Alexandra Zharchuk
SPR 2024 BSTA 662

A Survival Analysis of "Overfished" Status of Fishing Stock in Baja California for Years 2013-2018
Alexandra Zharchuk
BSTA 662
Dr. Kelly Fan

Alexandra Zharchuk
SPR 2024 BSTA 662

A Survival Analysis of "Overfished" Status of Fishing Stock in Baja California for Years 2013-2018

I. Introduction

Fishing stock has been on the decline for the last several decades at an alarming rate. Ecologists have developed a method of monitoring stock status by utilizing biomass measurements and theoretical mximum sustainable yield to create an overfishing scale, denotes with the source equation of $\frac{Biomass_{annual}}{Biomass_{MSY}}$. Annual bionass is thus divided by estimates annual biomass at maximum sustainable yield to rate overfishing at a scale from 0 to 2, where respective indices mark stock status: ≤ 0.5 = "Overfished" , ~1 = "Neither overfished or experiencing overfishing", 2 = "Pristine".

As a final element of information, it should be outlined that there is a difference between "overfishing" and "overfished". "Overfishing" is a measurement of fish mortality that indeed can lead to a stock being "overfished". However, unlike "overfished", "overfishing" is a rate of fish mortality of any body of water at a given time. In contrast, "overfished" is a an average value of stock availability compared to maximum sustainable yield, and is measured in units of biomass.

II. Intent

The concept of the "overfished" ratio (biomass ratio) and general knowledge of current seafood stock decline will be used as a basis for this analysis. This analysis will attempt to identify the survival function of stock and its dependence on covariates of a specially selected dataset. Data exploration will further highlight annual trends in commercial fishing in relation to stock sttus estimates and production metrics (catch in kilograms, number of observations per year).

III. Dataset and Variable Summary

Data was compiled by Erica M. Ferret, Alfredo Giron-Naca, Octavio Aburto-Oropeza from University of California San Diego for publication: "Overfishing Increases the Carbon Footprint of Seafood Production from Small-Scale fisheries." Data hs been sourced by authors from two databases: Gulf of California Marine Program (GCMP) Fisheries Monitoring Network , Comision Nacional de Acuacultura y Pesca (CONAPESCA). GCMP data is observerc through March 2018, CONAPESCA data is observed through December 2019 - cutoff date for both databases is 2018.

The finalized dataset contains 26 variables, with a total of n=4307 rows N=111,982 observations. There are 8 discrete variables and 18 continuous variables. In total, the variable summary for the variables used in the analysis is represented in Figure 1.

ElapsedTime = Years until "overfished" status occurs for any stock	Status Variable= B/Bmsy; B <0.6 = 0 B >0.6 = 1	WetWeight_kg = prepared weight * (0.40^-1 ; raw catch and/or prepared catch together	Catch_kg = raw catch in kg	TripDistance_km = trip distance in km
Scientific_Name = Species name for stock observation	Gas_Lt = Gas consumption per trip in liters	TripDuration_hr = trip duration in hours	fillet_TAG = TRUE, FALSE Whether catch was previously tagged or not	GeneralGear_category = Category of gear used for catch

Figure 1. Variables in use.

IV. Methodology

In the United States, the threshold for a stock to be considered overfished is a Biomass ratio of 0.5 or lower. This measurement will be used as the status indicator for this analysis.
Data exploration will achieve a visual and introductory analysis to the normality of covariates and their association to eachother. A Kaplan-Meier estimation will be performed to assess the survival function of stocks until being overfished.
A Cox proportional hazards model will be performed to assess the proportional hazards assumption with the Supremum test, and will also output the relationships between survivorship and covariates.
To achieve efficient computation, observations will be randomly sampled by SAS to achieve a reduced sample size of n=100.

V. Data Exploration

About 23.19% of the data is censored. The correlation matrix shows moderate relationship between time and all the covariates. The only variables that do not have correlation are trip duration, catch, and wet weight.

Pearson Correlation Coefficients, N = 4307
Prob > |r| under H0: Rho=0

	TimeElapsedMonths	Catch_kg	WetWeight_kg	Gas_Lt	TripDuration_hr	TripDistance_km
TimeElapsedMonths	1.00000	0.04786 0.0017	0.06429 <.0001	-0.33087 <.0001	0.12602 <.0001	-0.14327 <.0001
Catch_kg	0.04786 0.0017	1.00000	0.95114 <.0001	0.08076 <.0001	0.01114 0.4648	0.14913 <.0001
WetWeight_kg	0.06429 <.0001	0.95114 <.0001	1.00000	0.10964 <.0001	-0.01511 0.3216	0.15185 <.0001
Gas_Lt	-0.33087 <.0001	0.08076 <.0001	0.10964 <.0001	1.00000	0.32021 <.0001	0.37705 <.0001
TripDuration_hr	0.12602 <.0001	0.01114 0.4648	-0.01511 0.3216	0.32021 <.0001	1.00000	0.21957 <.0001
TripDistance_km	-0.14327 <.0001	0.14913 <.0001	0.15185 <.0001	0.37705 <.0001	0.21957 <.0001	1.00000

