Paige Duffin Data Analysis Project

Insert Descriptive Title

Paige Duffin

2019-09-13

# Summary/Abstract

*Write a summary of your project.*

# Illustrating setup

*This section is only there to show how to insert results from other places in the project and how to cite figures and other references. Delete this whole section at some point.*

This paper (Leek & Peng, 2015) discusses types of analyses.

Figure 1 shows a result figure from the analysis.

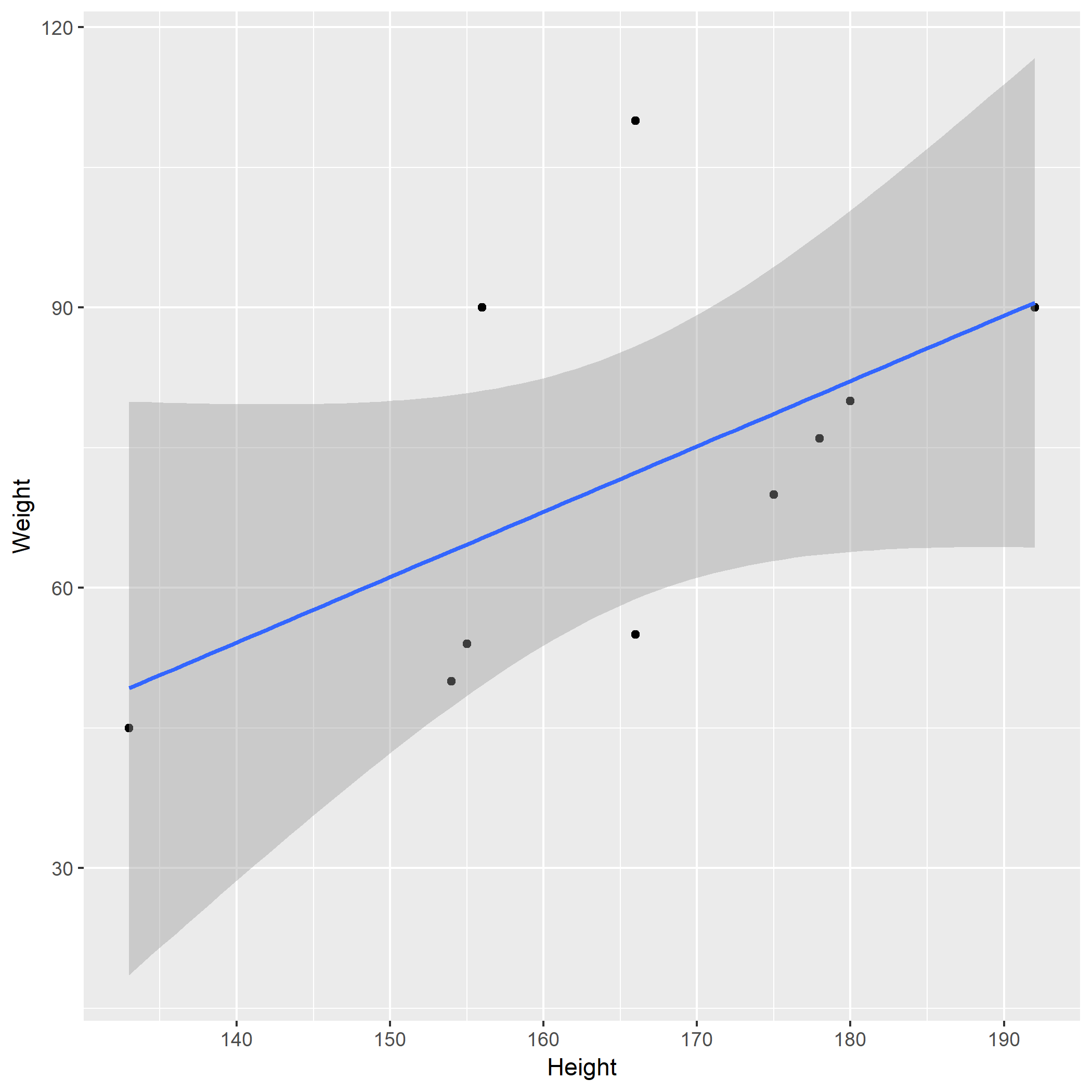


Figure 1: Analysis figure.

