ENVS 193DS Homework 3

Evelyn Bermudez

2024-06-04

Link to repo: https://github.com/evelynbermudez/bermudez-evelyn\_homework-03.git

# Problem 1

## Set up

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

### 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

## Model construction

The null model is:

1. **total biomass is not predicted by the water treatment, SLA, and species**

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

The saturated model is:

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

model1 <- lm(total\_g ~ sla + water\_treatment + species\_name,   
 data = drought\_exp\_clean)

In class we constructed the following models:

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

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

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

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

For homework, I will construct one more model:

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

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

## a. Table of models

modelsummary::modelsummary(list("Null" = model0,  
 "Model 1" = model1,  
 "Model 2" = model2,  
 "Model 3" = model3,  
 "Model 4" = model4)) # table showing each predictor and their associated coeffficients and s.e.

|  | 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 |

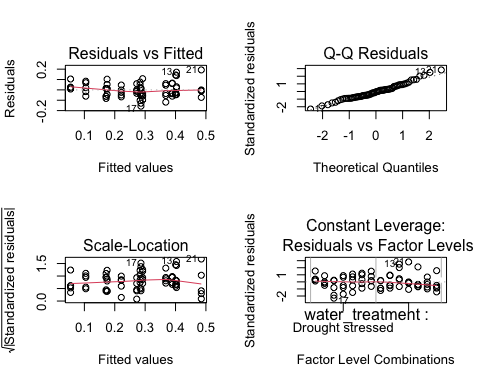
**Table 1: Models predicting plants’ total biomass** This table shows the coefficients, and their associated standard error, for the variables relevant to each model.

## b. Statistical methods

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)

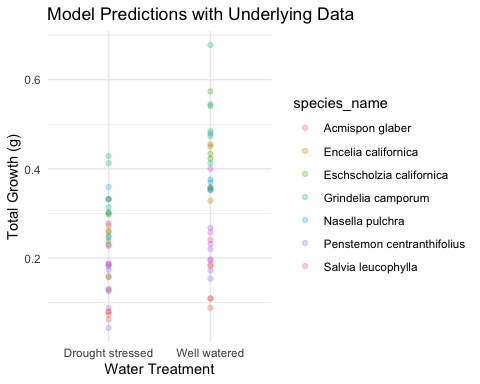
par(mfrow = c(2, 2))  
plot(model4)



**To examine the influence of specific leaf area (SLA), water treatment, and species on total plant biomass, I ran five linear regression models with different combinations of the three predictor variables. In order to compare the five models and select the best one, I used Akaike’s Information Criterion (AIC) and found water treatment and species to be the best variables to include in a model to predict plant biomass (model4 had the lowest AIC value and its delta = 0). Finally, the diagnostics for model4 confirm that the model conforms to the assumptions of linear models. The residuals and square root of the standardized residuals are evenly distributed across the model predictions, indicating homoscedastic residuals. We can also see that we have normality from the QQ plot since the relationship between the theoretical values and standardized residuals is roughly linear. Finally, there seems to be no outliers that significantly influence the model as visualized in the fourth plot.**

## c. Visualize model predictions with underlying data for “best” model

ggplot(drought\_exp\_clean,  
 aes(x = water\_treatment,  
 y = total\_g,   
 color = species\_name)) + # Add underlying data with transparency  
 geom\_point(aes(color = species\_name), alpha = 0.3) +  
 theme(legend.position = "none") + # Remove the legend  
 # finalize the plot w labs and theme  
 labs(  
 title = "Model Predictions with Underlying Data",  
 x = "Water Treatment",  
 y = "Total Growth (g)"  
 ) +  
 theme\_minimal()



## d. Visualization caption

## e. Results

summary(model4) # model summary

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 combination of predictors of total plant biomass were species and water treatment (F(7, 62) = 27.08, p < 0.001, = 0.05, R^2 = 0.73). On average, we see that well watered plants tend to be 0.12 %+-% 0.02g heavier than drought stressed plants. On average, Encelia californica, Eschscholzia californica, Grindelia camporum, Nasella pulchra, Penstemon centranthifolius, and Salvia leucophylla plants were 0.22 %+-% 0.03 g, 0.23 %+-% 0.03 g, 0.31 %+-% 0.03 g, 0.23 %+-% 0.03 g, 0.05 %+-% 0.03 g, and 0.12 %+-% 0.03 g, respectively, greater than Acmispon glaber plants

