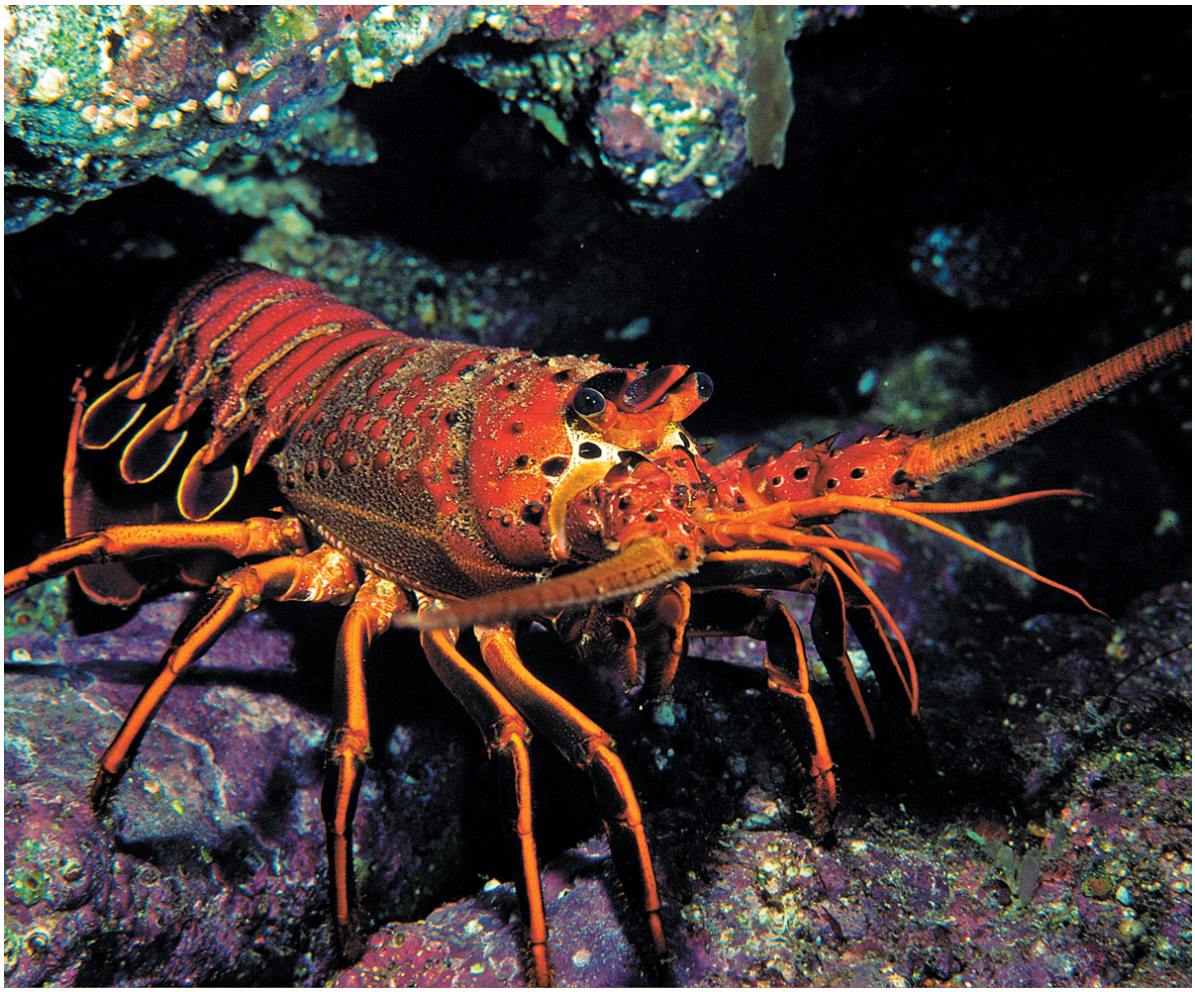
Assignment 1

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: 1/20**)

2026-01-08



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): Leela Dixit

```
library(tidyverse)
library(here)
library(janitor)
library(estimatr)
library(performance)
library(jtools)
library(gt)
library(gtsummary)
library(interactions)
library(ggridges)
```

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.

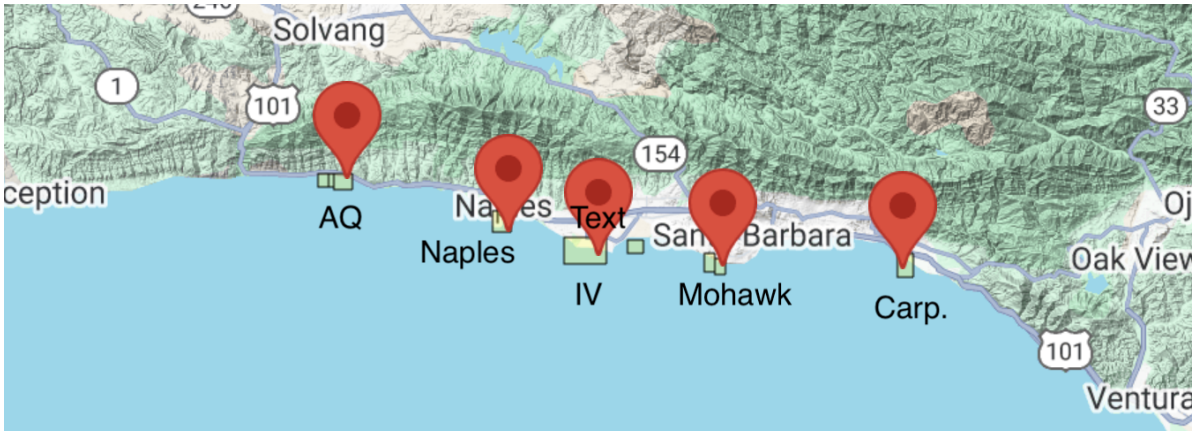
Introduction

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!).

We cannot investigate a true counterfactual of either site, as we cannot observe an instance of either site not as MPAs at the time frame of this experiment. The next best realistic counterfactual are non-MPA sites in the same area as our MPA sites. Researchers could have chosen sites as close as possible to our treatment sites, but ecologically we know processes like spillover effects would not make these control sites independent of our treatment. Likewise,

the control sites themselves are not close enough together for these ecological processes to eliminate independence. Visually, the control sites look are far enough from the treatments and each other, but not so far they are experiencing different oceanographic conditions to deem them incomparable, so I would argue these sites provide a sufficient comparison under ceteris paribus.

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"`)

rawdata <- read_csv(here("data", "spiny_abundance_sb_18.csv")) %>%
    clean_names() %>% # clean column names
    mutate(across(size_mm, na_if, -99999)) # transform -99999 values to NA
```

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"
```

```
# create new column of full site names
tidydata <- rawdata %>%
    # Add column of full names
    mutate(reef = case_when(site == "IVEE" ~ "Isla Vista",
                            site == "NAPL" ~ "Naples",
                            site == "MOHK" ~ "Mohawk",
                            site == "CARP" ~ "Carpenteria",
                            site == "AQUE" ~ "Arroyo Quemado")) %>%
    mutate(reef = as.factor(reef))

# re-order site names in the following order
levels = c("Arroyo Quemado", "Carpenteria", "Mohawk", "Isla Vista", "Naples")
tidydata <- tidydata %>%
    mutate(reef = fct_relevel(reef, levels))
```

```r
# check order is correct
levels(tidydata$reef)
```

```
[1] "Arroyo Quemado" "Carpenteria"    "Mohawk"         "Isla Vista"
[5] "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`

```r
#HINT(d): Use `group_by()` & `summarize()` to provide the total number of lobsters observed a

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

spiny_counts <- tidydata %>%
    group_by(site, year, transect) %>%
    summarize(counts = sum(count, na.rm = TRUE),
              mean_size = mean(size_mm, na.rm = TRUE)) %>%
    ungroup() %>%
    # designate MPA and non-MPA sites
    mutate(mpa = case_when(site == "IVEE" ~ "MPA",
                           site == "NAPL" ~ "MPA",
                           site == "MOHK" ~ "non_MPA",
                           site == "CARP" ~ "non_MPA",
                           site == "AQUE" ~ "non_MPA")) %>%
    # assign 0 and 1 for treatment
    mutate(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

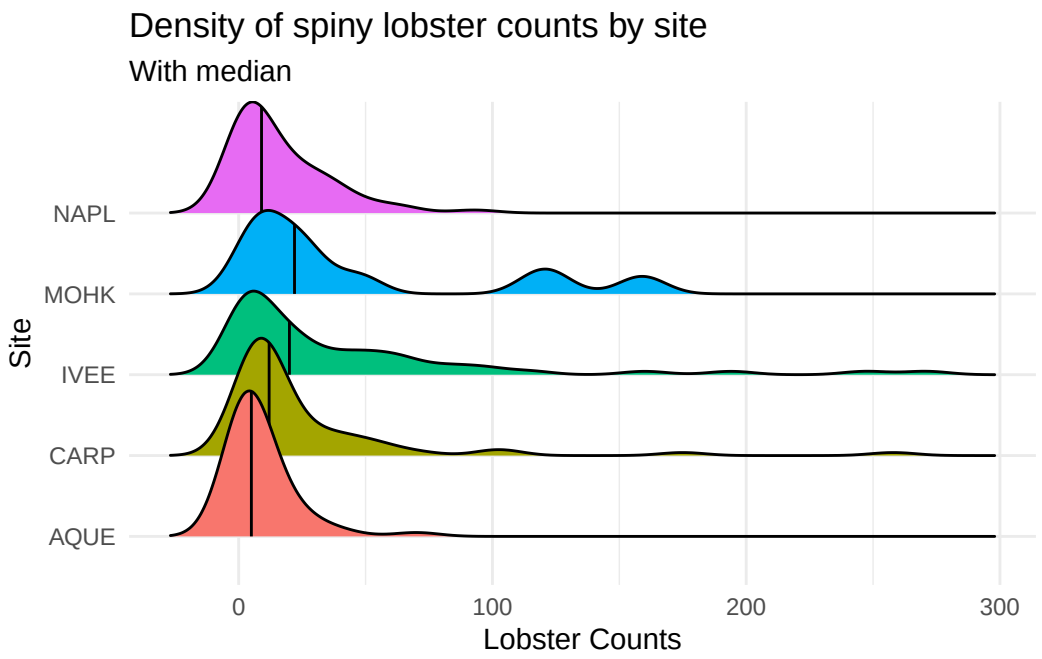
Create plots displaying the distribution of lobster **counts**:

