Stat 450: Case Studies in Statistics

Investigating Post-release Mortality of Coho Salmon in a Marine Recreational Fishery

Emma Lunzmann-Cooke
Doctoral student, Forest and Conservation Sciences

Present by Weihao Qiu, Andy Lin, Olivia Liu, Richard Hou

Introduction

Background: Coho salmon mortality rates after catch-and-release fishery are unknown.

Motivation: Protect wild salmon population in BC.

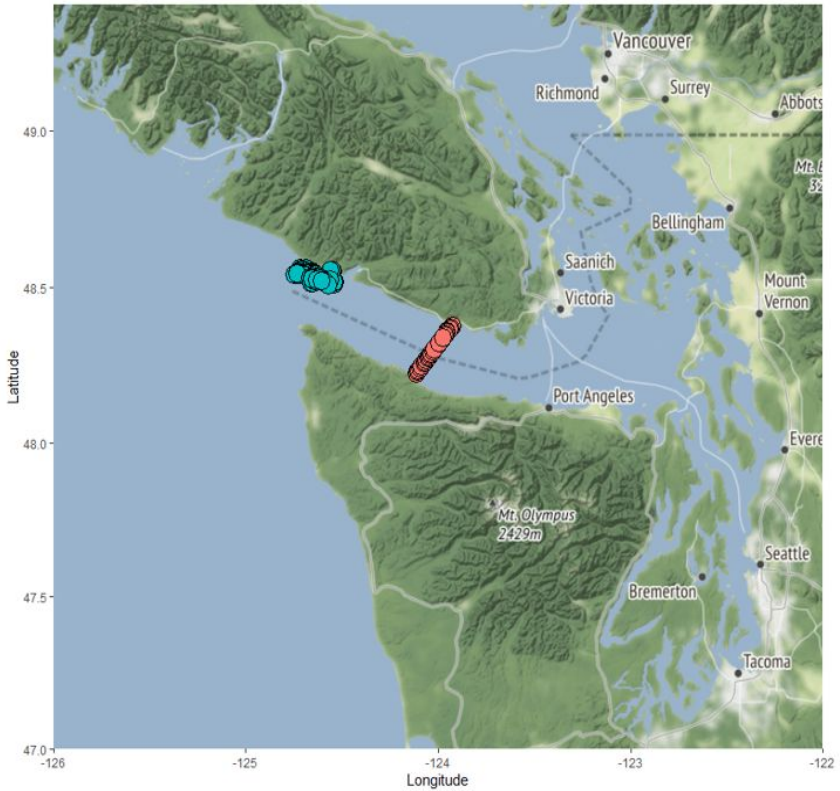


Figure 1. Signal Receiver Location Map
Blue = release locations. Red = signal receiver locations

Objectives

- **Main purpose**: identify the factors that influence coho salmon survival after a catch-and-release event.
- **Statistical problems:**
  - Assess logistic regression model & AIC, as well as alternative models and criteria.
  - Whether condensing some measurements into scores is the optimal strategy.

Data

Overview: 320 coho salmon were randomly captured and tagged around BC.

Among those 281 valid data will be used. Missing values will be removed (241).

Response variable: Detection.status (binary)

Data

Explanatory variables (14):

Categorical:

- Population
- Sex
- Hook locations

.....

Numerical:

- Length
- Air exposure time
- Mean.fat

.....

Exploratory Data Analysis

- Data cleaning & wrangling
  - Check the missing data
- Variable Exploration and visualization
  - Predictors vs. Response variable

