hw\_3

Lauren Berry

# Set up

## reading in packages

# Problem 1. Multiple linear regression: model selection and construction (52 points)

## 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. (8 points)

model\_types <- read\_xlsx(path = here("data",   
 "hw3-models.xlsx"))  
models <- flextable(model\_types)  
models <- add\_header\_lines(models, values = "Table 1: Model Predictive Variables  
 which variables each model includes")  
models <- align(models, align = "center", part = "all")  
models <- bold(models, part = "header")  
models

| **Table 1: Model Predictive Variables  which variables each model includes** | | | |
| --- | --- | --- | --- |
| **Model Number** | **surface leaf area** | **water treatment** | **species** |
| 0 |  |  |  |
| 1 | x | x | x |
| 2 | x | x |  |
| 3 | x |  | x |
| 4 |  | x | x |

## b. Write a 5-6 sentence “statistical methods” section. (8 points)

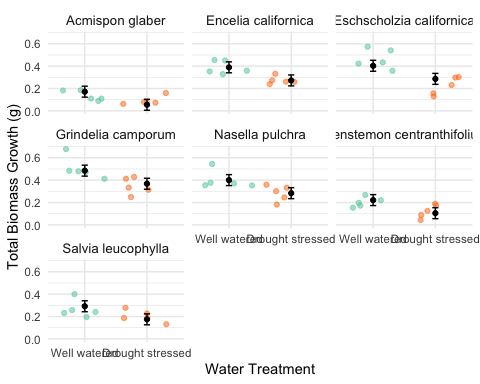
To examine the influence of surface leaf area, water treatment, and species on plant biomass I created a null model and 4 other models to analyze the difference. Model 1 looked at all potential influences, while model 2 only looked at water treatment and surface leaf area. Model 3 looked at surface leaf area and species, and model 4 analyzed water treatment and species. To determine the model that best described biomass, I first visually looked at all of the models and compared them. Then, I ran a model selection test (model.sel), and found that model 4 has the lowest AICc (-156.2).To evaluate linear model assumptions, I visually looked at the 4 plots created by the lm modelling assumption, and confirmed that model 4 meets linear model assumptions, meaning that the variables species and water treatment best describe biomass

## c. Make a visualization of the model predictions with underlying data for your “best” model. (20 points)

Show and annotate all your code. For full credit:

make the underlying data more transparent than the model predictions display species names in full (not as species codes like ENCCAL or ESCCAL) display water treatment types in full (not as WW or DS) represent well-watered and drought stressed treatments with different colors use colors that are not the default ggplot() colors facet your plot by species remove the legend finalize the plot Note Make sure that the only output is the visualization!

library(ggplot2)  
  
#load data  
drought\_exp <- read\_xlsx(path = here("data",   
 "Valliere\_etal\_EcoApps\_Data.xlsx"),  
 sheet = "First Harvest")  
# 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  
  
# create model 4  
model4 <- lm(total\_g ~ water\_treatment + species\_name,  
 data = drought\_exp\_clean)   
  
# add predicted total weight in grams to existing df  
model\_preds <- ggpredict(model4,   
 terms = c("water\_treatment",   
 "species\_name"))  
  
# change column names for future facet wrap so code doesn't get confusing   
colnames(model\_preds)[colnames(model\_preds) == "group"] <- "species\_name"  
  
  
#model\_preds <- model\_preds %>% # choosing data frame to edit  
# mutate(species\_name = group) # renaming group   
  
# plot  
ggplot()+  
 geom\_jitter(data = drought\_exp\_clean, # choose data for jitter poitts  
 aes(x = water\_treatment, # treatment x axis  
 y = total\_g, # plot total biomass  
 color = water\_treatment, # color by water treatmet   
 alpha = 0.3 #opaque background points  
 )) +  
 geom\_point(data = model\_preds, # choose data for avg dots  
 aes(x = x, #choose data  
 y = predicted, #choose data  
 )) +  
 geom\_errorbar(data = model\_preds, # add confidence intervals from predicted df  
 aes(ymin = conf.low,  
 ymax = conf.high,  
 x = x),  
 width = 0.1)+  
 scale\_color\_manual(values = c("Well watered" = "aquamarine3", "Drought stressed" = "chocolate1")) + # Custom colors  
 facet\_wrap(~ species\_name) + # Facet by species  
 theme\_minimal() + # simple theme  
 theme(legend.position = "none") + # Remove legend  
 labs(x = "Water Treatment", y = "Total Biomass Growth (g)") +  
 theme(strip.text = element\_text(size = 10)) # Adjust facet labels



## d. Write a caption for your visualization. (6 points)

Figure 1: Model Predictions by Species and Water Treatment. Data from Valliere, Justin; Zhang, Jacqueline; Sharifi, M.; Rundel, Philip (2019). Data from: Can we condition native plants to increase drought tolerance and improve restoration success? [Dataset]. Dryad. https://doi.org/10.5061/dryad.v0861f7 Points represent biomass (g), of different plant species on separate graphs. Colors represent the water treatment (drought stressed = orange, well watered = blue) these plants experienced.

