ENV S 193DS - Homework 3

Andrew Liang

2024-06-02

Table of contents

## Link to Forked Repository: https://github.com/andrewjliang/liang\_andrew-homework-03

# Setup

rm(list = ls())  
getwd()

[1] "/Users/andrewliang/Desktop/andrew/ENVS-193DS/liang\_andrew-homework-03/code"

library(tidyverse)  
library(readxl)  
library(here)  
library(janitor)  
library(GGally)  
library(MuMIn)  
library(ggeffects)  
library(gtsummary)  
library(flextable)  
library(modelsummary)  
library(car)  
  
drought\_exp <- read\_xlsx(path = here("data",   
 "Valliere\_etal\_EcoApps\_Data.xlsx"),  
 sheet = "First Harvest")  
  
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. Table

model0 <- lm(total\_g ~ 1, # formula  
 data = drought\_exp\_clean) # data frame  
model1 <- lm(total\_g ~ sla + water\_treatment + species\_name,  
 data = drought\_exp\_clean)  
model2 <- lm(total\_g ~ sla + water\_treatment,  
 data = drought\_exp\_clean)  
model3 <- lm(total\_g ~ sla + species\_name,  
 data = drought\_exp\_clean)  
model4 <- lm(total\_g ~ water\_treatment + species\_name, drought\_exp\_clean)  
  
selectiontable <- model.sel(model0,  
 model1,   
 model2,   
 model3,  
 model4)  
  
selectiontable

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)

## b. Statistical Methods

model.sel(model0, # checking models from above for AIC and Delta.   
 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)

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

|  |
| --- |
|  |

vif(model4) # Verifying VIF of predictor variables.

GVIF Df GVIF^(1/(2\*Df))  
water\_treatment 1 1 1  
species\_name 1 6 1

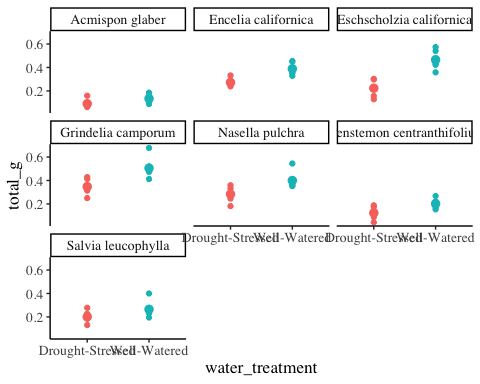
par(mfrow = c(2, 2))  
plot(model4) # checking residuals.

|  |
| --- |
|  |

We attempted to determine the relationship between the total biomass based off of three variables: the tyep of species, whether the area was treated with water, and the specific leaf area. In order to choose the most appropriate model, I looked at the Akaike Information Criterion and delta levels for each model. We found that the fifth model (model4), where we regressed total mass on water treatment and species was the most appropriate model because it has the lowest AIC and delta AIC. After checking the conditions for linearity (graphs shown above), we found that the assumptions for the multiple linear regression model (no heteroskedasticity, relatively normal residuals, no patterns in residuals versus leverage) were met, and thus concluded that our fifth model was the most appropriate model, and total mass was the better dependent variable.

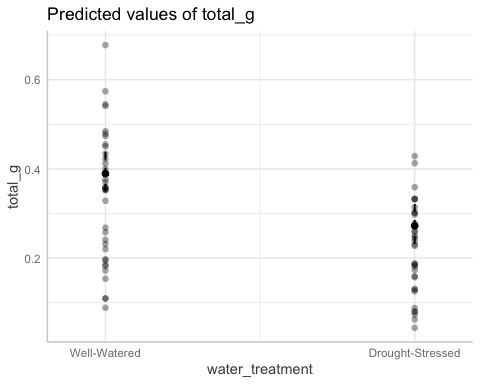
## c.

drought\_exp\_clean %>%  
ggplot(aes(x = water\_treatment, y = total\_g, group = water\_treatment, color = water\_treatment)) +  
 geom\_point() +  
 facet\_wrap(~species\_name) +  
 theme\_classic() +  
 stat\_summary(FUN = mean\_cl\_normal) +  
 theme(text = element\_text(size = 13, family = "serif"),   
 panel.grid.major = element\_blank(),   
 panel.grid.minor = element\_blank(),  
 panel.background = element\_blank(),   
 axis.line = element\_line(colour = "black"),   
 plot.title = element\_text(hjust=0.5),  
 legend.title = element\_blank(),  
 legend.position = "none")

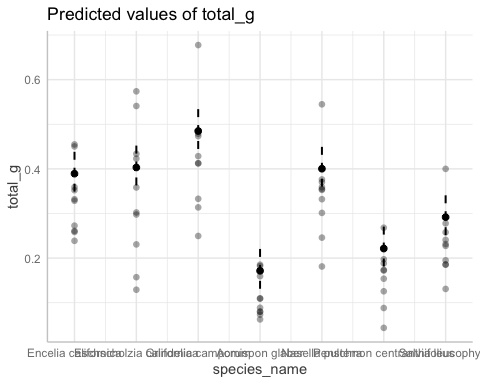


ggpredict(model4) %>%  
 plot(show\_data = T)

$water\_treatment



$species\_name



## d.

Figure 1.