1) grouped by reef site
2) grouped by MPA status
3) grouped by year

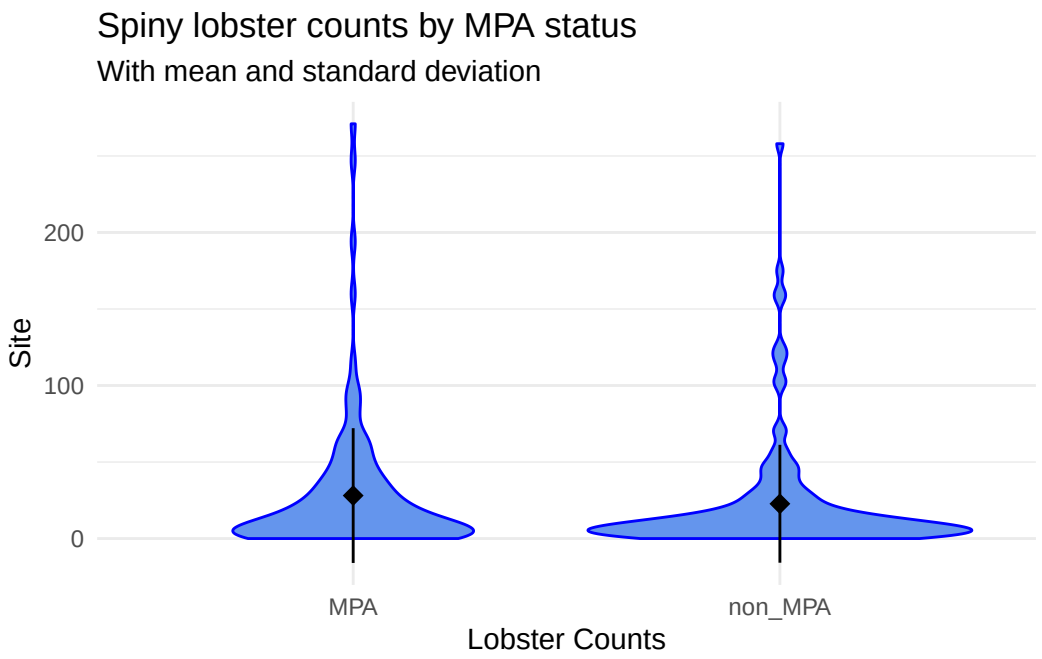
Create a plot of lobster **size** :

4) You choose the grouping variable(s)!

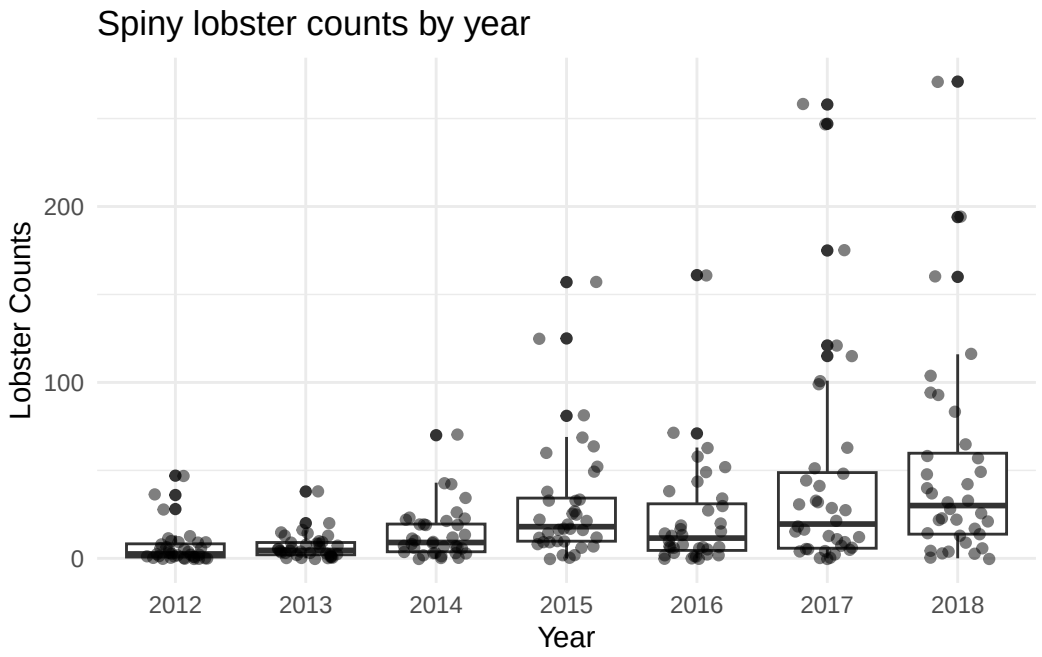
```
# plot 1: lobster counts grouped by reef site

spiny_counts %>%
    ggplot(aes(x = counts, y = site, fill = site)) +
    geom_density_ridges( # Add mean lines
        quantile_lines = TRUE,
        quantile_fun = function(x, ...) median(x)) +
    labs(x = "Lobster Counts",
         y = "Site",
         title = "Density of spiny lobster counts by site",
         subtitle = "With median") +
    theme_minimal() +
    theme(legend.position = "none")
```



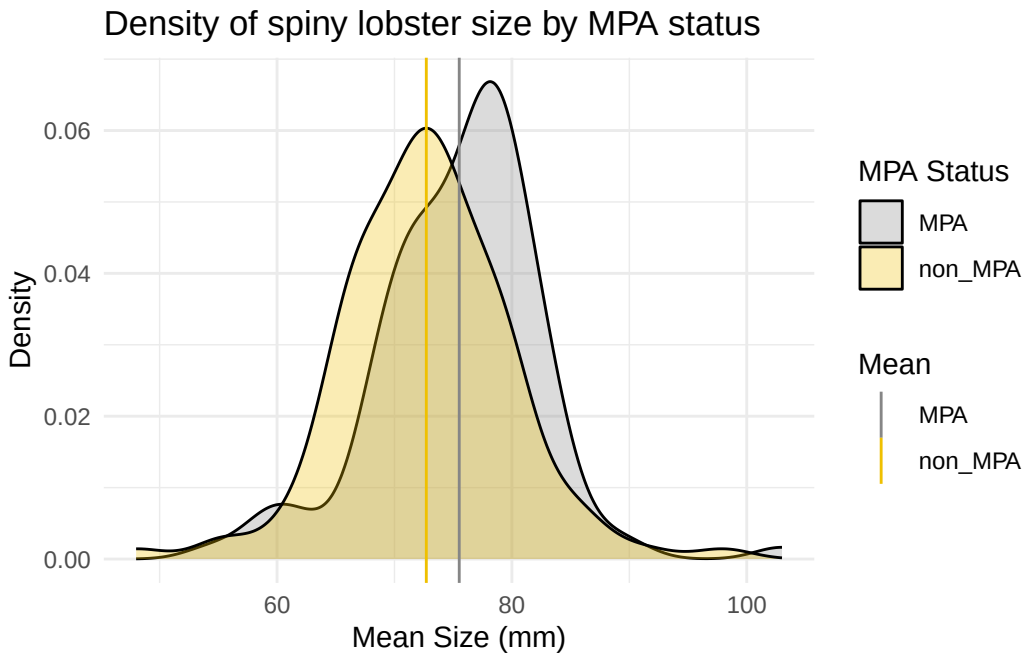

```
# plot 2: lobster counts grouped by MPA status

spiny_counts %>%
    ggplot(aes(x = mpa, y = counts)) +
    geom_violin(color = "blue", fill = "cornflowerblue") +
    labs(x = "Lobster Counts",
         y = "Site",
         title = "Spiny lobster counts by MPA status",
         subtitle = "With mean and standard deviation") +
    theme_minimal() +
    stat_summary(fun.data=mean_sdl, fun.args = list(mult = 1),
                 geom="pointrange", color="black", shape = 18, size = 0.75)
```




```
# plot 3: lobster counts grouped by year

spiny_counts %>%
    ggplot(aes(x = factor(year), y = counts)) +
    geom_boxplot() +
    geom_jitter(width = 0.25, alpha = 0.5) +
    labs(x = "Year",
         y = "Lobster Counts",
         title = "Spiny lobster counts by year") +
    theme_minimal()
```




```
# plot 4: lobster size grouped by mpa treatment
meansize <- spiny_counts %>%
    group_by(mpa) %>%
    summarise(mean = mean(mean_size, na.rm = TRUE)) %>%
    ungroup()

spiny_counts %>%
    ggplot(aes(x = mean_size)) +
    geom_density(aes(fill=factor(mpa)), size=0.5, alpha=0.3) +
    geom_vline(aes(xintercept = mean, color = mpa), data = meansize, linetype = "solid") +
    scale_color_manual(values = c("#868686FF", "#EFC000FF")) +
    scale_fill_manual(values = c("#868686FF", "#EFC000FF")) +
    labs(x = "Mean Size (mm)",
         y = "Density",
         title = "Density of spiny lobster size by MPA status",
         color = "Mean",
         fill = "MPA Status") +
    theme_minimal()
```



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

```
# USE: gtsummary::tbl_summary()

spiny_counts %>%
    gtsummary::tbl_summary(by = mpa,
                           include = c(counts, mean_size),
                           statistic = list(all_continuous() ~ "{mean} ({sd})"),
                           label = list(counts = "Lobster Abundance",
                                        mean_size = "Lobster Size")) %>%
    modify_header(label ~ "**Variable**") %>%
    modify_spanning_header(c("stat_1", "stat_2") ~ "**Reef Site**")
```

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

Variable	Reef Site	
	MPA N = 119[1]	**non_MPA** N = 133[1]
Lobster Abundance	28 (44)	23 (39)
Lobster Size	76 (7)	73 (7)
Unknown	12	15

[1] Mean (SD)

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.

