Homework 03

Sarah Tate

Table of contents

## GitHub Repo Link:

## Setup for Homework

### Loading in libraries/data from workshop

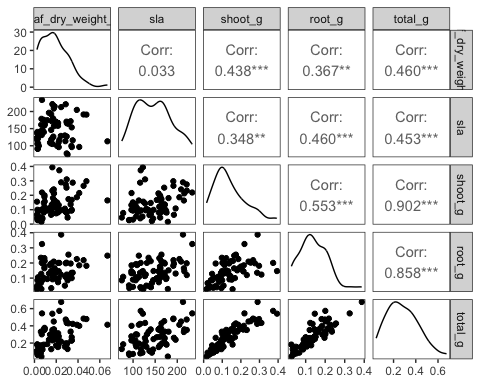
# general use  
library(tidyverse)  
library(readxl)  
library(here)  
library(janitor)  
  
# visualizing pairs  
library(GGally)  
  
# model selection  
library(MuMIn)  
  
# model predictions  
library(ggeffects)  
  
drought\_exp <- read\_xlsx(path = here("data",   
 "Valliere\_etal\_EcoApps\_Data.xlsx"), # reading in the data from an excel spreadsheet  
 sheet = "First Harvest") # choosing which sheet to read within the xlsx

## Cleaning data frame

# 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

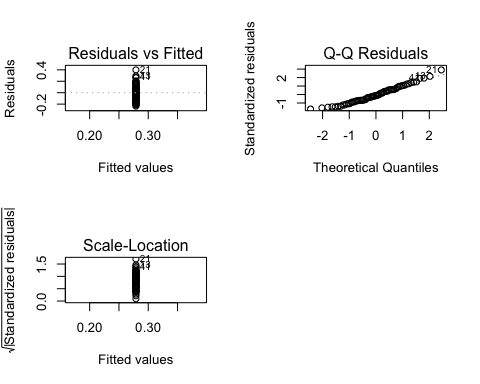
## Correlation plots for comparing variable relationships

ggpairs(drought\_exp\_clean, # data frame  
 columns = c("leaf\_dry\_weight\_g", # columns to visualize  
 "sla",   
 "shoot\_g",   
 "root\_g",   
 "total\_g"),   
 upper = list(method = "pearson")) + # calculating Pearson correlation coefficient  
 theme\_bw() + # cleaner theme  
 theme(panel.grid = element\_blank()) # getting rid of gridlines

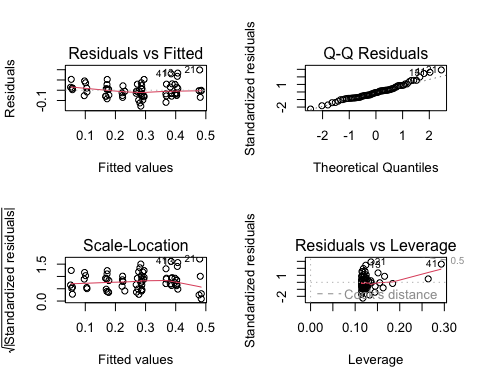


## Assigning models with different variables

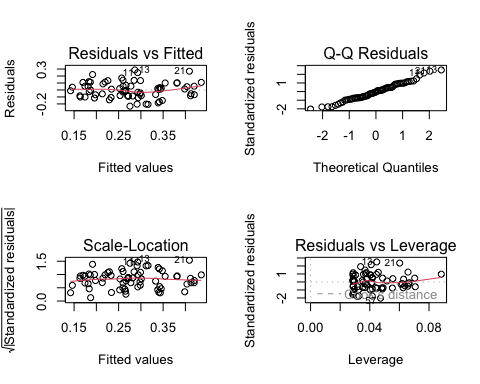
# null model  
model0 <- lm(total\_g ~ 1, # formula  
 data = drought\_exp\_clean) # data frame  
  
#diagnostics for model 0  
par(mfrow = c(2, 2))  
plot(model0)



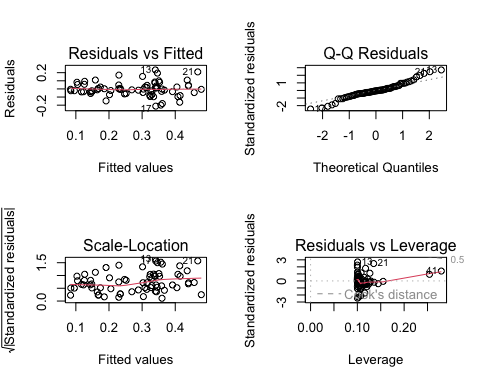
# saturated model (model 1)  
model1 <- lm(total\_g ~ sla + water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
  
#diagnostics for model 1  
par(mfrow = c(2, 2))  
plot(model1)



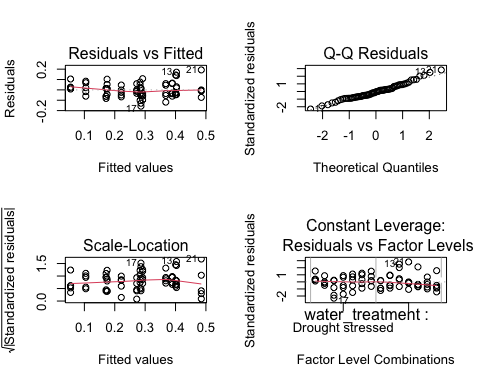
# model 2: total mass as predicted by specific leaf area and water treatment  
model2 <- lm(total\_g ~ sla + water\_treatment,  
 data = drought\_exp\_clean)  
  
