Homework 3

# Set up

## Reading in packages

# general use  
library(tidyverse)  
library(readxl)  
library(here)  
library(janitor)  
  
# visualizing pairs  
library(GGally)  
  
# model selection  
library(MuMIn)  
  
# model predictions  
library(ggeffects)  
library(ggplot2)  
  
# model tables  
library(gtsummary)  
library(flextable)  
library(modelsummary)  
  
drought\_exp <- read\_xlsx(path = here("data",   
 "Valliere\_etal\_EcoApps\_Data.xlsx"),  
 sheet = "First Harvest")  
  
# quick look at data   
str(drought\_exp)

tibble [70 × 13] (S3: tbl\_df/tbl/data.frame)  
 $ Species : chr [1:70] "ENCCAL" "ENCCAL" "ENCCAL" "ENCCAL" ...  
 $ Water : chr [1:70] "WW" "WW" "WW" "WW" ...  
 $ Rep # : num [1:70] 1 2 3 4 5 1 2 3 4 5 ...  
 $ Height (cm) : num [1:70] 5.8 4.9 8.4 6.5 7.1 3.2 4.4 4.2 4.5 3.9 ...  
 $ Leaf # : num [1:70] 11 8 11 12 10 7 7 10 8 6 ...  
 $ Leaf dry weight (g): num [1:70] 0.0294 0.0185 0.0177 0.0178 0.0164 0.017 0.0193 0.0153 0.0159 0.0133 ...  
 $ Leaf area (cm2) : num [1:70] 5.01 3.98 3.69 3.84 3.63 3.06 3.1 2.94 2.73 2.61 ...  
 $ SLA : num [1:70] 170 215 209 216 222 ...  
 $ Total LA : num [1:70] 55.1 31.8 40.6 46.1 36.3 ...  
 $ Shoot (g) : num [1:70] 0.253 0.164 0.241 0.213 0.232 ...  
 $ Root (g) : num [1:70] 0.202 0.165 0.209 0.146 0.12 ...  
 $ Total (g) : num [1:70] 0.455 0.329 0.45 0.359 0.352 ...  
 $ R:S : num [1:70] 0.8 1 0.9 0.7 0.5 0.8 1.2 3.1 0.9 1.2 ...

class(drought\_exp)

[1] "tbl\_df" "tbl" "data.frame"

## Data cleanup

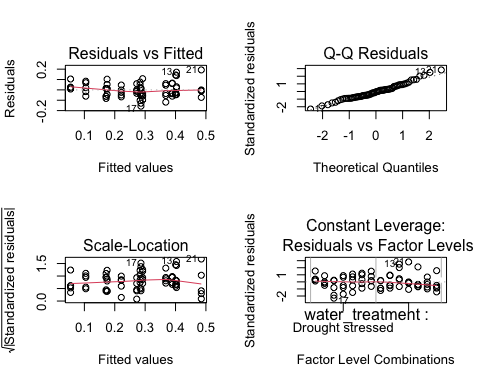
# cleaning  
drought\_exp\_clean <- drought\_exp %>%   
 clean\_names() %>% # nicer column names  
 mutate(species\_name = case\_when( # adding column with species scientific names  
 species == "ENCCAL" ~ "Encelia californica", # bush sunflower  
 species == "ESCCAL" ~ "Eschscholzia californica", # California poppy  
 species == "PENCEN" ~ "Penstemon centranthifolius", # Scarlet bugler  
 species == "GRICAM" ~ "Grindelia camporum", # great valley gumweed  
 species == "SALLEU" ~ "Salvia leucophylla", # Purple sage  
 species == "STIPUL" ~ "Nasella pulchra", # Purple needlegrass  
 species == "LOTSCO" ~ "Acmispon glaber" # deerweed  
 )) %>%   
 relocate(species\_name, .after = species) %>% # moving species\_name column after species  
 mutate(water\_treatment = case\_when( # adding column with full treatment names  
 water == "WW" ~ "Well watered",  
 water == "DS" ~ "Drought stressed"  
 )) %>%   
 relocate(water\_treatment, .after = water) # moving water\_treatment column after water

# Model code

# Null model  
model0 <- lm(total\_g ~ 1, # formula  
 data = drought\_exp\_clean) # data frame  
  
# Model 1: total biomass as a function of SLA, water treatment, and species  
model1 <- lm(total\_g ~ sla + water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
par(mfrow = c(2, 2))  
  