Spearman Correlation Coefficients, N = 4307
Prob > |r| under H0: Rho=0

	TimeElapsedMonths	Catch_kg	WetWeight_kg	Gas_Lt	TripDuration_hr	TripDistance_km
TimeElapsedMonths	1.00000	0.32807 <.0001	0.26051 <.0001	-0.32648 <.0001	0.12394 <.0001	-0.23897 <.0001
Catch_kg	0.32807 <.0001	1.00000	0.91363 <.0001	-0.21219 <.0001	0.10731 <.0001	0.04023 0.0083
WetWeight_kg	0.26051 <.0001	0.91363 <.0001	1.00000	-0.05829 0.0001	0.17073 <.0001	0.13046 <.0001
Gas_Lt	-0.32648 <.0001	-0.21219 <.0001	-0.05829 0.0001	1.00000	0.33667 <.0001	0.40888 <.0001
TripDuration_hr	0.12394 <.0001	0.10731 <.0001	0.17073 <.0001	0.33667 <.0001	1.00000	0.23983 <.0001
TripDistance_km	-0.23897 <.0001	0.04023 0.0083	0.13046 <.0001	0.40888 <.0001	0.23983 <.0001	1.00000

Figure 2. Spearman and Pearson Correlation Matrices

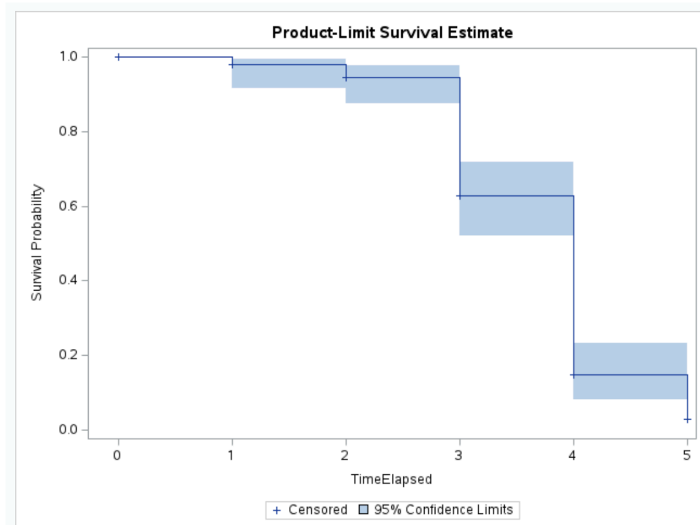

Figure 3. Kaplan-Meier Survival Curve

All covariates are normally distributed and usable in the context of this analysis. Analysis of the Kaplan Meier curve for all the data shows a general survivorship probability of 1 for 1 year, after which it gradually declines until 3 years. At 3 years, survival probability significantly decreases to 0.6. At 4 years, the survivorship drops to 0.1.

VI. Statistical Analysis

The Cox proportional hazards model is the main element of analysis to determine correlation of both categorical and numerical covariates pertaining to the survival time. The initial Cox model is fitted with the response variable elapsed time, with covariates scientific name, fillet tag, general gear category, trip distance, trip duration, gas, catch, and wet weight. The likelihood estimate analysis (Figure 4) shows the

