Alaskan Eastern Bering Sea Snow Crab Geospatial Abundance Analysis

Executive Summary

Jessica Schmidt

Fall 2024

Issue and Hypothesis

Research Question: To what extent does snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude affect the catch per unit effort of snow crab (measure of abundance)?

Null hypothesis: Snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude do not statistically significantly affect the catch per unit effort of snow crab.

Alternate Hypothesis: Snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude do statistically significantly affect the catch per unit effort of snow crab.

Context and Justification: Snow Crab Fishermen of the Bering Sea in Alaska need to maximize the size of their hauls of snow crab. The National Oceanic and Atmospheric Association (NOAA) has tracked the size of snow crab hauls (a count of snow crab per haul) as well as other relevant factors (crab gender, bottom depth of haul, surface temperature, bottom temperature, year specimen was collected, latitude, and longitude) from 1975-2018. Therefore, it would be useful for snow crab fishermen to know which factors are associated with the size of a haul of snow crab. This Multiple Linear Regression seeks to determine which of these factors are statistically significant and therefore relevant in assisting and informing snow crab fishermen when it comes time to fish.

Data-Analysis Process

Data Collection

The data for this analysis comes from the National Oceanic and Atmospheric Association (NOAA) and is called "Snow Crab Geospatial Data (1975-2018)" with a subtitle of "Alaskan Snow Crab Eastern Bering Sea Geospatial Data (1975-2018)". The data is publicly available on Kaggle and comes from the U.S. Government. "The dataset contains catch per unit effort data of commercial snow crab landings in the Alaskan Eastern Bering Sea. The catch per unit effort is an indirect measure of the abundance of a target species. The data was collected from NOAA then cleaned for data analysis." (Source: NOAA Snow Crab Data)

Please note the following variables:

Dependent variable: Catch per unit effort ("cpue", quantitative numeric variable): Catch number per area the net swept in number/square nautical mile. This is an "indirect measure of the abundance of a target species".

Independent variables:

- Year ("year", numeric): The year the specimen was collected.
- Gender of snow crab ("sex", categorical)
- Bottom Depth ("bottom_depth", numeric): In Meters.

- Surface temperature ("surface_temperature", numeric): In tenths of a degree of Celsius.
- Bottom temperature ("bottom_temperature", numeric): Average temperature in tenths of a degree of Celsius.
- Latitude ("latitude", numeric): Decimal degrees at start of haul.
- Longitude ("longitude", numeric): Decminal degrees at start of haul.

Advantage: The advantage of this data gathering methodology is that it is easy to use and prepared for data analysis. The dataset is publicly available via Kaggle and has been collected by the NOAA. The file itself is in csv format, and, therefore, it will be easy to create the dataframe for analysis by using the read_csv() function. This is very useful in that it is one of the most common data analytics tools and therefore highly accessible to anyone seeking to use this dataset. This dataset is also from a highly respectable source (the NOAA) and therefore should allow for high value analysis. Finally, my favorite aspect of this dataset is that it is from 1975-2018 and will therefore provide a lot of data over a long time. There are 17,927 records in this dataset.

Disadvantage: The main disadvantage of this data methodology is that it does not provide a lot of different variables for our analysis. This may lead to issues down the line further into the analysis if we find that there aren't many statistically significant variables. Thus, since our null hypothesis suggests that the independent variables do not statistically significantly affect the catch per unit effort of snow crab, we may have a greater chance of accepting the null hypothesis as the result since we do not have a lot of variables to use.

Challenges Overcome: The biggest challenge to overcome with this data methodology is in choosing which variables to select for the analysis. Here, we should use at least five independent variables to have a good analysis. In this case, it is likely that we will be using snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude.

Data Extraction and Preparation

The tools and techniques that will be used for this analysis are as follows: a multiple linear regression via the Python programming language in a Jupyter Notebook.

I have chosen to use Python for this Multiple Linear Regression because Python has many useful packages/libraries specifically for data science and machine learning processes (i.e., scikit-learn). As Multiple Linear Regression is a known “Structured Learning” process, Python’s machine learning and data science packages will be highly appropriate. Also, Python’s syntax is simple and concise when performing regression analyses, making it highly readable and easy to understand. One disadvantage of using Python is that python can use a lot of memory which can cause issues in particularly big projects.(Source: D208 Course Textbook)

Herein, I will be using the following Python Libraries:

```
# Importing Relevant Packages

# Standard imports

import numpy as np

import pandas as pd

from pandas import Series, DataFrame



# Vizualization libraries

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline



# Statistical packages

import pylab
```

```
from pylab import rcParams
import statsmodels.api as sm
import statistics
from scipy import stats

# Scikit-learn
import sklearn
from sklearn import preprocessing
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
from sklearn.metrics import classification_report

# Chisquare from SciPy.stats
from scipy.stats import chisquare
from scipy.stats import chi2_contingency
```

```
# Model Reduction/VIF

from statsmodels.stats.outliers_influence import variance_inflation_factor

from sklearn import preprocessing
```

Data Cleaning

Before beginning the MLR, it is imperative that we have cleaned and prepared the data for analysis. To do so, we will detect and treat duplicates, missing values, and outliers, as well as make necessary changes as appropriate such as re-expression of categorical variables and normalization of data.

