Daniel-Notcovich-homework-03

2024-06-06

Link: https://github.com/danielnotc/notcovich-daniel-homework-03.git

Table of contents

reading in packages

```r
# 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)

Valliere_etal_EcoApps_Data <- read_excel("Valliere_etal_EcoApps_Data.xlsx")

drought_exp <- read_xlsx("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"
```

cleaning

```
# 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
```

Problem 1. Multiple linear regression: model selection and construction

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.

```r
models <- data.frame(
  Model_numbers = c("Model 0", "Model 1", "Model 2", "Model 3", "Model 4"),
  Model = c("null model", "saturated model", "two predictors", "two
predictors", "two predictors"),
  Predictors = c("none", "SLA, water treatment, and species", "SLA and water
treatment", "SLA and species", "water treatment and species")
)

modeldatatable <- flextable(models) %>% # create flextable
set_header_labels(Model_numbers = "Model numbers", Model = "Model",
Predictors = "Predictors") %>% # rename headers
  align(align = "center", part = "all") %>% # formatting
  autofit() %>%
  theme_vanilla() %>%
  bold(part = "header") #bold everything in the header
```

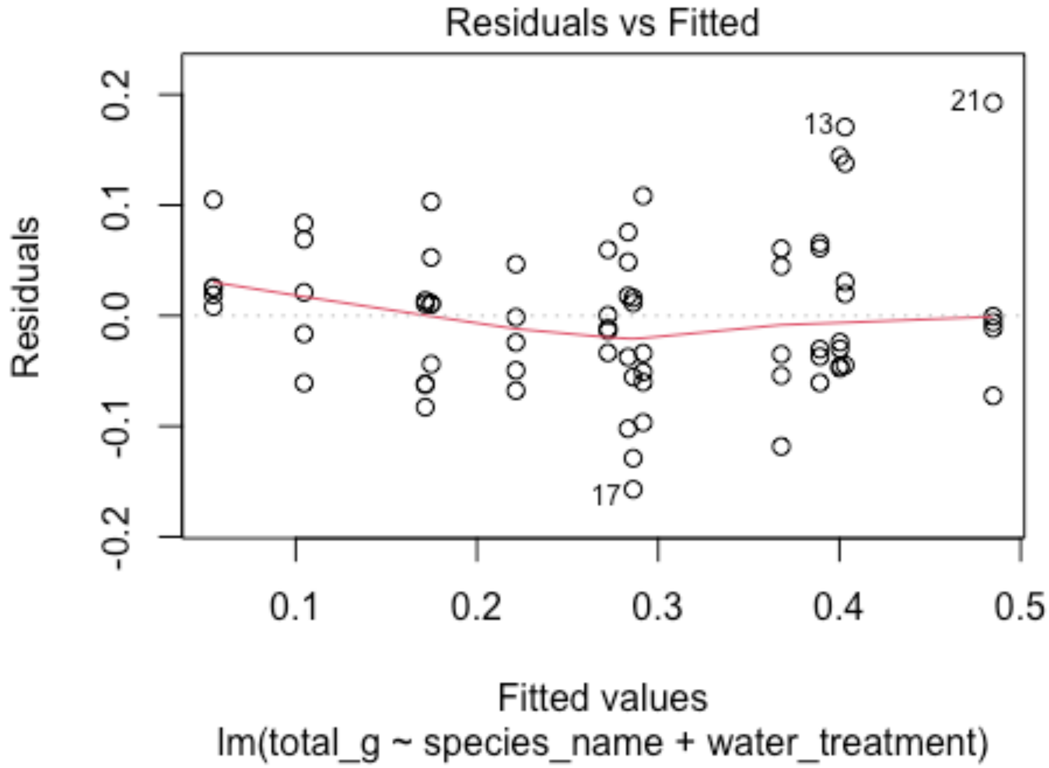
Table 1 showcases a set of five models for predicting plant species' total biomass. Each model is listed in a separate row, with columns indicating the model number, model description, and its predictors. The term 'SLA' stands for specific leaf area, which is calculated by dividing a leaf's surface area by its dry weight, representing the plant's life strategy (mm^2/g)