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

m1_ols <- lm(counts ~ treat, 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
treat	5.36	5.20	1.03	0.30

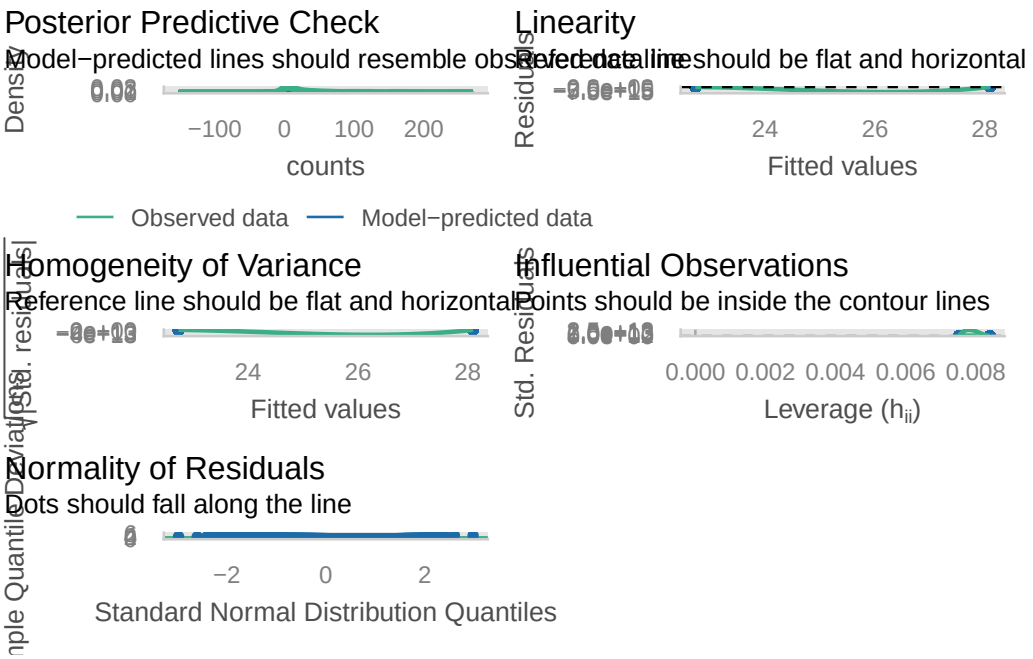
Standard errors: OLS

```
# pull out p-values
intercept_p_ols <- summary(m1_ols)$coefficients[1,4]
treat_p_ols <- summary(m1_ols)$coefficients[2,4]
```

Our model predicts there will be more spiny lobsters in reefs that are MPAs compared to reefs that are not MPAs. Based on our model, a non-MPA reef will have an estimated average of ~23 spiny lobsters, whereas an MPA reef will have an estimated average of ~27 spiny lobsters.

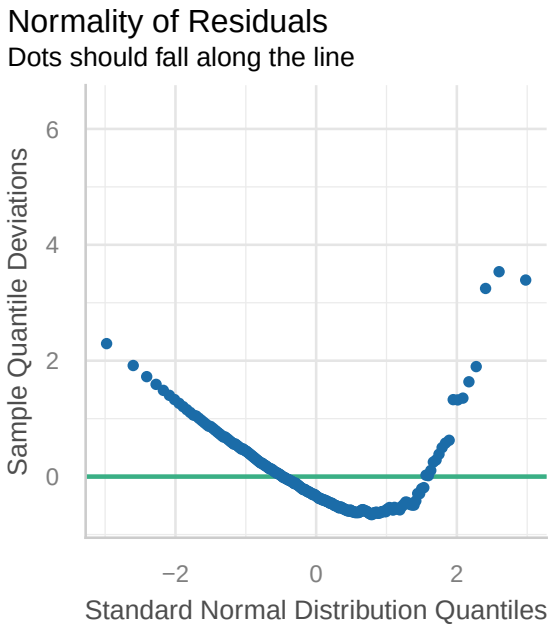
c. Check the model assumptions using the `check_model` function from the `performance` package

```
performance::check_model(m1_ols)
```



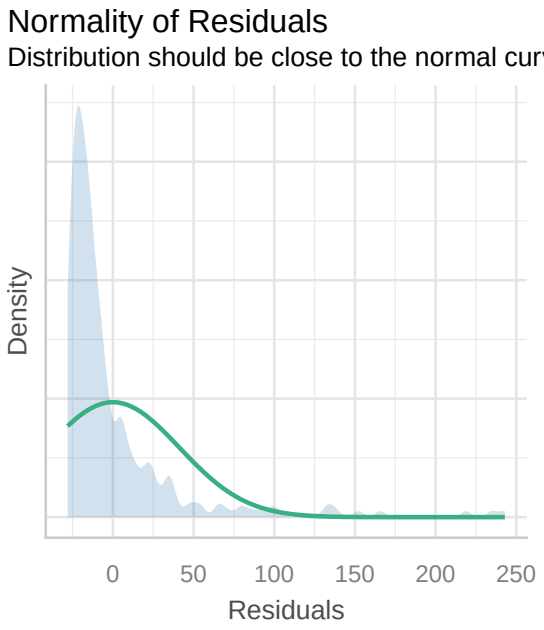
d. Explain the results of the 4 diagnostic plots. Why are we getting this result?

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



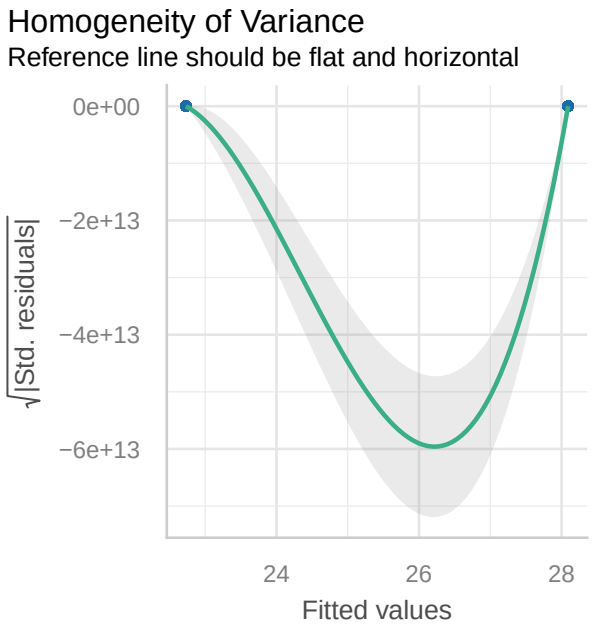
The residuals are the difference between the true value and model fitted value, which ideally is very close to 0. An assumption of an OLS model is that our data are linear and normally distributed. Because our residuals fall far from 0 (the normal distribution quantiles), we suspect our data do not satisfy these assumptions and are not normally distributed.

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



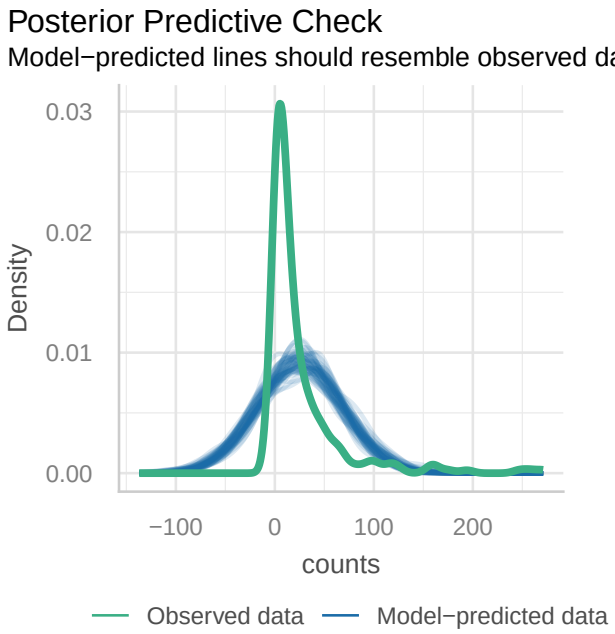
When looking at density distributions of the residuals, we would expect normal data to fall very close to the estimated residual density distribution line. As we can see here, our data's residuals do not match the estimated residual's under normal assumptions, indicating again that our data are likely not normally distributed and fail the assumptions of the OLS regression.

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



Another assumption of OLS regression is heteroscedasticity, that residuals have a constant variance. If heteroscedasticity is present, we would expect a straight horizontal line of the residuals against the predicted values from the model. We can see here a non linear pattern, indicating heteroscedasticity is not present.

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



A posterior predictive check simulates data under our model and assumptions, and compares many simulations to our observed data. This test is good for understanding if your model makes sense for your data, and if your model is good at retrodicting your data. Our model does not predict our observed data well as these lines do not match. This indicates this model is likely not the right choice, especially when looking in combination with all 4 model checks together.

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.

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

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

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

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

# a. fit a poisson regression
m2_pois <- glm(counts ~ treat, data = spiny_counts,
               family = poisson(link = "log"))

# b. interpret results
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
treat	0.21	0.03	8.44	0.00

Standard errors: MLE

```
# pull out p-values
intercept_p_pois <- summary(m2_pois)$coefficients[1,4]
treat_p_pois <- summary(m2_pois)$coefficients[2,4]

# calculate the percent change
intercept_pois <- exp(summary(m2_pois)$coefficients[1,1])
change_pois <- round((exp(summary(m2_pois)$coefficients[2,1]) - 1)*100)
```

Our model predicts non-MPA reefs will have an average of 23 spiny lobsters, with MPA reefs having an estimated average of 24% more lobsters when compared to non-MPA reefs.