continuous ⟶ boxplot

categorical ⟶ frequency plot

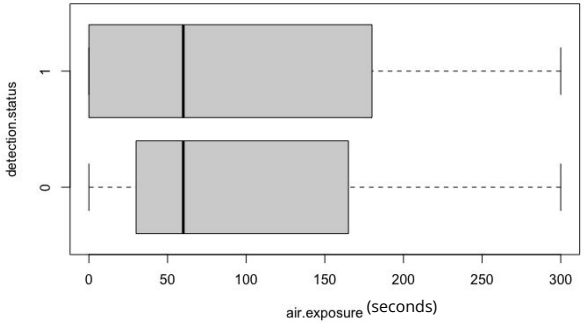


Figure 2. Air exposure time vs. Detection Status

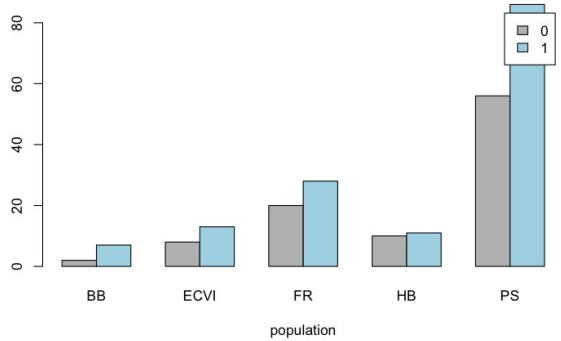


Figure 3. Freq. of Salmon With Different Population

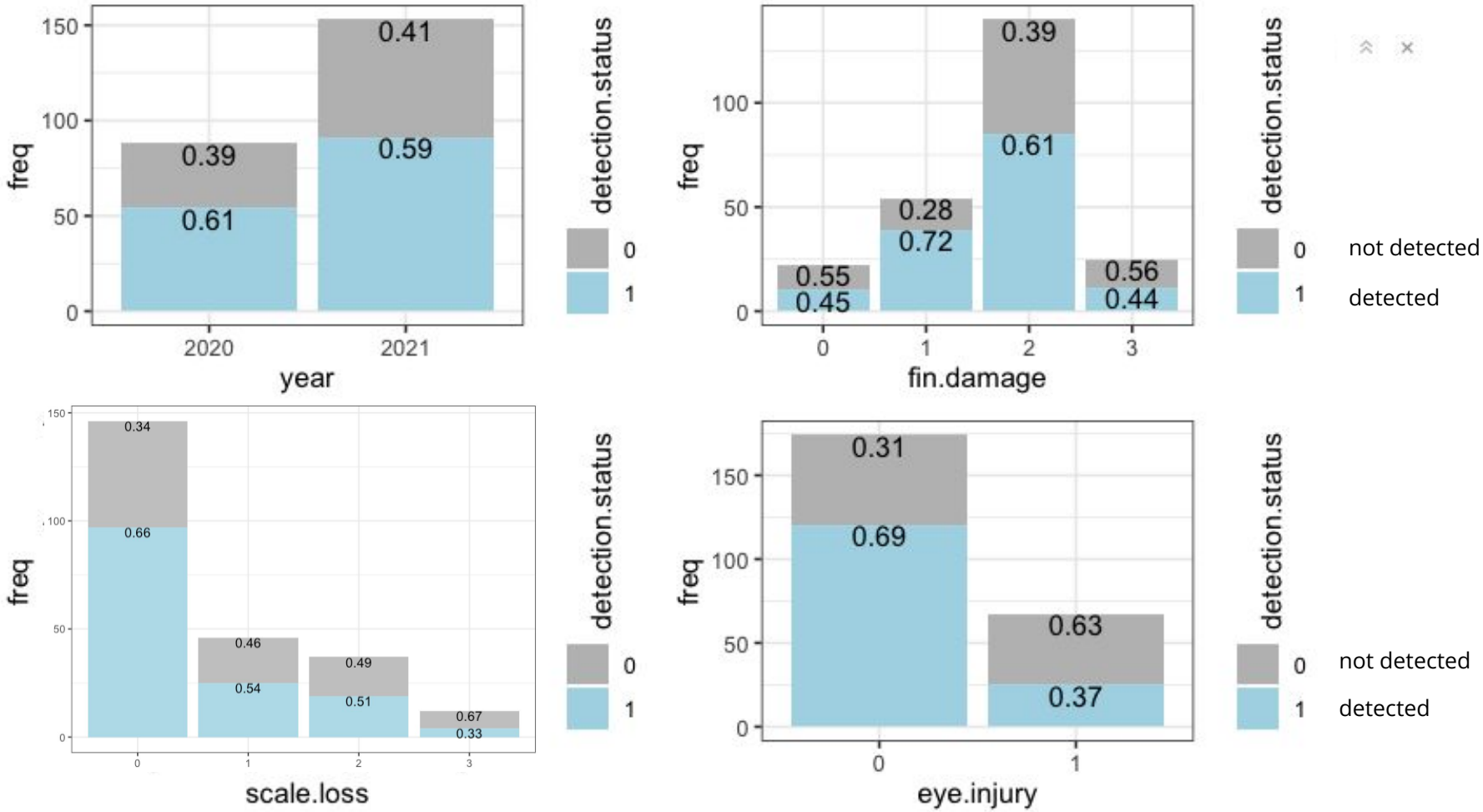
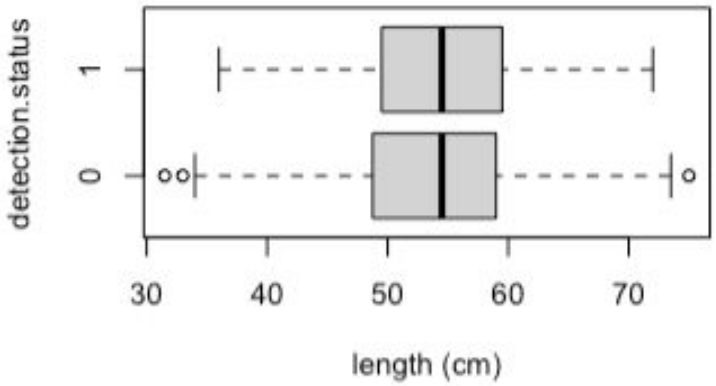


