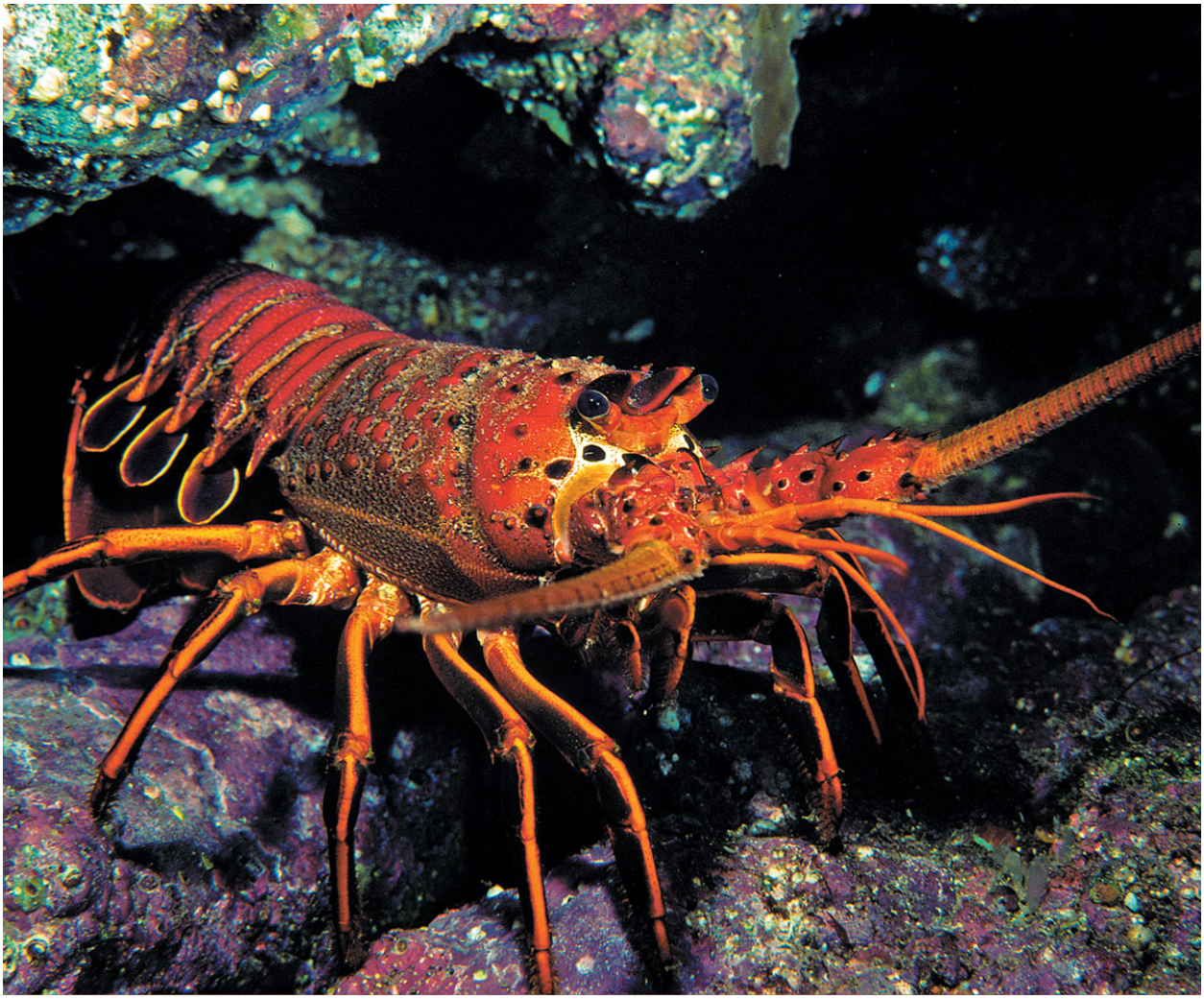
Assignment 1: California Spiny Lobster Abundance (*Panulirus Interruptus*)

