

# Assignment 1

California Spiny Lobster (*Panulirus Interruptus*): Assessing the Impact of Marine Protected Areas (MPAs) at 5 Reef Sites in Santa Barbara County

EDS 241 / ESM 244 (Due: 1/17)

1/8/26



---

## 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 Richard Montes Lemus: \_\_\_\_\_

```
# Load in libraries
library(tidyverse)
library(here)
library(janitor)
library(estimatr)
library(performance)
library(jtools)
library(gt)
library(gtsummary)
library(interactions)
```

## 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. Data accessed 11/17/2019.

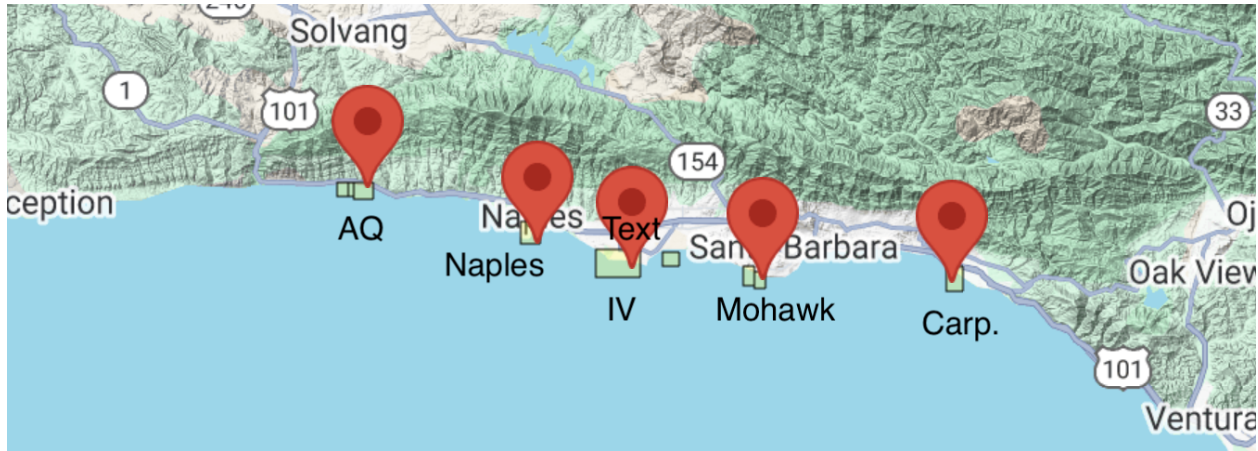
## Introduction

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 *ceteris paribus* or whether selection bias is likely (be specific!).

I do not think the control sites provide a strong counterfactual for our treatment sites. While all of these sites are relatively close, there are some major differences that might cause selection bias. For example, Mohawk and Carp are located in very urban areas. IV and Naples are relatively less urban and are either right next to undeveloped land or located in undeveloped land. This could mean California spiny lobsters are less common in Mohawk and Carp because of run off pollution from urbanized areas - not necessarily because they lack an MPA designation. Including AQ in the control group does balance this out more since it is located in undeveloped land that likely has little run off, but I don't think it should be used to offset the effect of pollution in the urban sites. It would be better to just choose three sites for the control group that are equally as close as possible to Naples and IV.

**Step 2: Read & wrangle data** a. Read in the raw data from the "data" folder named `spiny_abundance_sb_18.csv`. Name the data.frame `rawdata`

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

```
# HINT: check for coding of missing values (`na = "-99999"`)
# Turn column names into snake_case and replace -99999 with na
rawdata <- read_csv(here("data", "spiny_abundance_sb_18.csv"), na = "-99999") %>%
  clean_names()
```

c. Create a new df named `tidydata`. 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"

```
# Use refactoring to match complete site name to abbreviation and to order site names for plotting
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`.

```
spiny_counts <- tidydata %>%  
  group_by(site, year, transect, reef) %>%  
  summarise(counts = sum(count, na.rm = TRUE), mean_size = mean(size_mm, na.rm = TRUE)) %>% ungroup()
```

- 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 site*

*#HINT(e): Use `case\_when()` to create the 3 new variable columns*

```
spiny_counts <- spiny_counts %>%  
  mutate(  
    mpa = case_when(  
      site == "AQUE" ~ "non_MPA",  
      site == "CARP" ~ "non_MPA",  
      site == "IVEE" ~ "MPA",  
      site == "MOHK" ~ "non_MPA",  
      site == "NAPL" ~ "MPA"  
    ),  
    mpa = factor(mpa, levels = c("non_MPA", "MPA")),  
    treat = case_when(  
      mpa == "MPA" ~ 1,  
      mpa == "non_MPA" ~ 0  
    )  
  )
```

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).

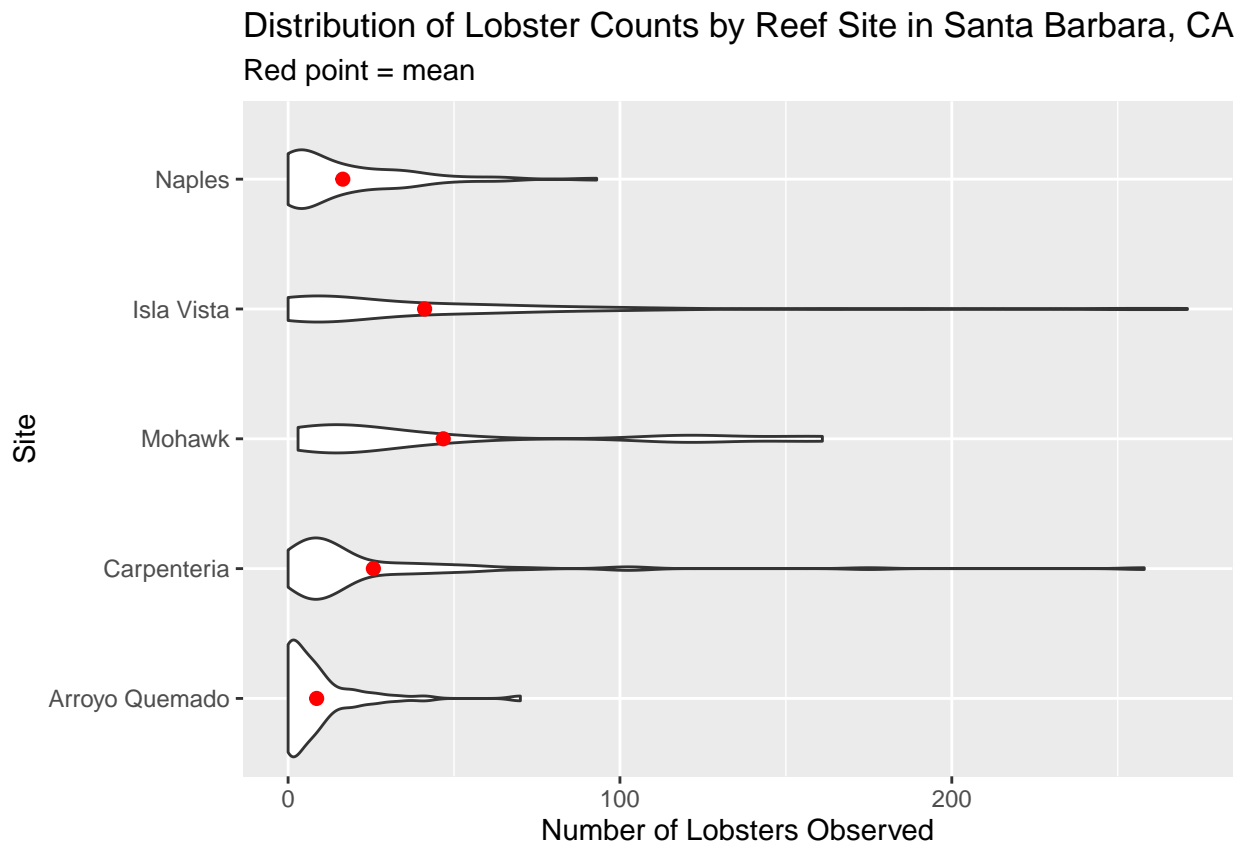
- Density plot
- Ridge plot
- Jitter plot
- Violin plot
- Histogram
- Beeswarm

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

1) grouped by reef site