Figure 4. Frequency of Detection Status vs. Different Categorical Predictors

Salmon Length vs. Whether survived



Salmon Mean Fat vs. Whether survived

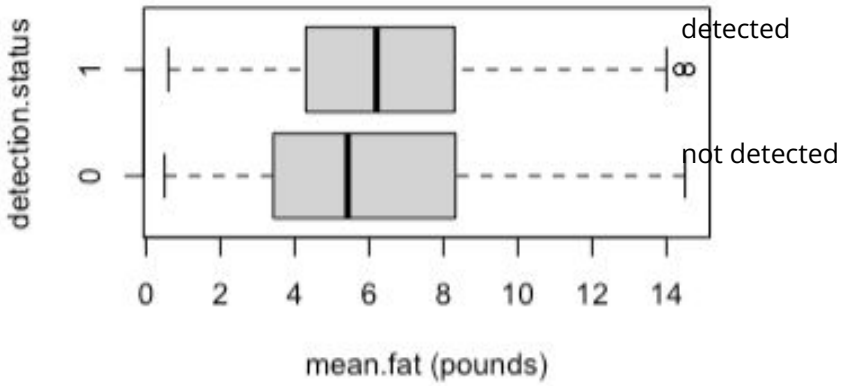


Figure 5. Salmon Length and mean.fat vs. Detection Status (for comparison)

Exploratory Data Analysis

- Visualize the correlation using correlation matrix to check variable multicollinearity

→ **No significant correlation between variables**

Note: reflex.score & air.exposure correlation ~0.6, ignore for now

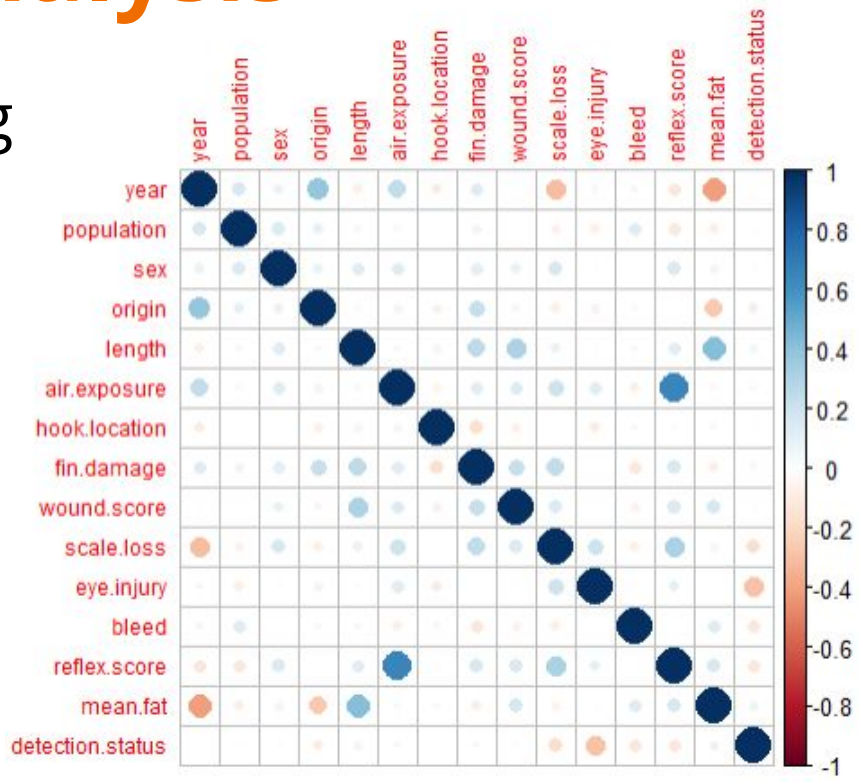


Figure 6. Correlation Plot of Variables

Statistical Analysis

(1) Models and methods:

- Data splitting, according to the fraction of survival
- Logistic regression with variable selection
- Forward, backward, best subset (AIC)
- BIC, (adjusted) R^2, Cp
- Test Accuracy & Cross validation

Summary of Some Representative Models Trained

Models	AIC	Train accuracy	Test accuracy
Full logistic	227.49	0.767	0.574
Stepwise logistic	213.02	0.772	0.623
*Best logistic (hypo. test)	218.45	0.728	0.672
*Random Forest		0.883	0.656
Regularized RF.		0.906	0.623

Variables selected: eye.injury, hook.location, fin.damage, air.exposure, reflex.score

Statistical Analysis

(2) Condensed score vs. separate variables:

- Condense into injury score → worse (0.639)
- Expand reflex score → better (0.705)
- New best model: eye.injury, hook.location, fin.damage, air.exposure, orientation reflex

Conclusions

- Variables selected and interpretation
- Logistic regression (& RF.), AIC + cross validation
- Usually not to condense variables
- Limitations: Imbalanced predictors, sample size
- Further suggestions: remove imbalance, more records

References

Previous Study:

Nelson, K. L. (1998). Catch-and-release mortality of striped bass in the Roanoke River, North Carolina. *North American Journal of Fisheries Management, 18*(1), 25-30.


Thank you!