ENVS 193DS Homework 3

Brooke Ryan

# link to forked repository:

https://github.com/Brookepryan/ryan\_brooke\_homework-03.git

# reading in packages/data

#suppressing messages/warnings  
suppressMessages({  
 suppressWarnings({  
# general use  
library(tidyverse)  
library(readxl)  
library(here)  
library(janitor)  
library(dplyr)  
library(ggplot2)  
   
# visualizing pairs  
library(GGally)  
  
# model selection  
library(MuMIn)  
  
# model predictions  
library(ggeffects)  
  
# 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)  
class(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 ...

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

#suppressing messages/warnings  
suppressMessages({  
 suppressWarnings({  
   
# 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  
  
 })  
})

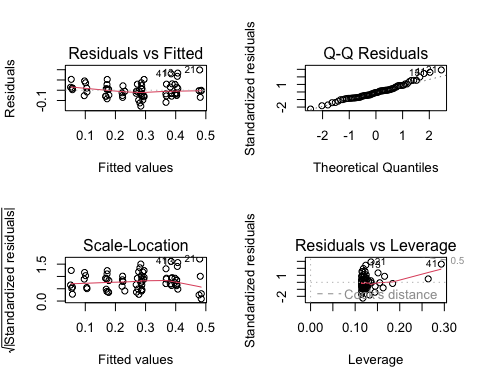
# a. Make a table or list of all the models from class and the last one you constructed on your own. Write a caption for your table

## 0. Null model

model0 <- lm(total\_g ~ 1, # formula  
 data = drought\_exp\_clean) # data frame

## 1. total biomass as a function of SLA, water treatment, and species

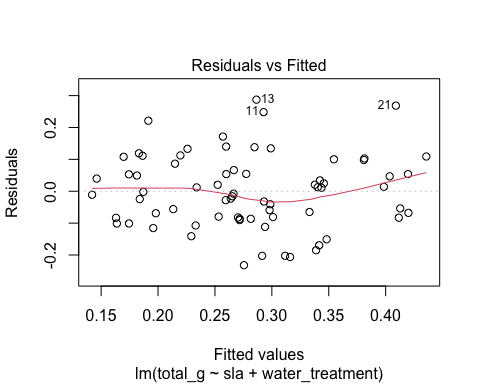
# saturated model  
model1 <- lm(total\_g ~ sla + water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
  
par(mfrow = c(2, 2))  
plot(model1)

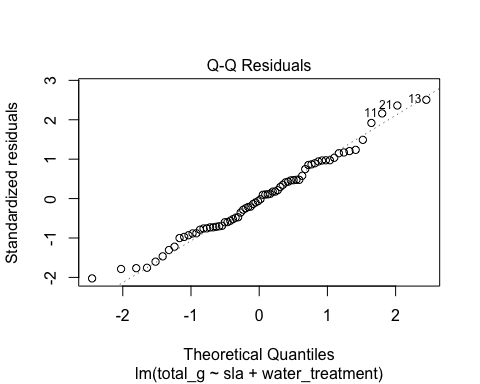


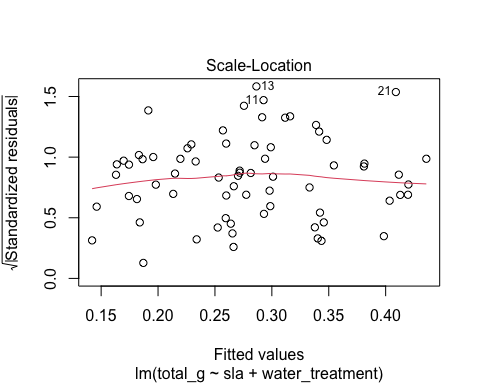
# you might get a warning when you run this code - that is ok!  
#diagnostics for sat model look good

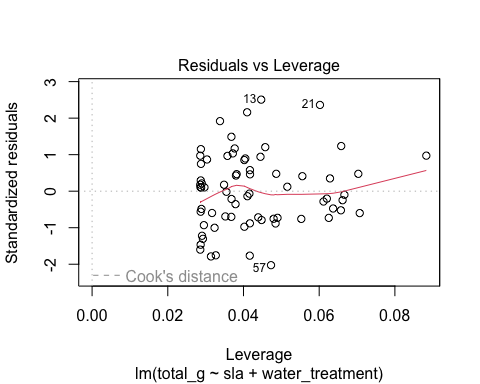
## 2. total biomass as a function of SLA and water treatment

model2 <- lm(total\_g ~ sla + water\_treatment,  
 data = drought\_exp\_clean)  
  
plot(model2)