```
spiny_counts %>%
  ggplot(aes(x = counts, y = reef)) +
  geom_violin() +
  stat_summary(fun = mean, geom = "point", color = "red", size = 2) +
  labs(title = "Distribution of Lobster Counts by Reef Site in Santa Barbara, CA",
        subtitle = "Red point = mean",
        x = "Number of Lobsters Observed",
        y = "Site")
```



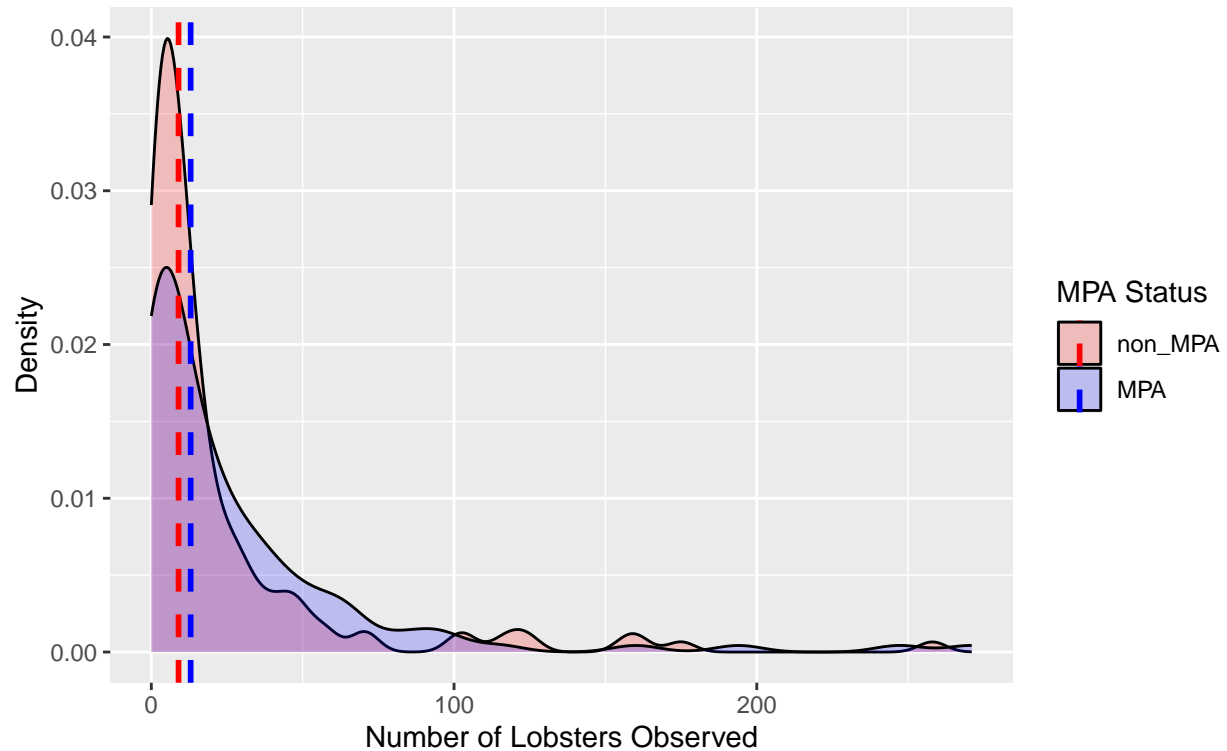
2) grouped by MPA status

```
spiny_counts %>%
  ggplot(aes(x = counts, fill = mpa)) +
  geom_density(alpha = 0.2) +
  geom_vline(data = spiny_counts %>%
    group_by(mpa) %>%
    summarize(median_counts = median(counts)),
    aes(xintercept = median_counts, color = mpa),
    linetype = "dashed", linewidth = 1) +
  scale_fill_manual(values = c("red", "blue")) +
  scale_color_manual(values = c("red", "blue")) +
  labs(title = "Distribution of Lobster Counts by MPA Status in Santa Barbara, CA",
        subtitle = "Dashed lines show median",
        x = "Number of Lobsters Observed",
        y = "Density",
        fill = "MPA Status",
```

