ENVS-193DS\_spring-2025\_final.git

Marlaina Lascano

2025-06-11

## Quarto

Git Hub Repo - <https://github.com/marlainalascano/ENVS-193DS_spring-2025_final>

#load packages   
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
✔ purrr 1.0.4   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(DHARMa) #3f

This is DHARMa 0.4.7. For overview type '?DHARMa'. For recent changes, type news(package = 'DHARMa')

library(gt)  
library(MuMIn) #3g  
library(lubridate)  
library(viridis)

Loading required package: viridisLite

library(janitor)

Attaching package: 'janitor'  
  
The following objects are masked from 'package:stats':  
  
 chisq.test, fisher.test

library(ggplot2)  
library(readxl)  
library(ggpubr)

Registered S3 methods overwritten by 'broom':  
 method from   
 nobs.fitdistr MuMIn  
 nobs.multinom MuMIn

library(dplyr)  
  
sst <- read\_csv("data/SST\_update2023.csv") #read in problem 2 data

Rows: 291384 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): site  
dbl (3): latitude, longitude, temp  
date (1): date  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data <- read\_csv("data/occdist.csv") #read in problem 3 data

Rows: 227 Columns: 10  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (4): box, box occupant, eventID, repeated use  
dbl (6): edge distance, season, sp, cs, e, tm  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Problem 1

### a)

In part 1, the co-worker likely used a correlation analysis or a linear regression . The reported p-value of 0.03 suggests they were testing whether the slope of the relationship between the two variables is significantly different from zero.

In part 2, the co-worker might have used a one-way ANOVA (Analysis of Variance). The p-value of 0.02 suggests that at least one group has a significantly different mean nitrogen load compared to the others.

### b)

1. A post-hoc test like Tukey’s Honest Significant Difference (HSD) could be performed to determine which specific nitrogen sources differ significantly from each other. Tukey’s HSD would allow you to tell whether fertilizer sources have significantly higher nitrogen loads than atmospheric deposition or grasslands, which is crucial for making management recommendations.
2. Including a table or summary statistic that reports the mean nitrogen load and standard deviation for each source would help someone understand the variability of nitrogen contributions from each source. Providing descriptive statistics gives context to the ANOVA results and helps audiences grasp which sources are contributing the most nitrogen on average and know that there is more than just a difference.

### c)

1. As th distance from the headwater increases, there is a statistically significant corelation with the annual total nitrogen load (kg year-1), which suggests a possible pattern in water contamination like nutritional build up or runoff. (Statistical test: Pearson correlation; r = [correlation coefficient], p = 0.03, α = 0.05)
2. Through our tests we observed a significant variation in average nitrogen load across different sources, meaning that some sources may contribute more to nutritional pollution and contamination than others. (Statistical test: ANOVA; F = (F stat), df = (deg freedom), p = 0.02, α = 0.05)

## Problem 2

### a)

#cleaning data  
sst\_clean <- sst|>   
 select(date, temp) |> #get rid of long and lat  
 mutate(  
 year = year(date), #extract year  
 month = month(date, label = TRUE) #extract month  
 ) |>   
   
 group\_by(year, month) |> #group by year and month  
 summarise(mean\_monthly\_sst = mean(temp, na.rm = TRUE)) |> #calculate monthly sst  
   
 mutate(  
 year = as.factor(year),  
 month = factor(month, levels = month.abb, ordered = TRUE) #order months in correct order  
 ) |>   
 ungroup()

`summarise()` has grouped output by 'year'. You can override using the  
`.groups` argument.

sst\_clean |>   
 slice\_sample(n = 5) #showing 5 random rows of cleaned data

# A tibble: 5 × 3  
 year month mean\_monthly\_sst  
 <fct> <ord> <dbl>  
1 2021 Aug 17.7  
2 1987 Aug 17.3  
3 1984 Jan 14.9  
4 2020 Jun 15.9  
5 2016 Feb 15.0

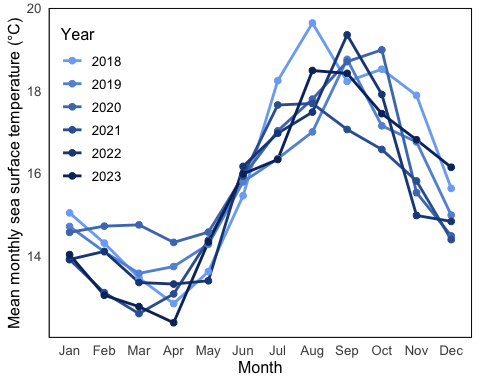
str(sst\_clean) #show the requested structure

tibble [504 × 3] (S3: tbl\_df/tbl/data.frame)  
 $ year : Factor w/ 42 levels "1982","1983",..: 1 1 1 1 1 1 1 1 1 1 ...  
 $ month : Ord.factor w/ 12 levels "Jan"<"Feb"<"Mar"<..: 1 2 3 4 5 6 7 8 9 10 ...  
 $ mean\_monthly\_sst: num [1:504] 13.1 13.5 13.3 12.9 14.4 ...

