Homework 3

Rachel Brody

2024-06-02

# ENVS 193 DS Homework 3

## Set Up

#read 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 3.5.0 ✔ 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/rachel/Desktop/ENVS193DS/ENVS193DS\_workshop-08

library(janitor)

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

library(ggsci)  
library(NatParksPalettes)  
  
  
# 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",   
 "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"

Clean 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

## Problems

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

Use the information from the homework-starter-doc.qmd to do this problem.

#### 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)

**Caption:** table captions typically go above the table. Number the table and provide a title. Describe what is in the table (columns and rows).

**Table:** In your table, each row should be a model with the model number (1, 2, 3, etc.) and the predictors for each model.

|  |
| --- |
| Double check your work! |
| There should be 5 models total (null model + 4 models with predictors). |

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

Your answer should be in paragraph form and include:

* how you addressed the central question(s) (i.e. to examine the influence of \_\_\_\_, \_\_\_\_, and \_\_\_\_ on \_\_\_\_\_, I…)
* how you chose the final model (i.e. to determine the model that best described \_\_\_\_, I…)
* how you visually determined that your final model conformed to the assumptions of a linear model (i.e. to evaluate linear model assumptions, I…)

**To examine the influence of specific leaf area, species, and water treatment on the total biomass of plants, I constructed five different liner models in R to predict total biomass based on a different combination of three predictors, specific leaf area, species, and water treatment. The structure of each model is shown in Table 1. Then, four diagnostic plots were generated for each model to check that the models fit the normality and variance assumptions of linear models. Models that have homostedasticity showed residuals randomly distributed with no distinct pattern in a Residual v Fitted plot and scale-location plot. Normality was confirmed by seeing that the residuals fell along the fitted line in a QQ plot. A constant leverage plot was used to identify any significant outliers. Once the normality and variance assumptions were confirmed, I compared the AIC’s and of all the models and selected the model that had the lowest AIC, while also making logical sense within the biology of the system and conforming to the assumptions of a linear model.**

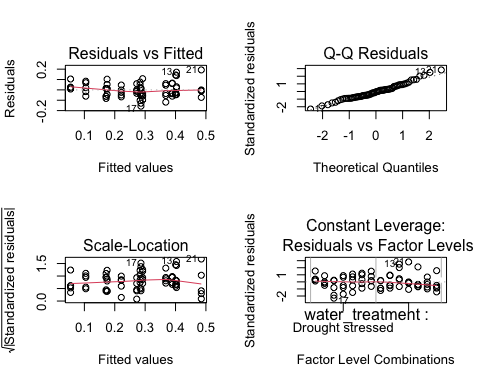
#### c. Make a visualization of the model predictions with underlying data for your “best” model. (20 points)

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! |

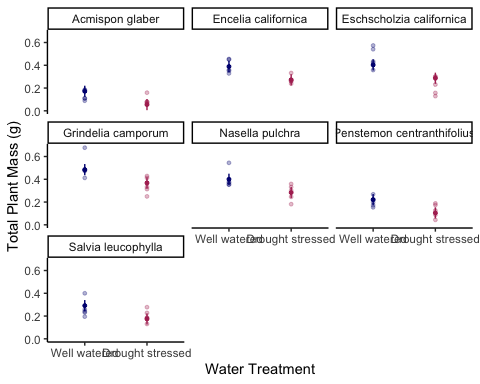
model4 <- lm(total\_g ~ water\_treatment + species\_name, #construct linear model using lm(y~x+x)  
 data = drought\_exp\_clean) #data from drought\_exp\_clean  
  
par(mfrow = c(2, 2)) #2x2 configuration of plots  
plot(model4) #check diagnostics



model\_preds <- ggpredict(model4, #make predictions using best model  
 terms = c("water\_treatment",   
 "species\_name")) #model includes these predictors  
  
model\_preds\_for\_plotting <- as.data.frame(model\_preds) %>% #create another data frame with easier column names to use for plotting  
 rename(water\_treatment = x, #rename columns  
 species\_name = group,  
 total\_g = predicted)

