homework-03

Owen Choy

2024-06-02

Link to forked repository: https://github.com/owenchoy/choy_owen-homework-03

Setup

```
# general use
library(tidyverse)
library(readxl)
library(here)
library(janitor)
library(lterdatasampler)

# visualizing pairs
library(GGally)

# model selection
library(MuMIn)

# model predictions
library(ggeffects)

# model tables
library(gtsummary)
library(flextable)
library(modelsummary)

drought_exp <- read_xlsx("/Users/owenchoy/Downloads/ENVS 193DS Statistics for
 ↪  Environmental
 ↪  Science/ENVS-193DS/git/choy-owen_homework-03/code/data/Valliere_etal_EcoApps_Data.xlsx",
                         sheet = "First Harvest")
```

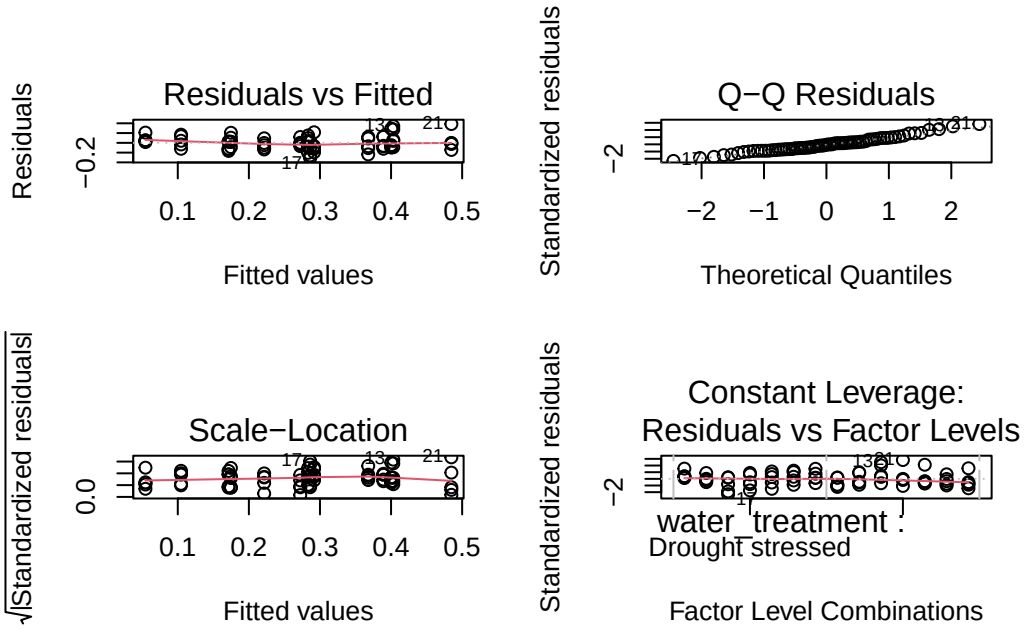
```r
# cleaning
drought_exp_clean <- drought_exp %>%
  clean_names() %>% # nicer column names
  mutate(species_name = case_when( # adding column with species scientific
   ↪  names
    species == "ENCCAL" ~ "Encelia californica", # bush sunflower
    species == "ESCCAL" ~ "Eschscholzia californica", # 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
```

```r
# Null model
model0 <- lm(total_g ~ 1,
             data = drought_exp_clean)
# total biomass as a function of SLA, water treatment, and species
model1 <- lm(total_g ~ sla + water_treatment + species_name,
             data = drought_exp_clean)
# total biomass as a function of SLA and water treatment
model2 <- lm(total_g ~ sla + water_treatment,
             data = drought_exp_clean)
# total biomass as a function of SLA and species
model3 <- lm(total_g ~ sla + species_name,
             data = drought_exp_clean)
# total biomass as a function of water treatment and species
model4 <- lm(total_g ~ water_treatment + species_name,
             data = drought_exp_clean)

# comparing all 5 models
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)
```

```
# model4 is best

#diagnostic test
par(mfrow = c(2, 2))
plot(model4)
```



```
# conslusion: diagnostics test for model 4 looks good: homoscedastic, normal,
 ↪  and no outliers, so model4 is good to use
```

```
# summary data for best model which includes slope and intercept
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

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.**

Table 1. Predictor variables of different models. The table shows the predictor variables of five models that were analyzed for the best ability of predicting plant biomass. The null

model (0) incorporates no predictor variables, while the saturated model (1) involves all three variables. SLA is the specific leaf area in mm^2/g. Water treatment varied between "well watered" and "drought stressed". "AIC" is the Akaike Information Criterion and a lower value represents a model that predicts the data best while remaining not too complex. A lower "delta" value also corresponds with a better predictive model based on predictive capacity and complexity.