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
Scientific_Name	Callinectes bellicosus	1	-14.26276	4485	0.0000	0.9975	0.000	Scientific_Name Callinectes bellicosus
Scientific_Name	Caulolatilus princeps	1	-19.36972	4485	0.0000	0.9966	0.000	Scientific_Name Caulolatilus princeps
Scientific_Name	Chione californiensis	1	-1.05210	38.06638	0.0008	0.9780	0.349	Scientific_Name Chione californiensis
Scientific_Name	Cynoscion othonopterus	1	-29.32051	4402	0.0000	0.9947	0.000	Scientific_Name Cynoscion othonopterus
Scientific_Name	Farfantepenaeus californ	1	3.58950	2.78459	1.6617	0.1974	36.216	Scientific_Name Farfantepenaeus californ
Scientific_Name	Hyporthodus niphobles	1	26.47851	11116	0.0000	0.9981	3.158E11	Scientific_Name Hyporthodus niphobles
Scientific_Name	Litopenaeus stylirostris	1	-0.69039	3.02687	0.0520	0.8196	0.501	Scientific_Name Litopenaeus stylirostris
Scientific_Name	Micropogonias megalops	1	51.56145	8523784	0.0000	1.0000	2.471E22	Scientific_Name Micropogonias megalops
Scientific_Name	Mugil spp	1	-15.97766	5789	0.0000	0.9978	0.000	Scientific_Name Mugil spp
Scientific_Name	Mustelus californicus	1	-8.08462	2.35892	11.7461	0.0006	0.000	Scientific_Name Mustelus californicus
Scientific_Name	Panopea generosa	1	-18.49438	5651	0.0000	0.9974	0.000	Scientific_Name Panopea generosa
Scientific_Name	Scomberomorus concolor	1	-1.78316	16.02453	0.0124	0.9114	0.168	Scientific_Name Scomberomorus concolor
fillet_TAG	FALSE	1	-4.03898	2.18712	3.4103	0.0648	0.018	fillet_TAG FALSE
GeneralGear_Category	Gillne	1	3.24103	2.48537	1.7005	0.1922	25.560	GeneralGear_Category Gillne
GeneralGear_Category	HookLi	0	0	.	.	.	.	GeneralGear_Category HookLi
GeneralGear_Category	Hookah	0	0	.	.	.	.	GeneralGear_Category Hookah
GeneralGear_Category	Trap	1	19.85378	4485	0.0000	0.9965	4.1917E8	GeneralGear_Category Trap
TripDistance_km		1	0.03628	0.01442	6.3341	0.0118	1.037	
TripDuration_hr		1	-0.25700	0.09656	7.0842	0.0078	0.773	
Gas_Lt		1	0.02400	0.01900	1.5959	0.2065	1.024	
Catch_kg		1	0.01457	0.09214	0.0250	0.8743	1.015	
WetWeight_kg		1	-0.0009227	0.03702	0.0006	0.9801	0.999	

Figure 4. Analysis of likelihood estimates for initial model.

p-values for the effect of covariates on the response in the initial model with significance level $\alpha = 0.05$. Covariates with insufficient p values are several types of fish, catch, gear category, wet weight, and gas. Generally, covariates with high significance imply an effect on the survivorship. In contrast, covariates with p values greater than 0.05 imply lack of effect on survivorship. Therefore, catch, gear category, wet weight, fillet tag, gas, and species will be removed to comply with backwards selection.

Proceeding with the final model, the results are showcased in Figure 5 and 6. The general consensus is the increase in survivorship by 0.23835 log units for every hour of a commercial fishing trip. The hazard ratio is 1.269 which suggests that every additional hour of trip duration is associated with a 26.9% increase in hazard. The global null hypothesis test shows sufficiency in the model's efficacy. The analysis of likelihood estimates shows insignificance for trip distance, and significance for trip duration on survivorship at significance level $\alpha = 0.05$.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15.2521	2	0.0005
Score	13.7377	2	0.0010
Wald	10.9858	2	0.0041

Figure 5. Likelihood ratio tests for final model

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TripDistance_km	1	0.00653	0.00908	0.5171	0.4721	1.007
TripDuration_hr	1	0.23835	0.08035	8.7997	0.0030	1.269

Figure 6. Analysis of MLE for final model.

Supremum Test for Proportionals Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
TripDistance_km	6.9372	1000	1139866958	0.7790
TripDuration_hr	11.9870	1000	1139866958	0.3410

Figure 7. Supremum Test for final model

The final Cox proportional hazards model is:
$h(y|x) = h_0(TimeElapsed) * e^{(0.23835*TripDuration)}$[1]

Holding the notion that this is the appropriate final model, the proportional hazards supremum test (Figure 7) furthermore confirms the legitimacy of trip durations conformance to the proportional hazards assumption. Since the value for trip duration in the PHA Supremum is greater than 0.05, it passes the assumption.
A final assessment of the model's validity is confirmed in the Schoenfeld residual plot; the scattering shows modest linearity.

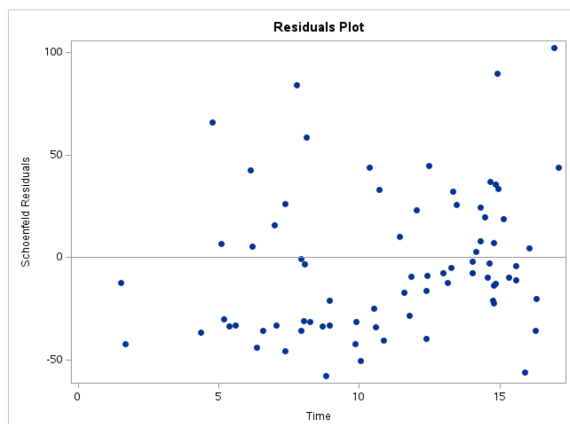

Figure 8. Final model residual plot

Conclusion & Future Considerations

After fitting a Cox model on the data, there seems to be a significant relationship between trip duration and survivorship. From this analysis, it can be interpreted that for each hour of additional fishing time, there is a 26.9% increase in risk of a stock reaching a biomass level that is considered overfished.