#diagnostics for model 2  
par(mfrow = c(2, 2))  
plot(model2)



# model 3: total mass as predicted by specific leaf area and species name  
model3 <- lm(total\_g ~ sla + species\_name,  
 data = drought\_exp\_clean)  
  
#diagnostics for model 3  
par(mfrow = c(2, 2))  
plot(model3)



# model 4: total mass as predicted by species name and water treatments  
model4 <- lm(total\_g ~ water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
  
#diagnostics for model 4  
par(mfrow = c(2, 2))  
plot(model4)



## Model Selection to determine model of best fit

model.sel(model0,  
 model1,   
 model2,   
 model3,  
 model4) # code for comparing models to each other to find model of best fit

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)

## Model 4 Summary Statistics

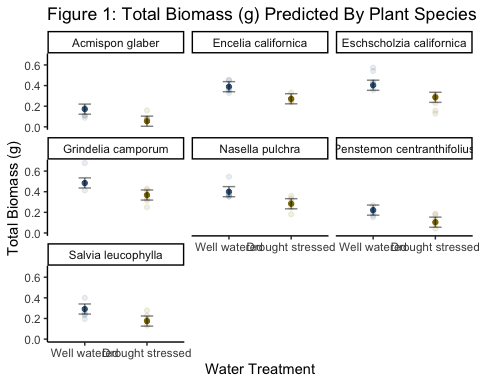
summary(model4)

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

## Problem 1: Multiple linear regression - model selection and construction

1. Table 1: Model Selection Data. The first column represents the model number, the second column represents which variables are present in each model, and the AIC and delta AIC values for each model are in the last two columns. Model 4 is the “best” model (AIC = -156.2, delta AIC = 0).
2. To address to central question of if total biomass (g) was influenced by specific leaf area, plant species, and/or water treatment, I assigned different models to a combination of the variables to determine which model contained the best predictor for total biomass. Model 0 included none of the variables, model 1 included all of them, and models 2-4 included combinations of 2 of the 3 variables. I used the model selection function in R to determine which model best explained the data found during experiment. It was shown that model 4 was the “best” model, which contained the variables of plant species and water treatment (AIC = -156.2, delta AIC = 0). With the largest AIC and a delta AIC of 0, it was concluded that model 4 contained the predictor variables which best explained the trends in the data. To finalize this conclusion, I determined that this model conformed to the assumptions of a linear model by creating 4 diagnostic plots to evaluate homoscedasticity, normality, and potential outliers. Through visual analysis of these plots, I determined that model 4 does conform to the assumptions of the linear model and is therefore the “best” model to predict total biomass.

model\_preds <- ggpredict(model4,   
 terms = c("water\_treatment",   
 "species\_name")) # assigning a model for prediction data based on "best" model  
  
model\_preds\_for\_plotting <- model\_preds %>% # new data frame for plotting prediciton data   
 rename(water\_treatment = x, # renaming columns to make this easier to use  
 species\_name = group)  
  
  
ggplot() +  
 # underlying data  
 geom\_point(data = drought\_exp\_clean, # choosing data frame  
 alpha = 0.1, # transparency of points  
 aes(x = water\_treatment,  
 y = total\_g,  
 color = water\_treatment)) + # setting x and y variables, and choosing color for variable  
 geom\_point(data = model\_preds\_for\_plotting, # choosing data frame to plot prediction data points  
 aes(x = water\_treatment,   
 y = predicted,  
 color = water\_treatment), # setting x and y axis variables, and color  
 alpha = 1) + # more bold point for prediction data  
 geom\_errorbar(data = model\_preds\_for\_plotting, # choosing data frame  
 aes(x = water\_treatment,   
 y = predicted, # setting x and y variables  
 ymin = conf.low,   
 ymax = conf.high, # setting bounds of 95% CI  
 fill = water\_treatment),  
 width = 0.2, # smaller width of errorbars  
 alpha = 0.5) + # more transparent bars so underlying data is easier to see  
 theme\_classic() + # cleaner theme, no gridlines  
 facet\_wrap(~species\_name)+ # creating different panels for species  
 theme(legend.position = "none")+ # getting rid of legend  
 labs(x = "Water Treatment",  
 y = "Total Biomass (g)",  
 title = "Figure 1: Total Biomass (g) Predicted By Plant Species And Water Treatment") + # changing x and y labels and adding title  
 scale\_color\_manual(values = c("Well watered" = "steelblue4",  
 "Drought stressed" = "gold4")) # setting colors for each water treatment different from original ggplot colors



1. Figure 1: Total Biomass (g) Predicted By Plant Species and Water Treatment. The underlying data represents experimental results for total biomass of different species based on different water treatments. Bold points represent model predictions for the relationship between variables. Error bars represent a 95% confidence interval for the model predictions. (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)

## Problem 2: Affective visualization

## Problem 3: Statistical critique