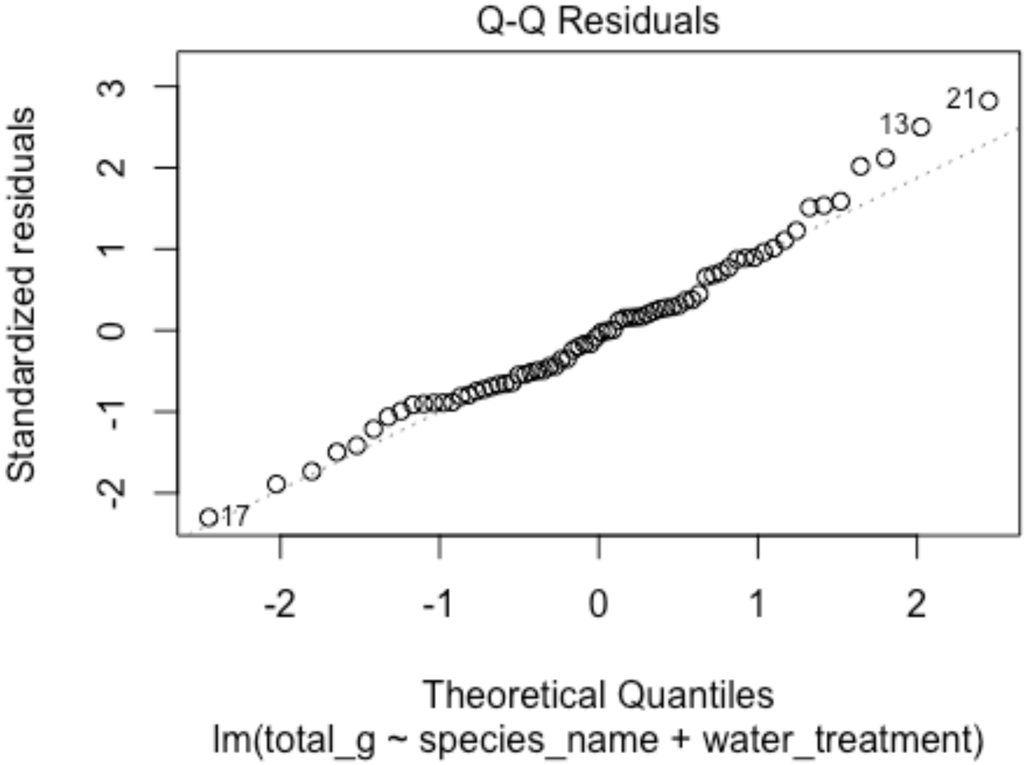
```
print(modeldatatable)

a flextable object.
col_keys: `Model_numbers`, `Model`, `Predictors`
header has 1 row(s)
body has 5 row(s)
original dataset sample:
  Model_numbers           Model                        Predictors
1       Model 0      null model                              none
2       Model 1 saturated model SLA, water treatment, and species
3       Model 2  two predictors          SLA and water treatment
4       Model 3  two predictors                   SLA and species
5       Model 4  two predictors       water treatment and species
```

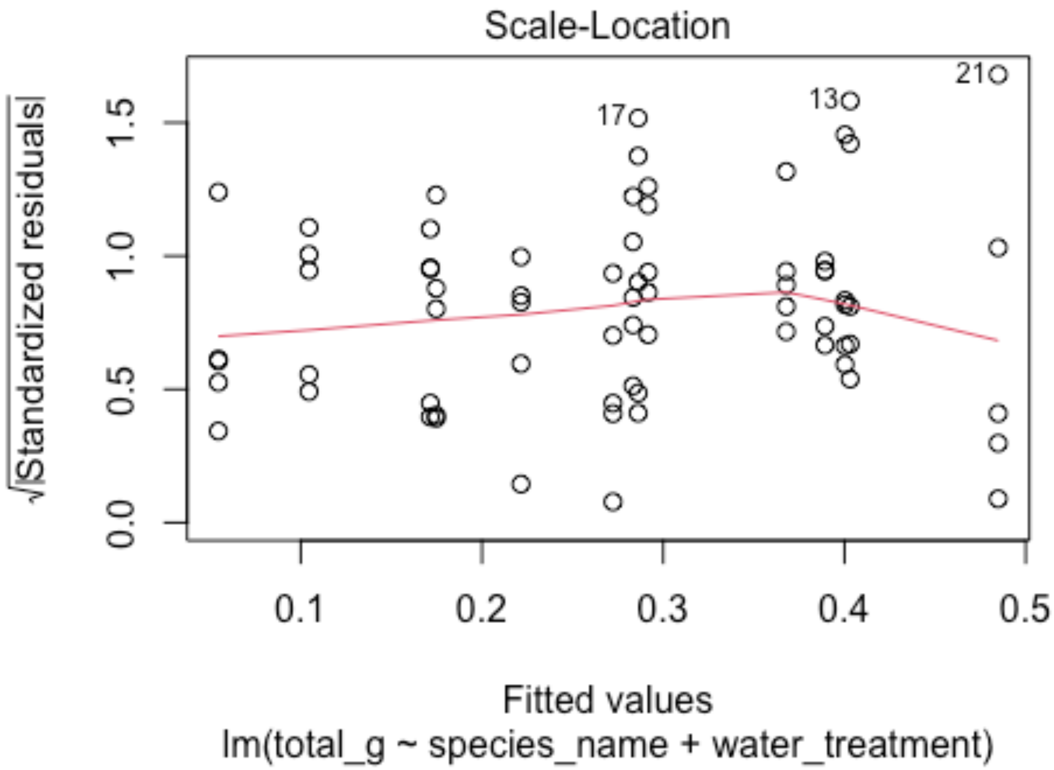
b. In this study, I explored how species type (categorical), water treatment (categorical), and specific leaf area (SLA, numeric) impact total biomass. To analyze these effects, I constructed five different models using these three predictors: a null model, a saturated model, and three models each containing two predictors. To determine the most effective model, I used the Akaike Information Criterion (AIC), which balances model fit and complexity. Among the five models, the combination of water treatment and species produced the lowest AIC value (AICc = -158.8), indicating it was the best predictor of biomass. I confirmed that the final model met the assumptions of linear regression by examining diagnostic plots for randomness, homoscedasticity, normality of residuals, and the absence of significant outliers.