In a poisson distribution, the variance equals the mean, both represented by the model's sole parameter λ. Dispersion means the variance = mean, which is a perfect fit and an assumption of the model. Overdispersion occurs when the observed variance > mean, which often happens with data that are "clumped", similar to zero-inflation where data fall closely in one area of the distribution instead of normally, or if important predictors are missing in the model.

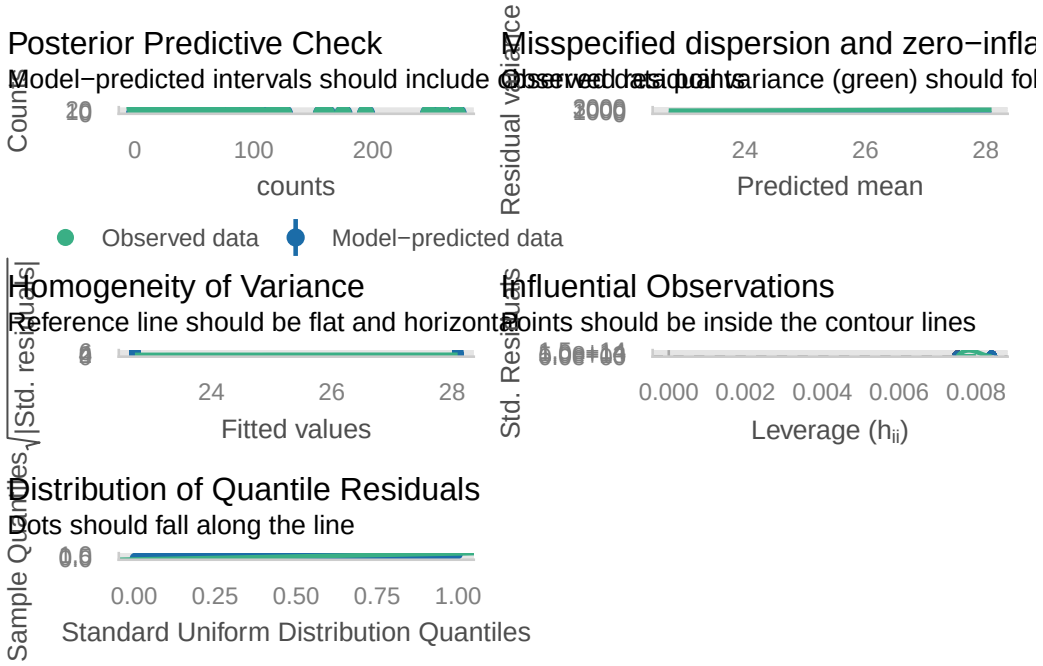
	OLS	Poisson
Treatment p-value	0.30	0.000

The p-value goes from p > 0.05 (insignificant) to p « 0.01 (significant) when changing the model from an OLS to poisson. This is a change from a insignificant effect of MPA status on lobster abundance, to a significant effect just based on the statistical model. If we look at the model assumptions and they seem appropriate for a poisson, this p-value is more likely to be accurate and we can assume a significant effect of MPA status on lobster abundance.

e. Check the model assumptions. Explain results.

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

```
check_model(m2_pois)
```



Similar to the qq-test and PPC for the OLS regression, the poisson regression did not pass these tests. The poisson regression's residuals fall far from a standard normal distribution line, and the model predicted data did not match our observed data in the PPC test.

The test for misspecified overdispersion and zero-inflation shows our observed variance falls extremely far from our model's predicted variance, indicating there is likely overdispersion and/or zero-inflation.

Homogeneity of variance tests for the assumption that each treatment group has the same variance. We have two columns of points indicating the MPA vs non-MPA reefs, and we can see a fairly horizontal line between them, indicating variance is similar between these groups.

Influential observations checks for outliers in our data, and as we can see our data points do not fall between the contour lines, indicating many outliers. This could potentially be the zero-inflation the overdispersion test indicated.

```
check_overdispersion(m2_pois)
```

```
# Overdispersion test

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

Our dispertion ration is ~67, which is much greater than an acceptable dispersion ratio of around 0-3, indicating overdispersion.

```
check_zeroinflation(m2_pois)
```

```
# Check for zero-inflation

   Observed zeros: 27
  Predicted zeros: 0
            Ratio: 0.00
```

The observed zeros are 27, whereas the predicted amount of zeros for this model is 0, which falls outside of the tolerance range for zero-inflation indicating this model is underfitting zeros (probable zero-inflation).

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

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

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

# fit a negative binomial model
m3_nb <- glm.nb(counts ~ treat, data = spiny_counts)

summ(m3_nb, model.fit = "none")
```

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
treat	0.21	0.17	1.23	0.22

Standard errors: MLE

```r
(exp(0.21) - 1)*100
```

```
[1] 23.36781
```

```r
exp(3.12)
```

```
[1] 22.64638
```

We decided to use a negative binomial regression based on the model's main two assumptions as follows:

- Outcome data are non-negative counts
- Overdispersion is probable (the data's variance > mean)

We know are data are count data, and we know based on our poisson regression that overdispersion is likely. This makes negative binomial a good next model to test.

	Poisson	Negative Binomial
Treatment p-value	0.000	0.22

	Poisson	Negative Binomial
Treatment Change (%)	24	23

In the negative binomial regression, our model predicts an average of 23 lobsters in non-MPA reefs, and predicts an average of ~23% increase in lobster abundance in MPA reefs compared to control non-mpa reefs. In a poisson regression, the models predicts an average 23 lobsters in non-MPA reefs and an average of ~24% increase in abundance in MPA reefs, which is very similar to the negative binomial. The negative binomial p-value is 0.22, which means an insignificant effect of MPA status on lobster abundance, but the poisson regression has a p-value $p<0.01$, which is a significant effect.

```
check_overdispersion(m3_nb)
```

```
# Overdispersion test

 dispersion ratio = 1.398
          p-value = 0.088
```

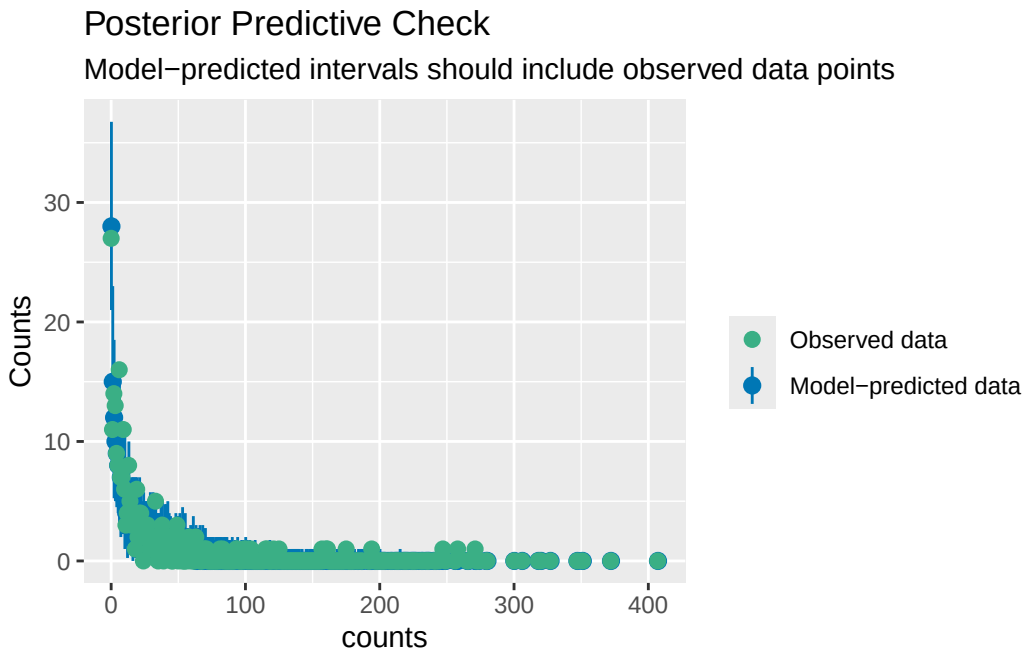
The dispersion ratio for the negative binomial is ~ 1.4, which is within the tolerance for no overdispersion.

```
check_zeroinflation(m3_nb)
```

```
# Check for zero-inflation

   Observed zeros: 27
  Predicted zeros: 30
            Ratio: 1.12
```

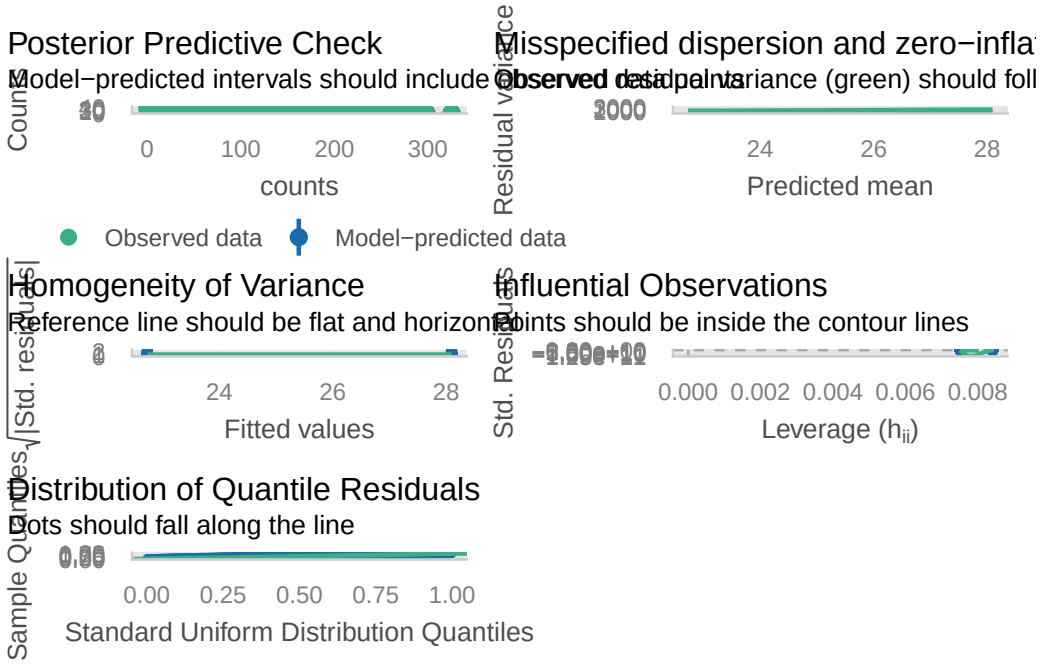
Our observed zero's are still 27, but this negative binomial regression predicts 30 zeros, meaning this model is closer to appropriately fitting zeros. This test creates a p-value and if it is not $p > 0.999$, it does not pass the check for zero inflation. According to this test, our p-value is $p = 0.600$ and thus would indicate our model is still overfitting zeros.

```
check_predictions(m3_nb)
```



The PPC shows our model's predicted data falling very close in the curve of our observed data, so it is likely our model is more appropriate for this data.

```
check_model(m3_nb)
```



The variance between groups is the same, and still passes the test for homogeneity of variance between MPA and non-MPA groups. Our residuals (QQ plot) fall almost exactly on the distribution line, indicating this model is a better fit than all previous models. The misspecified dispersion plot does show a discrepancy between the observed residual variance and predicted residual observance, although both lines now follow the same trend but are not overlapped. This could indicate zero-inflation, in accordance with the influential observations plot that shows our data fall outside the contour lines, indicating outliers.