model_number	predictors
0	none
1	SLA, water treatment, species
2	SLA, water treatment
3	SLA, species
4	water treatment, species

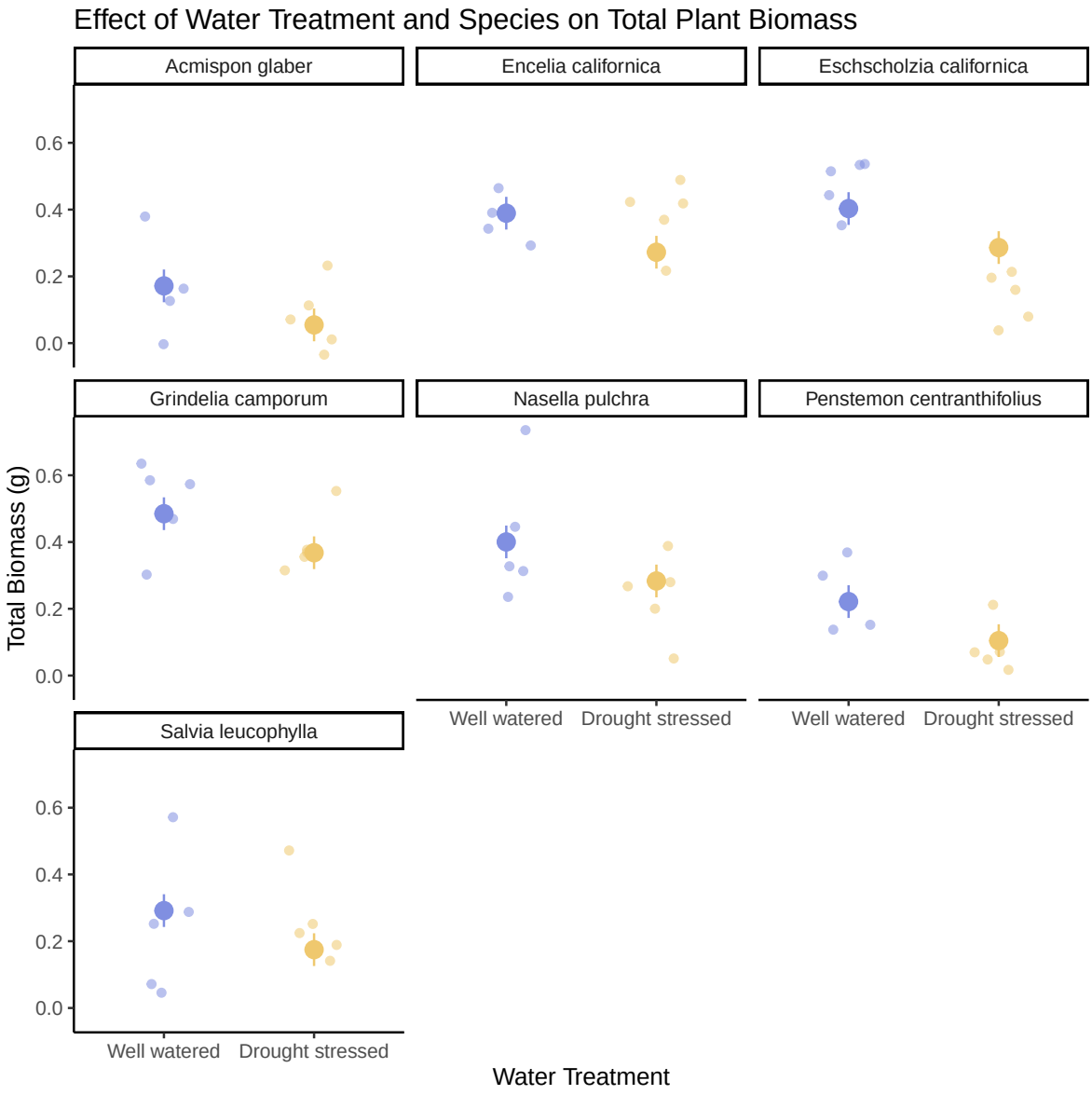
b. **Write a 5-6 sentence "statistical methods" section.**

To examine the influence of specific leaf area, water treatment, and species on the total plant mass, I constructed and compared several multiple linear regression models with these predictor variables. I established five different models to determine which best described total plant biomass: the null model with zero predictor variables, the saturated model with all three predictor variables, and three other models with different combinations of two predictor variables. To determine the best model, I assessed Akaike Information Criterion (AIC) of all five models and determined that Model 4 (with water treatment and species as predictor variables) was the best predictive linear model because it had the lowest AIC value. I then ran diagnostic tests to assess homoscedasticity, normality, and the potential influence of outliers for the model to evaluate the conformity of the model to linear model assumptions. All tests demonstrated comformity to linear model assumptions because the Residuals vs Fitted and Scale-Location plots had no visual pattern which suggests homoscedastic residuals, data followed the linear trend in the QQ plot, and no outliers were outside of Cook's distance in the Residuals vs Factor Levels plot. This process led me to select Model 4 as the best linear regression model.