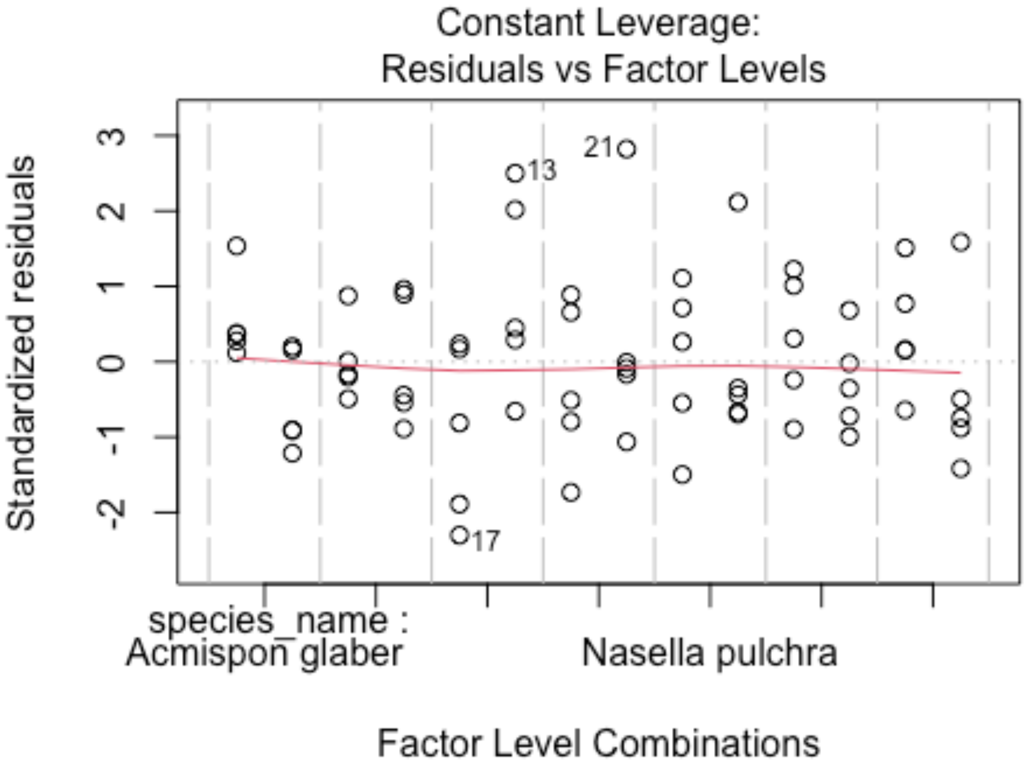
```r
# model of total biomass as a function of water treatment and species
model4 <- lm(total_g ~ species_name + water_treatment, # formula for model
             data = drought_exp_clean) # data frame
```

```r
plot(model4) # printing plots for model 4
```




Q-Q Residuals
Standardized residuals
3
2
1
0
-1
-2
-2
-1
0
1
2
Theoretical Quantiles
lm(total_g ~ species_name + water_treatment)
21
13
17


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
17
13
21
0.1
0.2
0.3
0.4
0.5
Fitted values
lm(total_g ~ species_name + water_treatment)