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.

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

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

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

We first need to calculate the exponent for the poisson and negative binomial regression, as they are in link space and not in response space like the OLS regression.

	OLS	Poisson	NB
Intercept	22.73	22.65	22.65
Treatment Change (%)	24%	23%	23%
Treatment p-value	p>0.05	p<0.001	p>0.05

As we can see from the table above, both intercept coefficients and treatment effect results from all three regressions are very similar to each other, despite assumption violations in some

models. However, the poisson regression is the only model with a significant treatment effect ($p < 0.001$). While the coefficients themselves are robust across models, the significance does not stay stable between models and thus are not robust. Despite wanting a significant result, it is unwise to choose a model that does not pass assumption checks. All 3 models fail at least one assumption check, and so it would be difficult to trust these p-values and models for explaining any patterns we see in our data or for trying to hypothesize causal effects.

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)

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

```
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

	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

We designed this model with the assumption that year has an influence on lobster abundance. We also believe that the effect of a reef's MPA status on lobster abundance will also be influenced by the year (represented by the interactive term). We don't necessarily want to know these effects directly in terms of coefficient values, we are including them in our model so that the effect of MPA status (treatment) is correctly accounting for these other potentially confounding variables.

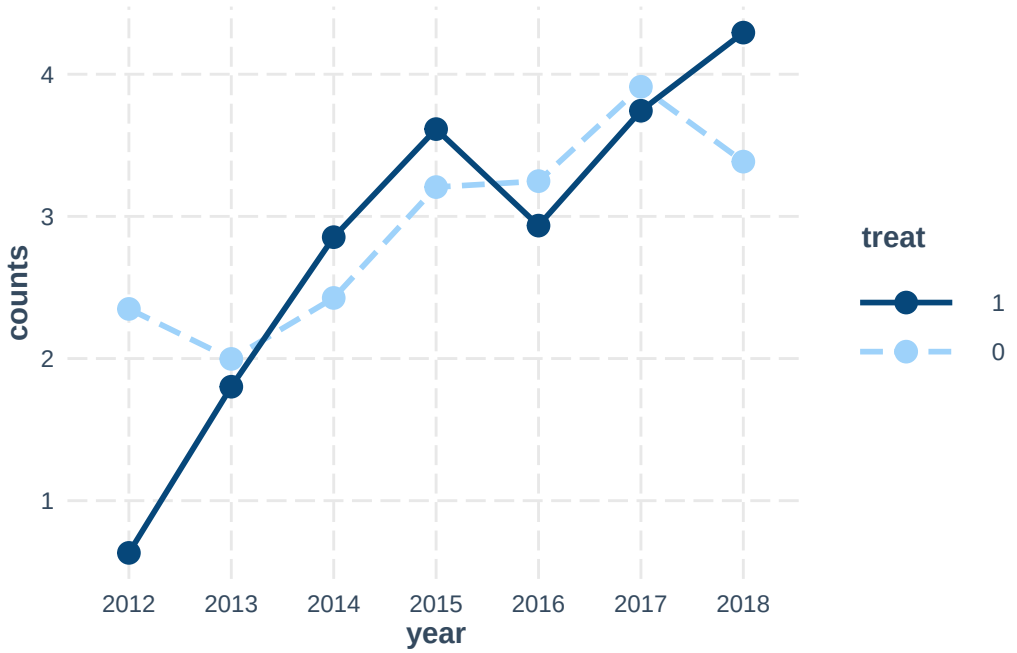
This model has estimated the counts in non-MPA reefs (exponentiated intercept coefficient) and MPA reefs (change in exponentiated intercept and treatment coefficients) after controlling for both year and the interaction between MPA status and year.

Treatment has a coefficient of -1.72, and after exponentiating that leaves a treatment effect of an average -82% lobster abundance in MPA reefs when compared to control non-MPA reefs. My first instinct is to come to the conclusion that a negative coefficient for treatment does not make sense, as ecologically MPA's are meant to protect and perpetuate higher biodiversity and healthier populations, and a difference of -82% is much larger than I would expect.

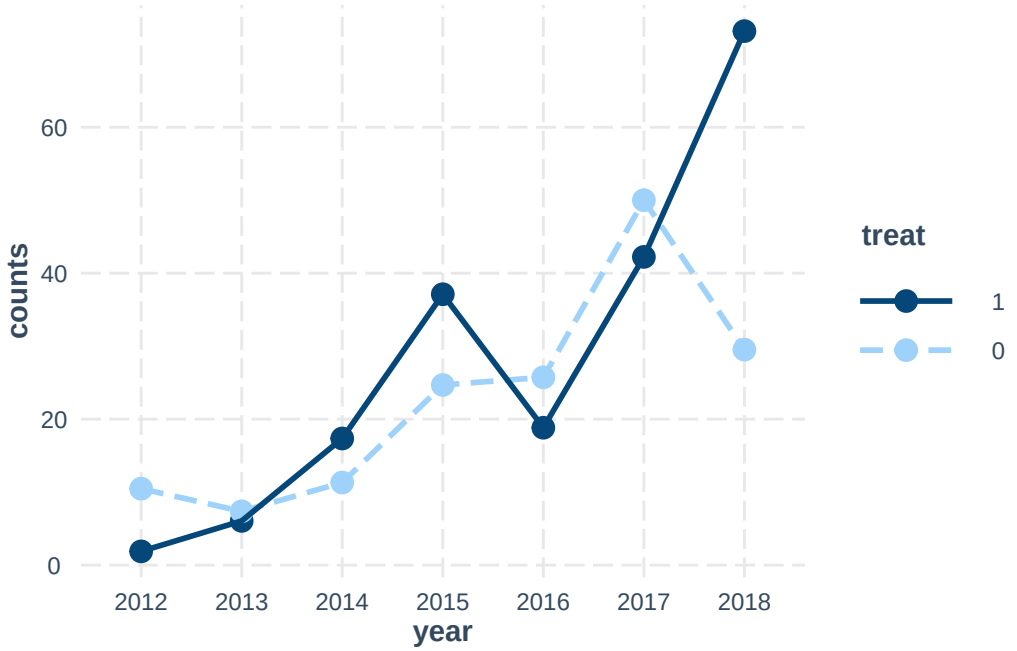
e. Look at the model predictions: Use the `interact_plot()` function from package `interactions` to plot mean predictions by year and treatment status.

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

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




```
# HINT: Change `outcome.scale` to "response" to convert y-axis scale to counts
interact_plot(m5_fixedeffs, pred = year, modx = treat,
              outcome.scale = "response")
```

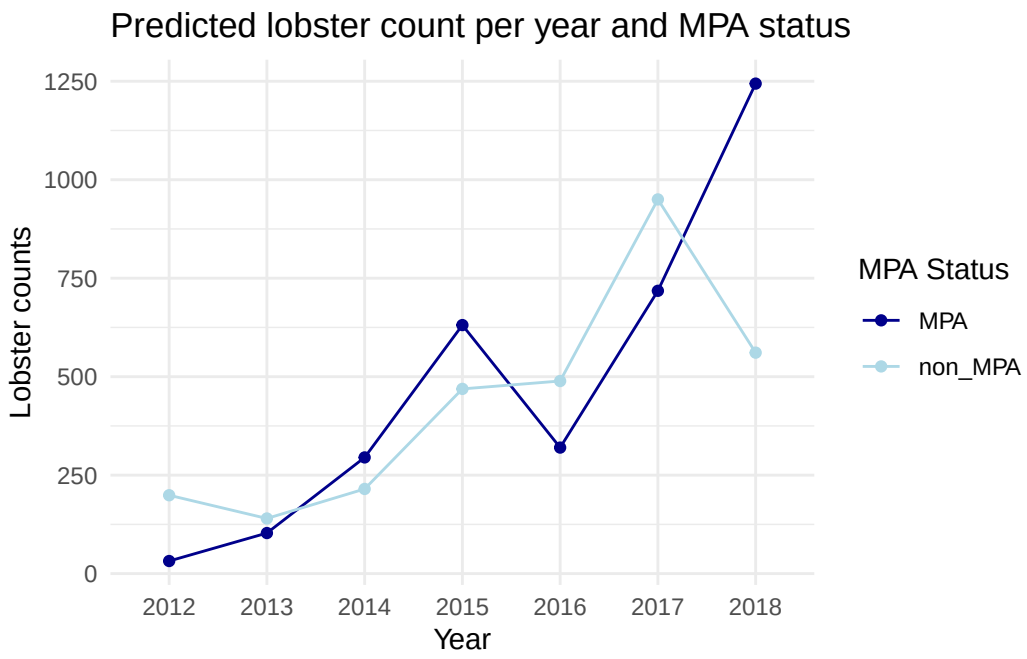


After plotting predictions, I can see from our data set that our model predicted MPA's had a higher lobster abundance in only 3 of the 7 years, and the other 4 years actually had lower abundances. While that still does not make sense ecologically, it brings up the possibility that there could be other things at play such as spillover effects, selection bias (3 non-MPA sites vs 2 MPA sites), unequal sampling year to year, or even other oceanographic or biological process affecting each site differently.

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 %>%
    mutate(year = factor(year)) %>%
    group_by(year, mpa) %>%
    summarise(counts = sum(counts))

ggplot(plot_counts, aes(x = year, y = counts, group = mpa)) +
    geom_point(aes(color = mpa)) +
    geom_line(aes(color = mpa)) +
    scale_color_manual(values = c("darkblue", "lightblue")) +
    labs(x = "Year",
         y = "Lobster counts",
         title = "Predicted lobster count per year and MPA status",
         color = "MPA Status") +
    theme_minimal()
```



