Homework 3

Peter Nuñez

2024-06-02

https://github.com/peternunezz/nunez-peter\_homework-3.git

RStudio/code steps:  
Throughout the course of doing your homework, make at least 10 commits and pushes to the remote.

## 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)

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 packages   
  
drought\_exp <- read\_xlsx(path = here("data",   
 "Valliere\_etal\_EcoApps\_Data.xlsx"),  
 sheet = "First Harvest") #creating data frame  
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

model0 <- lm(total\_g ~ 1, # formula  
 data = drought\_exp\_clean) # data frame

model1 <- lm(total\_g ~ sla + water\_treatment + species\_name, #which variables we want to be considered  
 data = drought\_exp\_clean) #dataframe  
  
par(mfrow = c(2, 2))  
plot(model1)

model2 <- lm(total\_g ~ sla + water\_treatment, #what variables we wanna consider  
 data = drought\_exp\_clean) #data frame  
  
plot(model2)

model3 <- lm(total\_g ~ sla + species\_name, #variables we want to consider  
 data = drought\_exp\_clean)  
plot(model3)

model4 <- lm(total\_g ~ water\_treatment + species\_name, #variables we wanna consider  
 data = drought\_exp\_clean) #what data we wanna consider  
  
par(mfrow = c(2, 2)) #format for the plot  
plot(model1) #creating the plot

**Caption: Table 1-Model Selection Table: Comparing 5 models that describe total biomass. The table compares how well each model and it’s variables predict total biomass. Model 4 best describes total biomass because it has the lowest AIC, and conforms to assumptions of linear models.** table captions typically go above the table. Number the table and provide a title. Describe what is in the table (columns and rows).

model.sel(model0,  
 model1,   
 model2,   
 model3,  
 model4) #comparing the 4 models using an Akaike Information Criterion table

**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)

**To examine how specific leaf area, water treatment, and species influences total biomass, I first plotted how each variable individually affected total biomass. From this initial analysis, I determined that SLA, water treatment, and species all have some effect on total biomass. I then created 5 models; one null model(no variables considered), one saturated model(all three variables considered), and three models with two of the three variables considered each. To determine the model that best described total biomass, I compared all 5 models using an AIC (Akaike Information Criterion) table. From this table, I chose the final model to be Model 4 (the model that analyzed total biomass as a function of water treatment and species) because it had the lowest AIC (-156). I then looked at the diagnostic plots and concluded that Model 4 adequately follows the linear model assumptions.**

#### 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

model\_predictions <- ggpredict(model4, #view the predictions as a data frame  
 terms = c("water\_treatment",   
 "species\_name")) #specifing what terms we want in the plot

plot(model\_predictions, # model predictions  
 limit\_range = TRUE, # limit the range of predictions to the range of predictor values  
 show\_data = TRUE) + # show the underlying data  
 facet\_wrap(~group) + #grouped the plot by species  
 theme\_classic() + # classic theme  
 theme(legend.position = "none") + #removing the legend  
 labs(title = "Model 4 visualization", # plot title  
 x= "Water Treatment",  
 y= "Total Biomass") + #renaming the axes  
 theme(panel.grid = element\_blank()) + #removing the gridlines  
 scale\_color\_manual(values = c("red", "blue", "darkgreen", "purple", "pink", "orange", "gray")) #using colors that aren't default ggplot colors

|  |
| --- |
| Note |
| Make sure that the only output is the visualization! |

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

**Figure 1: Visualization of Model 4 is an prediction of total biomass (y-axis) as a function of species and water treatment (x-axis). Each species is distinguished by a different color.**

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

**The predictors that best described total biomass were species and water treatment, which we know by analyzing the model statistics (AIC = -156.2, detla = 0). We verified this by analyzing this model’s diagnostics plot, where the factors followed what we would expect from a linear model’s. Generally, the well watered treatments had a higher biomass than drought stressed treatments. The species with the highest biomass were Great Valley Gumweed, the California Poppy, and Purple Needlegrass. And the species with the lowest biomass were Purple Sage, Scarlet Bugler, and Deerweed.**

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?

### 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.

**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)

**My affective visualization for my personal data is a psuedo tile plot. I would lay out all of my clothes, organized by rows and grouped by color. I would then take a video of the clothes layed out on the floor, for a visualization of the diversity of color in my closet.**

#### 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 outside of R, include a photo of your visualization in your submission.

