homework03JC

Jayne Campbell

link: https://github.com/jayne-campbell/Campbell-Jayne\_homework-03.git

Homework 3

Loading in packages and data

# general use  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(readxl)  
library(here)

here() starts at /Users/jayne/Desktop/ENVS-193DS/Campbell-Jayne\_homework-03

library(janitor)

Attaching package: 'janitor'  
  
The following objects are masked from 'package:stats':  
  
 chisq.test, fisher.test

# visualizing pairs  
library(GGally)

Registered S3 method overwritten by 'GGally':  
 method from   
 +.gg ggplot2

# model selection  
library(MuMIn)  
  
# model predictions  
library(ggeffects)  
  
# model tables  
library(gtsummary)  
library(flextable)

Attaching package: 'flextable'  
  
The following objects are masked from 'package:gtsummary':  
  
 as\_flextable, continuous\_summary  
  
The following object is masked from 'package:purrr':  
  
 compose

library(modelsummary)

`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing  
 backend. Learn more at: https://vincentarelbundock.github.io/tinytable/  
  
Revert to `kableExtra` for one session:  
  
 options(modelsummary\_factory\_default = 'kableExtra')  
 options(modelsummary\_factory\_latex = 'kableExtra')  
 options(modelsummary\_factory\_html = 'kableExtra')  
  
Silence this message forever:  
  
 config\_modelsummary(startup\_message = FALSE)

drought\_exp <- read\_xlsx(path = here("data", # reading in 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"

cleaning the data

# 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

Problem 1. Multiple linear regression: model selection and construction

1. 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.

models <- data.frame( # creating a data frame for the table  
 Model\_numbers = c("Model 0", "Model 1", "Model 2", "Model 3", "Model 4"),  
 Model = c("null model", "saturated model", "two predictors", "two predictors", "two predictors"),  
 Predictors = c("none", "SLA, water treatment, and species", "SLA and water treatment", "SLA and species", "water treatment and species"),  
 AIC = c("-75.0", "-153.8", "-95.8", "-124.1", "-156.2"),  
 delta\_AIC = c("81.22", "2.44", "60.37", "32.12", "0.00")# creating rows and columns  
)  
  
modeldatatable <- flextable(models) %>% # create flextable  
set\_header\_labels(Model\_numbers = "Model numbers", Model = "Model", Predictors = "Predictors", AIC = "AIC", delta\_AIC = "delta AIC") %>% # rename headers   
 align(align = "center", part = "all") %>% # formatting   
 autofit() %>%  
 theme\_vanilla() %>%  
 bold(part = "header") #bold everything in the header

Table 1. Five models that predict the total biomass of plant species. Each row represents one of the models. The columns represent the model number, the model, and the predictors for that model. SLA stands for the specific leaf area, which is the surface area of a leaf divided by its dry mass and it represents life strategy (mm2/g).

print(modeldatatable)

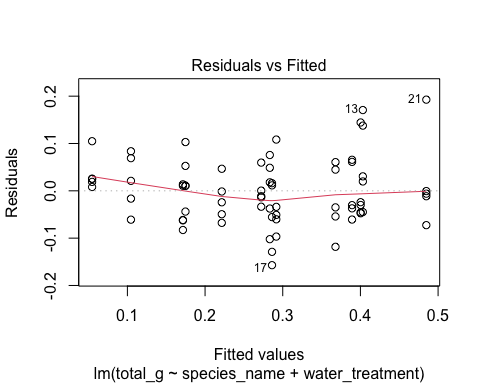
a flextable object.  
col\_keys: `Model\_numbers`, `Model`, `Predictors`, `AIC`, `delta\_AIC`   
header has 1 row(s)   
body has 5 row(s)   
original dataset sample:   
 Model\_numbers Model Predictors AIC  
1 Model 0 null model none -75.0  
2 Model 1 saturated model SLA, water treatment, and species -153.8  
3 Model 2 two predictors SLA and water treatment -95.8  
4 Model 3 two predictors SLA and species -124.1  
5 Model 4 two predictors water treatment and species -156.2  
 delta\_AIC  
1 81.22  
2 2.44  
3 60.37  
4 32.12  
5 0.00