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)

It is possible spillover effects were present in this research study. For example, it is possible when lobsters reach a certain capacity within an MPA, they begin to migrate outside of it due to pressures such as food availability or predators (this is a good thing in the larger ecological/conservation context, but bad for independence). From our predictions plot, we can see an example of when counts for lobsters in MPA's drop, the counts for lobsters in non-MPA's increases between 2015 and 2016. After a few years of increasing populations, lobsters in non-MPA's decrease possibly due to pressures like human fishing, while lobsters in MPA's continue to increase until, I would hypothesize, that spillover effect is noticed in the predictions again.

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

One of the assumptions of our models is that our data are independent of each other. If spillover was present in our system and the data that was collected, then our data are no longer

independent and do not satisfy Ceteris paribus. It would be very difficult to attribute any patterns to causal effects when all data are dependent on each other.

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

As mentioned above, spillover in this context would mean the abundance of lobsters in an MPA directly influences the abundance of lobsters in non-MPA's, specifically areas neighboring MPAs. This is great in terms of conservation and is actually one of the main targeted benefits of MPAs. While California spiny lobsters are not an IUCN listed species, it could indicate other species that are of conservation concern could be experiencing the same spillover effects.

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

The stable unit treatment value assumption (SUTVA) requires the response/outcome to be directly dependent on only the treatment assigned. In this example, this would mean lobster abundance in non-MPAs would be solely dependent on the non-MPA status, and have no influence from MPAs. From our assumption tests and prediction plots, we suspect that our data/model do not satisfy SUTVA as we suspect spillover effects.

`2) Excludability assumption`

Similar to SUTVA, the excludability assumption assumes that treatment assignment has no effect on outcomes beyond the treatment effect itself. As we suspect spillover effects to be present, this assumption would be invalid as we suspect that the MPA treatment is inadvertently affecting lobster abundance in non-MPA sites through these spillover effects.

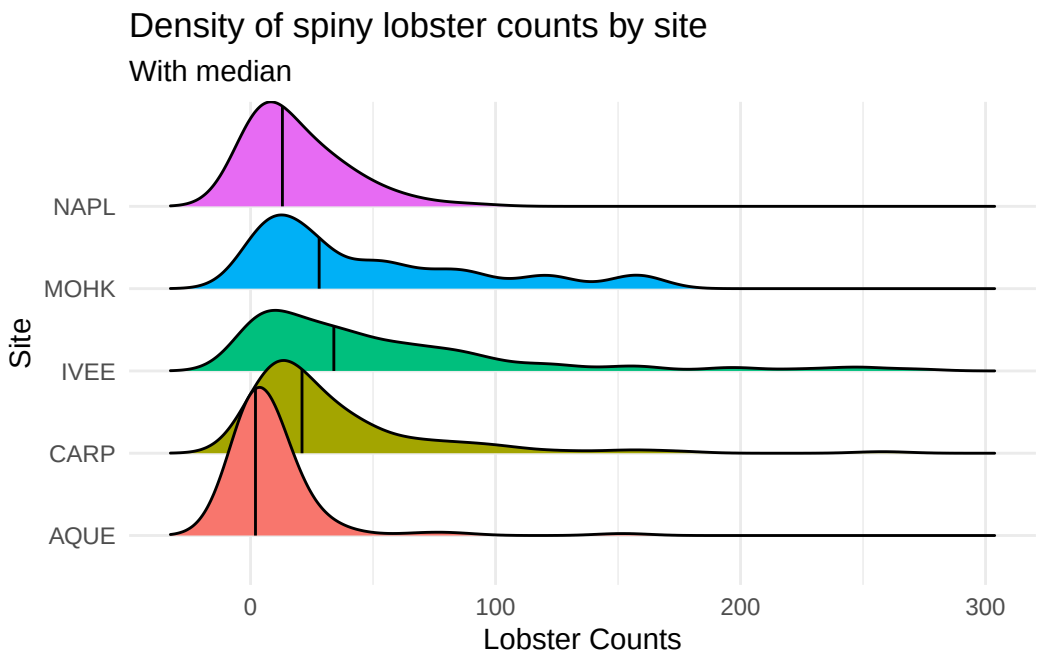
EXTRA CREDIT

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.

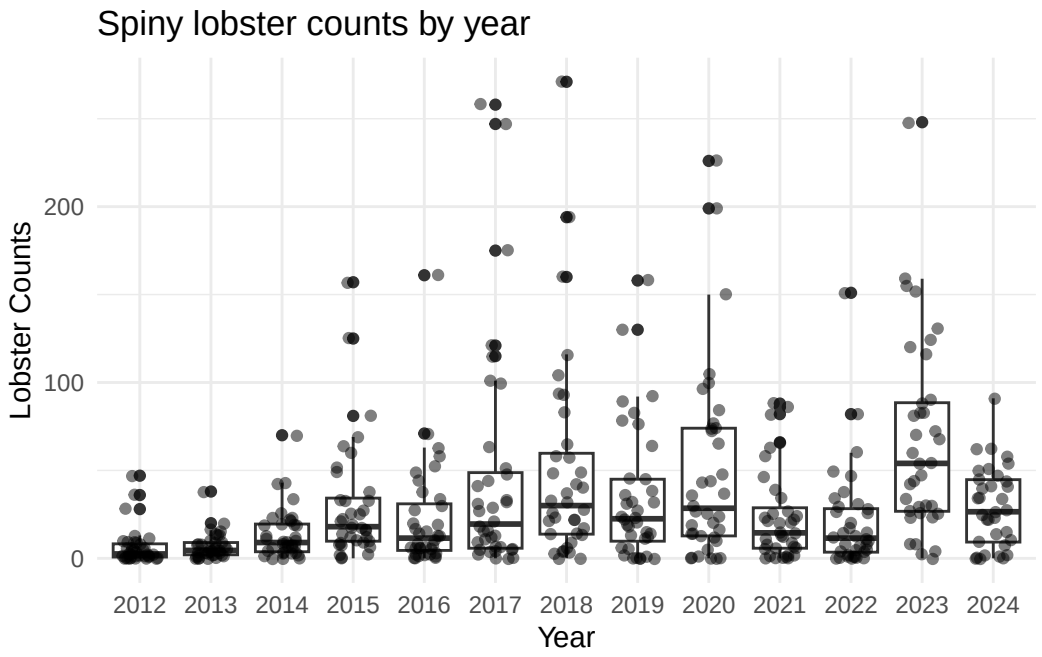
a. Create a new script for the analysis on the updated data

```r
# read in data
extracred <- read_csv(here("data", "extracredit_sblobstrs24.csv")) %>%
    clean_names() %>%
    mutate(across(size_mm, na_if, -99999))

# clean and prepare data for analysis
tidyextracred <- extracred %>%
    mutate(reef = case_when(site == "IVEE" ~ "Isla Vista",
                            site == "NAPL" ~ "Naples",
                            site == "MOHK" ~ "Mohawk",
                            site == "CARP" ~ "Carpenteria",
                            site == "AQUE" ~ "Arroyo Quemado")) %>%
    mutate(reef = as.factor(reef))

extracred_counts <- tidyextracred %>%
    group_by(site, year, transect) %>%
    summarize(counts = sum(count, na.rm = TRUE),
              mean_size = mean(size_mm, na.rm = TRUE)) %>%
    ungroup() %>%
    mutate(mpa = case_when(site == "IVEE" ~ "MPA",
                           site == "NAPL" ~ "MPA",
                           site == "MOHK" ~ "non_MPA",
                           site == "CARP" ~ "non_MPA",
                           site == "AQUE" ~ "non_MPA")) %>%
    mutate(treat = case_when(mpa == "MPA" ~ 1,
                           mpa == "non_MPA" ~ 0))
```

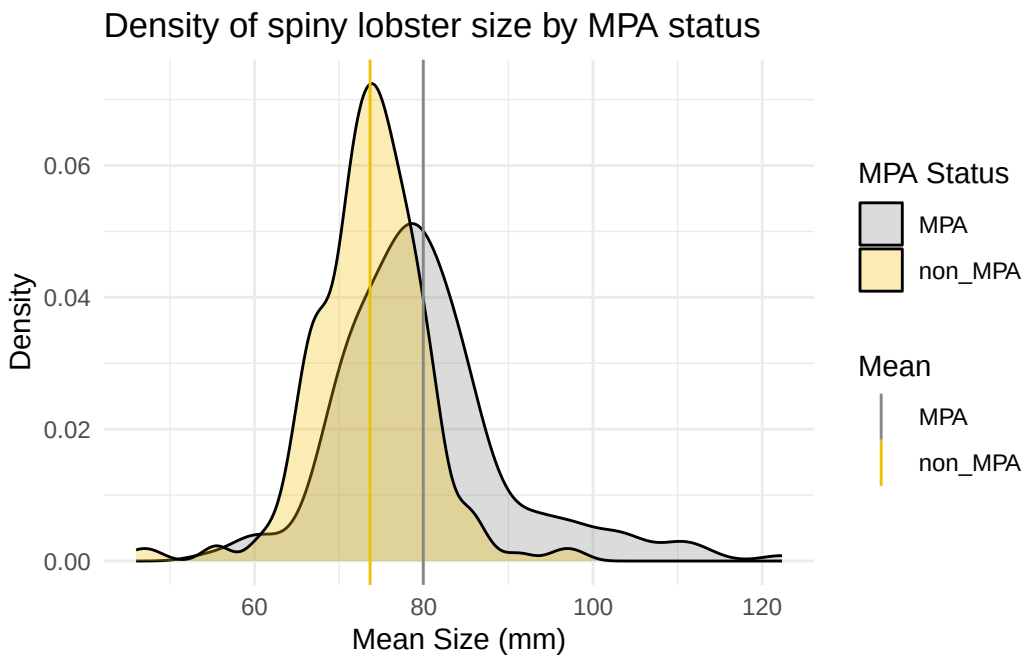
```r
# plot counts by site
extracred_counts %>%
    ggplot(aes(x = counts, y = site, fill = site)) +
    geom_density_ridges( # Add mean lines
        quantile_lines = TRUE,
        quantile_fun = function(x, ...) median(x)) +
    labs(x = "Lobster Counts",
         y = "Site",
         title = "Density of spiny lobster counts by site",
         subtitle = "With median") +
    theme_minimal() +
    theme(legend.position = "none")
```



