ENV S 193DS Homework 3

Aidan Robertson

2024-05-31

Forked repository

Set-up

reading in data & packages

```
# general use
library(tidyverse)
library(readxl)
library(here)
library(janitor)

# visualizing pairs
library(GGally)

# model selection
library(MuMIn)

# model predictions
library(ggeffects)

# model tables
library(gtsummary)
library(flextable)
library(modelsummary)
```

loading in models

```r
# loading in data
drought_exp <- read_xlsx(path = here("data",
                        "Valliere_etal_EcoApps_Data.xlsx"),
                         sheet = "First Harvest")
# cleaning up data
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

# making models!
# model 0 - null
model0 <- lm(total_g ~ 1, # formula
             data = drought_exp_clean) # data frame

# model 1 - saturated
model1 <- lm(total_g ~ sla + water_treatment + species_name,
             data = drought_exp_clean)

# model 2 - sla-water treatment
model2 <- lm(total_g ~ sla + water_treatment,
             data = drought_exp_clean)

# model 3 - sla-species
model3 <- lm(total_g ~ sla + species_name,
             data = drought_exp_clean)
```

```
# model 4 - water treatment-species
model4 <- lm(total_g ~ water_treatment + species_name,
             data = drought_exp_clean)
```

creating dataframes for plotting

```
# model 4 predictors
model_preds <- ggpredict(model4,
                         terms = c("water_treatment",
                                   "species_name"))

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

Problem 1. Multiple linear regression: model selection & construction

a. table of models

```
modelsummary::modelsummary( # list of all models (null, saturated, sla/water treatment, sla/
  list(
    "null" = model0,
    "model 1" = model1,
    "model 2" = model2,
    "model 3" = model3,
    "model 4" = model4),
  output = "flextable", # display as flextable
  gof_omit = "^(?!.*IC)") # keep only AIC and BIC statistics
```

	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	

	null	model 1	model 2	model 3	model 4
		(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)
AIC	-75.2	-157.5	-96.4	-127.1	-159.2
BIC	-70.7	-135.0	-87.4	-106.8	-139.0

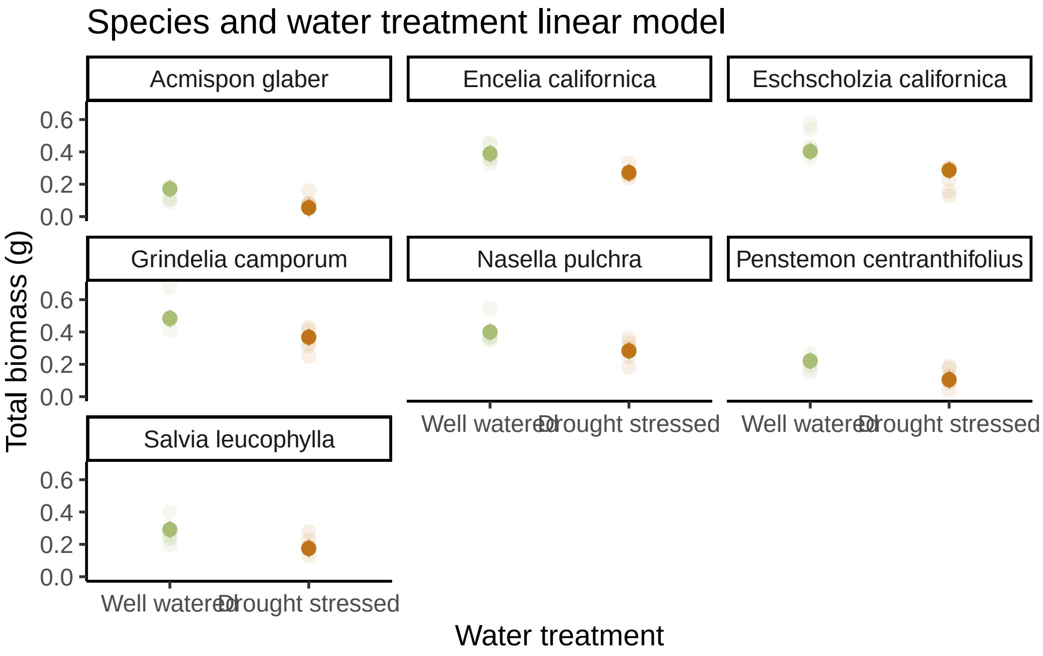
Table 1. Comparing models predicting total biomass of plants. This table depicts the predictor values for 5 different models with different combinations of the potential predictor values of specific leaf area, species, and water treatment.

b. statistical methods

In exploring how to best predict total biomass of plants, I created 5 linear models with different combinations of potential influential variables - specific leaf area (sla), water treatment, and species. To determine the model that best described total plant biomass, I first created models for each combination of variables: none (null model), all (saturated, model 1), sla-water treatment (model 2), sla-species (model 3), and species-water treatment (model 4). Then, I compared the models using the Akaike Information Criterion (AIC) to determine which had the closest fit to the data. The model with the lowest reported AIC was one that used species and water treatment to predict plant biomass. To make sure this model conformed to linear model assumptions, I looked at it's diagnostic plots which confirmed that this data's residuals are homoscedastic and relatively normal, without any major outliers. Because this model meets linear model assumptions and has the lowest AIC in comparison to other models, I determined that species and water treatment are the best variables in a model prediction of total biomass of plants.

c. best model visualization