```
color = "MPA Status")
```

## Distribution of Lobster Counts by MPA Status in Santa Barbara, CA

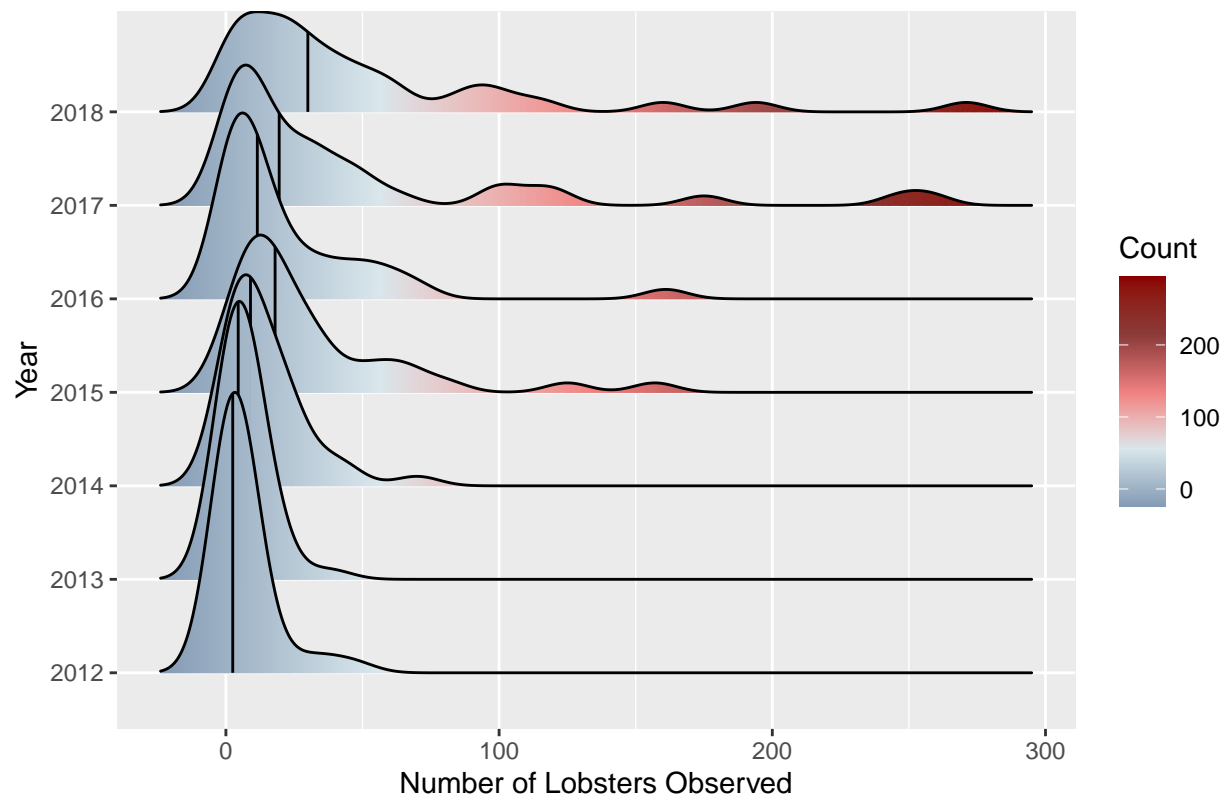
Dashed lines show median



3) grouped by year

```
spiny_counts %>%
  ggplot(aes(x = counts, y = factor(year), fill = after_stat(x))) +
  ggridges::geom_density_ridges_gradient(rel_min_height = 0,
                                         scale = 3,
                                         quantile_lines = TRUE,
                                         quantiles = 2 ) +
  scale_fill_gradientn(colors = c("#849BB4", "#D9E7EC", "#EF8080", "#8B3A3A", "darkred")) +
  labs(title = "Distribution of Lobster Counts by Year in Santa Barbara, CA",
       x = "Number of Lobsters Observed",
       y = "Year",
       fill = "Count")
```

Distribution of Lobster Counts by Year in Santa Barbara, CA



Create a plot of lobster **size** :

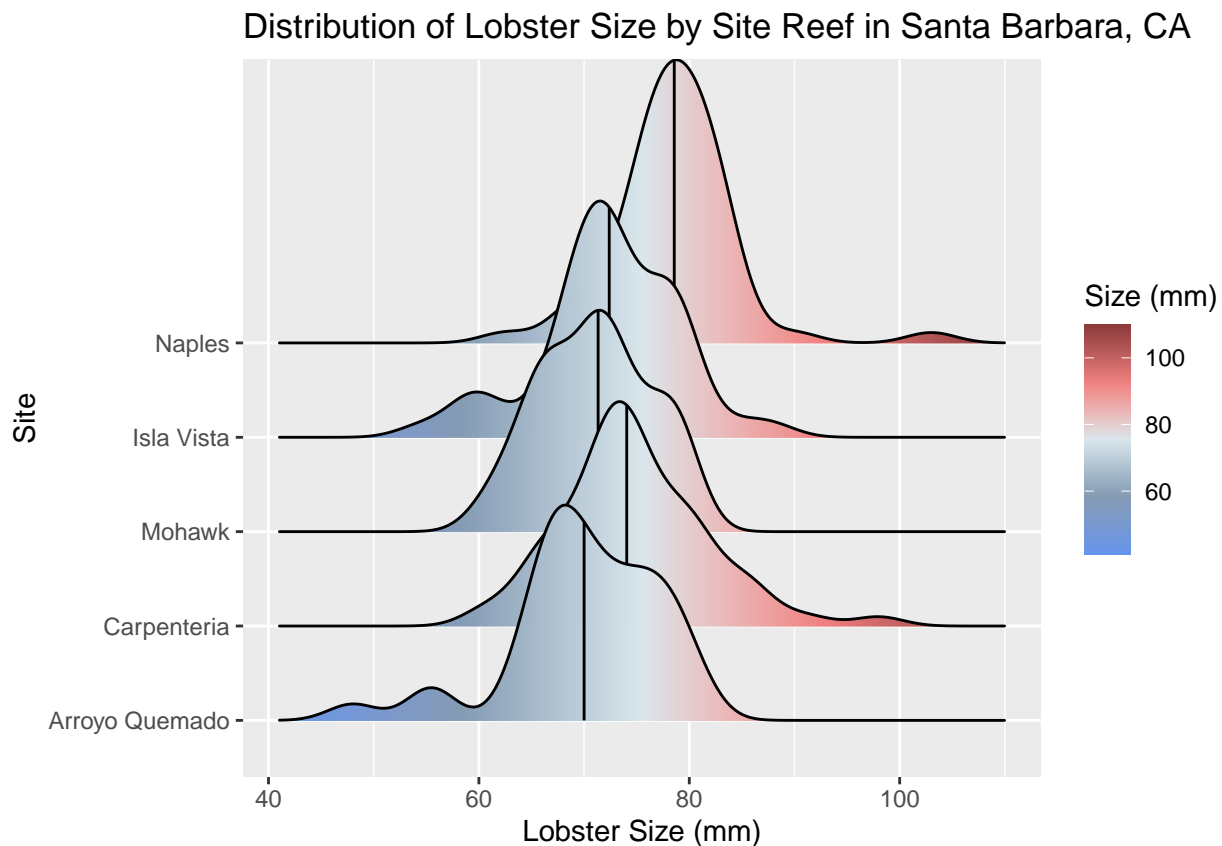
4) You choose the grouping variable(s)!