Data source: Can we condition native plants to increase drought tolerance and improve restoration success?, Valliere JM, Zhang J, Sharifi MR, Rundel PW.

## e.

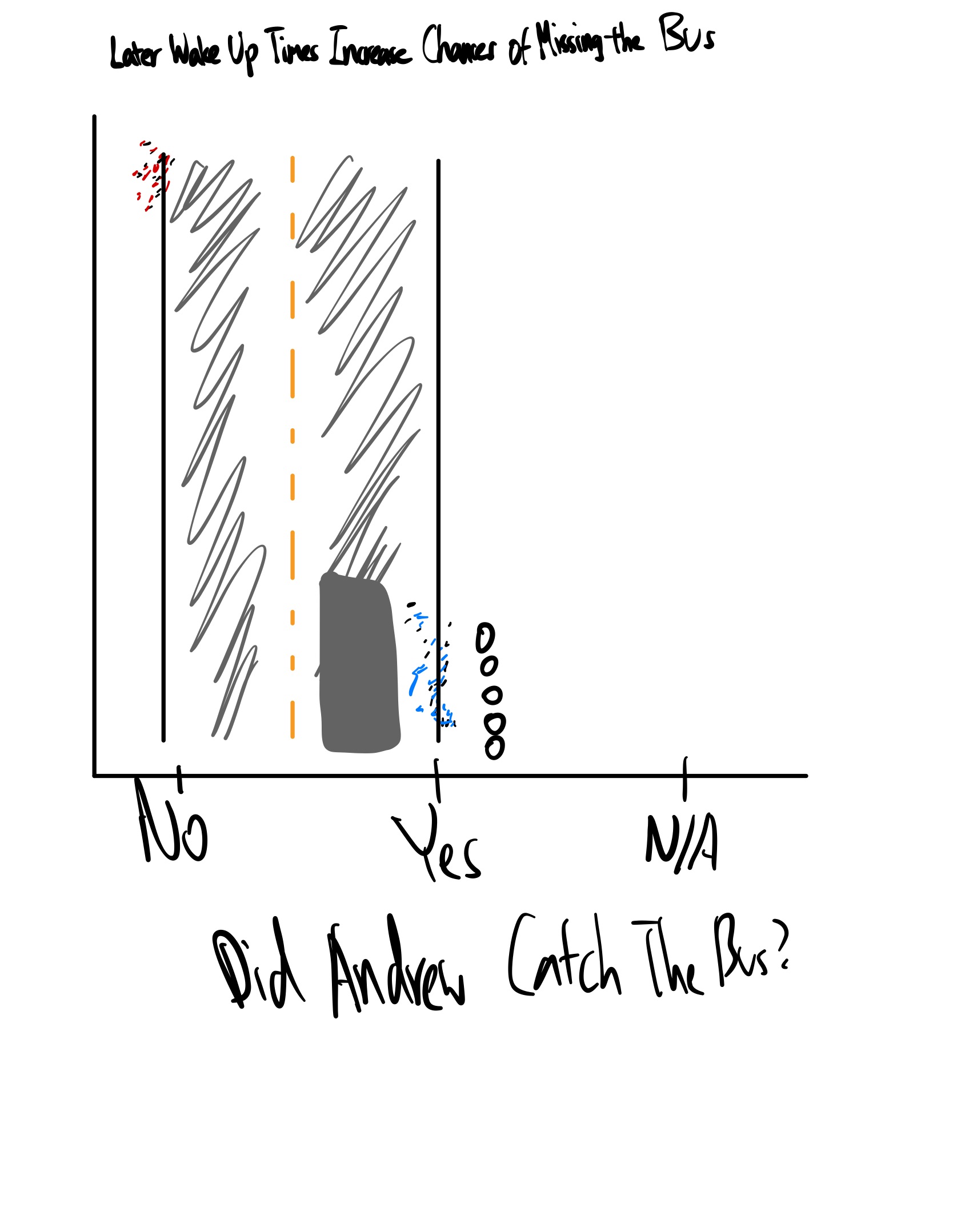
*We argue that the*

# Problem 2 - Affective Visualization

## a.

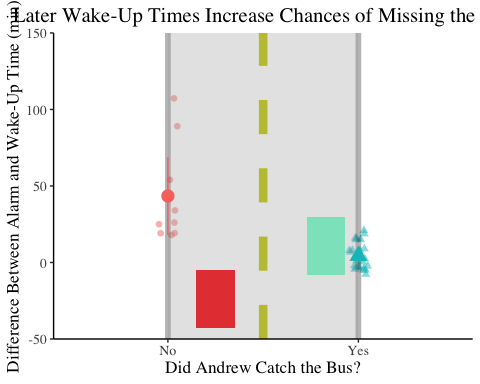
*My thought for an affective visualization for my data would include some sort of reference to sleep or trying to get on the bus. The main reason I may have trouble getting on the bus is oversleeping or taking too long to prepare for the day, so I believe that a visualization showing the consequences of doing so would be appropriate for my data. Because my dependent variable is binary (whether I got on the bus), I could create a “road” with the observations that I could include a “bus” on for my art.*