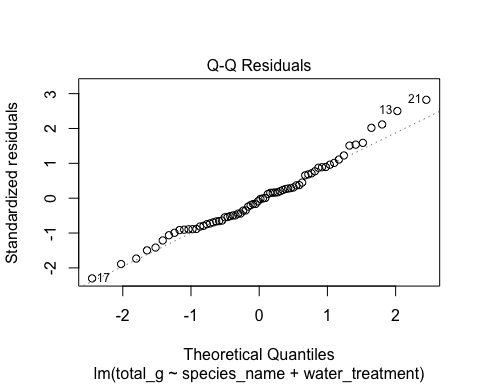
1. Write a 5-6 sentence “statistical methods” section

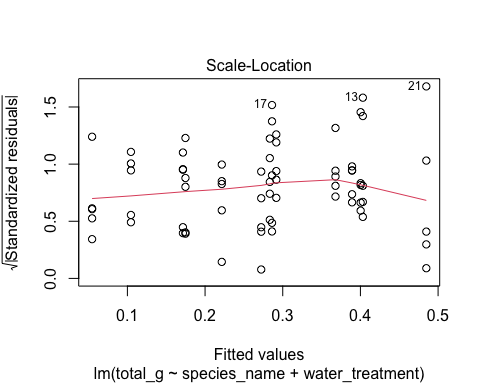
In this scenario, we’re interested in how three variables affect total plant biomass: species (categorical), water treatment (categorical), and specific leaf area (SLA, numeric). In order to examine the influence of these predictors on plant biomass, I created five models based on the three potential predictors. These included a null model, a saturated model, and three models with two predictors each. To determine which linear model best predicted the biomass, I used the Akaike Information Criterion (AIC), which takes into account how well the model predicts the data as well as increasing complexity. Out of the five models we tested, the best model was the one that included water treatment and species because it produced the lowest AIC. I visually determined that my final model conformed to the assumptions of a linear model by looking at the diagnostics and ensuring there was random and even dispersion, homoscedasticity, normality of variance, and no excessive outliers.

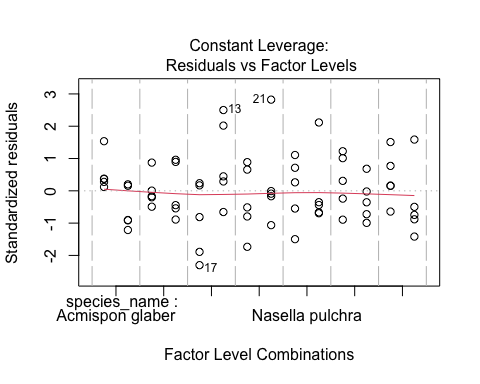
1. Make a visualization of the model predictions with underlying data for you “best” model

# model of total biomass as a function of water treatment and species  
model4 <- lm(total\_g ~ species\_name + water\_treatment, # formula for model  
 data = drought\_exp\_clean) # data frame  
  
plot(model4) # printing plots for model 4









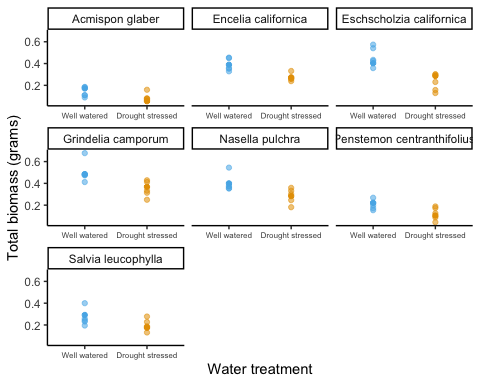
summary(model4) # summary of this model

Call:  
lm(formula = total\_g ~ species\_name + water\_treatment, 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 \*   
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 \*\*\*  
water\_treatmentWell watered 0.11695 0.01733 6.746 5.90e-09 \*\*\*  
---  
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

model\_preds <- ggpredict(model4, # create model predictions   
 terms = c("water\_treatment",   
 "species\_name"))  
  
View(model\_preds) # View the predictions as a data frame

Creating visualization for best model (predictors = water treatment and species)