c. **Make a visualization of the model predictions with underlying data for your "best" model.**

```
# creating new data frame of model predictions for plotting
model_preds <- ggpredict(model4,
                         terms = c("water_treatment",
                                   "species_name"))
```

```r
model_preds_for_plotting <- model_preds %>%
  rename(water_treatment = x,
         species_name = group)

# plot
ggplot() +
  geom_point(data = drought_exp_clean, # underlying data
             position = position_jitter(width = 0.2,
                                        height = 0.2,
                                        seed = 1),
             aes(x = water_treatment,
                 y = total_g,
                 group = species_name,
                 color = water_treatment, # color by water treatment
                 alpha = 0.1), # transparency
             size = 1.5) +
  geom_pointrange(data = model_preds_for_plotting, # mean/prediction data
                  aes(x = water_treatment,
                      y = predicted,
                      ymin = conf.low, # 95% CI
                      ymax = conf.high,
                      color = water_treatment),
                  size = 0.7) +
  labs(x = "Water Treatment", # axis and title labels
       y = "Total Biomass (g)",
       title = "Effect of Water Treatment and Species on Total Plant
        ↪  Biomass") +
  scale_color_manual(values = c("Well watered" = "#808FE1FF", # color by
   ↪  water treatment
                                "Drought stressed" = "#EFC86EFF")) +
  theme_classic() + # white theme
  theme(legend.position = "None") + # remove legend
  facet_wrap(~species_name) # separate plots by species
```



d. **Write a caption for your visualization.**

Figure 1. Effect of Water Treatment and Species on Total Plant Biomass. Plot shows the difference in total biomass (in grams) between well watered and drought stressed plants across species. Blue points represent well-watered treatments and yellow points represent drought stressed treatments. The larger opaque points represent model predictions as means with whiskers displaying the 95% confidence interval. (Data source: Valliere, J. M., Zhang, J., Sharifi, M. R., & Rundel, P. W. (2019). Can we condition native plants to increase drought tolerance and improve restoration success? *Ecological Applications*, 29(3), e01863. https://doi.org/10.1002/eap.1863.)

