren't drawn due to fluke or erroneous data and, more generally, that other researchers are able to check the validity of the experiment and analytical conclusions. Stratification is the main element of an experiment–without differences between test or observational groups, the experiment isn't actually testing any hypotheses. Randomization ensures that no bias has entered into the experiment, allotting more trust to the researchers' analyses.

Exercise 6 (group)

How is uncertainty in data handled? For example, is there a likelihood? If so, what is it? If not, how do you think uncertainty affects estimates and their interpretation?

α-value uncertainty for all models (except for the extinction model EX) is given by propagating errors and therefore obtaining natural variation of K over the six 1-species replicates. For the EX model, the 95% CI is obtained by bootstrapping over the six replicates. However, there are no distributions or likelihoods associated with these techniques.

Are observations independent? If not, how does this affect the estimates and their interpretation?

Observations are independent across replicates, but not between samples drawn at time $t_0 = 0$ and $t^* = 21$ days. The authors have taken this into account, however, and in fact the basis of these experiments is to mark the change in population density and biomass over time.

Are there fixed and random parameters? How might the differences affect the model?

All model parameters, in this context, are fixed. Species interactions ($\alpha_{i,j}$), inherent growth rate (r), and carrying capacity (K) are all estimated from the data. Random, in this context, would indicate a parameter was drawn from a probability distribution or can be described by a probability distribution, which none are. However, there is stochasticity in the community stability analysis; the authors ran simulations randomly drawing from uncertainty intervals to perturb r, K, and α to analyze stability.

What is the role of computation in each model? Do Jordan's concerns apply to any of these methods?

No computation is used in any of the deterministic models (EX, RY, EQ, and LVD), as the parameters are simply fitted to the model, with LVD fitting using OLS. However, the stochastic community model does include computation, adding demographic noise and experimental uncertainties of α to the simulation. We conclude that Jordan's concerns don't necessarily apply to any of the five methods outlined above, because 6 replicates and 55 interactions is not nearly enough to caution against erroneous causal inference. Jordan's concerns were more applicable where hundreds of variables were tested for causality.

Appendix

```
#-----------------------------------------------------------------
# Set-up
#-----------------------------------------------------------------
knitr::opts_chunk$set(echo=F, eval=T, message=F)
library(GGally)
library(ggplot2)
library(gjam)
library(knitr)
source('clarkfunctions2020.R')

#-----------------------------------------------------------------
# Exercise 2
#-----------------------------------------------------------------

# Load data and grab only Carya tomentosa (mockernut hickory)
d <- "https://github.com/jimclarkatduke/gjam/blob/master/forestTraits.RData?raw=True"
invisible( repmis::source_data(d) )
tmp <- gjam::gjamReZero(forestTraits$treesDeZero) # decompress data
data <- forestTraits$xdata
data$biomass <- tmp[,"caryTome"]

# First, let's plot for C. tomentosa in the same way; with
# winter temperature anomaly as x-axis and moisture & deficit
# as responses.

xlab <- 'Winter temperature'
g1 <- ggplot(data=data, aes(y=deficit, x=temp)) +
  geom_point(size=0.5) +
  labs(x=xlab, y='Moisture deficit', tag='a')
g2 <- ggplot(data=data, aes(y=moisture, x=temp)) +
  geom_point(size=0.5) +
  labs(x=xlab, y='Site moisture', tag='b')
gridExtra::grid.arrange(g1, g2, nrow=1)

# Our three linear models, but for Carya tomentosa now.
fit1 <- lm(biomass ~ moisture, data)
fit2 <- lm(biomass ~ moisture + I(moisture^2), data)
fit3 <- lm(biomass ~ soil + moisture*stdage + temp + deficit, data)

data$fit1 <- predict(fit1)
data$fit2 <- predict(fit2)
data$fit3 <- predict(fit3)

p1 <- ggplot(data=data, aes(y=biomass, x=moisture)) +
  geom_point() +
  geom_line(aes(y = fit1), size = 1, color='green') +
  geom_line(aes(y = fit2), size = 1, color='blue') +
  labs(tag='a')
p2 <- ggplot(data=data, aes(y=biomass, x=deficit)) +
  geom_point() +
  geom_point(aes(y = fit3), color="green", alpha=0.2) +
  labs(tag='b')
```

```
gridExtra::grid.arrange(p1, p2, nrow=1)

#-----------------------------------------------------------------
# Exercise 3
#-----------------------------------------------------------------
data <- forestTraits$xdata
data$biomass <- tmp[,"sassAlbi"]

# choose the formula
form <- as.formula(biomass ~ soil + moisture + deficit, data)

# Table for Exercise 3
# Note: I had to fill these in manually since the actual regression is
# in the appendix, as I couldn't get the fits to not print summaries here.
t <- data.frame(deficit=c(-0.034, -0.034, -0.326),
                moisture=c(0.078, 0.078, 0.433),
                soil.SpodHist=c(-0.206, -0.204, -4.758),
                soil.reference=c(0.152, 0.152, 0.881))
row.names(t) <- c('Linear Regression', 'Standard Bayes', 'Tobit')
kable(t)

#-----------------------------------------------------------------
# Appendix
#-----------------------------------------------------------------
summary(fit3)
data <- forestTraits$xdata
data$biomass <- tmp[,"sassAlbi"]
form <- as.formula(biomass ~ soil + moisture + deficit, data)
fit4 <- lm(form, data)
fitb <- bayesReg(form, data)
fitTobit <- bayesReg(form, data, TOBIT=T)
summary(fit4)
```