### b)

sst\_subset <- sst\_clean %>% #use clean data  
 filter(year %in% c("2018", "2019", "2020", "2021", "2022", "2023")) #only want 2018-2023  
  
#creating plot  
ggplot(data = sst\_subset, aes(x = month, y = mean\_monthly\_sst, group = year, color = year)) +  
 geom\_line(linewidth = 1) + #size of connecting lines  
 geom\_point(size = 2) + #size of dots  
 scale\_color\_manual(  
 values = colorRampPalette(c("#7cadf5", "#08306b"))(6) #blue gradient  
 ) +  
 labs(  
 x = "Month", #x axis title  
 y = "Mean monthly sea surface temperature (°C)", #y axis title  
 color = "Year" #legend title  
 ) +  
 theme\_minimal(base\_size = 12) +  
 theme(  
 legend.position = c(0.1, 0.7), #legend on the top left  
 legend.background = element\_rect(fill = "white", color = "white"), #no border on table of contents  
 panel.border = element\_rect(color = "black", fill = NA), #outline of box in graph  
 panel.background = element\_rect(fill = "white"), #background color is white  
 panel.grid.major = element\_blank(), #remove grid lines  
 panel.grid.minor = element\_blank(), #remove grid lines  
 axis.title = element\_text(size = 12), #size of axis title  
 axis.text = element\_text(size = 10), #size of axis text  
 legend.title = element\_text(size = 12), #size of legend title  
 legend.text = element\_text(size = 10) #size of legend text internal  
 )

Warning: A numeric `legend.position` argument in `theme()` was deprecated in ggplot2  
3.5.0.  
ℹ Please use the `legend.position.inside` argument of `theme()` instead.



## Problem 3

### a)

Biologically the 1’s and 0’s in this data set represent what bird, if any, is in the nest box during the given breeding season. Depending on which column you look at 1 means there is a Swift Parrot, Tree Martin, Common Starling, or no bird.

### b)

Swift Parrots are a critically endangered species while tree Martins and Common Starlings are more common and considered competition for different recources.

### c)

The ‘seasons’ they are talking about are the 2016 and 2019 summer breeding seasons when parrots bred at the study area. These seasons differ because a lot can change in 3 years, such as, average global temperature, management practices, weather patterns, populations of different supporting or threatening species, and increased or decreased compettition if other populations change over the years.

### d. Table of Models

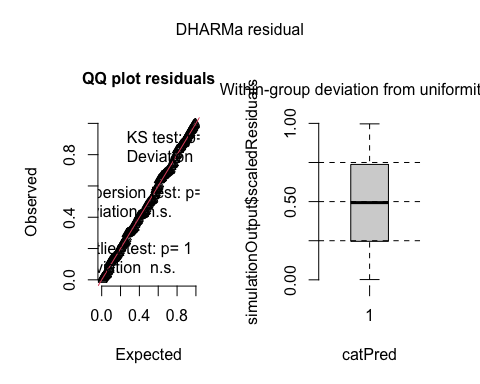
| Model # | Season Included | Distance from Edge Included | Model Description |
| --- | --- | --- | --- |
| 1 | No | No | Null model |
| 2 | Yes | Yes | Saturated model |
| 3 | Yes | No | Season model |
| 4 | No | Yes | Distance model |

### e)

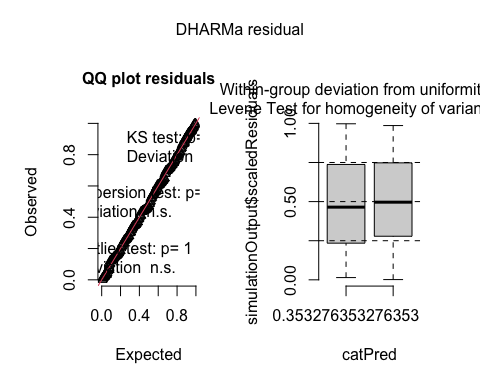
data <- data %>%  
 rename(edge\_distance = `edge distance`) %>%  
 mutate(season = as.factor(season)) #cleaning data  
  
model1 <- glm(sp ~ 1, data = data, family = "binomial") #intercept only null model  
  
model2 <- glm(sp ~ season + edge\_distance, data = data, family = "binomial") #sason and edge saturated model  
  
model3 <- glm(sp ~ season, data = data, family = "binomial") #season only  
  
model4 <- glm(sp ~ edge\_distance, data = data, family = "binomial") # distance only