```
# plot 1: ....
spiny_counts %>%
  ggplot(aes(x = mean_size, y = reef, fill = after_stat(x))) +
  ggribes::geom_density_ridges_gradient(rel_min_height = 0,
                                        scale = 3,
                                        quantile_lines = TRUE,

                                        quantiles = 2 ) +
  scale_fill_gradientn(colors = c("cornflowerblue", "#849BB4", "#D9E7EC", "#EF8080", "#8B3A3A")) +
  labs(title = "Distribution of Lobster Size by Site Reef in Santa Barbara, CA",
       x = "Lobster Size (mm)",
       y = "Site",
       fill = "Size (mm)")
```

Characteristic	non_MPA N = 133 <sup>1</sup>	MPA N = 119 <sup>1</sup>
counts	23 (39)	28 (44)
mean_size	73 (7)	76 (7)
Unknown	15	12

<sup>1</sup> Mean (SD)



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 %>%
  tbl_summary(by = mpa, include = c(counts, mean_size), statistic = list(all_continuous() ~ "{mean} ({SD})"))
```

**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.

Intercept: Without an MPA designation, we expect the lobster count to be 22.73 on average per transect.

Treatment Coefficient: Giving a site an MPA designation increased the average lobster count by approximately 5.36 per transect.



Percent change calculation:

$$(5.36 / 22.73) * 100 = 23.58\%$$

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

```
m1_ols <- lm(counts ~ mpa, data = spiny_counts)
```

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

Observations	252
Dependent variable	counts
Type	OLS linear regression

	Est.	S.E.	t val.	p
(Intercept)	22.73	3.57	6.36	0.00
mpaMPA	5.36	5.20	1.03	0.30

Standard errors: OLS

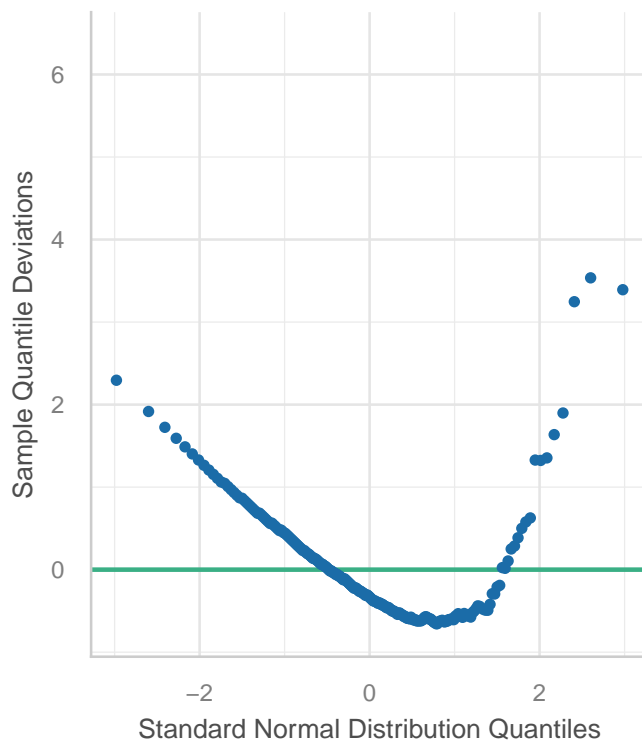
- c. Check the model assumptions using the `check_model` function from the `performance` package
- d. Explain the results of the 4 diagnostic plots. Why are we getting this result?

The four diagnostic plots show we did not use the correct model for this analysis. Our quantile quantile plot shows residuals did not fall along the line and therefore our data does not follow a normal distribution. The second plot also shows residuals do not follow a normal distribution. The homogeneity of variance plot shows that the standard deviation of residuals varies based on lobster count predicted - is it not constant. And the posterior predictive check plot shows lobster counts predicted by the model do not match the same concentration as the observed lobster data. We are getting these results because an OLS model is not the right model for our data, the misalignments we see from these checks show there are many violations to the OLS model's assumptions. An OLS model assumes the outcome variable is continuous and can take any value from negative infinity to positive infinity, with residuals that are normally distributed and have constant variance. This is not the case for our outcome variable, we have discrete count data, it has to be positive, and we don't have a normal distribution for residuals or constant variance. We should explore a different model.

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

## Normality of Residuals

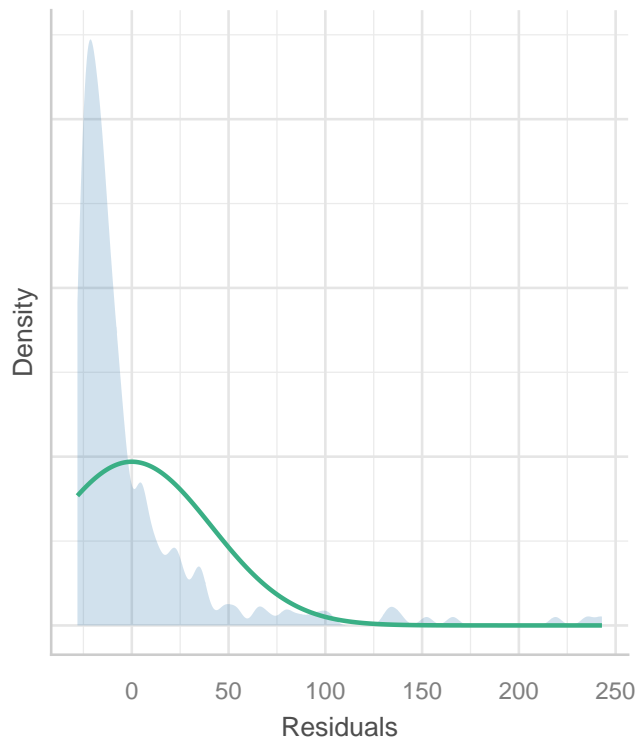
Dots should fall along the line



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

## Normality of Residuals

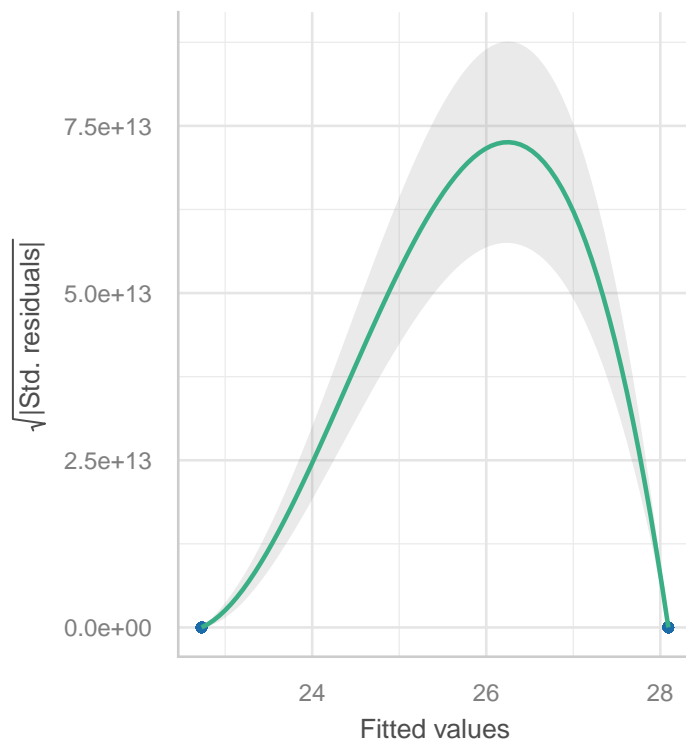
Distribution should be close to the normal curve