Assessing the Impact of Marine Protected Areas (MPAs) at 5 Reef Sites in Santa Barbara County

EDS 241

1/8/2024 (Due 1/26)



Assignment instructions:

- Working with partners to troubleshoot code and concepts is encouraged! If you work with a partner, please list their name next to yours at the top of your assignment so Annie and I can easily see who collaborated.
- All written responses must be written independently (**in your own words**).
- Please follow the question prompts carefully and include only the information each question asks in your submitted responses.
- Submit both your knitted document and the associated `RMarkdown` or `Quarto` file.
- Your knitted presentation should meet the quality you'd submit to research colleagues or feel confident sharing publicly. Refer to the rubric for details about presentation standards.

Assignment submission (YOUR NAME): _____Naomi___Moraes_____

```
library(tidyverse)
library(here)
library(janitor)
library(estimatr)
library(performance)
library(jtools)
library(gt)
library(gtsummary)
library(MASS) ## NOTE: The `select()` function is masked. Use: `dplyr::select()` ##
library(interactions)
library(ggplot2)
library(ggridges)
library(gtsummary)
library(ggbeeswarm)
```

DATA SOURCE: Reed D. 2019. SBC LTER: Reef: Abundance, size and fishing effort for California Spiny Lobster (Panulirus interruptus), ongoing since 2012. Environmental Data Initiative. https://doi.org/10.6073/pasta/a593a675d644fdefb736750b291579a0. Dataset accessed 11/17/2019.

Introduction

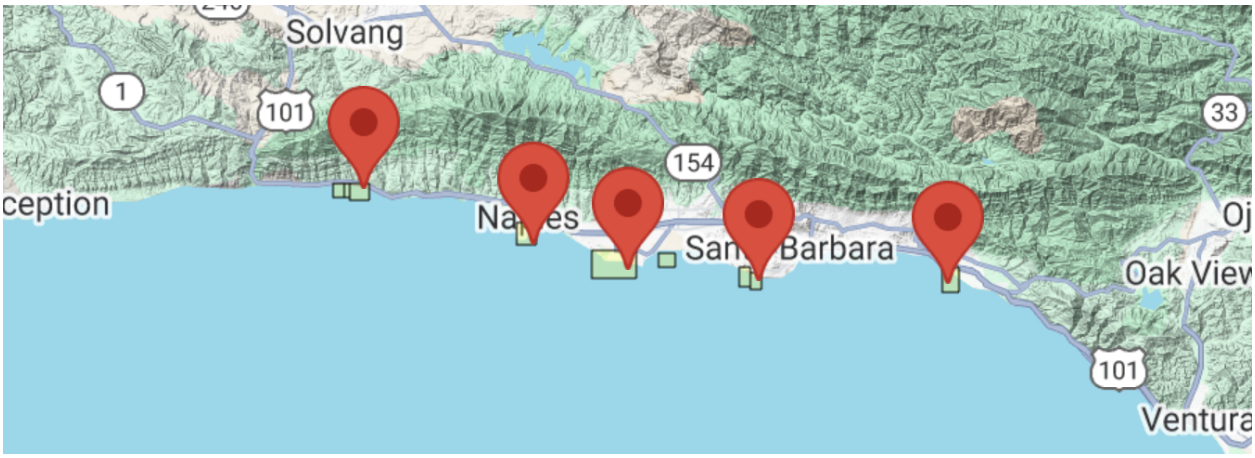
You're about to dive into some deep data collected from five reef sites in Santa Barbara County, all about the abundance of California spiny lobsters! Data was gathered by divers annually from 2012 to 2018 across Naples, Mohawk, Isla Vista, Carpinteria, and Arroyo Quemado reefs.