### f)

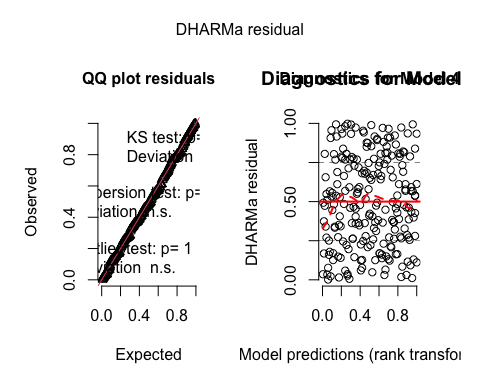
simulation1 <- simulateResiduals(model1)  
plot(simulation1, main = "Diagnostics for Model 1") #Model 1 diagnostics



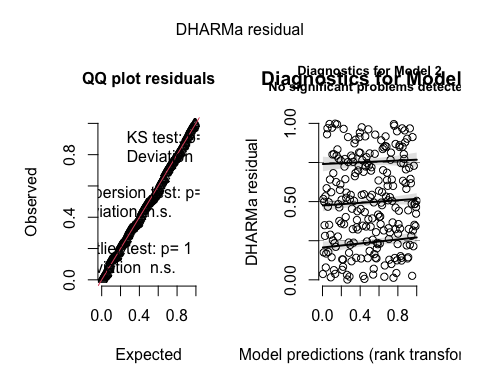
simulation3 <- simulateResiduals(model3)  
plot(simulation3, main = "Diagnostics for Model 3") #Model 3 diagnostics



simulation4 <- simulateResiduals(model4)  
plot(simulation4, main = "Diagnostics for Model 4", quantreg = FALSE) #Model 4 diagnostics



simulation2 <- simulateResiduals(model2)  
plot(simulation2, main = "Diagnostics for Model 2") #Model 2 diagnostics



### g)

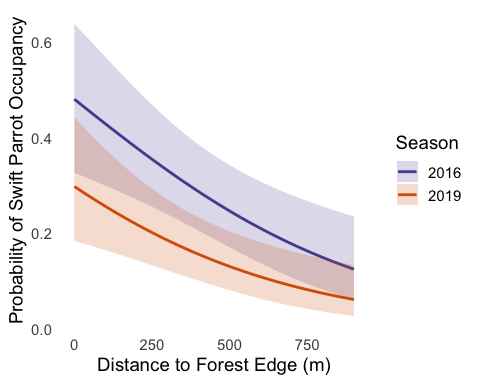
model.sel(model1, model2, model3, model4)

Model selection table   
 (Int) edg\_dst ssn df logLik AICc delta weight  
model2 -0.07749 -0.002078 + 3 -110.103 226.3 0.00 0.837  
model4 -0.50620 -0.001967 2 -112.809 229.7 3.36 0.156  
model3 -0.95020 + 2 -116.160 236.4 10.06 0.005  
model1 -1.29000 1 -118.407 238.8 12.52 0.002  
Models ranked by AICc(x)

The best model as determined by Akaike’s Information Criterion (AIC) was model2. This model had the lowest AICc value (226.3), meaning it has the balance of goodness-of-fit and model simplicity compared to the other models.

### h)

new\_data <- expand.grid(  
 edge\_distance = seq(0, 900, by = 10),  
 season = factor(c(2016, 2019), levels = levels(data$season)) #new dataframe for edge distances during both seasons  
)  
  
  
preds <- predict(model2, newdata = new\_data, type = "link", se.fit = TRUE) #Get predicted probabilities and confidence intervals  
new\_data$fit <- preds$fit  
new\_data$se <- preds$se.fit  
  
new\_data <- new\_data %>%  
 mutate(prob = plogis(fit),  
 lower = plogis(fit - 1.96 \* se),  
 upper = plogis(fit + 1.96 \* se)) #Convert from logit to probability scale  
  
  
ggplot(new\_data, aes(x = edge\_distance, y = prob, color = factor(season))) +  
 geom\_line(linewidth = 1) +  
 geom\_ribbon(aes(ymin = lower, ymax = upper, fill = factor(season)), alpha = 0.2, color = NA) +  
 labs(  
 x = "Distance to Forest Edge (m)", #x axis title  
 y = "Probability of Swift Parrot Occupancy", #y axis title  
 color = "Season",   
 fill = "Season" #legend title  
 ) +  
 scale\_color\_manual(values = c("#5b4d9f", "#d95f02")) +  
 scale\_fill\_manual(values = c("#5b4d9f", "#d95f02")) +  
 theme\_minimal(base\_size = 14) +  
 theme(panel.grid = element\_blank())