# Problem 2

## a.

**I would like to plot a bar graph with each bar being an individual run, and the y-axis representing the total distance in miles of that run. Each bar will be colored by the time of day the run occurred at with the goal of representing what time of day I prefer to run at given my weekly schedule. The three categories will be Morning (7AM - 11:59AM), Noon/Afternoon (12PM - 4:59PM), and Evening/Night (5PM - 10PM). The final plot will be completed on paper using watercolor paint.**

## b. Paper Sketch

|  |
| --- |
| Alt text |

## c. Visualization Draft

|  |
| --- |
| Alt text |

## d. Artist Statement

This bar plot is meant to highlight the patterns associated with what time I start a run at. In particular, it shows that I go on more runs after 12PM. This aligns with my true preference of running later in the day over the morning. The final draft of the plot will be in watercolor, as it is my second most practiced hobby after running. The significance behind finishing this plot in watercolor is rooted in the associated stress relief I experience when running as I do when painting watercolors. I made my draft on GoodNotes to obtain an idea of the color scheme I would like to use for the three time of day levels. The color of the sky during my runs is what inspired me to color the runs by a bright greyish blue in the morning, a light orange shade for the sun during noon/afternoon, and a purplish blue for the evening sky as the sun is setting.

# Problem 3

## a. Revisit and summarize (https://doi.org/10.1007/s11356-023-26621-1)

The authors use a baseline regression model (panel quantile regression) with the independent variables being a financial development index, income to represent economic development, trade openness to measure economic openness, and a human capital index, with a country-fixed constant and a time-fixed constant to predict multidimensional financial development. The goal of conducting this analysis is to understand the influences of climate risk on financial development. The

|  |
| --- |
| Alt text |

## b. Visual clarity

The authors did a good job representing their model predictions for the models they ran. Table 2 shows the model predictions for the main explanatory variables used by the authors. Some of the main explanatory variables were composed of other dummy variables or multiple financial indexes. All of the variables were defined in another table with their sample size, mean, standard deviation, minimum, and maximum. For Table 2, the authors showed the coefficients for the independent variables predicting financial development using a global sample across five quantiles of the outcome variables. They include the standard errors for the coefficients and denote their significance level with asterisks. Taking a look at one of the explanatory variables, the table effectively communicates that lack of coping capacity (LCop) to climate risk seems to have significant negative effects in the low quantiles (0.10, 0.25, and 0.50) and insignificant effects in high quantiles (0.75 and 0.90) on financial development. Thus, increases in lack of coping capacity toward climate risk looks to negatively affect financial development, especially in countries with low levels of financial development. The underlying data is not shown for any of the variables in the research paper.

## c. Aesthetic clarity

I would say the authors handled visual clutter well but there could have been larger spacing between the rows. The data:ink ratio is good as all the information on the table is necessary to understand the results of the model yet includes short hands to describe the standard errors and significance levels easily. I believe that the authors should have forgone the abbreviated variable names and used the complete variable names for on the table. The abbreviated variable names are described in another table in the paper, so I can see why the authors decided that the variables on Table 2 could be identified using this other table in order to reduce the text on Table 2.

## d. Recommendations

The first improvement I would make to this table is to enlarge the title and make it bold. I believe this would help readers quickly skim the paper faster to see if it includes an analysis of something they are interested in. I would also remove the numbers in parenthesis in the same row as “Dep. var: FD” as they do not seem to mean anything from the discussion of the results. If these numbers do have meaning, I would add an explanation for them. I would also add a bit more spacing between the rows grouped by the different independent variables in order to help readers identify what coefficients and standard errors belong to a variable. Lastly, I would change the independent variable names to be either all abbreviations or full names. Even though there is a reference table to the variable names, it bothers me that some of the names are full words, while others are not.