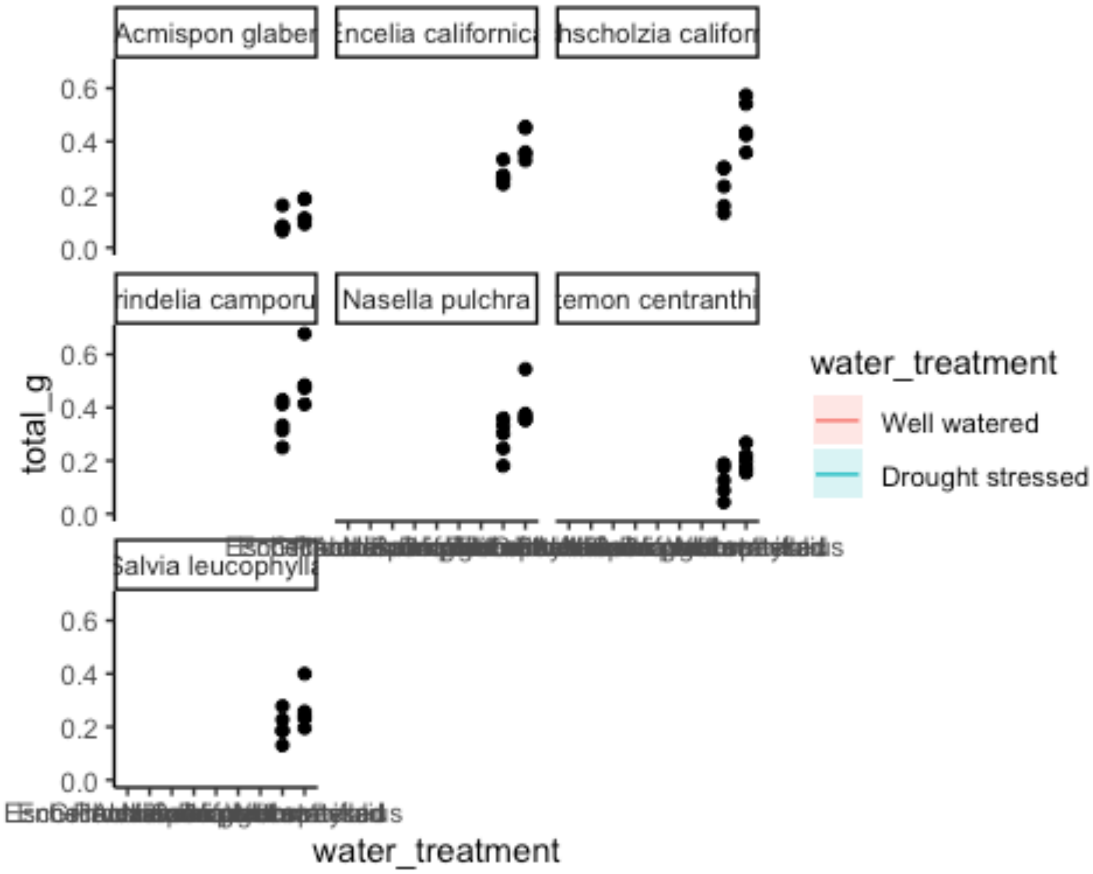
```
summary(model4)


Call:
lm(formula = total_g ~ species_name + water_treatment, 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 *
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
***
water_treatmentWell watered             0.11695    0.01733   6.746 5.90e-09
***
---
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
```

```
model_preds <- ggpredict(model4, # create model predictions
                         terms = c("water_treatment",
                                   "species_name"))
```

c. Make a visualization of the model predictions with underlying data for your “best” model.

```
# Creating a new data frame of model predictions for plotting
model_preds_for_plotting <- model_preds %>%
  rename(
    water_treatment = x,
    species_name = group
  )

ggplot() +
  # Underlying data
  geom_point(data = drought_exp_clean,
             aes(x = water_treatment,
                 y = total_g)) +
  # Model prediction 95% CI ribbon
  geom_ribbon(data = model_preds_for_plotting,
              aes(x = species_name,
                  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 = species_name,
                y = predicted,
                color = water_treatment)) +
  # Cleaner theme
  theme_classic() +
  # Creating different panels for species
  facet_wrap(~species_name)
```