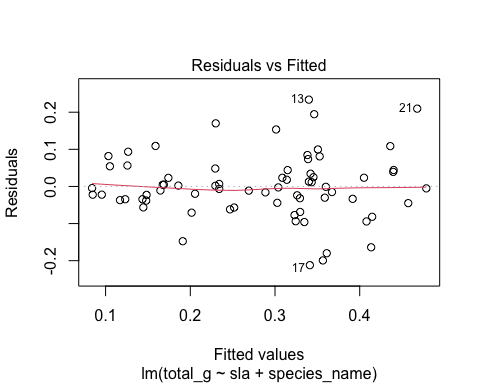


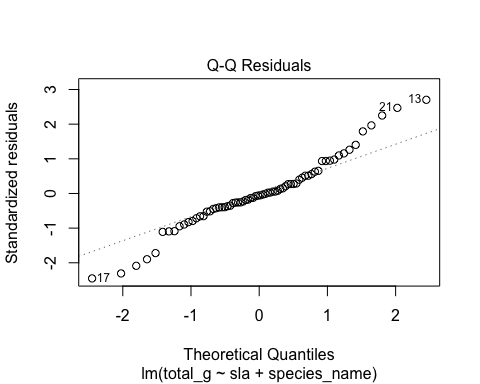


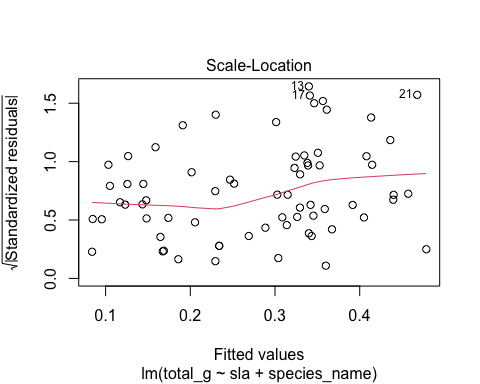
**look at diagnostics later**

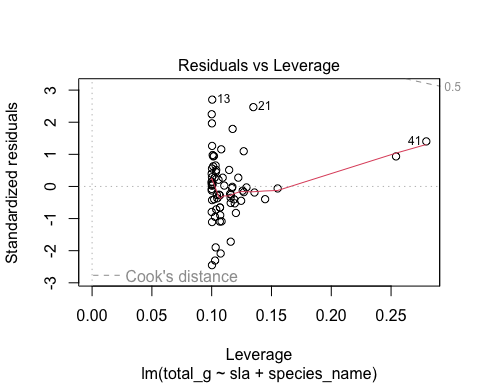
## 3. total biomass as a function of SLA and species

model3 <- lm(total\_g ~ sla + species\_name,  
 data = drought\_exp\_clean)  
  
plot(model3)



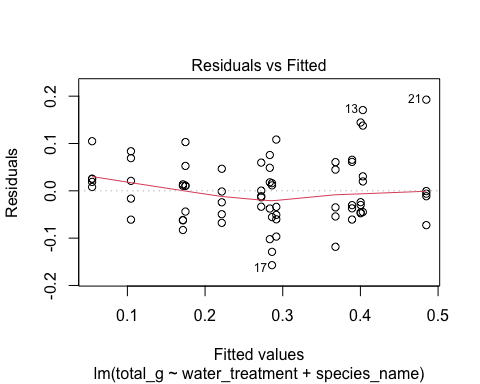


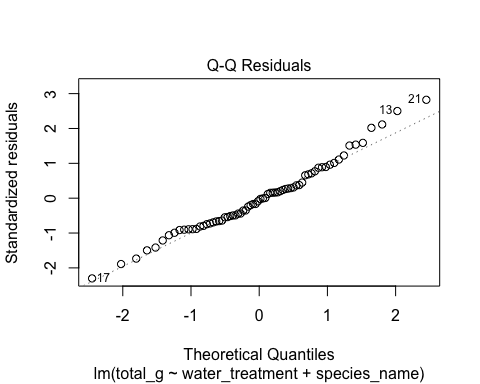


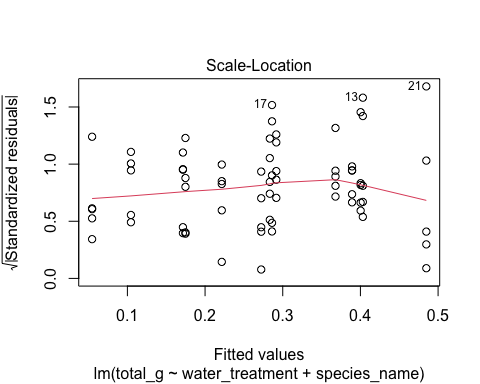


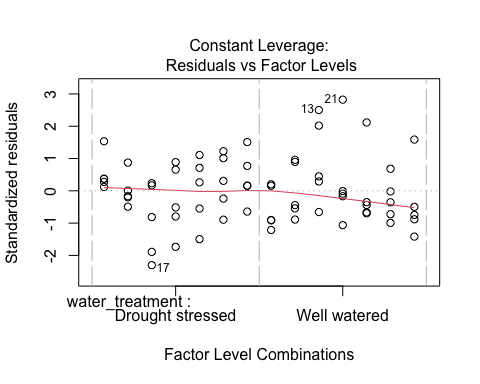
## 4. total biomass as a function of water treatment and species