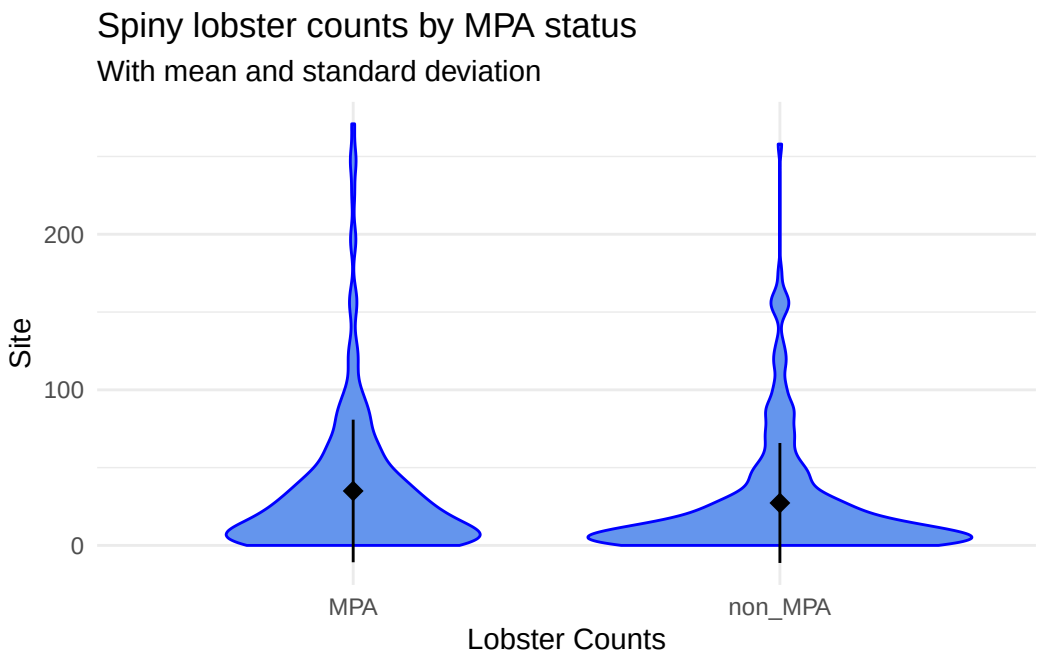

```
# plot counts by year
extracred_counts %>%
    ggplot(aes(x = factor(year), y = counts)) +
    geom_boxplot() +
    geom_jitter(width = 0.25, alpha = 0.5) +
    labs(x = "Year",
         y = "Lobster Counts",
         title = "Spiny lobster counts by year") +
    theme_minimal()
```




```
# plot distribution of size
meansize_extra <- extracred_counts %>%
    group_by(mpa) %>%
    summarise(mean = mean(mean_size, na.rm = TRUE)) %>%
    ungroup()

extracred_counts %>%
    ggplot(aes(x = mean_size)) +
    geom_density(aes(fill=factor(mpa)), size=0.5, alpha=0.3) +
    geom_vline(aes(xintercept = mean, color = mpa), data = meansize_extra, linetype = "solid
    scale_color_manual(values = c("#868686FF", "#EFC000FF")) +
    scale_fill_manual(values = c("#868686FF", "#EFC000FF")) +
    labs(x = "Mean Size (mm)",
         y = "Density",
         title = "Density of spiny lobster size by MPA status",
         color = "Mean",
         fill = "MPA Status") +
    theme_minimal()
```




```
# plot counts by MPA
extracred_counts %>%
    ggplot(aes(x = mpa, y = counts)) +
    geom_violin(color = "blue", fill = "cornflowerblue") +
    labs(x = "Lobster Counts",
         y = "Site",
         title = "Spiny lobster counts by MPA status",
         subtitle = "With mean and standard deviation") +
    theme_minimal() +
    stat_summary(fun.data=mean_sdl, fun.args = list(mult = 1),
                 geom="pointrange", color="black", shape = 18, size = 0.75)
```

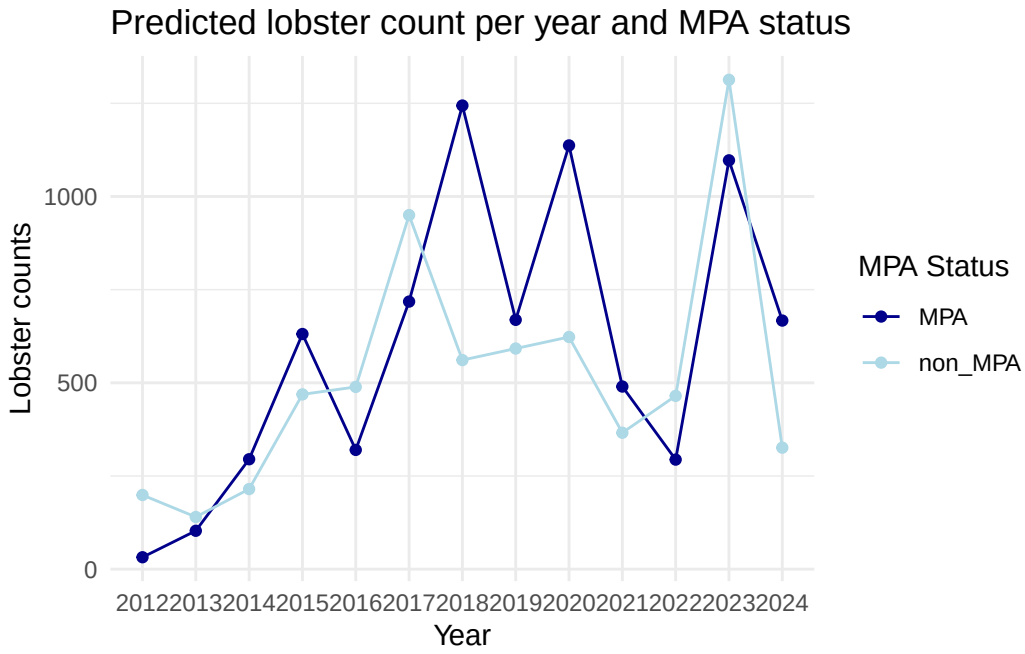


b. Run at least 3 regression models & assess model diagnostics

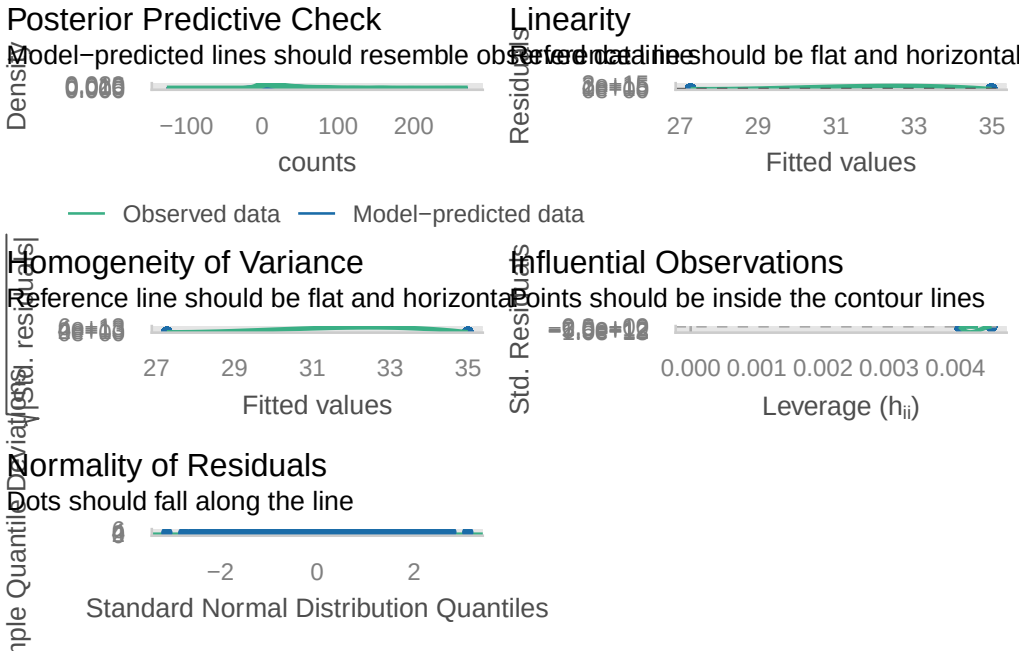
```
m1_extra <- lm(counts ~ treat, extracred_counts)
m2_extra <- glm(counts ~ treat, data = extracred_counts,
                family = poisson(link = "log"))
m3_extra <- glm.nb(counts ~ treat, data = extracred_counts)
m4_extra <- glm.nb(
    counts ~
        treat +
        year +
        treat*year,
    data = ff_counts)

# compare coefficients
export_summs(m1_extra, m2_extra, m3_extra, m4_extra,
             model.names = c("OLS","Poisson", "NB", "NB w/ controls"),
             statistics = "none")
```

```
# plot predictions
plot_extra <- extracred_counts %>%
    mutate(year = factor(year)) %>%
    group_by(year, mpa) %>%
```