Table 1 shows a result table from the analysis.

Table 1: Result Table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -43.7883068 | 61.1150617 | -0.7164896 | 0.4940713 |
| Height | 0.6996272 | 0.3675692 | 1.9033889 | 0.0934786 |

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# Introduction (required for part 1)

## General Background Information

Outbreaks of infectious disease are an important, yet poorly understood, driving force in population biology and, while prevalent in both terrestrial and marine systems, the manner by which outbreaks influence biotic relationships, age demographics, community structure and function, and trophic interactions in the aquatic realm consistently lags behind that of terrestrial disease ecology. The growing body of evidence in this field, however, supports the observation that marine disease epidemics are increasing in frequency and severity (Orth et al., 2006; Waycott et al., 2009). Elucidating the relationship between aquatic species and their pathogenic diseases is pertinent in the field of marine ecology because infection outbreaks have the potential to drastically alter ecosystem functionality (Blakesley et al., 2002; Burge et al., 2013). Many of the organisms which have faced chronic and/or severe outbreaks of disease (such as sea urchins, scleractinian corals and seagrasses) are also considered keystone species and/or ecosystem engineers, meaning that they contribute highly valuable services or functions to their surrounding community and ecosystem. Thus, it comes as no surprise that disease-driven mass mortalities of these species often generate waves of ecological permutation, ranging from temporary local disruptions to permanent phase shifts (Burge et al., 2013), such as the transformation from a coral to macroalgal dominated Caribbean reef structure that accompanied the massive *Diadema antillarum* (black sea urchin) die-off of the early 1980s (Lessios et al., 1984; Lessios, 1988). This event, likely caused by a biological pathogen (Schultz et al., 2016), reduced Diadema populations to less than 1% of their original size (Lessios, 1988). Mass mortality events impacting critical foundation species, ecosystem engineers, or keystone species such as the black sea urchin have been coined ‘marine disease emergencies’ due to the detrimental cascade of events that often succeeds them (Miner et al., 2018).

Sea Star Wasting Disease (SSWD) is another epizootic crisis facing modern day coastal ecosystems. Outbreaks of asteroid wasting have been documented periodically since the late 1970s and SSWD describes a suite of symptoms observed across a broad range of sea star species, most of which play integral parts in shaping their community structure. Generally speaking, the wasting disease events which punctuated the past four decades were relatively brief, in localized areas, and largely failed to capture the attention of the scientific community. Beginning in summer 2013, however, mass mortalities of sea stars due to wasting disease have caused unprecedented damage, owing to the geographical and temporal extent of impact. This ongoing epizootic, which has killed millions of asteroids across over 20 taxa, is widely referred to as the largest disease event sweeping through a wildlife marine species in documented history (Hewson et al. 2014, Gudenkauf & Hewson, 2015, Eisenlord et al., 2016).

## Description of data and data source

*Describe what the data is, what it contains, where it is from, etc.*

The vast majority of disease documentation and surveys cover the west coast of the United States. While it is almost certainly true that SSWD is most prevalent in this region, there is a strong bias towards the discovery of diseased organisms owing to the emphasis of long-term survey networks in the area, such as the extensive [Multi-Agency Rocky Intertidal Network](https://marine.ucsc.edu/index.html) (MARINe for short). The vast majority of what we know about SSWD dynamics comes from studies along the Pacific Coast of North America using data produced through the combined effort of dedicated scientists, government agencies, and local citizens. The purpose of this investigation is to make use of this data monitoring system to investigate spatial and temoporal changes in sea star abundance along the west coast of the United States.