#plot observed data  
ggplot(drought\_exp\_clean, aes(x = water\_treatment, y = total\_g)) +  
 geom\_point(aes(color = water\_treatment), alpha = 0.3, size = 1) + #make each water treatment a different color #set low alpha for high transparency #set size of points to 1  
 labs(x = "Water Treatment", y = "Total Plant Mass (g)") + #change axis titles  
 facet\_wrap(~species\_name) + #facet by species\_name  
 scale\_color\_manual(values = c("Well watered" = "navyblue", "Drought stressed" = "maroon")) + # set color for each water treatment  
 theme\_classic()+ #set theme to remove gridlines and look more professional  
 theme(legend.position = "none")+ #remove legend  
  
#add predicted data  
geom\_point(data = model\_preds\_for\_plotting, aes(x = water\_treatment, y = total\_g, color = water\_treatment), alpha = 1, size = 1)+#plot the predicted data from model\_preds\_for\_plotting data frame #set same aesthetics as observed data #set alpha as 1 to make opaque and set size to 1  
 geom\_errorbar(data = model\_preds\_for\_plotting, aes(x = water\_treatment, ymin = conf.low, ymax = conf.high, color = water\_treatment), width = 0, size = 0.5)

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



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

**Figure 1. This figure displays the total plant mass in grams predicted by Model 4 based on water treatment and species underlaid with the actual observed data. Predicted values are represented by opaque points, while the transparent points represent the observed data. The predicted and observed data is separated into species, each with their own graph, and all points representing well watered plants are navy blue, while drought stressed data is maroon. The bars are showing the 95% confidence interval for the predictions.**

**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**

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

Your answer should be in paragraph form and address the following points:

* what predictors “best” described total mass (include model statistics here)?
* on average, what differences did you find between water treatments?
* on average, what differences did you find between species?

**Out of the five models tried, Model 4 best predicted the total plant mass. (Multiple Linear Regression, Multiple R-squared = 0.75, Adjusted R-squared = 0.73, F(7,26) = 27.08, alpha = 0.05, p-value< 2.2e-16) This suggests that species and water treatment together are significant predictors of total plant mass. On average, plants that were well watered had a higher total mass. Additionally, plants of the species Grindelia camporum and Eschscholzia californica averaged the highest total mass and plants of the species Penstemon centranthifolius tended to have the lowest total mass.**

### Problem 2. Affective visualization (24 points)

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

In lecture, we talked about the three vertices of data visualization: 1) exploratory, 2) affective, and 3) communicative. We’ve done a lot of exploratory and communicative visualization, but have yet to think about affective visualization.

When thinking of affective visualization, you can expand your ideas of what data visualization *could* be. Some examples of affective visualizations include:

* [Jill Pelto’s paintings](https://www.jillpelto.com/gallery-1)
* [Lorraine Woodruff-Long’s warming strips quilt](https://www.instagram.com/p/C689Kybxget/?utm_source=ig_web_copy_link&igsh=MzRlODBiNWFlZA==)
* [Stefanie Posavec and Giorgia Lupi’s Dear Data project](https://www.dear-data.com/all)

**Before starting, update your spreadsheet of observations.**

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

**For my personal data collected about my cat’s scratching behavior, I could create a pie chart with the proportions of total scratches that occurred during each activity, playing, petting, picking up, and unprovoked. I could further divide each portion of the chart by subcategories. So, the playing section is divided into slices depending on what toy was being played with, the petting section divided into slices depending on what part fo his body I was petting and picking up and unprovoked dont have subcategories. I would make the visualization more affective by illustrating each portion of the pie with visuals that represent that activity (I would draw the different toys for each toy section, draw me picking im up for the picking up portionetc.) Furthermore, I am considering having the pieces of the pie floating around within a larger picture that depicts lucas doing those activities.**

#### b. Create a sketch (on paper) of your idea. (2 points)

Include a photo of this sketch in your submission.

#### c. Make a draft of your visualization. (12 points)

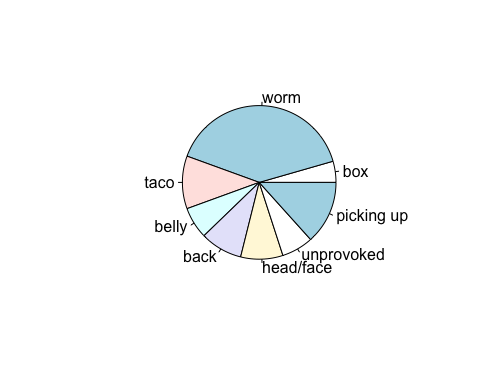
Feel free to be creative with this! *You do not have to do this in R.* You could create a sculpture, painting, textile object, etc.