Exercise #2

Summary for fit3:

```
##
## Call:
## lm(formula = biomass ~ soil + moisture * stdage + temp + deficit,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2006 -0.6685 -0.1670  0.2938 12.0523
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.151413   0.119122   1.271    0.204
## soilMol          0.167196   0.230649   0.725    0.469
## soilreference    0.771300   0.124733   6.184 7.93e-10 ***
## soilSpodHist     0.234462   0.141639   1.655    0.098 .
## soilUltKan       1.338344   0.205114   6.525 9.10e-11 ***
## moisture        -0.039899   0.030664  -1.301    0.193
```

```
## stdage           0.211320   0.031491   6.711 2.68e-11 ***
## temp             0.502547   0.039042  12.872  < 2e-16 ***
## deficit         -0.027411   0.036164  -0.758    0.449
## moisture:stdage -0.002018   0.033047  -0.061    0.951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.22 on 1607 degrees of freedom
## Multiple R-squared:  0.2327, Adjusted R-squared:  0.2284
## F-statistic: 54.16 on 9 and 1607 DF,  p-value: < 2.2e-16
```

Exercise #3

Standard Bayesian regression summary:

```
##
## Coefficients:
##                 median std error     0.025    0.975 not zero
## intercept        0.194   0.07241   0.05323   0.3386        *
## soilMol        0.04059    0.1391   -0.2361   0.3099
## soilreference    0.153   0.07609 0.0006874    0.298        *
## soilSpodHist   -0.2037   0.08416   -0.3757 -0.03878        *
## soilUltKan     -0.1048    0.1225   -0.3393   0.1304
## moisture       0.07758   0.01899   0.03891   0.1141        *
## deficit       -0.03382   0.01951  -0.07031 0.005319
##
##  * indicates that 95% predictive interval does not include zero
##
## Residual standard error 0.7445, with 1610 degrees of freedom,
##  root mean sq prediction error 0.7419.
```

Tobit regression summary:

```
##
## Coefficients:
##                median std error  0.025   0.975 not zero
## intercept       -3.01     0.491 -4.013  -2.102        *
## soilMol         0.185    0.7737 -1.268   1.783
## soilreference  0.8793    0.4468 0.0432   1.822        *
## soilSpodHist   -4.763     1.158 -7.605  -2.847        *
## soilUltKan    -0.9264     0.813 -2.537  0.6248
## moisture       0.4288    0.1016 0.2285  0.6262        *
## deficit       -0.3282     0.114 -0.551 -0.1082        *
##
##  * indicates that 95% predictive interval does not include zero
##
## Residual standard error 2.72, with 1610 degrees of freedom,
##  root mean sq prediction error 0.7997.
```

Standard linear regression summary:

```
##
## Call:
## lm(formula = form, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.5389 -0.3296 -0.2156  0.0050 14.5444
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.19541    0.07252   2.694  0.00712 **
## soilMol        0.03899    0.13985   0.279  0.78042
## soilreference  0.15170    0.07587   2.000  0.04572 *
## soilSpodHist  -0.20568    0.08459  -2.432  0.01514 *
## soilUltKan    -0.10541    0.12365  -0.853  0.39406
## moisture       0.07754    0.01864   4.159 3.36e-05 ***
## deficit       -0.03383    0.01972  -1.715  0.08649 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7439 on 1610 degrees of freedom
## Multiple R-squared:  0.04715,    Adjusted R-squared:  0.0436
## F-statistic: 13.28 on 6 and 1610 DF,  p-value: 9.901e-15
```