More specifically, I will be using the following data file:

library(readr)  
raw\_SS\_count <- read.csv("~/Documents/GitHub/PaigeDuffin-Project/data/raw\_data/seastarkat\_size\_count\_totals\_download.csv")  
head(raw\_SS\_count, 3)

## group\_code site\_code marine\_site\_name marine\_sort\_order latitude  
## 1 UCLA ALEG Alegria 6420 34.46714  
## 2 UCLA ALEG Alegria 6420 34.46714  
## 3 UCLA ALEG Alegria 6420 34.46714  
## longitude marine\_common\_season marine\_season\_code marine\_common\_year  
## 1 -120.2774 85 SP02 2002  
## 2 -120.2774 85 SP02 2002  
## 3 -120.2774 85 SP02 2002  
## season\_sequence season\_name target\_assemblage method\_code species\_code  
## 1 1 Spring sea\_star IP PISOCH  
## 2 1 Spring sea\_star IP PISOCH  
## 3 1 Spring sea\_star IP PISOCH  
## size\_sort\_order size\_bin total mpa\_designation mpa\_region georegion  
## 1 3 20 1 reference South Coast CA South  
## 2 4 30 3 reference South Coast CA South  
## 3 5 40 16 reference South Coast CA South  
## bioregion state\_province island last\_updated  
## 1 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 2 Government Point to Mexico California Mainland 2019-04-05 20:07:59  
## 3 Government Point to Mexico California Mainland 2019-04-05 20:07:59

levels(raw\_SS\_count$georegion)

## [1] "AK" "CA Central"   
## [3] "CA Channel Islands South" "CA North"   
## [5] "CA North Central" "CA South"   
## [7] "OR" "WA Olympic Coast"   
## [9] "WA Salish Sea"

levels(raw\_SS\_count$species\_code)

## [1] "EVATRO" "KATTUN" "PISOCH"

This is the raw download from a link I recieved after submitting a request for the data.

Notably, it contains the following information (some parameters removed from this overview):

* groupcode: The unique code for each monitoring group
* marine\_site\_name: The name of the site where the survey was conducted.
* latitude and longitude: (self explanatory)
* marine\_season\_code: A four-character code to identify the Sampling season. The first two characters indicate the season and the last 2 characters indicate the year.
* method\_code: The method used for sampling the plot.
* size\_bin: The size of the species being counted, binned to the nearest 5 or 10 millimeter.
* total: Total number of individuals counted in a given size\_bin
* mpa\_designation: Describes whether the referenced site is located within a Marine Protected Area (MPA) or is a reference site. If this field is blank, the referenced site is not located within an MPA, or is not a reference site for an existing MPA site.
* georegion/ bioregion: geographic/biogeographic region in which site is located
* state\_province: The State or Province and Country where the referenced site is located.
* island: The name of the island where the referenced site is located. Sites not on islands are designated as mainland.

Note: I believe I have more data coming in to work with. I recieved an email today from a representative of the Multi-Agency Rocky Intertidal Network. This email said: > “You should have been able to download our sea star size/count data from the data download page that you were directed to after filling out the data request form. We also have associated disease category data, but because we only sample 1x per year at most sites, this additional info will not give you an idea of disease emergence or progression within a site or region. You might also be interested in our observation data set, which would be better for these types of questions. We can have this data to you within about a week.”

I’m very interested in this disease data, but don’t know what it will look like. Hopefully, I can incorporate this into my analysis, but for right now I am going to focus on outlining the questions I can answer with the data I have in hand.

## Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis*

Primary Questions: 1.

Secondary Questions: 1. \*\* 2. *Is there evidence of sampling bias?*

Questions I may be able to answer with disease data: 1.

# Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what each file does. The files themselves should be commented well so everyone can follow along.*

## Univariate analysis

*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for most of your variables. You definitely need to do it for the important variables, i.e. if you have main exposure or outcome variables, those need to be explored. Depending on the total number of variables in your dataset, explore all or some of the others.*

## Bivariate analysis

*Create plots or tables and compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) to look for associations between your outcome(s) and each individual predictor variable*

## Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You can then load the results produced by this code*

# Discussion

## Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

# References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>