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

## Homogeneity of Variance

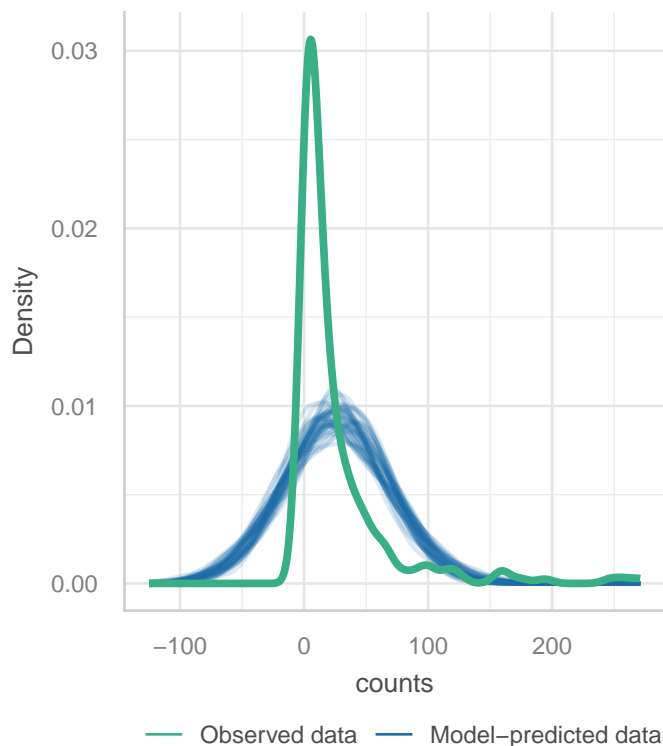
Reference line should be flat and horizontal



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

## Posterior Predictive Check

Model-predicted lines should resemble observed data line



**Step 5: Fitting GLMs** a. Estimate a Poisson regression model using the `glm()` function

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.

Intercept: On average, the model estimates that transects in non-MPA areas have approximately 22.6 lobsters.

Treatment Coefficient: The model estimates that the MPA designation increased lobster counts by 23.36% in transects within MPAs.

Calculation:

Intercept:  $(\exp(3.12))$

Treatment Coefficient:  $((\exp(0.21) - 1) * 100) = 23.36\%$

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

Dispersion refers to a ratio between the variance in our model and mean. If this ratio is equal to 1, this signals good model fit. For example, we would receive a dispersion ratio of 1 for this study if the average lobster count was 25 and the variance was also 25.

Overdispersion refers to an instance in which variance is greater than the mean, so the ratio between them is greater than 1. This signals bad model fit. We might expect that for this study since there is high variability between lobster transects. They are not distributed evenly and randomly, lobster populations tend to cluster around sites with optimal conditions. Therefore, high variance in lobster abundance between transects is expected.

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

The OLS model estimated a similar intercept (22.73 count vs 22.6 count) for the control group and a similar coefficient (23.58% vs 23.36%) for the treatment when compared to the poisson model.

The significance of the treatment effect was not significant for the OLS model, but it was significant for the poisson model.

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

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

```
m2_pois <- glm(counts ~ mpa, family = poisson(link = "log"), data = spiny_counts)

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

Observations	252
Dependent variable	counts
Type	Generalized linear model
Family	poisson
Link	log

	Est.	S.E.	z val.	p
(Intercept)	3.12	0.02	171.74	0.00
mpaMPA	0.21	0.03	8.44	0.00

Standard errors: MLE

e. Check the model assumptions. Explain results.

The posterior predictive check plot shows the predicted outcomes from the model do not match the observed outcomes well.

The misspecified dispersion and zero inflation plot shows the observed residual variance line to be much higher than the predicted residual variance line. This may indicate there is overdispersion.

The influential observations test plot shows there aren't any significant outliers for our residuals since they are within the contour lines.

The distribution of quantile residuals show they do not fall along the line, this may indicate overdispersion.

These model checks demonstrate a poisson model is also not the right model to predict lobster count outcomes given an MPA status.

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

The overdispersion check showed an extremely high dispersion ratio. This indicates there is overdispersion.

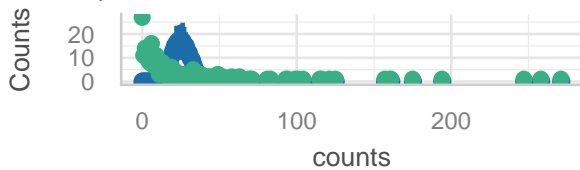
The zero inflation test indicated the model is not accounting for the variance coming from having so many zeros in our lobster count data set. We know this because the test shows we have a lot more zeros in our observed data than in what our model is able to predict. This signals there is zero inflation and our model can't account for the variance due to zeros. This causes our model to falsely gives us statistical significance and generally hinders our model's predictive ability.