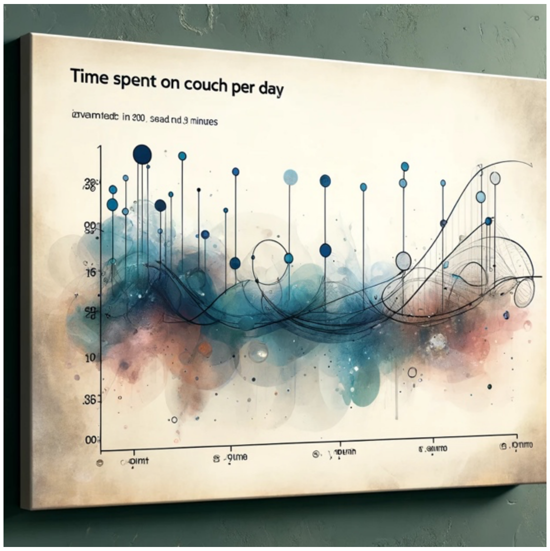
d. Caption for table Visualization of plant biomass predictions across different water treatments and species. The points represent observed data, while the shaded ribbons indicate the 95% confidence intervals of model predictions. The lines show the predicted biomass values based on the model, with different colors representing different water treatments. Each panel corresponds to a different species, facilitating comparison across species and treatments

e. Results

The predictors, water treatment, and species best described the total biomass, as evidenced by the model with the lowest Akaike Information Criterion (AICc = -158.8). On average, biomass was lower for drought-stressed plants compared to well-watered plants, which aligns with biological expectations. Additionally, plant species with larger ranges tended to have larger biomasses. This variation highlights the significant differences in biomass responses among species under different water treatments.

Problem 2. Affective visualization

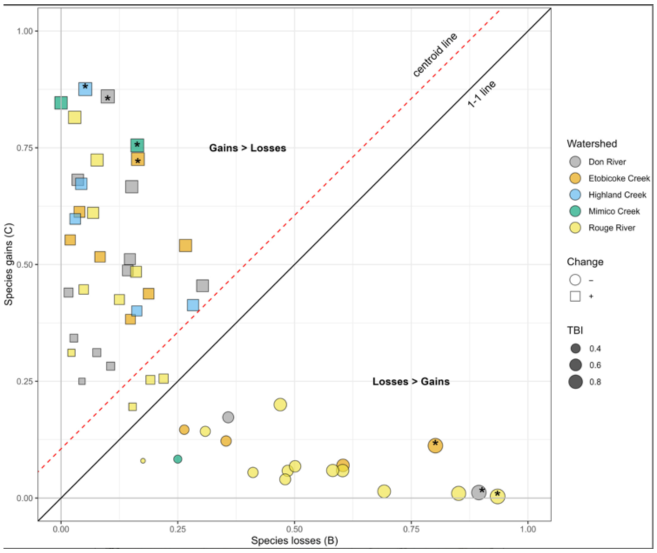
a. For my personal data representing the amount of time I spend on my couch per day I wanted an artistic representation that shows the asthetic of what my data represents. In this case I want my data to show the randomness that my data represents, showing that I never have a consistent schedule of the amount of time I am at home. This is also fitting with representing how my ADHD brain works, having no consistent schedule, alawys having variability in my life.



b. My piece showcases the amount of time I spend on my couch per day, visualizing the randomness and inconsistency in my daily routine. I wanted an absract representation of my data to further the feeling of randomness that i think abstract artworks do very well. I chose to use DALE to create my digital representation of my data, integrating real data with watercolor textures and soft gradients.

Problem 3. Statistical critique

https://doi.org/10.1002/ece3.10845



a. The authors employed the Temporal Beta Diversity Index (TBI) to investigate changes in species diversity over time within various watersheds. They used this index to quantify species gains and losses across different sites, determining whether diversity changes were statistically significant. Additionally, they applied Holm correction to adjust p-values for multiple comparisons, ensuring the robustness of their statistical findings.

b. The authors effectively represented their statistical results through clear and logically positioned axes, with the x-axis showing species losses and the y-axis showing species gains. They included both summary statistics and underlying data, using different shapes and colors to differentiate between sites with more gains than losses and vice versa. The figure also clearly marked high TBI sites with asterisks, enhancing the interpretability of significant findings.

c. The figure maintains a high data-to-ink ratio by using simple shapes and minimal text. Colors are used effectively to distinguish between watersheds without overwhelming the viewer. The inclusion of the centroid line and the 1-1 line provides a clear visual reference for interpreting the results, and the overall layout avoids unnecessary elements that could detract from the primary data.

d. To improve the figure, I would suggest the following changes:

Axis Labels and Titles: Enhance axis labels with more descriptive titles, such as "Species Losses (B)" and "Species Gains (C)," and include units if applicable. Adding a main title could also help viewers quickly understand the context of the figure.

Legend Clarity: The legend could be expanded to provide more details on the symbols and colors used. For instance, explicitly stating what the asterisks represent within the legend can avoid confusion.

Data Points: Increase the size of data points slightly for better visibility, especially for those with high TBI values. This change would ensure that significant points are easily identifiable even at a glance.