```r
ggplot() +
  # model prediction
  geom_point(
    data = model_preds_for_plotting,
    aes(
      x = water_treatment,
      y = predicted,
      group = species_name,
      color = water_treatment),
    size = 2) +
  # 95% confidence interval
  geom_errorbar(
    data = model_preds_for_plotting,
    aes(
      x = water_treatment,
      y = predicted,
      ymin = conf.low,
      ymax = conf.high,
      color = water_treatment),
    width = 0) +
  # underlying data
  geom_point(
    data = drought_exp_clean,
    aes(x = water_treatment, y = total_g, color = water_treatment),
    alpha = 0.1,
    size = 2) +
  # creating different panels for species
  facet_wrap( ~ species_name) +
  # cleaner theme & finalizing
  theme_classic() +
  scale_color_manual(values = c("#A8BE74FF", "#BF7417FF")) +
  labs(title = "Species and water treatment linear model", # title of plot
       x = "Water treatment", # x-axis name
       y = "Total biomass (g)") + # y-axis name
  theme(legend.position = "none") # no legend
```



d. caption

Figure 1. Linear model to predict total plant biomass uses water treatment and species. Data from: 'Can we condition native plants to increase drought tolerance and improve restoration success?' (Dryad, Valliere J, et al., 2019). This figure uses a linear model to predict the total plant biomass (g) for each plant species based on water treatment. Each graph depicts the predicted biomass, its corresponding 95% confidence interval, and underlying data for its titled species. Color distinguishes different water treatments (green: well watered, orange: drought stressed).

e. results

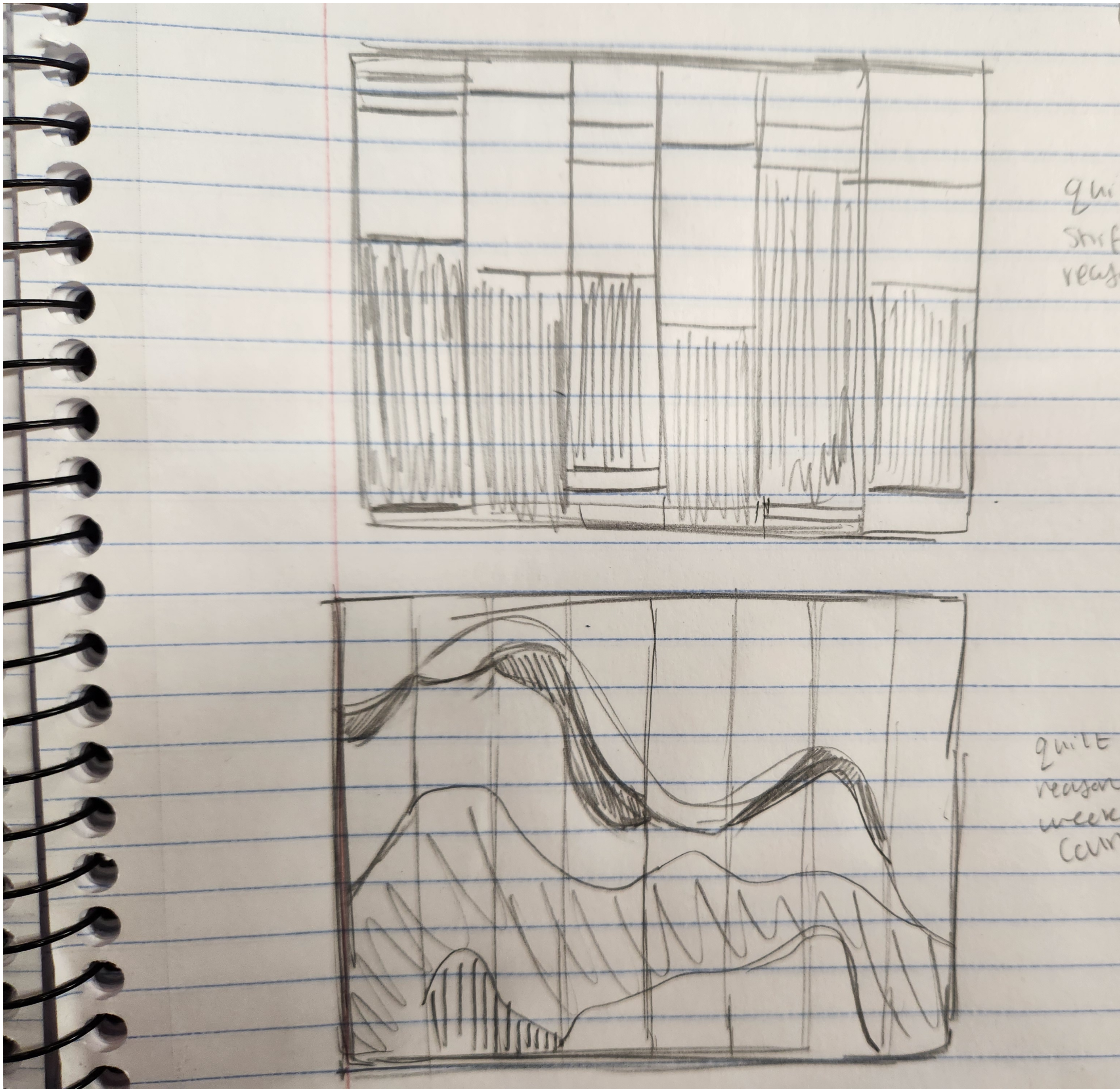
I found that species and water treatment best describe the total plant biomass using a multiple linear regression model (F(7, 62) = 27.08, $p < 0.001$, = 0.05, adjusted $R^2 = 0.726$). In this model, total plant biomass = 0.055 + 0.117(water treatment) + 0.218(encelia californica) + 0.232(eschscholzia californica) + 0.313(grindelia camporum) + 0.229(nasella pulchra) + 0.050(penstemon centranthifolius) + 0.120(salvia