```
check_model(m2_pois)
```



### Posterior Predictive Check

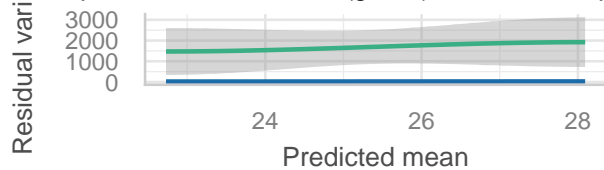
Model-predicted intervals should include observed data points



● Observed data ● Model-predicted data

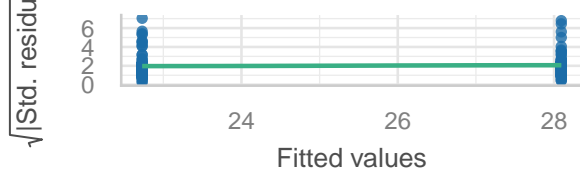
### Misspecified dispersion and zero-inflation

Observed residual variance (green) should follow pre



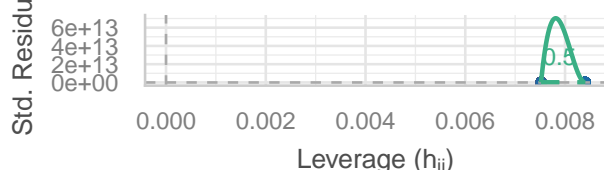
### Homogeneity of Variance

Reference line should be flat and horizontal



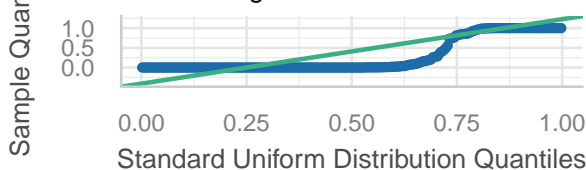
### Influential Observations

Points should be inside the contour lines



### Distribution of Quantile Residuals

Dots should fall along the line



```
check_overdispersion(m2_pois)
```

```
## # Overdispersion test
##
##      dispersion ratio =    67.033
##      Pearson's Chi-Squared = 16758.289
##      p-value =    < 0.001
```

```
check_zeroinflation(m2_pois)
```

```
## # Check for zero-inflation
##
##      Observed zeros: 27
##      Predicted zeros: 0
##      Ratio: 0.00
```

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.

I wanted to fit this GLM model because it includes a dispersion parameter to account for overdispersion and because it is more flexible than a poisson regression.

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

Treatment Coefficient: The model estimates that the MPA designation increased lobster counts by 23.36% in transects within MPAs.

Calculation:

Treatment Coefficient: :  $((\exp(0.21) - 1) * 100) = 23.36\%$

Comparison between negative binomial and poisson model treatment coefficients:

Both of these models have the same treatment coefficient, however, the negative binomial model has a p value of 0.22 while the poisson model has a p value of 0.00. This is because the poisson model did not account for zero inflation so it falsely gave statistical significance to the poisson model's treatment coefficient.

```
library(MASS) ## NOTE: The `select()` function is masked. Use: `dplyr::select()` ##
```

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

```
m3_nb <- glm.nb(counts ~ mpa, data = spiny_counts)
```

```
summ(m3_nb, model.fit = FALSE)
```

Observations	252
Dependent variable	counts
Type	Generalized linear model
Family	Negative Binomial(0.55)
Link	log

	Est.	S.E.	z val.	p
(Intercept)	3.12	0.12	26.40	0.00
mpaMPA	0.21	0.17	1.23	0.22

Standard errors: MLE

```
check_overdispersion(m3_nb)
```

```
## # Overdispersion test
```

```
##
```

```
## dispersion ratio = 1.398
```

```
## p-value = 0.088
```

```
check_zeroinflation(m3_nb)
```

```
## # Check for zero-inflation
```

```
##
```

```
## Observed zeros: 27
```

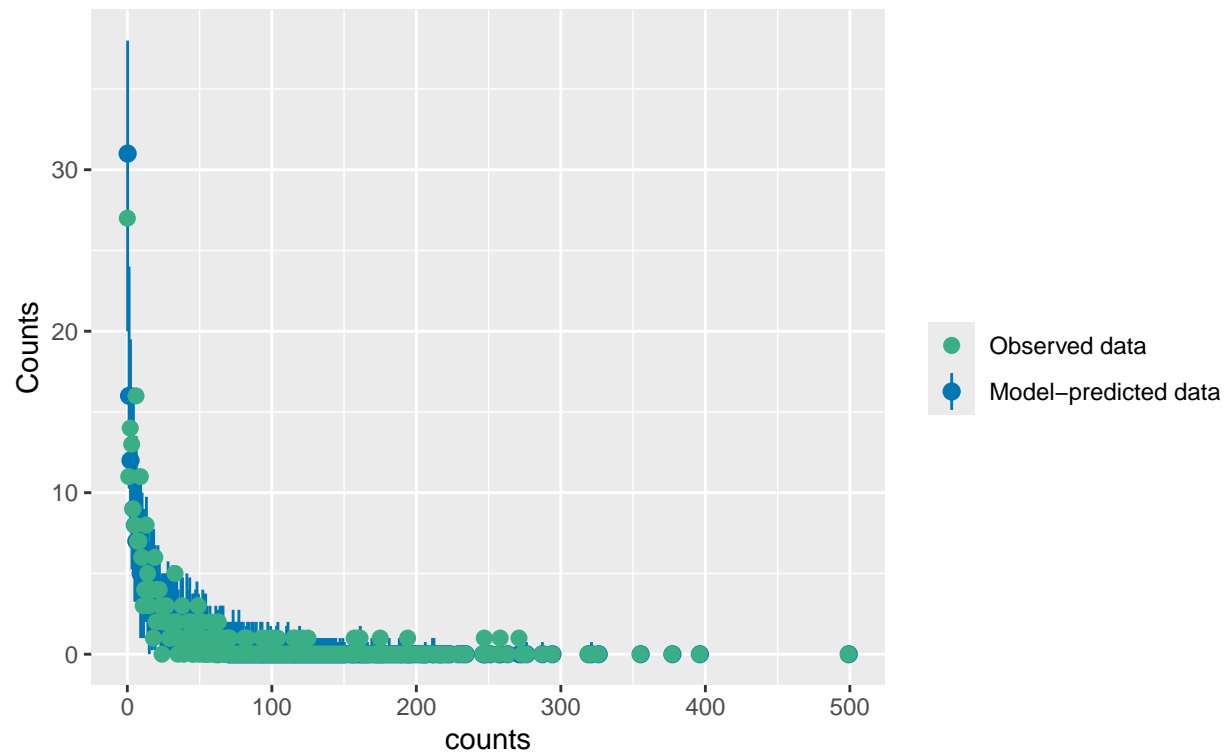
```
## Predicted zeros: 30
```

```
## Ratio: 1.12
```

```
check_predictions(m3_nb)
```

## Posterior Predictive Check

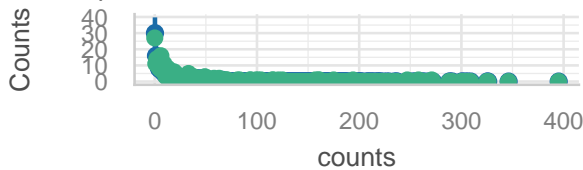
Model-predicted intervals should include observed data points



