Homework3

Yicong Zeng

Set up

```r
library(tidyverse)
library(readxl)
library(here)
library(janitor)
library(GGally)
library(MuMIn)
library(ggeffects)
library(gtsummary)
library(flextable)
library(modelsummary)
library(gt)
drought_exp <- read_xlsx(path = here("data",
                                     "Valliere_etal_EcoApps_Data.xlsx"),
                         sheet = "First Harvest")
```

Problem 1: Multiple linear regression: model selection and construction

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.

Cleaning

```r
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
```

Table

“Table 1. List of five Models. Each column in the table represent a different linear model that used to predict the total biomass. The row of the table include different combinations of predictors: specific leaf area, water treatment, and species name. Also, the second part of table shows the statistics of each models, include R2, R2 Adj., AIC, delta AIC, etc.”

```
model0 <- lm(total_g ~ 1, # formula
             data = drought_exp_clean) # data frame

model1 <- lm(total_g ~ sla + water_treatment + species_name, # formula
             data = drought_exp_clean) # data frame

model2 <- lm(total_g ~ sla + water_treatment, # formula
             data = drought_exp_clean) # data frame

model3 <- lm(total_g ~ sla + species_name, # formula
             data = drought_exp_clean) # data frame

model4 <- lm(total_g ~ water_treatment + species_name, # formula
             data = drought_exp_clean) # data frame

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
  )
)
```

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

“To examine the influence of specific leaf area(SLA), water treatment, and species on total biomass, I construct 5 different models. Specifically, the null model is not predicted by any of these variables, and saturated model utilize total biomass as a function of SLA, water treatment, and species. Other models include some combination of other predictors. To determine the model that best described total biomass, I create the model selection table to select the model with lowest AIC value. To evaluate linear model assumptions, I will look at the diagnostics for the model with the lowest AIC, including residual vs fitted, QQ residuals, scale-location, and residuals vs leverage.”

c. Make a visualization of the model predictions with underlying data for your “best” model.

model selection

```
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)
```

“The best model is model4 with lowest AICc”

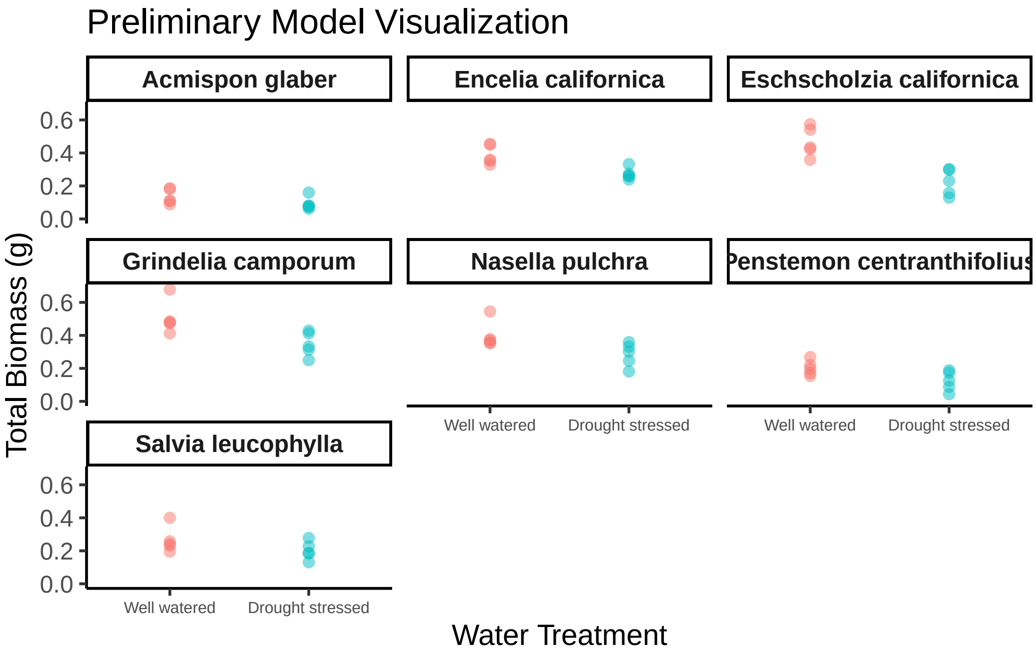
	null	model 1	model 2	model 3	model 4
(Intercept)	0.279	0.080	0.047	−0.033	0.055
	(0.017)	(0.056)	(0.054)	(0.067)	(0.025)
sla		0.000	0.001	0.001	
		(0.000)	(0.000)	(0.001)	
water_treatmentWell watered		0.122	0.090		0.117
		(0.020)	(0.029)		(0.017)
species_nameEncelia californica		0.238		0.115	0.218
		(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