Why lobsters? Well, this sample provides an opportunity to evaluate the impact of Marine Protected Areas (MPAs) established on January 1, 2012 (Reed, 2019). Of these five reefs, Naples, and Isla Vista are MPAs, while the other three are not protected (non-MPAs). Comparing lobster health between these protected

and non-protected areas gives us the chance to study how commercial and recreational fishing might impact these ecosystems.

We will consider the MPA sites the `treatment` group and use regression methods to explore whether protecting these reefs really makes a difference compared to non-MPA sites (our control group). In this assignment, we'll think deeply about which causal inference assumptions hold up under the research design and identify where they fall short.

Let's break it down step by step and see what the data reveals!



Step 1: Anticipating potential sources of selection bias

a. Do the control sites (Arroyo Quemado, Carpenteria, and Mohawk) provide a strong counterfactual for our treatment sites (Naples, Isla Vista)? Write a paragraph making a case for why this comparison is centris paribus or whether selection bias is likely (be specific!).

I believe that the control sites, of Arroyo Quemado, Carpenteria, and Mohawk, do prove a sufficient counter factual for the treatment sites of Naples and Isla Vista. This is because geographically, the sites are similar leading to similar conditions of the sites (e.g. in temperature, initial lobster population, surrounding species, etc.) Thus data from the treatment sites, ceteris paribus, should provide a robust analysis of the post-experiment differences and similarities between the two.

Step 2: Read & wrangle data

a. Read in the raw data. Name the data.frame (`df`) `rawdata`

b. Use the function `clean_names()` from the `janitor` package

```
# HINT: check for coding of missing values (`na = "-99999"`)

rawdata <- rawdata <- read_csv(here("data", "spiny_abundance_sb_18.csv"), na = "-99999") %>%
    clean_names()
```

c. Create a new `df` named `tidyata`. Using the variable `site` (reef location) create a new variable `reef` as a `factor` and add the following labels in the order listed (i.e., re-order the `levels`):

```
"Arroyo Quemado", "Carpenteria", "Mohawk", "Isla Vista",  "Naples"
```

```
tidydata <- rawdata %>%
    mutate(reef = factor(site,
                         levels = c("AQUE", "CARP", "MOHK", "IVEE", "NAPL"),
                         labels = c("Arroyo Quemado", "Carpenteria", "Mohawk", "Isla Vista",  "Naples"))
```

Create new `df` named `spiny_counts`

d. Create a new variable `counts` to allow for an analysis of lobster counts where the unit-level of observation is the total number of observed lobsters per `site`, `year` and `transect`.

- Create a variable `mean_size` from the variable `size_mm`
- NOTE: The variable `counts` should have values which are integers (whole numbers).
- Make sure to account for missing cases (`na`)!

e. Create a new variable `mpa` with levels `MPA` and `non_MPA`. For our regression analysis create a numerical variable `treat` where MPA sites are coded `1` and non_MPA sites are coded `0`

```
#HINT(d): Use `group_by()` & `summarize()` to provide the total number of lobsters observed at each sit

#HINT(e): Use `case_when()` to create the 3 new variable columns

spiny_counts <- tidydata %>%
    group_by(site, year, transect) %>%
    summarise(count = sum(count, na.rm = TRUE), mean_size = mean(size_mm, na.rm = TRUE)) %>%
    mutate(mpa = case_when(site %in% c("IVEE", "NAPL") ~ "MPA", .default = "non_MPA")) %>%
    mutate(treat = case_when(mpa == "MPA" ~ 1, .default = 0)) %>%
    ungroup()
```

> NOTE: This step is crucial to the analysis. Check with a friend or come to TA/instructor office hours to make sure the counts are coded correctly!

Step 3: Explore & visualize data

a. Take a look at the data! Get familiar with the data in each `df` format (`tidydata`, `spiny_counts`)

b. We will focus on the variables `count`, `year`, `site`, and `treat`(`mpa`) to model lobster abundance. Create the following 4 plots using a different method each time from the 6 options provided. Add a layer (`geom`) to each of the plots including informative descriptive statistics (you choose; e.g., mean, median, SD, quartiles, range). Make sure each plot dimension is clearly labeled (e.g., axes, groups).