```
check_model(m3_nb)
```

### Posterior Predictive Check

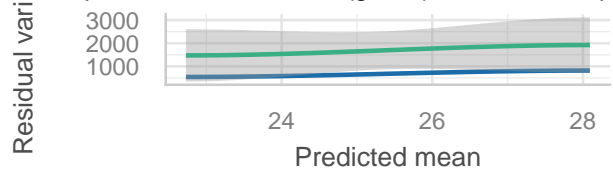
Model-predicted intervals should include observed data points



● Observed data ● Model-predicted data

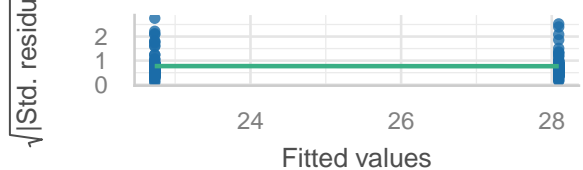
### Misspecified dispersion and zero-inflation

Observed residual variance (green) should follow predicted mean



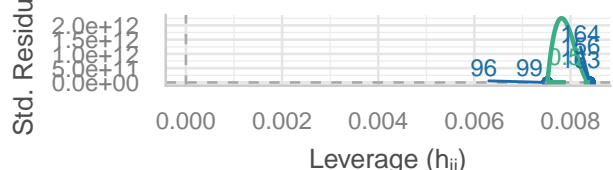
### Homogeneity of Variance

Reference line should be flat and horizontal



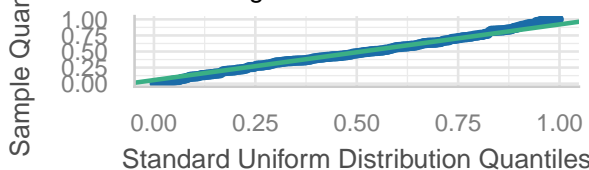
### Influential Observations

Points should be inside the contour lines



### Distribution of Quantile Residuals

Dots should fall along the line



**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.

Across all three models, the treatment effect is stable. MPAs are consistently associated with ~ 23% average increase in lobster count. The treatment effect's statistical significance, however, is not robust. It is statistically significant in the poisson model due to zero inflation and not statistically significant in the OLS and negative binomial model.

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

```
# m1_change = (5.36 / 22.73) * 100 # Change in OLS 23.58%
# m2_change = ((exp(0.21) - 1)*100) # Change in POIS 23.36%
# m3_change = ((exp(0.21) - 1)*100) = # Change in Neg Bi 23.36%
```

**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`)

	OLS	Poisson	NB
(Intercept)	22.73 ***	3.12 ***	3.12 ***
	(3.57)	(0.02)	(0.12)
mpaMPA	5.36	0.21 ***	0.21
	(5.20)	(0.03)	(0.17)

\*\*\* p < 0.001; \*\* p < 0.01; \* p < 0.05.

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 has estimated the impact of an MPA designation per year, allowing the model to capture the changing impact of an MPA designation on lobster counts over time. All interaction terms between the treatment and year are positive and statistically significant. This indicates the MPA's effect on higher lobster counts increased over time since MPAs are increasingly associated with higher lobster counts as years progress.

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

The main effect for the treatment is negative because this coefficient corresponds to the reference year of 2012, when the MPAs were first established. It is not surprising that MPAs had lower lobster abundance initially because these sites may have been chosen specifically as new MPAs because they were depleted and needed recovery. Furthermore, the effect of an MPA takes time, the lobster counts in these MPAs do not accurately reflect the impact of an MPA designation yet in 2012.

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

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

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

Observations	252
Dependent variable	counts
Type	Generalized linear model
Family	Negative Binomial(0.8129)
Link	log

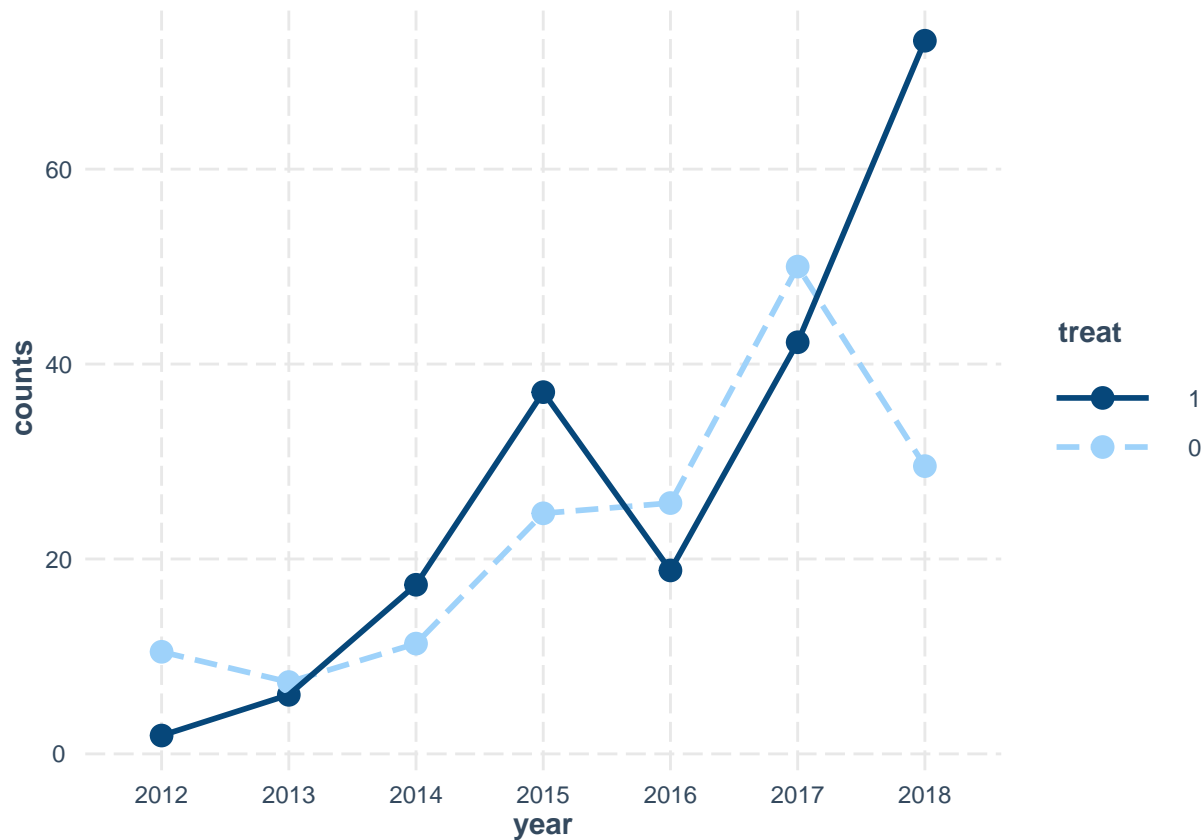
e. Look at the model predictions: Use the `interact_plot()` function from package `interactions` to plot mean predictions by year and treatment status.

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

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

	Est.	S.E.	z val.	p
(Intercept)	2.35	0.26	8.89	0.00
treat	-1.72	0.42	-4.12	0.00
year2013	-0.35	0.38	-0.93	0.35
year2014	0.08	0.37	0.21	0.84
year2015	0.86	0.37	2.32	0.02
year2016	0.90	0.37	2.43	0.01
year2017	1.56	0.37	4.25	0.00
year2018	1.04	0.37	2.81	0.00
treat:year2013	1.52	0.57	2.66	0.01
treat:year2014	2.14	0.56	3.80	0.00
treat:year2015	2.12	0.56	3.79	0.00
treat:year2016	1.40	0.56	2.50	0.01
treat:year2017	1.55	0.56	2.77	0.01
treat:year2018	2.62	0.56	4.69	0.00

Standard errors: MLE



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

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).

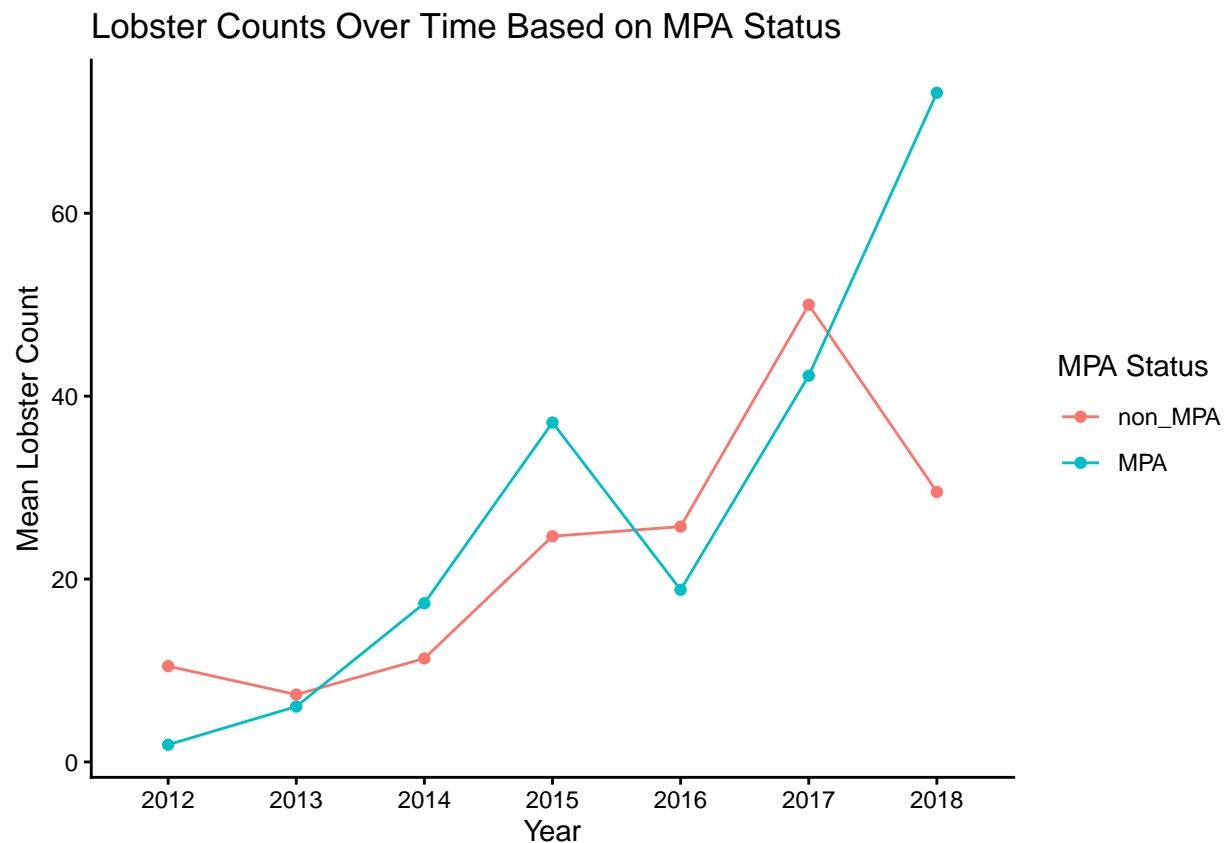
The plot should have... - **year** on the x-axis - **counts** on the y-axis - **mpa** as the grouping variable



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