## b.



## c.

setwd("/Users/andrewliang/Desktop/andrew/ENVS-193DS/busdata")  
busdata <- read\_csv("BusData.csv") %>%  
 clean\_names()  
view(busdata)  
  
busdata %>%  
 ggplot(aes(x = got\_on\_bus, y = difference\_in\_wake\_up\_time\_minutes, shape = got\_on\_bus, color = got\_on\_bus)) +  
 geom\_point(position = position\_jitter(width = 0.05), size = 2, na.rm = T, alpha = 0.5) + # Adding points based on whether Andrew got on the bus. Shapes and color differ. Points where the estimated bus time is unavailable are removed.   
 labs(title = "Later Wake-Up Times Increase Chances of Missing the Bus", x = "Did Andrew Catch the Bus?", y = "Difference Between Alarm and Wake-Up Time (min)") + # adding titles  
 scale\_y\_continuous(expand = c(0, 0)) + # making sure that the graph starts at 0  
 geom\_vline(xintercept = 1, linewidth = 2, alpha = 0.3) + # first road line  
 geom\_vline(xintercept = 1.5, linewidth = 3, linetype = "dashed", color = "yellow3") + # median line for the road.  
 geom\_vline(xintercept = 2, linewidth = 2, alpha = 0.3) + # second road line  
 annotate("rect", xmin = 1.15, xmax = 1.35, ymin = -43, ymax = -5, fill = "red") + # this rectangle represents the buses that drive away with those who miss it (points in red).  
 annotate("rect", xmin = 1.73, xmax = 1.93, ymin = -8, ymax = 30, fill = "aquamarine") + # this "bus" is at the bus stop where on-time passengers are present.  
 annotate("rect", xmin = 1, xmax = 2, ymin = -50, ymax = 150, alpha = 0.3, fill = "darkgray") + # road texture  
 stat\_summary(fun.data = mean\_cl\_normal,  
 geom = "pointrange", size = 1) +  
 theme(text = element\_text(size = 13, family = "serif"),   
 panel.grid.major = element\_blank(),   
 panel.grid.minor = element\_blank(),  
 panel.background = element\_blank(),   
 axis.line = element\_line(colour = "black"),   
 plot.title = element\_text(hjust=0.5),  
 legend.title = element\_blank(),  
 legend.position = "none") +   
 guides(fill=guide\_legend(title="Got on Bus"))



## d.

*The Bus Stop*

*This visualization displays data indicating the amount of time used to prepare each day versus whether Andrew was able to make it to the bus, based off of a model road and “buses” on both sides of the road. The image represents data points where Andrew did not catch the bus as people who miss the bus (the red rectangle) because they are late, while the data points where Andrew was able to catch the bus represent the people who are able to get on the bus (aquamarine rectangle) on time. Jill Pelto’s paintings influenced my decision to incorporate the data within the image itself. I created my image through R and ggplot; I used various line and rectangle geoms in order to create the road and the “buses,” while the data and confidence interval for means were directly added to the roads.*

# Problem 3 - Statistical Critique

## a.

*The authors are using t-tests, linear regression, and difference-in-difference models in order to test the following questions: whether a retrofit effectively reduced energy consumption in homes, whether differences in energy usage could be seen between subsidized and non-subsidized retrofits, and whether higher amounts of subsidies influenced the difference in energy usage*.

|  |
| --- |
| T-Test |

|  |
| --- |
| First Linear Regression Model |

|  |
| --- |
| Second Linear Regression Model |

|  |
| --- |
| Third Linear Regression Model |

|  |
| --- |
| Fourth Linear Regression Model |

|  |
| --- |
| Fifth Linear Regression Model |

|  |
| --- |
| Difference-in-Differences Model |

## b.

*The authors of the paper did not visualize any of their models, except one graph (Figure 6) that seemed to indicate an inverse relationship between the amount of subsidies given towards a retrofit of a building and changes in a building’s energy consumption. Rather, the results of the regression models, t-tests, and difference-in-differences model were all incorporated into the tables above.*

## c.

*For the graphs/figures that WERE in the article, the data:ink ratios were relatively low, and pertinent information was definitely distinguished (for some graphs, the fitted curves were colored while other curves were grayscale). However, the figure where a linear regression model was fitted onto data had all of its data colored with similar tones of green, which could definitely confuse the reader if there was not a legend. Figures where means of groups were compared saw data for those groups distinguished sufficiently, usually through different colors.*

## d.

*Because none of the statistical tests (besides one regression model) were represented in graphs, I would suggest that the authors depict more of their*