### i)

**Figure 1. Predicted Probability of Swift Parrot Nest Box Occupancy by Distance to Forest Edge**  
This figure shows the predicition of probability of the Swift Parrot occupancy in nest boxes in corelation to its distance to the forrest’s edge during the breeding season 2016 (purple) and 2019 (pink). The sahded regions represent 95% confidence intervals and the lines represent the predicted trend. These values are predicted based on logistic regression model with both season and edge distance as predictors. **Data source**: Stojanovic et al. (2021). Dryad. <https://doi.org/10.5061/dryad.83bk3j9sb>

### j)

pred\_points <- data.frame( #new data frame  
 edge\_distance = c(0, 0, 900, 900), #distance  
 season = factor(c(2016, 2019, 2016, 2019), levels = levels(data$season)) #seasons  
)  
  
pred\_results <- predict(model2, newdata = pred\_points, type = "link", se.fit = TRUE) #getting predictions for model 2  
  
pred\_points$fit <- pred\_results$fit #predicted log odds  
pred\_points$se <- pred\_results$se.fit #se of log odds  
  
pred\_points <- pred\_points %>%  
 mutate(  
 prob = plogis(fit), #predicted probability  
 lower = plogis(fit - 1.96 \* se), #lower 95%ci  
 upper = plogis(fit + 1.96 \* se) #upper 95%ci  
 )  
  
pred\_points #viewable output

edge\_distance season fit se prob lower upper  
1 0 2016 -0.07749021 0.3293209 0.48063714 0.32674169 0.6382966  
2 0 2019 -0.85668874 0.3205975 0.29803163 0.18466437 0.4431673  
3 900 2016 -1.94788877 0.3926047 0.12478375 0.06195475 0.2353430  
4 900 2019 -2.72708730 0.4374157 0.06139379 0.02700329 0.1335704

### k)

The probability predicted for the Swift Parrot nest box occupancy is higher when the distance to forest edge was (0m) compared to (900m) in both seasons. It was found that in 2016, the probability at 0m was 48% while in 2019 it was 30% and as we move to the 900m mark those percents drop to 12% and 6%. In Figure1, the negative relationship between distance and occupancy probibility is shown when both lines have a clear downward trend. THis pattern is most likely due to the Swift Parrots’ habitual preference for food availability, and or animal communities found at the 0m mark which may change or diminish as you move to larger distances.

## Problem 4

### a)

The Homework 3 graph is an effective bar chart, using color and categorical grouping (sleep and protein levels) to emphasize mood differences. It also compares multiple variables at one time to find the optimal mix. The Homework 2 visualizations are more straightforward: There is a Q-Q plot with confidence intervals, used to check normality, a box plot showing distributions and outliers of PM2.5 levels by location, and a time series line plot comparing PM2.5 levels over time across multiple locations. These graphs focus on data exploration and distribution rather than emotional storytelling.

All the graphs in Homework 2 and 3 have a clean look to them, including an axis title to help the view interpret the trends and differences. Each plot involves comparisons from 2 or more categories whether that be location or conditions. They also use individual data points to show the variability.

In the Homework 3 graph, we can see a pattern of a higher mood score when more protein is consumed and more sleep is had, showing a pattern of better well being with these factors. In the Homework 2 visuals, the Q-Q plot shows deviation from the normal in the tails, most likely due to PM2.5 outliers. The box plot shows a higher median and more extreme outliers, and the line plot shows higher PM2.5 levels over time, especially in the start of December. We can see these differences arise because the Homework 3 graph is emphasizing categorical interactions while the Homework 2 graphs are exploring distribution and variation in environmental trends.

Some Feedback I got in week 9 was to make sure my visuals are clear and readable, and tell the story I am trying to tell. I implemented this in my Homework 3 graph by pairing a bar chart and a scatter plot, so you can see 2 things at once without getting confused at what is happening. I was also told to be careful with my placement of things like the legend and titles so that they don’t overlap. I implemented this in my Homework 3 as well, making sure my visual was perfect and professional before turning it in.