#### 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?)

**For this affective visualization, I aimed to show the diversity of color in the clothes from my closet through a psuedo tile plot. My inspiration for this visualization was to do something fun and personal to me, and to take the idea of a tile plot and to make it my own. I chose this project to get a sense of what colors I typically like to wear, and what factors influence the micro decisions I make every day. Since coming to University, fashion has been a method of creativity and expression, and I chose this as my project to combine statistics with this passion of mine. I chose this format of visualization to see this diversity in a real life medium, all layed out in front of me. I often forget about clothes in my closet, so doing this visualization reminded me of pieces I love, but don’t wear often. My process involved picking my favorite pieces of each color, and laying them out in a color scheme that flowed. I then took a video going across the rows, and took pictures of the tile plot in it’s entirety. My space to lay out the clothes was limited, or I would have included even more clothes.**

### 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?

**The statistical test that this figure describes is a generalized linear model with a binomial distribution. This test is used to compare the probability of mortality three years postfire of Sequioa trees with and without burn scars, specifically at the Black Mountain Grove. Earlier in the paper, the authors also used a generalized linear mixed model (with a binomial family and logit link) to test the probability of mortality in all groves in the park by topographic and tree characteristics.**

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

#### 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 statistics is visually represented well. The models described in table 4 of this paper are confusing, but the corresponding figure is simple and adequately represents the statistics. The response variable on the y-axis (probaility of mortality) makes sense, and the predictor variable (total crown damage %) on the x-axis makes sense as well. The underlying data is visible, which makes the comparison of fire scar verses no fire scar clear and understandable. Also, a casual viewer doesn’t need to understand the statistics to understand this figure. This figure ties back nicely to the hypothesis, “what first order fire effects best predict mortality by three years postfire at Black Mountain Grove?” by showing that crown damage and the presence of a fire scare signficantly predicts the mortality rate of the trees.**

#### c. Aesthetic clarity (10 points)

How well did the authors handle “visual clutter”? How would you describe the the data:ink ratio?  
**The aesthetics of this figure are quite nice. It is simple, clean, and easy to understand. There is no visual clutter, especially when comparing this figure to other figures in the paper. The underlying data is a nice size, adding necessary information without creating clutter. Every part of the figure is important and necessary, and there is no extraneous information.**

#### 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 add a title to this figure. Something like, “3 Years Post Fire Data at BLMO.” Although there’s a description of the figure in the caption, a title would allow for an even easier interpretation. I would also add the actual name of the BLMO acronym (Black Mountain Grove) in the caption. I would do this so a reader doesn’t have to go back to the introduction of the paper to understand what grove this figure is actually describing. The only thing I would alter is changing the x-axis from increments of 25 to increments of 20. The paper mentions that 80% crown damage is a significant number, so having that specific value on the x-axis would make it easier to visualize at what point crown damage influences probability of mortality.**

### Checklist

Your homework should

* Include your name, the title (“Homework 3”), and the date you turned in the assignment **(3 points)**
* Include for Problem 1:
  + a table for part a (if you made it in code, show the code and annotations)
  + written response for part b
  + full work (code and annotations) and output for part c
  + written response for part d
  + written response for part e
* Include for Problem 2:
  + written response for part a
  + photo for part b
  + code and annotations or photo for part c
  + written response for part d
* Include for Problem 3:
  + written responses for part a-d
  + written responses for parts g-h
* be uploaded to Canvas as a single PDF **(1 point)**
* be organized and readable **(10 points)**, meaning that
  + you are submitting a rendered PDF
  + all messages and warnings are hidden
  + all figures and tables are displayed correctly in the rendered PDF
  + all code annotations are visible in the rendered PDF
  + you have a set up chunk at the beginning and each section only has the code corresponding to that specific section (for example, no reading in packages/data in the chunk to create a visualization)
* be associated with a GitHub repository **(15 points)**, meaning that
  + you have a link to your GitHub repository at the top of your document
  + you have committed/pushed at least 10 times over the course of completing your homework
  + at least 10 of your commits/pushes contain a descriptive, concise commit message (a few words describing what changes you are committing)

**141 points**