leucophylla). On average, if species stays the same, well watered plants have a plant biomass that is 0.117 ± 0.017 grams greater in comparison to drought stressed plants. If water treatment stays constant, then compared to acmispon glaber: total plant biomass is 0.218 ± 0.032 grams greater for encelia californica, 0.232 ± 0.032 grams greater for eschscholzia californica, 0.313 ± 0.032 grams greater for grindelia camporum, 0.229 ± 0.032 grams greater for nasella pulchra, 0.050 ± 0.032 grams greater for penstemon centranthifolius, and 0.120 ± 0.032 grams greater for salvia leucophylla.

Problem 2. Affective visualization

a. description of an affective visualization

An affective visualization for my data could display my observations in the form of a quilt in some fashion. Quilts are a motif found throughout the SRB, both in decoration and design, so incorporating a quilt would be able to reflect those ideas. Since I don't have quilting experience, I could either draw this idea or create a quilt mock-up of what it could potentially look like.

b. sketch of idea



c. draft of visualization

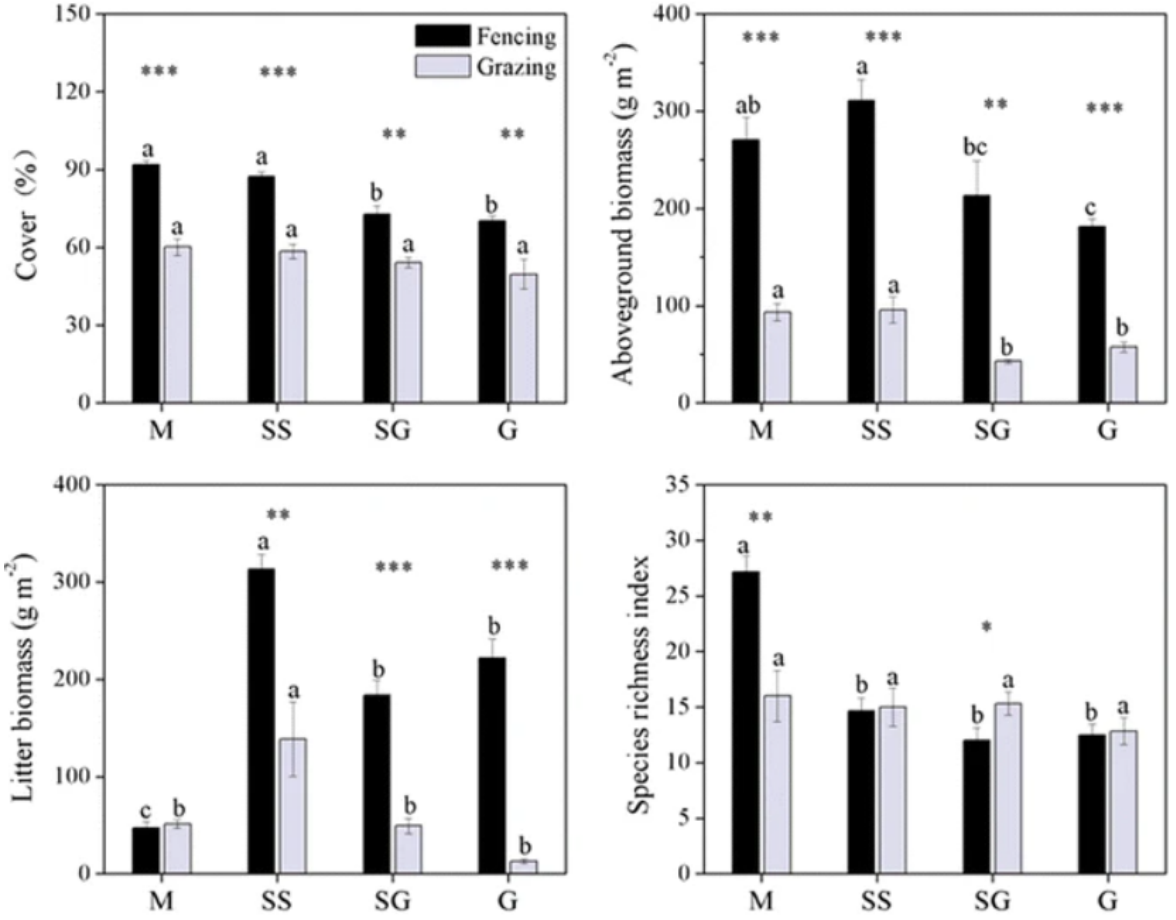
d. artist statement

This is a mock-up of a quilt depicting the reasons people visited the Student Resource Building during times I was working throughout the quarter, acting as a timeline of students' needs. It is modified percentage bar chart, with each square on the bottom of the quilt representing a week, and each color representing a different reason for approaching the desk. This was inspired by the creation of the SRB, which was built around the idea of "quilting together" different communities and cultures on campus. This is a motif that is present in the construction of the building, as well as the decorations, as there are many quilts on the walls of the building which I was inspired by. To create my work, I used a quilt mock-up website, where I followed the general form of a bar chart that I created in R, connecting "weeks" to one another to create a somewhat seamless quilt.

Problem 3. Statistical critique

a. revisit and summarize