- Density plot
- Ridge plot
- Jitter plot
- Violin plot
- Histogram
- Beeswarm

Create plots displaying the distribution of lobster **counts**:

1) grouped by reef site

2) grouped by MPA status
3) grouped by year

Create a plot of lobster **size** :

4) You choose the grouping variable(s)!

```
# plot 1: Jitter plot grouped by reef site
jitter_spiny <-  ggplot(data = spiny_counts , aes(x = site, y = count)) +
    geom_boxplot(width = 0.3, outliers = FALSE) +
    geom_jitter(fill = "darkblue",
                alpha = 0.4) +
    xlab("California Spiny Lobster Count") +
    ylab("Reef Site") +
    labs(title = "Jitter plot of California Spiny Lobster Counts by Reef Site")

jitter_spiny
```

```
# Plot 2: Violin plot of counts grouped by MPA status
violin_spiny <- ggplot(data = spiny_counts, aes(y = count, x = mpa)) +
    geom_boxplot(width = 0.3, outliers = FALSE) +
    geom_violin(width = 0.3, color = "darkblue", alpha = 0.5) +
    xlab("MPA Status") +
    ylab("California Spiny Lobster Count") +
    labs(title = "Violin Plot of California Spiny Lobster Counts by MPA status") +
    theme_minimal()

violin_spiny
```

```
# Plot 3: Beeswarm Grouped by year
bees_spiny <- ggplot(data = spiny_counts, aes(x = factor(year), y = count)) +
    geom_beeswarm() +
    stat_summary(fun.y = mean, geom = "point", color = "hotpink", size = 3, alpha = 0.7)  +
    xlab("California Spiny Lobster Count") +
    ylab("Year") +
    labs(title = "Beeswarm plot California Spiny Lobster Counts by Year & Yearly Mean Counts") +
    theme_minimal()

bees_spiny
```

```
# Plot 4: Density Plot of lobster size
density_spiny <- ggplot(data=spiny_counts, aes(x = mean_size, fill = mpa)) +
    geom_density(alpha = 0.75) +
    scale_fill_manual(values = c("hotpink2", "cornflowerblue")) +
    theme_classic() +
    xlab("California Spiny Lobster Mean Size (mm)") +
    ylab("Density") +
    labs(title = "Density Plot of CA Spiny Lobster Size by MPA Status") +
    geom_vline(xintercept = mean(spiny_counts$mean_size, na.rm = TRUE), color = "black") +
    geom_label(aes(x = 80,
                   y = 0.05,
                   label = paste("Lobster Mean Size = 74 mm")),
```

```
                  size = 3,
                  show.legend = FALSE)

density_spiny
```

c. Compare means of the outcome by treatment group. Using the `tbl_summary()` function from the package `gt_summary`

```
# USE: gt_summary::tbl_summary()
 spiny_counts %>%
    dplyr::select(count, mean_size, mpa) %>%
    tbl_summary(by = mpa) %>%
    modify_caption("**Comparing CA Spiny Lobster mean counts and sizes between MPA and non-MPA sites**")
```

Step 4: OLS regression- building intuition

a. Start with a simple OLS estimator of lobster counts regressed on treatment. Use the function `summ()` from the `jtools` package to print the OLS output

b. Interpret the intercept & predictor coefficients *in your own words*. Use full sentences and write your interpretation of the regression results to be as clear as possible to a non-academic audience.