model4 <- lm(total\_g ~ water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
  
plot(model4)









summary(model4)

Call:  
lm(formula = total\_g ~ water\_treatment + species\_name, data = drought\_exp\_clean)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.157087 -0.046953 -0.003733 0.041244 0.192657   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.05455 0.02451 2.225 0.02973 \*   
water\_treatmentWell watered 0.11695 0.01733 6.746 5.90e-09 \*\*\*  
species\_nameEncelia californica 0.21774 0.03243 6.714 6.70e-09 \*\*\*  
species\_nameEschscholzia californica 0.23164 0.03243 7.143 1.22e-09 \*\*\*  
species\_nameGrindelia camporum 0.31335 0.03243 9.662 5.53e-14 \*\*\*  
species\_nameNasella pulchra 0.22881 0.03243 7.055 1.72e-09 \*\*\*  
species\_namePenstemon centranthifolius 0.05003 0.03243 1.543 0.12799   
species\_nameSalvia leucophylla 0.12020 0.03243 3.706 0.00045 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.07252 on 62 degrees of freedom  
Multiple R-squared: 0.7535, Adjusted R-squared: 0.7257   
F-statistic: 27.08 on 7 and 62 DF, p-value: < 2.2e-16

# Making a table

modelsummary::modelsummary( # this function takes a list of models  
 list(   
 "null" = model0, # "model name" = model object  
 "model 1" = model1,  
 "model 2" = model2,  
 "model 3" = model3,  
 "model 4" = model4))

|  | 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 |

# Caption:

### Table 1: Comparison of Linear Regression Models for Predicting Total Growth

This table shows the coefficients, standard errors, and performance metrics for four linear regression models that predict total growth biomass (total\_g) based on water treatment and species name. Model 4, which includes both predictors, shows the best performance with the highest R² (0.754) and lowest AIC (-159.2), indicating the most reliable fit for the data.

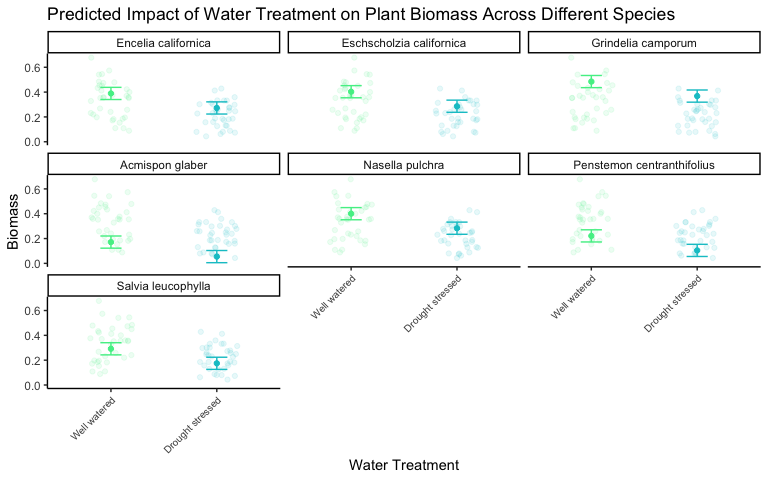
1. Statistical Methods

To examine the influence of specific leaf area (SLA), species, and water treatment on total plant biomass of plants species native to southern California, I conducted a series of linear regression models. I used a null model to establish a baselines, and then four subsequent models comparing different combinations of predictors (SLA, species, water treatment). To determine the model that best described total biomass, I used Akaike Information Criterion (AIC) and R² values. I decided that the model that best described variation in plant biomass is model 4, because it had the highest R² of 0.754 and lowest AIC of -159.2. To evaluate linear model assumptions, I visually inspected diagnostic plots for normality, homoscedasticity, linearity, and outliers.