## e. Write a 3-4 sentence results section. (10 points)

The best predictors for described total mass were found to be both water treatment and species (F(7,62) = 27.08, p < 0.001, alpha = 0.05, adjusted R^2 = 0.7257). Well watered plants tended to have a greater biomass than drought stressed plants. Species such as Salvia leucophylla, Acmispon glaber, and Penstemon centranthifolius tend to have less biomass, while Nasella pulchra, Encelia californica, and Eschscholzia californica have far more biomass growth, and Grindelia camporum has the most biomass growth out of all species.

# Problem 2. Affective visualization (24 points)

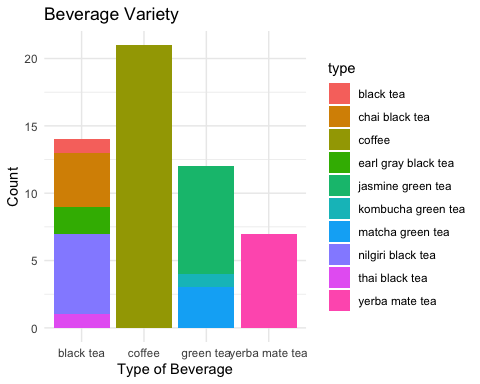
## a. Describe in words what an affective visualization could look like for your personal data (3-5 sentences). (2 points)

An affective visualization for my personal data on caffeine and different types of beverage consumption might feature a dynamic, interactive dashboard. This could use a color-coded timeline, with warm colors like red and orange representing high caffeine intake and cool colors like blue and green indicating lower or no caffeine levels. Instead of a timeline it could be the time of day each is common. Affective visualization could also look like cups of drinks in different sizes based on how often i drink them.

## b.

## c.

# Create graph for drawing reference  
# Load excel  
beverages <- read\_xlsx(path = here("data", "caffeinatedbeverages.xlsx"))  
  
# clean df   
beverages\_clean <- beverages %>%   
 mutate(broad\_type = case\_when(  
 grepl('green tea', type) ~ 'green tea', # change various green teas to just say green tea  
 grepl('black tea', type) ~ 'black tea', # change various black teas to just say black tea  
 TRUE ~ type  
 )) %>%   
 mutate(coffee\_or\_tea = case\_when( # create new column tea or coffee  
 grepl('tea', type) ~ "tea",  
 TRUE ~ type  
 ))  
  
# create graph of amounts for drawing reference  
ggplot(beverages\_clean,   
 aes(x = broad\_type,  
 fill = type)) +  
 geom\_bar() +  
 labs(title = "Beverage Variety",  
 x = "Type of Beverage",  
 y = "Count") +  
 theme\_minimal()



## d.

I am showing the amount of different caffeinated beverages I have consumed throughout the past month and a half. The most caffeinated beverage begins on the left and the amount of caffeine decreases to the right, and the rainbow background is in reference to the quantity getting higher as the drinks are taller. I was influenced by my oma’s watercolor paintings, and decided to use watercolor because of this. It was created by first creating a graph with beverage count, sketching a count onto the paper, and then painting.

# Problem 3. Statistical critique (36 points)

## a. Revisit and summarize (6 points)

What are the statistical tests the authors are using to address their main research question?

Multivariate analysis methods such as FAMD and HCPC were used to categorize farms. They also used the Kruskal Wallis test to determine the significance of variabilityand Wilcoxon-Mann-Whitney test.

## b. Visual clarity (10 points)

How clearly did the authors visually represent their statistics in figures? For example, are the x- and y-axes in a logical position? Do they show summary statistics (means and SE, for example) and/or model predictions, and if so, do they show the underlying data?

The authors visually represented their statistics in figures fairly well, but the underlying data wasn’t clear. There are black dots present in some graphs, but it’s unclear what these represent. The legend in the caption was also slightly confusing because it leaves out certain elements of the graph such as the black dots.

## c. Aesthetic clarity (10 points)

How well did the authors handle “visual clutter”? How would you describe the the data:ink ratio?

I think the authors handled visual clutter very weak because the amount of ink is minimal compared to the amount of data. However, in cutting out some clutter the graph lacks some in-depth information and perspective.

## d. Recommendations (can be longer than 4 sentences, 10 points)

I believe the figure needs either greater explanation or more data. If the black dots represent outliers then this needs to be addressed in the caption. The significance thresholds using different numbers of “\*\*” are also confusing, and I believe would be better represented by listing the actual significance numbers.