At the non-MPA designated reef sites, there are approximately 23 CA spiny lobsters. At MPA designated reef sites, there are approximately 6 more CA Spiny Lobsters.

```
# NOTE: We will not evaluate/interpret model fit in this assignment (e.g., R-square)

m1_ols <- lm(count ~ treat,
             data = spiny_counts)

summ(m1_ols, model.fit = FALSE)
```

c. Check the model assumptions using the `check_model` function from the `performance` package

d. Explain the results of the 4 diagnostic plots. Why are we getting this result?

```
check_model(m1_ols,  check = "qq" )
```

This 'qq' plot allows us to observe that the residuals are not normally distributed, violating a key assumption of OLS. There may be better models for our data, than an OLS model.

```
check_model(m1_ols, check = "normality")
```

The above figure plots the distribution of our residuals - further confirming that they are not normally distributed in nature. Here, we can observe that the plot us unimodal, with a tail to the right.

```
check_model(m1_ols, check = "homogeneity")
```

The figure above shows us that the plotted residuals do not have constant variance, which violates another OLS assumption.

```
check_model(m1_ols, check = "pp_check")
```

In the figure above, we can observe that the observed data is not a good fit to our expected model output. This further confirms our belief that the OLS model is not the most appropriate model to fit our data.

Step 5: Fitting GLMs

a. Estimate a Poisson regression model using the `glm()` function

```
    #HINT1: Incidence Ratio Rate (IRR): Exponentiation of beta returns coefficient which is interpreted

#HINT2: For the second glm() argument `family` use the following specification option `family = poisson

# Estimate Poisson regression model
m2_pois <- glm(count ~ treat,
               family = poisson(link = "log"),
               data = spiny_counts)

summ(m2_pois, model.fit = FALSE)
```

b. Interpret the predictor coefficient in your own words. Use full sentences and write your interpretation of the results to be as clear as possible to a non-academic audience.

The estimated Poisson regression coefficient comparing the MPA to the Non-MPA sites, given the other variables are held constant in the model, is 0.21. This indicates that the percent of lobsters in the MPA (treated) areas is 0.21 times more than in the non-treated areas.

c. Explain the statistical concept of dispersion and overdispersion in the context of this model.

The Poisson regression model assumes that variance (or the dispersion) us proportional to the mean of the data. If overdispersion exists, than the variance is greater than the mean.

d. Compare results with previous model, explain change in the significance of the treatment effect

In the previous OLS model, we observed that the were more lobsters in the MPA sites than in the Non-MPA sites. (Analyzing the coefficients, we assumed that the MPA area had approximately 6 more lobsters than in the Non-MPA areas, or a 23.58% increase in the amount of lobsters.)

```
(28.09 - 22.73)/22.73
```

In the Possion regression model, we can observe that the increase is less than stated in the OLS model - 21%. The new model shows a slightly decreased impact of the treatment effect.

e. Check the model assumptions. Explain results.

```
check_model(m2_pois)
```

Checking the model assumptions above, it seems that the Poisson model does not provide the best fit for the data either. In the pp-check, observed data does not align with the predicted values which is mirrored in the zero-inflation check, where the residual variance does not follow the predicted. The residual values are also not normally distributed, as seen in the qq plot.

f. Conduct tests for over-dispersion & zero-inflation. Explain results.

```
check_overdispersion(m2_pois)
```

As the dispersion ratio is far greater than one, this model does not seem like a good fit to the data. The p-value of < 0.001 indicates that that the null_hypothesis of correct dispersion is false, thus we reject it in favour of the alternate hypothesis (that there is not correct dispersion). (We can conlcude that that the variance is most likely greater than the mean)

```
check_zeroinflation(m2_pois)
```

The amount of observed 0s (27) is greather than the amount of predicted 0s (0). The model is underfitting the 0s, thus indicating zero-inflation in the data.

g. Fit a negative binomial model using the function glm.nb() from the package `MASS` and check model diagnostics

h. In 1-2 sentences explain rationale for fitting this GLM model.

Negative binomial regression is a form of a Poisson regression, which loosens the assumption of variance-mean equivalency. We observed overdispersion in the previous test, indicating the necessity of a different model.

i. Interpret the treatment estimate result in your own words. Compare with results from the previous model.