# Model 2: total biomass as a function of SLA and water treatment  
model2 <- lm(total\_g ~ sla + water\_treatment,  
 data = drought\_exp\_clean)  
  
# Model 3: total biomass as a function of SLA and species  
model3 <- lm(total\_g ~ sla + species\_name,  
 data = drought\_exp\_clean)  
  
# Model 4: total biomass as a function of water treatment and species  
model4 <- lm(total\_g ~ water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
# Determines best model  
model.sel(model0,  
 model1,   
 model2,   
 model3,  
 model4)

Model selection table   
 (Int) sla spc\_nam wtr\_trt df logLik AICc delta weight  
model4 0.05455 + + 9 88.598 -156.2 0.00 0.772  
model1 0.07994 -0.0002475 + + 10 88.741 -153.8 2.44 0.228  
model3 -0.03315 0.0012900 + 9 72.538 -124.1 32.12 0.000  
model2 0.04670 0.0012810 + 4 52.220 -95.8 60.37 0.000  
model0 0.27900 2 39.580 -75.0 81.22 0.000  
Models ranked by AICc(x)

# plots model 4 to use in question 1b  
plot(model4)



# Problem 1

## a.

# table  
## puts models into a list  
models <- list(   
 "null" = model0, # "model name" = model object  
 "model 1" = model1,  
 "model 2" = model2,  
 "model 3" = model3,  
 "model 4" = model4)  
  
modelsummary(models, # this function takes a list of models  
 title = "Table 1: Multiple linear regression of the 5 models. The rows showcase the predictors for each model. The columns show each numbered model." # table title  
 )

Table 1: Multiple linear regression of the 5 models. The rows showcase the predictors for each model. The columns show each numbered model.

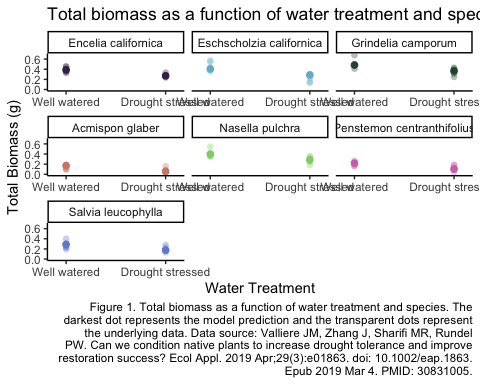
|  | null | model 1 | model 2 | model 3 | model 4 |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | 0.279 | 0.080 | 0.047 | -0.033 | 0.055 |
|  | (0.017) | (0.056) | (0.054) | (0.067) | (0.025) |
| sla |  | 0.000 | 0.001 | 0.001 |  |
|  |  | (0.000) | (0.000) | (0.001) |  |
| water\_treatmentWell watered |  | 0.122 | 0.090 |  | 0.117 |
|  |  | (0.020) | (0.029) |  | (0.017) |
| species\_nameEncelia californica |  | 0.238 |  | 0.115 | 0.218 |
|  |  | (0.051) |  | (0.059) | (0.032) |
| species\_nameEschscholzia californica |  | 0.234 |  | 0.222 | 0.232 |
|  |  | (0.033) |  | (0.041) | (0.032) |
| species\_nameGrindelia camporum |  | 0.330 |  | 0.226 | 0.313 |
|  |  | (0.047) |  | (0.054) | (0.032) |
| species\_nameNasella pulchra |  | 0.241 |  | 0.168 | 0.229 |
|  |  | (0.040) |  | (0.048) | (0.032) |
| species\_namePenstemon centranthifolius |  | 0.061 |  | -0.006 | 0.050 |
|  |  | (0.039) |  | (0.047) | (0.032) |
| species\_nameSalvia leucophylla |  | 0.117 |  | 0.139 | 0.120 |
|  |  | (0.033) |  | (0.041) | (0.032) |
| Num.Obs. | 70 | 70 | 70 | 70 | 70 |
| R2 | 0.000 | 0.755 | 0.303 | 0.610 | 0.754 |
| R2 Adj. | 0.000 | 0.722 | 0.282 | 0.566 | 0.726 |
| AIC | -75.2 | -157.5 | -96.4 | -127.1 | -159.2 |
| BIC | -70.7 | -135.0 | -87.4 | -106.8 | -139.0 |
| Log.Lik. | 39.580 | 88.741 | 52.220 | 72.538 | 88.598 |
| RMSE | 0.14 | 0.07 | 0.11 | 0.09 | 0.07 |