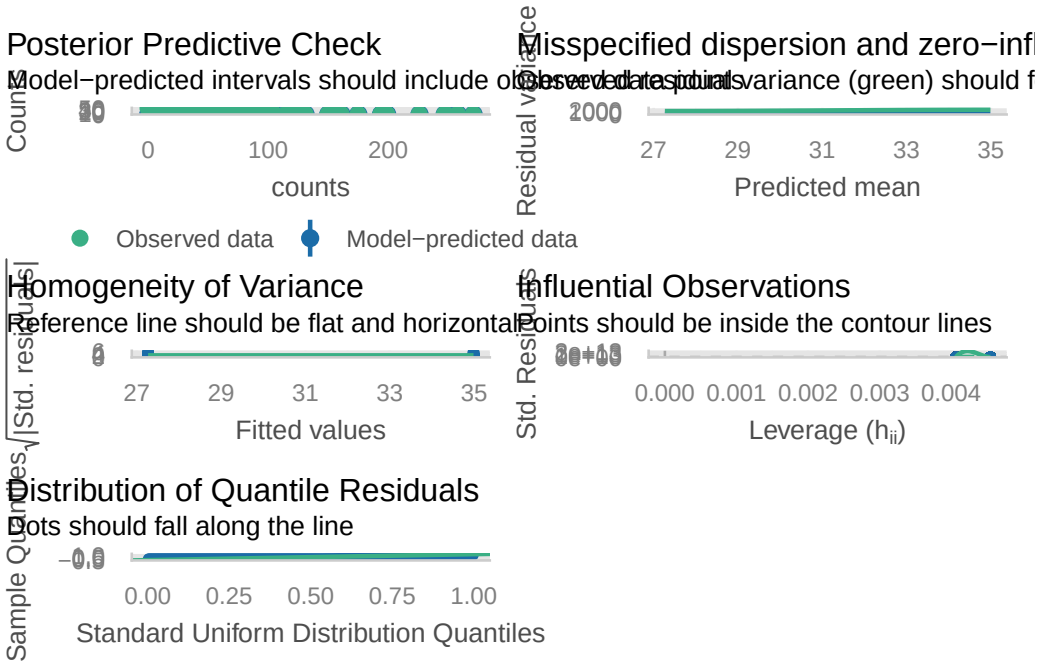
```
    summarise(counts = sum(counts))

ggplot(plot_extra, aes(x = year, y = counts, group = mpa)) +
    geom_point(aes(color = mpa)) +
    geom_line(aes(color = mpa)) +
    scale_color_manual(values = c("darkblue", "lightblue")) +
    labs(x = "Year",
         y = "Lobster counts",
         title = "Predicted lobster count per year and MPA status",
         color = "MPA Status") +
    theme_minimal()
```



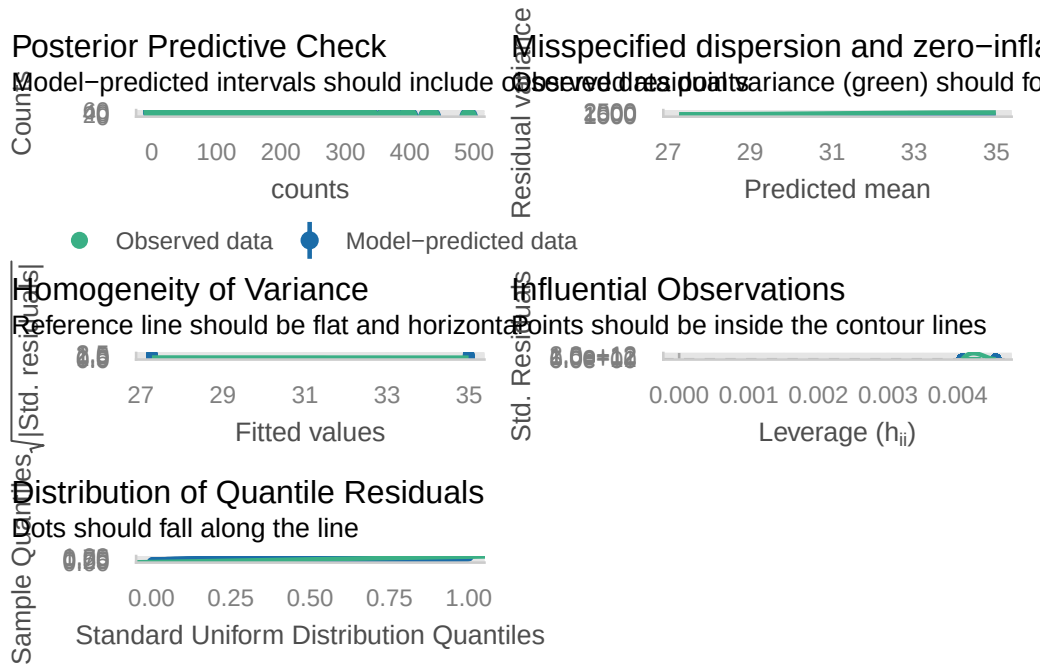

```
# check diagnostics for each model

check_model(m1_extra)
```




```
check_model(m2_extra)
```



```
check_model(m3_extra)
```




```
check_model(m4_extra)

# check overdispersion and zero inflation
check_overdispersion(m2_extra) #overdispersion
```

```
# Overdispersion test

       dispersion ratio =    57.103
  Pearson's Chi-Squared = 26496.023
                p-value =   < 0.001
```

```
check_overdispersion(m3_extra) # no overdispersion
```

```
# Overdispersion test

 dispersion ratio = 1.035
          p-value = 0.808
```

```
check_overdispersion(m4_extra) # no overdispersion
```

```
# Overdispersion test

 dispersion ratio = 1.032
          p-value =  0.72
```

```
check_zeroinflation(m2_extra) # zero-inflation
```

```
# Check for zero-inflation

   Observed zeros: 51
  Predicted zeros: 0
            Ratio: 0.00
```

```
check_zeroinflation(m3_extra) # no zero-inflation
```

```
# Check for zero-inflation

   Observed zeros: 51
  Predicted zeros: 47
            Ratio: 0.91
```

```
check_zeroinflation(m4_extra) # no zero-inflation
```

```
# Check for zero-inflation

   Observed zeros: 27
  Predicted zeros: 25
            Ratio: 0.93
```

c. Compare and contrast results with the analysis from the 2012-2018 data sample (~ 2 paragraphs)

2012 - 2024

	OLS	Poisson	NB	NB w/ Controls
Intercept	27.27	27.38	27.38	10.49
Treatment Change (%)	28	28	28	-82

	OLS	Poisson	NB	NB w/ Controls
Treatment p-value	p<0.05	p<0.001	p<0.05	p<0.001

2012 - 2018

	OLS	Poisson	NB	NB w/ Controls
Intercept	22.73	22.65	22.65	10.49
Treatment Change (%)	24	23	23	-82
Treatment p-value	p>0.05	p<0.001	p>0.05	p<0.001

I ran the same four models from the truncated (6) year data set: OLS, Poisson, negative binomial, and a negative binomial controlling for confounding variables. Overall, these four models from the 12 year data set trend very similarly to the models run on only 6 years of data. When checking model assumptions, each model is fairly consistent for which tests were validated and which tests were not. When checking for overdispersion, both OLS models were highly overdispersed, the poisson model changed from overdispersed to no overdispersion when adding the extra 6 years, and both negative binomials were consistently not overdispersed. Both OLS models had evidence of zero inflation. For the 6 year data set, the poisson and both negative binomials did not pass the test for zero-inflation, however in the 12 year data set they both showed evidence of no zero-inflation. This makes sense as the addition of more data points that are not zero help to alleviate the total ratio of zeros in the data set.

In the 12 year data set, coefficients and p-values follow the same pattern as the 6 year data set but with different values. In the 12 year data set, coefficients are quite robust between model types, but p-values fluctuate (though all treatment coefficients return with $p<0.05$ indicating treatment significance). However, all coefficients are higher in value in the 12 year data set compared to the 6 year, except for the negative binomial that controls for year which is the very similar between data sets. When looking at the predictions for the 12 year data, it makes sense that there are not too many changes in the model coefficients as the pattern of abundances between treatments stays relatively consistent over time. The difference in coefficients between data sets could be due to the omitted variable of year and interaction between year and treatment, and because we included that in our negative binomial we don't see as much variation between data sets.

There are 4 instances in the prediction plot of MPA and non-MPA lobster abundance switching trends (2013-2014, 2015-2016, 2021-2022) with abundance in MPA reefs often decreasing and abundance in non-MPA reefs increasing. Whether these trends in abundance are due to common fluctuations due to ecological or oceanographic pressures, or evidence of spillover effects, would likely require more experimentation, but it does lend evidence that these processes are worth exploring for species of conservation importance in MPAs.

	OLS	Poisson	NB	NB w/ controls
(Intercept)	27.27 ***	3.31 ***	3.31 ***	2.35 ***
	(2.69)	(0.01)	(0.08)	(0.26)
treat	7.72 *	0.25 ***	0.25 *	-1.72 ***
	(3.91)	(0.02)	(0.12)	(0.42)
year2013				-0.35
				(0.38)
year2014				0.08
				(0.37)
year2015				0.86 *
				(0.37)
year2016				0.90 *
				(0.37)
year2017				1.56 ***
				(0.37)
year2018				1.04 **
				(0.37)
treat:year2013				1.52 **
				(0.57)
treat:year2014				2.14 ***
				(0.56)
treat:year2015				2.12 ***
				(0.56)
treat:year2016				1.40 *
				(0.56)
treat:year2017				1.55 **
				(0.56)
treat:year2018				2.62 ***
				(0.56)

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