Analysis

For this analysis, we will be performing a Multiple Linear Regression (MLR).

The first assumption of a MLR is that the target/dependent variable is a continuous variable, whereas the independent variables can be continuous or categorical in nature. Second, for MLR, there are multiple independent variables (aka “predictor variables”) and one dependent variable (aka the “target variable”). Third, there is expected to be a linear relationship between the independent variables and the dependent variable. Finally, the independent variables should not be too highly correlated with one another. (Source: D208 Course Webinars)

The MLR is appropriate for the research question because the dependent variable (catch per unit effort) is a continuous numeric variable, which is required for MLR. Also, one advantage of MLR is it can be used to test the multiple independent variables regardless of whether they are numeric or categorical assuming they are correlated to the dependent variable. However, one disadvantage is that the MLR can have issues with multicollinearity wherein independent variables that are highly correlated can cause the model to be unstable (source: D208 Course Textbook).

Findings

Following the MLR, the remaining variables are Year, Sex, Bottom Depth, and Latitude. These are the statistically significant variables (with p-values below 0.05). Now, it is time to revisit the research question and hypothesis:

Research Question: To what extent does snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude affect the catch per unit effort of snow crab (measure of abundance)?

Null hypothesis: Snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude do not statistically significantly affect the catch per unit effort of snow crab.

Alternate Hypothesis: Snow crab gender, year of haul, bottom depth, surface temperature, bottom temperature, latitude, and longitude do statistically significantly affect the catch per unit effort of snow crab.

In this case, our final MLR model shows that Year, Sex, Bottom Depth, and Latitude are statistically significant, meaning that we reject the null hypothesis and accept the alternate hypothesis: Snow crab gender, year of haul, bottom depth, and latitude do statistically significantly affect the catch per unit effort of snow crab.

Limitations of Techniques and Tools

One limitation of this analysis was that there were multiple variables with a high Variance Inflation Factor and/or a high p-value, so they had to be removed - also, as aforementioned, we were not given many independent variables for the analysis and we may have had more robust findings if there were more.

Also, please note that the r-squared value of the final model is a bit on the low side, however, we can still accept the results of the model because all of the p-values of the remaining independent variables (following the backward-stepwise elimination) are statistically significant (within 5%, as per industry standard).

Proposed Actions

As for the recommended course of action, we turn to the interpretation of results:

Model Equation:

$\hat{y} = 0.0017 - 0.0041(Year) + 0.0033(Sex) - 0.0105(BottomDepth) + 0.0159(Latitude)$

Interpretation of Coefficients:

- Ceteris paribus, a one unit increase in Year results in 0.41% decrease in the Catch per Unit Effort.
- Ceteris paribus, a one unit increase in Bottom Depth results in 1.05% decrease in Catch per Unit Effort.

- Ceteris paribus, a one unit increase in Latitude results in 1.59% increase in Catch per Unit Effort.
- Ceteris paribus, female crab are 0.33% more likely to be in a Catch per Unit Effort.

Therefore, based on the interpretation of results, each year will result in a lower catch per unit effort, deeper bottom depths will result in a lower catch per unit effort, higher latitudes will result in a higher catch per unit effort, and female crab are more plentiful in terms of catch per unit effort.

With this in mind, snow crab fishermen can expect lower catch per unit effort of snow crab year-over-year and expect more female than male snow crab, and it is recommended that they fish at shallower depths and fish at higher latitudes for higher catch per unit effort.

For future courses of study, it would be most prudent to repeat this analysis with more independent variables (if available), and/or with more data over time. The data is from 1975 to 2018, and it is currently 2024. Consequently, I recommend collecting data from 2019-2024 and rerunning the analysis.

Secondly, it could be useful to perform a different type of analysis on the data with a different chosen dependent variable. For example, since the gender of snow crab is statistically significant, it could be a good next step to perform a KNN analysis on the categorical variable of gender of the snow crab to determine if there are any predictive independent variables in the dataset.

Expected Benefits

‘In 2022 alone, the commercial landings of Alaskan snow crab totaled 5.5 million pounds and were valued at $24.5 million, according to the NOAA Fisheries commercial fishing landings database.’ (source: NOAA) Therefore, the knowledge of snow crab geospatial abundance is critical to aid Bering Sea large-scale and small vessel snow crab fishermen alike. The findings of this analysis can aid this $24.5m industry in helping said crab fishermen more easily pinpoint and catch the hauls. The geospatial abundance findings of this analysis are crucial in this regard so that snow crab fishermen can catch as many crab as quickly as possible: there is a set harvest limit each season and “harvesting generally occurs from January to April in the Eastern Bering Sea” (NOAA). In other words, the faster the fishermen can harvest the crab, the better.

Sources

- NOAA Snow Crab Dataset: https://www.kaggle.com/datasets/mattop/snowcrab
- NOAA: https://www.fisheries.noaa.gov/species/alaska-snow-crab#:~:text=Generally%20harvested%20from%20January%20to,%2C%20but%20available%20year%2Dround.
- D208 Course Webinars
- D208 Course Textbook - recommended by Dr. Middleton: Bruce, Peter, et al. "Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python"