In this model, there are 0.21 times (21%) more lobsters in our treatment (MPA) group than our control (non-MPA) group. This effect is the same that we observed in the previous model.

```
# NOTE: The `glm.nb()` function does not require a `family` argument

m3_nb <- glm.nb(count ~ treat,
                data = spiny_counts)

# Print negative binomial model output
summ(m3_nb, model.fit = FALSE)
```

```
check_overdispersion(m3_nb)
```

As the p-value is 0.088 and the dispersion ratio is greater than 0.05 - no overdispersion is detected.

```
check_zeroinflation(m3_nb)
```

The observed zeros (27) is slightly less than the predicted zeros (30). The model is overfitting the zeros ($p = 0.600$).

```
check_predictions(m3_nb)
```

```
check_model(m3_nb)
```

Step 6: Compare models

a. Use the `export_summ()` function from the `jtools` package to look at the three regression models you fit side-by-side.

c. Write a short paragraph comparing the results. Is the treatment effect `robust` or stable across the model specifications.

In the OLS model, there are about 6 more lobsters in MPA sites than non-MPA sites (a 23.58% increase). In the Poisson and the negative binomial model, there is a 21% increase in lobsters in MPA sites. The treatment effect is stable across the model specifications, as the treatment does not impact the potential outcome of other subjects.

```
export_summs(m1_ols, m2_pois, m3_nb,
             model.names = c("OLS","Poisson", "NB"),
             statistics = "none")
```

Step 7: Building intuition - fixed effects

a. Create new `df` with the `year` variable converted to a factor

b. Run the following negative binomial model using `glm.nb()`

- Add fixed effects for `year` (i.e., dummy coefficients)
- Include an interaction term between variables `treat` & `year` (`treat*year`)

c. Take a look at the regression output. Each coefficient provides a comparison or the difference in means for a specific sub-group in the data. Informally, describe the what the model has estimated at a conceptual level (NOTE: you do not have to interpret coefficients individually)

The model estimates the impact that the treatment has on lobster counts for each year, from 2012 - 2018. It estimates the mean count in both the treatment and the control groups, for each year.

d. Explain why the main effect for treatment is negative? *Does this result make sense?

The coefficient for treat represents the difference in lobster counts in MPA sites ansd non-MPA sites in 2012. As it is negative (-1.72), the model estimates that there were less lobsters in MPA sites than non-MPA sites in 2012. Seeing that MPA sites were not designated until 2012, it is possible that we would not the impact of the established sites immediately (or even until 2014 as the coefficient was negative in 2013 as well).

```
ff_counts <- spiny_counts %>%
    mutate(year=as_factor(year))

m5_fixedeffs <- glm.nb(
    count ~
        treat +
        year +
        treat*year,
    data = ff_counts)

summ(m5_fixedeffs, model.fit = FALSE)
```

e. Look at the model predictions: Use the `interact_plot()` function from package `interactions` to plot mean predictions by year and treatment status.

f. Re-evaluate your responses (c) and (b) above.

The plot aligns with my previous responses. The model is estimating the treatment effect on lobster counts in MPA sites vs. non-MPA sites. The negative coefficients show us that in 2012 and 2013, there were less lobsters in the MPA sites than in the non-MPA sites. This changes over time until there are more lobsters in MPA sites than non-MPA sites as of 2018.

```
# HINT: Change `outcome.scale` to "response" to convert y-axis to counts

interact_plot(m5_fixedeffs, pred = year, modx = treat,
              outcome.scale = "response")  # NOTE: y-axis on log-scale
```

g. Using `ggplot()` create a plot in same style as the previous `interaction plot`, but displaying the original scale of the outcome variable (lobster counts). This type of plot is commonly used to show how the treatment effect changes across discrete time points (i.e., panel data).