# c. Visualization of model 4, the best predictor.

suppressMessages({  
 suppressWarnings({  
   
 model4\_preds <- ggpredict(model4, # getting predictions   
 terms = c("water\_treatment", "species\_name"))  
 # plotting predictions + underlying data  
ggplot(model4\_preds,   
 aes(x = x, y = predicted, color = x)) + #plotting predictions   
 geom\_jitter(data = drought\_exp\_clean, aes(x = water\_treatment, y = total\_g, color = water\_treatment), width = 0.2, alpha = 0.1) + # plotting underlying data   
 geom\_line() +   
 geom\_point() +   
 geom\_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.2) + # representing st error   
 theme\_classic() +   
 labs(title = "Predicted Impact of Water Treatment on Plant Biomass Across Different Species",x = "Water Treatment", y = "Biomass") + # adding and changing titles   
 scale\_color\_manual(name = "Water Treatment",   
 values = c("Well watered" = "seagreen2", "Drought stressed" = "turquoise3")) + # editing colors   
 theme(panel.grid = element\_blank(),  
 axis.text.x = element\_text(angle = 45, hjust = 1, size = 7.5), #changing orientation and size of x-axis text   
 legend.position = "none") + # remove legend  
 facet\_wrap(~ group) # facet by species name  
 })  
})

`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?



1. Figure 1. “Predicted Impact of Water Treatment on Plant Biomass Across Different Species” This visualization compares predicted biomass of various plant species under two water treatment conditions: well watered and drought stressed. The error bars represent the confidence intervals of the predicted values while he underlying data is represented through jitter. Data source: Valliere J. M., J. Zhang, M. R. Sharifi, and P. W.Rundel. 2019. Can we condition native plants to increase drought tolerance and improve restoration success? Ecological Applications 29(3): e01863.10.1002/eap.1863
2. Results Section

In conclusion, water treatment condition and species type are the most effective predictors of the total mass of native plants in Southern California, as demonstrated by the linear model with the highest R² (0.754) and lowest AIC (-159.2). On average plants in well watered treatment are 0.11695 ± 0.01733g greater than plants in drought stressed treatment. On average, Scarelet Bulgers (Penstemon centranthifolius) were the smallest and Great Valley Gumweed (Grindelia camporum) were the largest. Purple Needlegrass (Nasella pulchra) were, on average, the same size as California Poppy (Eschscholzia californica) were the same

# Problem 2. Affective visualization

## a. What would an affective visualization look like for my personal data?

An affective visualization of my data would include as many variables as possible. For this, I would probably be visualizing each observation individually and grouping them by day. I want to represent mg, form, pleasure/purpose, and hours of sleep at the very least. I think a drawing similar to the “dear data” project would be most fitting to accomplish this.

## b. Visualization sketch:

## c. Visualization draft:

## d. Write an artist statement.

An artist statement gives the audience context to understand your work. Write 4-5 sentences to address:

the content of your piece (what are you showing?) the influences (what did techniques/artists/etc. did you find influential in creating your work?) the form of your work (written code, watercolor, oil painting, etc.) your process (how did you create your work?)

My piece represents the amount of caffeine I consume per day in milligrams while representing variables such as the amount if sleep I got the night before, the form of caffeine I am consuming, and the reason I consumed the caffeine. I was largely influenced from the “Dear Data” project by Giorgia Lupi and Stefanie Posavec. My work is pen and colored pencil on paper, and I had to do many rough drafts in order to perfect it. I tried out various types of visual representations for each form of caffeine, and the amount of mg it represented, until I found a method that seemed clear.

# Problem 3. Statistical critique

## a. Revisit and summarize

The authors are using a one way analysis of variance (ANOVA) and two-way ANOVA both followed by Tukey’s post hoc test to address the question of better understand the shell calcification responses of P. fucata to ocean acidification. Specifically, they are testing the effects of lowered seawater pH on calcium content and shell hardness.

## b. Visual clarity

These visualizations are moderately clear representations to those who are skimming the article or do not know much about marine animal calcification. Some aspects could have been more defined, such as the lack of titles and the y-axis labeled “IcT”, considering I had never heard of that measuring unit so could not immediately understand the plot at first glance. The plots did show summary measurements including mean and standard error. The plots do not show underlying data.

## c. Aesthetic clarity

The plots are not cluttered at all, I would say that they are almost too bland and uninteresting. The data: ink ratio is good considering they are bar plots.

## d. Recommendations

I recommend doing a scatter plot to represent each variable instead of a box plot. This way, they could include the underlying data because it is a respectable amount of observations to have in a jitterplot, as well as an errorbarr to provide a clearer representation of mean and standard error. Also, I recommend specifying IcT as Growth Global Index of length and weight on the y-axis, and even possibly including the equation (IcT = (Ln At - Ln A0)/t]100) somewhere on the plot, or at least in the caption. Lastly, I recommend including a title for each of their plots, it is confusing to read without one.