Homework 3

Jackie Chen

2024-06-02

[Forked repository](https://github.com/jackiechen24/chen-jackie_homework-03)

# Set Up

## Reading in packages/data

# general use  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
✔ 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/jackiechen/Desktop/Academics/ENVS-193DS/GitHub/chen-jackie\_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)

#Uighur

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)

Warning: package 'modelsummary' was built under R version 4.3.3

`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) # table dataframe object

[1] "tbl\_df" "tbl" "data.frame"

## Cleaning

# cleaning the 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

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

## 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

1. **total mass is not predicted by any of the 3 predictors** (null model)

model1 <- lm(total\_g ~ 1, # formula  
 data = drought\_exp\_clean) # data frame

1. **total mass predicted by SLA, water treatment, and species** (saturated model)

# saturated model  
model2 <- lm(total\_g ~ sla + water\_treatment + species\_name,  
 data = drought\_exp\_clean)

1. **total biomass as a function of SLA and water treatment**

model3 <- lm(total\_g ~ sla + water\_treatment,  
 data = drought\_exp\_clean)

1. **total biomass as a function of SLA and species**

model4 <- lm(total\_g ~ sla + species\_name,  
 data = drought\_exp\_clean)

1. **total biomass as a function of water treatment and species**

model5 <- lm(total\_g ~ water\_treatment + species\_name,  
 data = drought\_exp\_clean)

Table of Models

# comparing models  
modelsummary::modelsummary( # this function takes a list of models  
 list(   
 "null" = model1, # "model name" = model object  
 "model 2" = model2,  
 "model 3" = model3,  
 "model 4" = model4,  
 "model 5" = model5  
 )  
)

|  | null | model 2 | model 3 | model 4 | model 5 |
| --- | --- | --- | --- | --- | --- |
| (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 |
| F |  | 23.439 | 14.572 | 13.855 | 27.079 |
| RMSE | 0.14 | 0.07 | 0.11 | 0.09 | 0.07 |

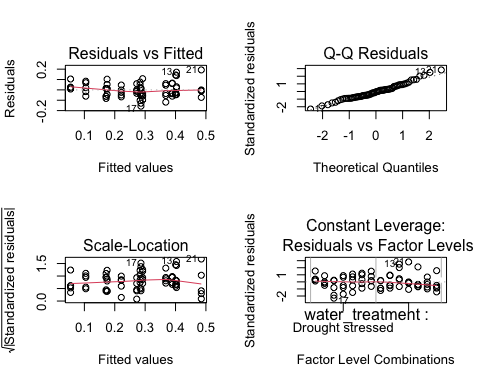
Model selection

model.sel(model1,  
 model2,   
 model3,   
 model4,  
 model5)

Model selection table   
 (Int) sla spc\_nam wtr\_trt df logLik AICc delta weight  
model5 0.05455 + + 9 88.598 -156.2 0.00 0.772  
model2 0.07994 -0.0002475 + + 10 88.741 -153.8 2.44 0.228  
model4 -0.03315 0.0012900 + 9 72.538 -124.1 32.12 0.000  
model3 0.04670 0.0012810 + 4 52.220 -95.8 60.37 0.000  
model1 0.27900 2 39.580 -75.0 81.22 0.000  
Models ranked by AICc(x)

Model diagnostics

par(mfrow = c(2, 2))  
plot(model5)



## 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 SLA, water treatment, and species on total biomass, I made a table of all of the possible models (the null model, the saturated model, and the models with combinations of some of the predictors). To determine the model that best described the total biomass, I looked at the AIC of all of the models. The model with the lowest AIC is the best model that describes total mass without sacrificing complexity, and out of the five models we have tried, the best model is the one that includes the predictors of water treatment and species. To evaluate linear model assumptions, I looked at the diagnostics of this model. The diagnostics for this model look good because the standardized residuals might flare out a bit but potentially is not a big deal and so the residuals are homoscedastic, the Q-Q Plot (Q-Q Residuals) shows a mostly straight line aside from a little bit of flare out at the tails and so the residuals are normally distributed, and there does not seem to be any outliers since there does not appear to be anything outside the dotted lines when looking at the Constant Leverage: Residuals vs Factor Levels Plot.

## 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:

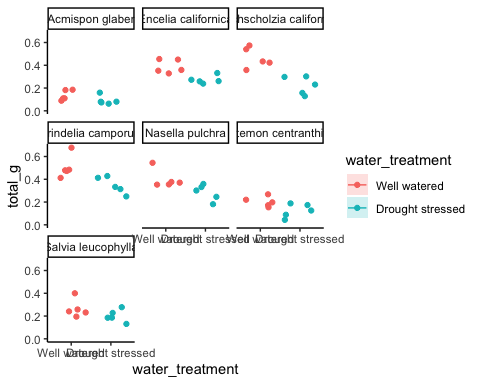
* make the underlying data more transparent than the model predictions
* display species names in full (not as species codes like ENCCAL or ESCCAL)
* display water treatment types in full (not as WW or DS)
* represent well-watered and drought stressed treatments with different colors
* use colors that are not the default ggplot() colors
* facet your plot by species
* remove the legend
* finalize the plot

Model predictions

model\_preds <- ggpredict(model5,   
 terms = c("water\_treatment",   
 "species\_name"))

# creating new data frame of model predictions for plotting  
model\_preds\_for\_plotting <- model\_preds %>%   
 rename(water\_treatment = x, # renaming columns to make this easier to use  
 species\_name = group)  
  
# use View(model\_preds\_for\_plotting)   
# to compare this to the original model\_preds data frame  
  
ggplot() +  
 # underlying data  
 geom\_jitter(data = drought\_exp\_clean,  
 aes(x = water\_treatment,  
 y = total\_g,  
 color = water\_treatment)) +  
 # 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),  
 alpha = 0.2) +  
 # model prediction lines  
 geom\_line(data = model\_preds\_for\_plotting,  
 aes(x = water\_treatment,   
 y = predicted,  
 color = water\_treatment)) +  
 # cleaner theme  
 theme\_classic() +  
 # creating different panels for species  
 facet\_wrap(~species\_name)

`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?  
`geom\_line()`: Each group consists of only one observation.  
ℹ Do you need to adjust the group aesthetic?



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

Include a data citation.

## 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?

# Problem 2. Affective visualization (24 points)

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

An affective visualization could

## 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.

## 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?)

# 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.

## 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?

## c. Aesthetic clarity (10 points)

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

## 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.

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.