```
plot_counts <- spiny_counts %>%
  group_by(year, mpa) %>%
  summarise(mean_count = mean(counts)) %>%
  ungroup() %>%
  mutate(year=as_factor(year))

plot_counts %>%
  ggplot(aes(x = year, y = mean_count,
             color = mpa,
             group = mpa)) +
  geom_line() +
  geom_point() +
  labs(x = "Year",
       y = "Mean Lobster Count",
       color = "MPA Status",
       title = "Lobster Counts Over Time Based on MPA Status") +
  theme_classic()
```



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>)

I definitely think “spillover effects” are likely in this research. There is no physical barrier between these MPA and non-MPA zones so the increase in lobster abundance in MPAs could likely produce more wandering lobsters that end up in non-MPA zones.

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

This is an issue for the identification of causal effects because it reduced the difference between the non-MPA group and MPA group means. This means we underestimated the true effect of an MPA designation on lobster abundance. We can't isolate the impact of an MPA designation on lobster abundance because our control (non-MPA designation) is impacted by the MPA designation as well.

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

Spillover makes it harder to estimate the true effect of an MPA designation on lobster abundance. Furthermore, it causes us to underestimate the true impact of an MPA designation.

- 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

1. The SUTVA assumption requires the MPA designation outcome to not affect another outcome from a different treatment or control.
2. The SUTVA assumption requires one standard version of the treatment.

The first SUTVA assumption is violated due to spillover effects. Increase in lobster abundance due to an MPA designation invariably causes nearby non-MPA sites to also have a higher lobster abundance due the lack of a barrier stopping lobsters in MPA sites from wandering into non-MPA sites.

The second SUTVA assumption is mostly met because there is one standard version of the treatment (MPA designation). Each MPA designation has the same law and protection, however, we can't be sure they were enforced in the same way.

- 2) Excludability assumption

The excludability assumption requires the MPA designation to only impact the lobster abundance because it stops overfishing of lobster.

This assumption is mostly met because that is the main impact of an MPA designation - limiting over-harvesting.

---

## EXTRA CREDIT

Use the recent lobster abundance data with observations collected up until 2024 (`extracredit_sblobstrs24.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)
-