If you are making your visualization in R, show the annotated code and the output.

If you are making your visualization outside of R, include a photo of your visualization in your submission.

**I just used R to get the right proportions for the pie chart. I traced them onto my paper.**

slices <- c(2, 18, 5, 3, 4, 4, 3, 6) #inpuring data, number of scratches per activity  
lbls <- c("box", "worm", "taco", "belly", "back", "head/face", "unprovoked", "picking up") #labelling each section  
pie(slices, labels = lbls) #create pie chart



#### d. Write an artist statement. (8 points)

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?)

**This piece depicts the different scenarios in which my cat, Lucas, has scratched me over the last six weeks. The four activities that we engage when he scratches me are playing together, me picking him up, me petting him, and him scratching me completely unprovoked (seemingly). Pieces of a deconstructed pie chart that depict the proportion of scratches that took place during that activity float next to an illustration of Lucas engaging in that activity. The pieces are further broken down into the toy being used for playing and where on his body I was petting him for petting.**

**I was inspired by Jill Pelto’s paintings in which she incorporated graphs seemlessly into her paintings. I was unable to merge my chart with my drawing quite as well, but I tried to follow a similar idea. This work was created on paper using pencil and colored pens. After recording every time Lucas scratched me since April 26th, and what activity directly preceded it, I used that data to create a pie chart in R and traced the slices onto the sheet to get the proportions correct, then sketched the rest of the drawing by hand.**

### Problem 3. Statistical critique (36 points)

At this point, you have seen and created a lot of figures for this class. Revisit the paper you chose for your critique and your homework 2, where you described figures or tables in the text. Address the following in full sentences (3-4 sentences each).

**For this section of your homework, you will be evaluated on the logic, conciseness, and nuance of your critique.**

#### a. Revisit and summarize (6 points)

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

Insert the figure or table you described in Homework 2 here.

**The authors are using generalized linear mixed models (GLMM) with a log-normal distribution to address their main research question which is to evaluate the effect of habitat restoration treatments on the travel speed of four predator and prey species, moose, bear, caribou, and wolf. They included treatment, microtopography, and land cover as predictors and several other factors as random intercepts. They constructed a model for each species for a total of 4.**

#### 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 represented their statistics fairly well. The x and y axis are logical, since the researchers want to test the effect of habitat restoration (treatment), on the movement speed of animals, making treatment the explanatory variable and movement speed the response variable. The axisis are labeled informatively with units when appropriate. The box and whisker plots show the median and the spread of the data. The underlying data is included. Each species is represented by its own separate graph. The results are clear and understandable. It is easy to see the difference between treated and untreated and between species.**

#### c. Aesthetic clarity (10 points)

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

**My main critique is of the color choice. It’s uninteresting and there isn’t much contrast between the dark grey used for the treated points and the lighter grey used for the untreated points. Replacing the greys with two distinct colors would make the figure more aesthetic and emphasize the two groups more, making not only clearer to understand, but more engaging and likely to be retained by the audience. Also, removing the gridlines would contribute to the overall more polished look, reducing the amount of clutter or unnecessary ink.**

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

What recommendations would you make to make the figure better? What would you take out, add, or change? Provide explanations/justifications for each of your recommendations.

**I would definitely change the greys to colors that are more contrasting and eye catching. The two shades of grey can be difficult to distinguish from one another, more contrasting colors would make differences between the treated and untreated data more visible. Reinforcing these differences is one of the main points of the figure. Removing the grid lines would also be helpful because it reduces visual clutter, making the data easier to focus on, and it makes the figure look more polished and finished so audiences will pay attention more.** **Since land cover was found to be a significant covariate, I would consider representing the effect of that variable as well. There would still be four plots separated by species, but instead of jitter plots, they would be scatterplots land cover on the x axis and travel speed on the y. The treated and untreated points would be different colors and a line with a confidence interval would be displayed for each set of data (treated and untreated). Since land cover is not the main interest of the researchers, I understand why they didn’t include it in the figure, but it would help more fully explain the model they produced. (They made a table for all the linear regression stats, which I think is enough, but just a suggestion).**

**Alternatively**, if they did not represent their statistics in a figure, what kind of figure would you recommend to them? Describe the x- and y-axes, with any geometries and aesthetics (colors, etc.). Provide enough explanation such that someone would be able to create the figure you describe in code. Provide explanations/justifications for the figure you would recommend.