## b.

**To examine the influence of specific leaf area, water treatment, and species on total biomass, I created 5 models. The null model was a function of none of the predictors on total biomass, while model 1 was a function of all of the predictors on total biomass. Models 2, 3, and 4 were functions of 2 different predictors on total biomass. To determine the model that best described the influences on total biomass, I used multiple linear regression and found that model 4 (total biomass as a function of water treatment and species) had the lowest AIC value (-159.2). To evaluate linear model assumptions, I plotted model 4 and found that it was linear, normally distributed, homoscedastic, and has no high leverage points.**

## c + d.

ggpredict(model4, #selects model 4 to be analyzed  
 terms = c("water\_treatment", # focuses on the water treatment and species name  
 "species\_name")) %>%   
 plot(show\_data = TRUE, # shows the underlying data  
 facets = TRUE, # facets the data based on species name  
 show\_legend = FALSE, # stops the legend from showing  
 colors = "ipsum") +  
 theme\_classic() + # adds color to the graph  
 labs(title = "Total biomass as a function of water treatment and species", # title  
 x = "Water Treatment", # x-axis title  
 y = "Total Biomass (g)", # y-axis title  
 caption = str\_wrap("Figure 1. Total biomass as a function of water treatment and species. The darkest dot represents the model prediction and the transparent dots represent the underlying data. Data source: Valliere JM, Zhang J, Sharifi MR, Rundel PW. Can we condition native plants to increase drought tolerance and improve restoration success? Ecol Appl. 2019 Apr;29(3):e01863. doi: 10.1002/eap.1863. Epub 2019 Mar 4. PMID: 30831005.")) # caption that wraps to prevent it from going off the page



## e.

**Water treatment and species were the predictors that best described total mass (AIC = -159.2). On average, well-watered plants had a higher total biomass than drought stressed plants. On average Grindelia camporum had the highest total biomass with Eschscholzia californica, Nasella pulchra, and Encelia californica slightly smaller than it. On average Acmispon glaber had the lowest total biomass wiht Penstemon centranthifolius and Salvia leucophylla slightly bigger than it.**

# Problem 2

## a.

**For my data, I could do a piece of art. I would first change my graph to be a line plot with number of times the water bottle was refilled on the y-axis and date on the x-axis. I am thinking that the line of data could be the water flowing into the bottle. I would then draw a bottle below it to catch the water. This would be affective visualization because it includes the data and art that resembles what the data is of.**

## b.

## c.

## d.

**This artpiece is a water cooler that has water in it and is dispensing it into the water bottle. This is to showcase the amount of times I filled my water bottle per day. I was influenced by Jill Pelto’s paintings used to represent her data. My piece of work is a printing of my data as a line plot and includes a drawing that uses crayon for color. I created this work by printing the piece, drawing a water cooler in pencil, and coloring it in using crayon.**

# Problem 3

## a.

**This study measured the probability of food insecurity among people with disabilities. There was a focus on number of disabilities and type of disabilities. The author uses a logistical regression to model the probability of any and severe food insecurity as a function of the number of disabilities and the type of disability, controlling for socio-demographic variables.** **INSERT FIGURE 1**

## b.

**The first graph is a line plot which shows the probability of food insecurity (y-axis) based on the number of disabilities (x-axis). The second graph is a histogram plot which shows the probability of food insecurity (y-axis) based on the category of disability (x-axis). The author’s choice to separate these graphs made sense since number of disabilities is continuous data and the category of disability is categorical data. Both graphs show the mean probability of food insecurity and the standard error, but do not show the underlying data.**

## c.

**The authors did a good job handling ‘visual clutter.’ They graphs are relatively simple and primarily focus on the probabilities for each x value. The data:ink ratio is relatively high. There data is portrayed in a simple matter and is not cluttered or overwhelmed by grid lines and extraneous elements.**

## d.

**One recommendation I have is to change the y-axis values. Instead of probability represented as a decimal (ex: 0.1), it should be represented as a percentage (ex: 10%). This would make it easier to see on the graphs and easier for people with little research experience to understand. Another recommendation I have is to make the histogram bars a lighter and more transparent color on the inside (fill) and keep a dark outline for the bars. This would make the graph to seem more simple and the coloration of the graph would be less aggressive. I think this could help reduce visual clutter.**