e. **Write a 3-4 sentence results section.**

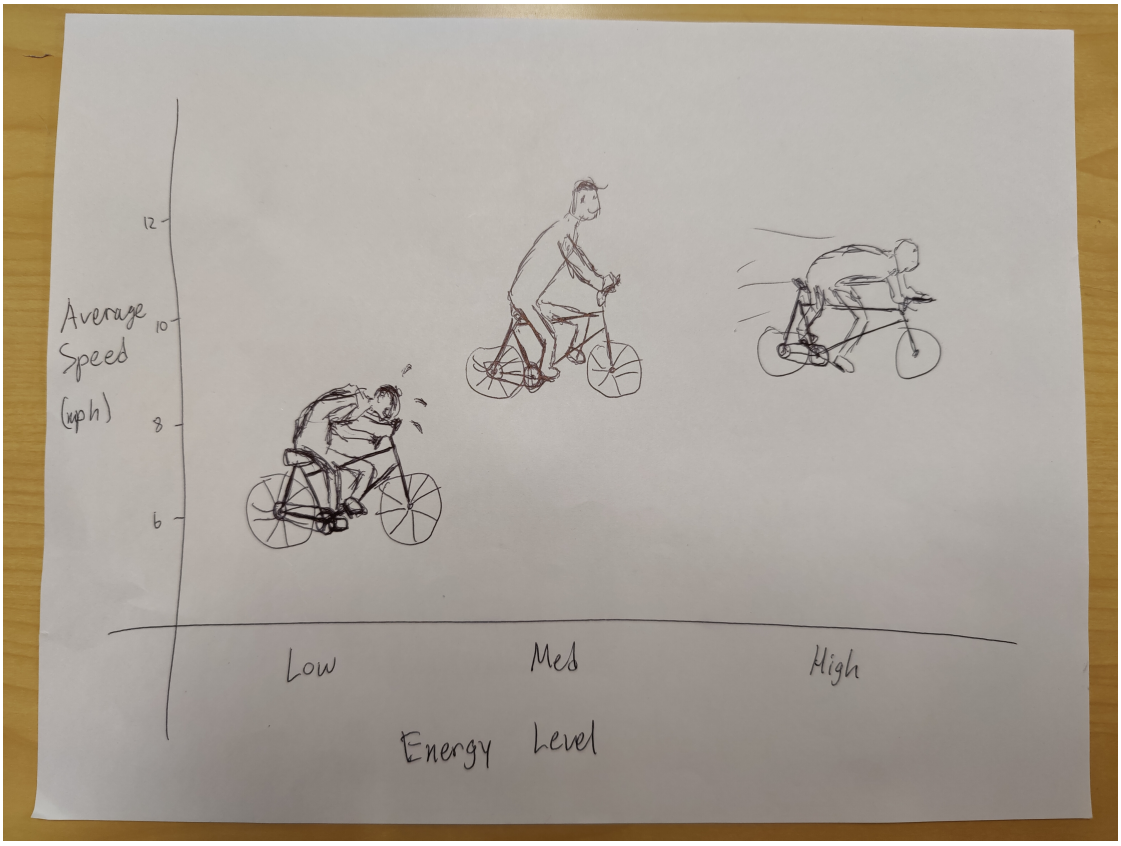
The results indicate that water treatment and species are the best predictors for total plant biomass (linear regression, $F(62,7) = 27.08$, $p < 0.001$, $= 0.05$, Multiple $R^2 = 0.7535$) and contribute to the best predictive model (multiple linear regression, normal distribution, AICc = -156.2, delta = 0.00). Holding species constant while using model predictions, well watered plants are expected to have 0.117 ± 0.017g more mass than drought stressed plants ($t = 6.746$, $p < 0.001$, $= 0.05$). On average, *Grindelia camporum* had the greatest difference in mass compared to the reference species *Acmispon glaber*, with a greater mass of 0.313 ± 0.032g ($t = 9.662$, $p < 0.001$, $= 0.05$).

Problem 2 Affective visualization

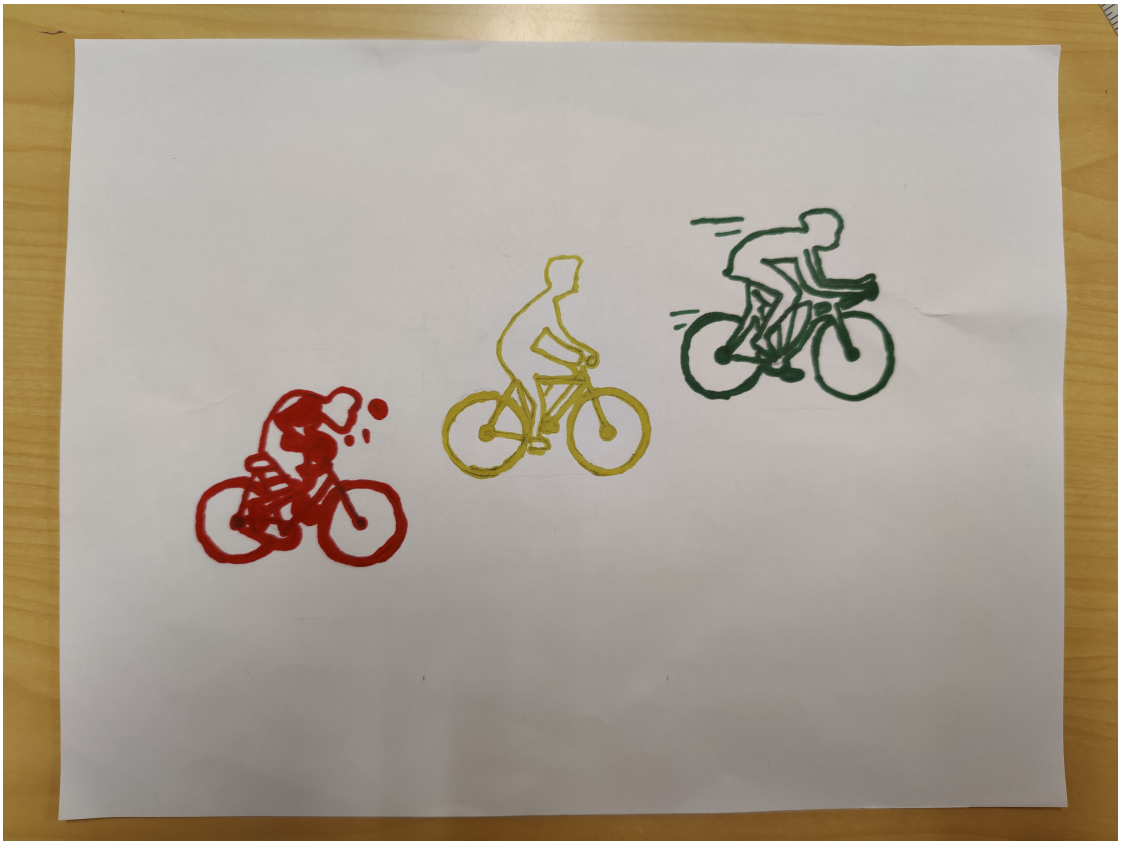
a. **Describe in words what an affective visualization could look like for your personal data (3-5 sentences).**

One affective visualization I can create from my personal data incorporates images onto a plot. On the x-axis of the plot is the energy levels (low, medium, high) and the y-axis is my average speed for each bike ride. Each category of energy level would visualize a biker expressing the appropriate energy level and associated with a color. For instance, the "low energy biker" would be in the low energy column and colored red. Ideally, the image of the biker would connect the data points of the respective category, but that may be quite difficult to do so I may just have a mean point on the biker.

b. **Create a sketch (on paper) of your idea.**



c. **Make a draft of your visualization.**



d. **Write an artist statement.**

In this piece of work, which is a drawing, I visualize the three different categories of "energy levels" that was one of the variables measured, from left to right: low, medium, high. The bikes are in placed in reference to each other based on their mean speed of bike trips to and from campus. I gained inspiration of this piece from Jill Pelto's paintings. I created this piece by tracing images of bikers over a grid to get the relative positions of each biker.

Problem 3 Statistical Critique

a. **Revisit and summarize.**