[1] (Note: Since some species are actually significant to survivorship - it might be worth considering a model that incorporates species and their appropriate parameter estimates. The following equation would be:
$h(y|x) = h_0(TimeElapsed) * e^{(0.23835*TripDuration)+(\beta_{ScientificName}*ScientificName)}$)

A further multi-factor analysis might suggest stratifying the data by species, so that each significantly associated species can be analyzed for its relationship with survival time. In other words, a multi-factor analysis would further identify the types of fish species that are at risk of being overfished. Additionally, a better model fit that has more than 3 variables would provide a better rounded summary of current overfishing statistics.

Appendix and Code:

Ferrer, E. M., Giron-Nava, A., & Aburto-Oropeza, O. (2022). Overfishing Increases the Carbon Footprint of Seafood Production From Small-Scale Fisheries. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.768784

Ferrer, E. M., Giron-Nava, A., & Aburto-Oropeza, O. (2022). Overfishing Increases the Carbon Footprint of Seafood Production From Small-Scale Fisheries. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.768784 (Supplementary Material)

Code

```
/************* Loading Dataset and Creating Status Variable ************/
data overfishing_final_wstatus;
    set work.overfishing_final;
    if Year then do;
        TimeElapsed = Year - 2013;
        TimeElapsedMonths = TimeElapsed * 12;
    end;
    else do;
        TimeElapsed = 2018 - 2013;
        TimeElapsedMonths = TimeElapsed * 12;
    end;
    /* Set Status Variable */
    if B.Bmsy >= 0.6 then Status = 1;
    else Status = 0;
run;

/* Sort by Year */

proc sort data=overfishing_final_wstatus out=overfishing_final_wstatus_sorted;
    by Year;
run;
/************ Data Exploration ***************/
proc freq data=overfishing_final_wstatus;
tables Status/nocum;
title "Proportion of Censored Data";
run;
title;

/* Looking at means of all variables */
proc means data=overfishing_final_wstatus_sorted N MEAN;
var TimeElapsedMonths Catch_kg WetWeight_kg Gas_Lt TripDuration_hr TripDistance_km;
run;
```

```
/* Take a random sample of 5000 observations from the dataset */
proc surveyselect data=overfishing_final_wstatus_sorted out=sample_data method=srs sampsize=100;
run;

proc corr data=sample_data spearman plots=matrix;
var TimeElapsedMonths Catch_kg WetWeight_kg Gas_Lt TripDuration_hr TripDistance_km;
run;

proc univariate data=overfishing_final_wstatus_sorted normaltest;
var TimeElapsedMonths Catch_kg WetWeight_kg Gas_Lt TripDuration_hr TripDistance_km;
HISTOGRAM/NORMAL;
RUN;

/*********** Model Fitting **************/
proc lifetest data=sample_data method=km plots=(survival(cl),ls,lls); /*ls = hazard, lls=proportional hazards */
    time TimeElapsed*Status(0);
run;

proc phreg data=sample_data;
   class Scientific_Name fillet_TAG GeneralGear_Category;
   model TimeElapsed*Status(0) = Scientific_Name fillet_TAG
   TripDistance_km TripDuration_hr TripDistance_km GeneralGear_Category/ties=discrete;
   strata Scientific_Name;
run;

data allcovals;
set overfishing_final_wstatus;
run;
title "Proportional Hazards Assumption Diagnostic";
title2 "Final Model";

proc phreg data=sample_data;
model TimeElapsed*Status(0) =
TripDistance_km TripDuration_hr/ties=discrete;
baseline out=pred2 covariates=allcovals survival=s lower=lcl upper=ucl
cumhaz=H /nomean;
run;

proc phreg data=sample_data;
    class Scientific_Name;
    model TimeElapsed*Status(0) = TripDistance_km TripDuration_hr/ ties=discrete;
    strata Year;
    assess PH/resample;
    output out=phreg_results RESSCH=resid;
run;

proc sgplot data=phreg_results;
    title Residuals Plot;
   scatter x=TripDuration_hr y=resid / markerattrs=(symbol=CircleFilled);
   refline 0 / axis=y;
   xaxis label='Time';
   yaxis label='Schoenfeld Residuals';
run;
```