The authors conducted many statistical analyses, including finding the mean and standard error for different plots and treatments, conducting two- and three-way ANOVAs, least significant differences, and a linear mixed model. The central question of this paper analyzed the impact of fencing compared to grazing in four grassland types in China. In Homework 2, I described a figure depicting these variables' impacts on percent cover, aboveground biomass, litter biomass, and species richness, using mean and standard error.



b. visual clarity

For the most part, this figure is visually clear, as it shows a clear comparison between the grazing-fencing treatment in each grassland type, and shows which grasslands these treatments were most effective in. The means and standard error are depicted in the figure, and the scale of each y-axis shows the difference between treatments. The underlying data is not present alongside the summary statistics.

c. aesthetic clarity

This figure was initially a little difficult for me to understand, as there is a lot of visual clutter, like with the asterisks and letters above each bar. The letters are supposed to represent statistically significant differences between treatments, but this attribute isn't very clearly explained. The separation between each figure, grassland type, and treatment within this figure are very clear, and easy to understand, which helps with the aesthetics of the figure.

d. recommendations

I would first recommend that the authors remove the letters from the tops of each bar. Removing this visual clutter will improve the data-ink ratio, and create a more seamless and easily understood figure. I think that noting significant differences can be done in a later analysis, or by keeping the asterisks. If the authors are willing to stray from the bar chart, I would recommend that they plot their data using a dot-whisker plot to show the mean and standard error, and include the underlying data. This will allow for data transparency, and show the trends in data alongside the calculated statistics.