The authors are using ANOVA in their analysis. They used ANOVA to compare diets of quokka populations across sites and seasons and compared with the location of food plants, dietary diversity (H') scores, and sex. Their main research question asked what the diet and dietary preferences of different populations of quokkas in the northern Jarrah Forest were. The table below compared sites and seasons with diet.

Comparison	d.f.	s.s.	F	P
Factor 1				
Season	3	1.77	0.70	0.558
Site	3	11.84	4.64	<0.001
Season × site	9	5.93	0.77	0.641
Residual	81	68.934		
Factor 2				
Season	3	47.02	189.58	<0.001
Site	3	36.40	146.77	<0.001
Season × site	9	73.68	99.03	<0.001
Residual	81	6.70		

b. **Visual clarity.**

Plots lacked an x-axis title and some plots had two y-axes, which provided too much information in a crammed space. A legend was shown in one plot but omitted in later plots with similar information but more complexity, so it was not easy to distinguish what some of the groups were. Labels used abbreviations so the viewer has to find a reference to refer to in order to understand the labels, which was not always located in the figure caption. Summary statistics (mean and standard deviation) were displayed and described in the captions, but error bars were only shown in the positive direction and underlying data and model predictions were not included.

c. **Aesthetic clarity.**

The authors handled visual clutter a bit poorly for the plots that packed more information and data. Information in plots were often crammed and cluttered, and difficult to interpret because of the abundance of bars, lines, boxes, and variables. Lack of color and the gray-scale color scheme makes the figures seem dull and uninteresting. However, the data-ink ratio is pretty solid—there are no unnecessary lines or shapes that distract from the (complexity of the) data, and all dots, lines, bars, and boxes are described in the figure caption.

d. **Recommendations.**

I would recommend including a legend in the more complex and condensed figures to add clarity to the colored elements of the figures. Separating the plots with multiple y-axes into separate plots would also be beneficial to increase clarity and reduce visual clutter.

Including an x-axis title would make it more clear on what the plots are comparing. Displaying x-axis labels at an angle would also help the viewer read the plot more easily without reducing aesthetic clarity.