model predictions

```r
model_preds <- ggpredict(model4,
                         terms = c("water_treatment",
                                   "species_name"))

model_preds_for_plotting <- model_preds %>%
  rename(water_treatment = x, # renaming columns to make this easier to use
         species_name = group)

ggplot() +
  # underlying data
  geom_point(data = drought_exp_clean,
             aes(x = water_treatment,
                 y = total_g,
                 color = water_treatment),
             alpha = 0.5, size = 1.5) +
  # model prediction 95% CI ribbon
  geom_ribbon(data = model_preds_for_plotting,
              aes(x = water_treatment,
                  y = predicted,
                  ymin = conf.low,
                  ymax = conf.high,
                  fill = water_treatment)) +
  # model prediction lines
  geom_line(data = model_preds_for_plotting,
            aes(x = water_treatment,
                y = predicted,
                color = water_treatment)) +
  # cleaner theme
  theme_classic() +
  labs(title = "Preliminary Model Visualization",
      x = "Water Treatment",
      y = "Total Biomass (g)") +
  theme(panel.grid = element_blank(),
        strip.text = element_text(size = 9, face = "bold"), # Facet labels
        axis.text.x = element_text(size = 6), # Rotate x-axis text
        legend.position = "none") +# getting rid of gridlines
  # creating different panels for species
  facet_wrap(~species_name)
```



d. Write a caption for your visualization.

"Figure 1: Model visualization of Total Biomass Predicted by Water Treatment and Species. The figure shows the predicted total biomass under well watered and drought stressed, and each facet represent different species. The rad and green dot indicate the well water and drought stress respectively. And The darkest dot in the plot represent the predicted value. Data source: 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"

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

```
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
```

"The best predictors of total mass are water treatment and species presented in Model 4 (df = 62, F = 27.08, p < 2.2e-16, alpha = 0.05, R2 = 0.754). On average, the well watered treatment typically will produce higher total biomass than drought stressed treatment. Furthermore, the difference species will also result in different total biomass. For instance, the estimate of Grindelia camporum (0.31335 g) shows a higher total biomass than that of Penstemon centranthifolius (0.05003 g).

Problem 2

In this problem, you will create an affective visualization using your personal data in preparation for workshop during week 10.

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

"In affactive visualization of my personal data, I want to include a line chart about date VS. exercise duration to recording my sport planning. I will also add some gym elements, such as dumbbell and weight bench, to display aesthetics. Finally, fitting the line chart to the background of the figure to make my own affective visualization."

b. Create a sketch (on paper) of your idea.

The sketch and draft of my idea will shows in the end of pdf.

c. Make a draft of your visualization.

The sketch and draft of my idea will shows in the end of pdf.

d. Write an artist statement.

"The work I create is a watercolor painting. It tried to display the relationship between date and excerise duration in my personal data, but also fitting the background element, such as dumbbell, human, and weight bench. In terms of my creation process, I first plot the date vs excerise duration in the coordinate axis. and then fitting the line chart with the weight bench to make it looks natural. However, because of my bed painting skill, when I paint the weight bench, the line chart looks not perfectly integrate with the gym element, making it quite strange."

Problem 3

a. Revise and summarize

"The authors trying to use the analysis of variance to explore how the two different shrimp (Atya lanipes and Xiphocaris elongata) affect the rate of size fractionation of leaf material, the localized nutrient concentrations in pools, and the rate of particulate export from the pools."

"The figure will show in the last page"

b. Visual clarity

"The table provide some summarized statistical outcomes, such as degree of freedom, F-value, and P-value. The F and P balue are explicitly provided, allowing uss to assess the statistical significance and predict the effects."

c. Aesthetic clarity

"The authors did a great job on addressing the"visual clutter", because the table is basically simple and clean (Each variable corresponds to its own statistical outcome), using the only necessary information and row-column relationship to avoid visual confusion. Furthermore, almost all of the ink used in table is presented in variable, effect and statistical outcomes without any redundant sections, so it looks efficnecy in ink ratio.

d. Recommendations

“1. It is better to use the different color of line styles for different treatment, because a single color make effect column a little bit compact.”

“2. I suggest to create a additional legend bar to explain the indicator of lead particle and nutrient chemistry, such as FPOM, MPOM, and CPOM”, which can make it much detailed.

“3. Add an additional line above the”Nutrient chemistry”, distinguishing the ANOVA result of two part.