The plot should have... - `year` on the x-axis - `counts` on the y-axis - `mpa` as the grouping variable

```
# Hint 1: Group counts by `year` and `mpa` and calculate the `mean_count`
# Hint 2: Convert variable `year` to a factor

spiny_counts_plot <- spiny_counts %>%
    group_by(year, mpa) %>%
    summarise(mean_count = mean(count, na.rm = TRUE)) %>%
    mutate(year = as_factor(year)) %>%
    ungroup()

plot_counts <- ggplot(data = spiny_counts_plot, aes(x = year, y = mean_count, group = mpa, color = mpa,
    geom_point(size = 3) +
    geom_line(size = 1) +
    scale_color_manual(values = c("#1D4670", "#A9D1F0" ),
                       labels = c("MPA", "Non-MPA")) +
    scale_linetype_manual(values = c("solid", "longdash"),
                          labels = c("MPA", "Non-MPA")) +
    theme_minimal() +
    labs(title = "Mean lobster counts by MPA designation (2012 - 2018)",
         color = "MPA designation",
         linetype = "MPA designation") +
    xlab("Year") +
    ylab("Mean Lobster Count")


plot_counts
```

Step 8: Reconsider causal identification assumptions

a. Discuss whether you think `spillover effects` are likely in this research context (see Glossary of terms; https://docs.google.com/document/d/1RIudsVcYhWGpqC-Uftk9UTz3PIq6stVyEpT44EPNgpE/edit?usp=sharing)

" Spillover specifically refers to the possibility that one unit's treatment affects a control unit's outcome." (EDS241) I believe that spillover effects are possible in this experiment - from the MPA sites to the Non-MPA sites. This is becuase as the lobsters are free to rom, they may go over the confines of the iste boundaries and impact the populations of lobsters in the other sites. (Secondary impacts of other species populations may spillover/cause a feedback loop for the control sites.) As the sites are geographically close - the lobsters could travel from a Non-MPA site to a MPA site feasibly. However, looking at the plot above of mean lobster counts, the difference in lobster populations in the 2017 - 2018 years, for both sites, might indicate that any spillover impacts were a non-issue.

b. Explain why spillover is an issue for the identification of causal effects

In the identification of causal effects, the treatment and control groups should be truely independent of each other. However, spillover impacts of the treatment group influencing the control group outcomes - may muddy what impacts of the experiment could be considered causal effects.

c. How does spillover relate to impact in this research setting?

In this research setting, through a pure lens of study - spillover may cause false or exaggerated identifications of causal effect. However, in reality this spillover hekps to bolster the biodiversity of lobster populations outside of the treatment zone - which is a positive secondary outcome as it may indicate that areas adjacent to MPA zones may receive the tangential benefits of the treament areas.

d. Discuss the following causal inference assumptions in the context of the MPA treatment effect estimator. Evaluate if each of the assumption are reasonable:

1) SUTVA: Stable Unit Treatment Value assumption

SUTVA implies that MPA treatment effects are applied equally to all lobsters within the treatment zones (furthermore, lobsters in the control groups did not recieve any impact). If there were spillover effects, this assumption would be violated; however, I believe that the assumption is reasonable to hold.

`2) Excludability assumption`

The excludability assumption implies that the MPA treatment is the only thing having a causal effect on the outcome of lobster populations in the area. In this experiment, I believe that there are extrinsic factors (environmental, policy, etc.), that will impact the lobster populations in the non-MPA and MPA sites. However, as the aras are close together, this extrinisic impacts might have s "zero-sum" impact, as it effects all areas equally. (So that comparisons between the non-MPA sites and MPA sites are still valid.) Thus, I believe thatexcludability is a reasonable assumption.

EXTRA CREDIT

Use the recent lobster abundance data with observations collected up until 2024 (`lobster_sbchannel_24.csv`) to run an analysis evaluating the effect of MPA status on lobster counts using the same focal variables.

a. Create a new script for the analysis on the updated data
b. Run at least 3 regression models & assess model diagnostics
c. Compare and contrast results with the analysis from the 2012-2018 data sample (~ 2 paragraphs)