# creating new data frame of model predictions for plotting  
model\_preds\_for\_plotting <- model\_preds %>%   
 rename( # renaming columns  
 water\_treatment = x,  
 species\_name = group)  
ggplot() +  
 # underlying data  
 geom\_point(data = drought\_exp\_clean,  
 aes(x = water\_treatment,  
 y = total\_g,   
 color = water\_treatment,  
 alpha = 0.1)) + # making underlying data more transparent  
 geom\_point(data = model\_preds\_for\_plotting, # making predictions opaque   
 aes(x = water\_treatment, y = predicted, color = water\_treatment), alpha = 1) +  
 theme\_classic() + # changing the theme  
 # creating different panels for species  
 facet\_wrap(~species\_name, scales = "free\_x") +  
 labs( # changing axis names  
 x = "Water treatment",  
 y = "Total biomass (grams)"  
 ) +  
 scale\_color\_manual(values = c("Drought stressed" = "#E69F00", "Well watered" = "#56B4E9")) + # changing the colors   
 theme(  
 axis.text.x = element\_text(size = 6), # making the text smaller so they don't overlap  
 legend.position = "none") # getting rid of the legend



1. Caption for your visualization (include a data citation)

Figure 1. Total biomass of plant species as a function of water treatment. The y-axis represents total plant biomass in grams and the x-axis represents the water treatment received; either well watered or drought-stressed. Each of the seven figures represents a specific plant species. The transparent points consist of the underlying data and the opaque points represent the model predictions.

1. Results section

The predictors, water treatment and species, best described the total mass based on model statistics… On average, the biomass was lower for drought stressed plants than for well watered plants, which aligns with biological knowledge. On average, plant species with larger ranges tended to have larger biomasses.

Problem 2. Affective visualization

1. Describe in words what an affective visualization could look like for your personal data

Since affective visualizations prioritize creativity and should convey emotion, I would like to draw images of the types of caffeinated beverages I have been drinking. I could do different shapes and colors for each beverage and make the sizes dependent on the amount of caffeine in each. I could also organize them by time of day by having a symbol for morning, afternoon, and night and clump the other symbols around them. I would create a key for each of the symbols and describe what they represent.

1. Create a sketch on paper of your idea
2. Make a draft of your visualization Painting
3. Write an artist statement

This visualization shows my consumption of caffeinated beverages in terms of the type, approximate caffeine content, and the time of day they were consumed at. I was very influenced by the Dear Data project done by Giorgia Lupi and Stefanie Posavec and their use of symbols and keys to convey information and messages. My work is in the form of acrylic paint and ink pen on a 8 by 12 inch canvas. I created my work by first creating an outline of where I wanted everything and using a ruler and pencil to map it out. Then I began to use acrylic paint and a pen to fill in and draw the symbols and write the key.

Problem 3. statistical critique

1. Revisit and summarize

The authors used a two-way ANOVA test to analyze variance in order to investigate the variation in both density and biomass of fish among different locations and sampling dates. This was done to evaluate whether or not and how a small marine reserve impacts the structure of temperature reef fish assemblages near the isthus at Santa Catalina Island in California. The authors visualized the results of the ANOVA test by creating species-specific figures with bar graphs that showed the mean density of the species.

\*insert figures

1. Visual clarity

I think that the authors did an effective job of clearly visually representing their statistics in figures. Firstly, the x and y axis are in logical positions because mean density is on the y axis and the x axis is used for the categories; total, adult, and juvenile, which each have a paired bar graphs. The bar graphs do show summary statistics as they represent the mean densities and have error bars that represent the standard error, but underlying data is not included. The figure captions describe how the ANOVA test information corresponds in terms of a lack of or significant differences between reserve and not reserve fish.

1. Aesthetic clarity

I think that authors did a great job in minimizing visual clutter as the figures are straightforward and easy to digest and understand when paired with the captions. The data to ink ratio is thus close to one because eliminating anything would hinder the effectiveness f its communication of data. Every aspect included is necessary to interpret its meaning and understand how the mean density is connected to fish being in or not in a reserve.

1. Recommendations

While I like the images of the fish, I think the size of the adult fish in the California sheephead figure is distracting and potentially misleading. Also, the images are confusing because there are three images in the first figure to represent what I am assuming is an adult male, adult female, and a juvenile, while the second figure only has one image. If the authors cannot find all the three images for both species, I would recommend taking out the pictures. I would also advise adding in the underlying data since the bar graphs themselves are only representing one value; the mean density. It would also be beneficial for the bar graphs to be colored and slightly transparent because in the black bars, you cannot see where the standard error line goes.