2020 Preseason Pink Salmon Forecast

Sara Miller and Rich Brenner

October 2019

Objective

To forecast the Southeast Alaska (SEAK) pink salmon harvest in 2020.

Analysis

Three hierarchical models were investigated. The full model was:

$$E(y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

where X_1 was cpue and X_2 was the average temperature in Icy Strait in May, June, and July. The regression coefficients cpue and temperature (ISTI_MJJ) are significant in the first two models. The interaction term in not significant (Table 1). Therefore, only the first two models will be considered.

Table 1: Parameter estimates

X1	term	estimate	std.error	statistic	p.value
1	(Intercept)	1.3427021	7.641121	0.1757206	0.8622804
2	CPUE	14.6533685	2.674148	5.4796395	0.0000231
3	(Intercept)	136.8505836	42.278874	3.2368550	0.0043400
4	CPUE	17.1659248	2.334149	7.3542529	0.0000006
5	ISTI_MJJ	-15.6825735	4.838540	-3.2411791	0.0042981
6	(Intercept)	-2.4558282	126.116094	-0.0194728	0.9846782
7	CPUE	69.5357104	44.781476	1.5527784	0.1378814
8	ISTI_MJJ	-0.2881339	13.992384	-0.0205922	0.9837975
9	CPUE:ISTI_MJJ	-5.7359519	4.898281	-1.1710132	0.2568628

The model summary results using the metrics AIC, BIC, mean absolute percent error (MAPE), median absolute percent error (MEAPE), and mean absolute scaled error (MASE)(Hyndman and Kohler 2006) are shown in Tables 2 and 3. For all these metrics, the smallest value is the preferred model. These metrics suggest that model two is the preferred model.

Table 2: Summary of model outputs

X1	model	AdjR2	AIC	AICc	BIC
1	CPUE	0.5802221	183.8258	185.1592	187.0990
2	CPUE+ISTI_MJJ	0.7154553	176.1430	178.4959	180.5072
3	CPUE+ISTI_MJJ+CPUE:ISTI_MJJ	0.7209090	176.5278	180.2778	181.9830

Table: Forecast error measures

|| || || ||

Model Diagnostics

Model diagnostics included residual plots, the curvature test, and influential observation diagnostics using Cook's distance (Cook 1977), the Bonferroni outlier test, and leverage plots. Model diagnostics were used to identify observations that were potential outliers, had high leverage, or were influential (Zhang 2016). These observations may have significant impact on model fitting and may need to be excluded. An observation that is distant from the average covariate pattern is considered to have high leverage. If an individual observation has a leverage value h_i greater than 2 or three times p/n, it may be a concern (where p is the number of parameters (i.e., 2) and n is the number of observations (i.e., 22); $p/n = 3/22 = 0.14$ for this study; Dobson 2002). Therefore, a leverage cut-off of 0.27 was used; observations with a leverage value greater than 0.14 were investigated further. Cook's distance is a measure of influence, or the product of both leverage and outlier. Cook's distance,

$$D_i = \frac{e_{PSi}^2}{p+1} * \frac{h_i}{1-h_i},$$

where e_{PSi}^2 is the standardized Pearson residuals, h_i are the hat values (measure of leverage), and p is the number of predictor variables in the model, is a measure of overall influence of the ith on all n fitted values (Fox and Weisburg 2019). A large value of Cook's distance indicates that the data point is an influential observation. Cook and Weisberg (1994) suggest using the median of the F-distribution with $(p+1)$ and $(n-p-1)$ degrees of freedom as a benchmark for identifying the subset of influential observations. Therefore, a Cook's distance cut-off of 0.87 was used; observations with a Cook's distance value greater than 0.87 were investigated further with the Bonferroni outlier test ($P<$ 0.05). Influential observations were removed and coefficients reexamined to determine the impact of the observation. A significant shift in the coefficient when influential observations are removed may indicate an unusual observation that needs to be investigated further. To determine if a variable has a relationship with residuals, a lack-of fit curvature test was performed. In this test, terms that are non-significant suggest a properly specified model. Statistical analyses were performed with the R Project for Statistical computing version 3.6.0 (R Core Team 2019).

Residuals vs. Fitted Plot

The characteristics of a well-behaved residual vs. fitted plot and what they suggest about the appropriateness of the simple linear regression model: 1)The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable. 2) The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal. 3) No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers. Taken directly from the source: https://newonlinecourses.science.psu.edu/stat462/node/117/

The one point that stands out is juvenile year 2012 when the SEAK catch was 94.7.

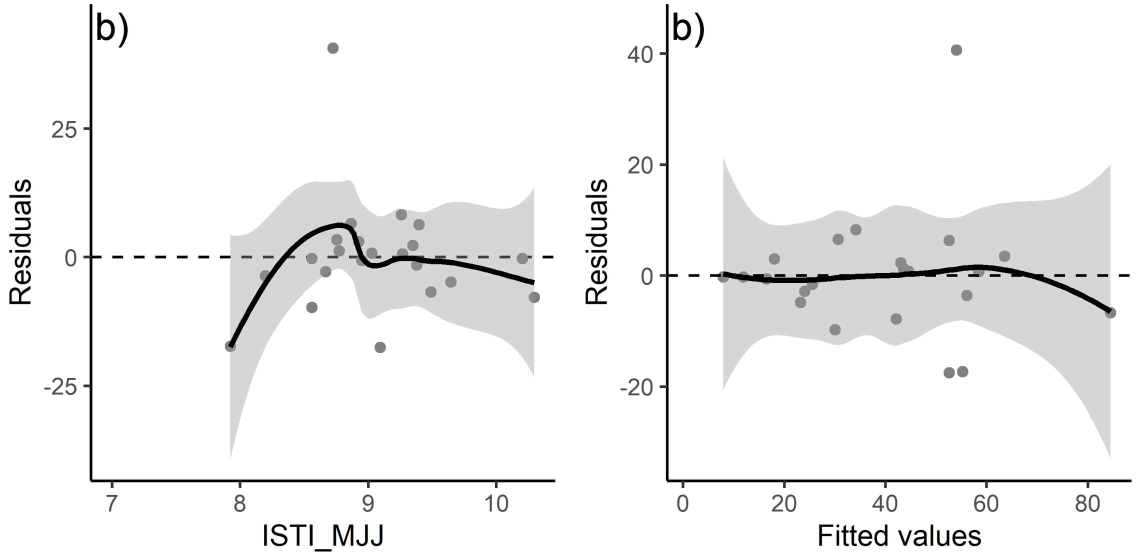

Figure 1: Residuals versus fitted plot for model 2.

Residuals vs. Predictor Plots

The interpretation of a “residuals vs. predictor plot” is identical to that for a “residuals vs. fits plot.” That is, a well-behaved plot will bounce randomly and form a roughly horizontal band around the residual = 0 line. And, no data points will stand out from the basic random pattern of the other residuals. Taken directly from the source: https://newonlinecourses.science.psu.edu/stat462/node/117/

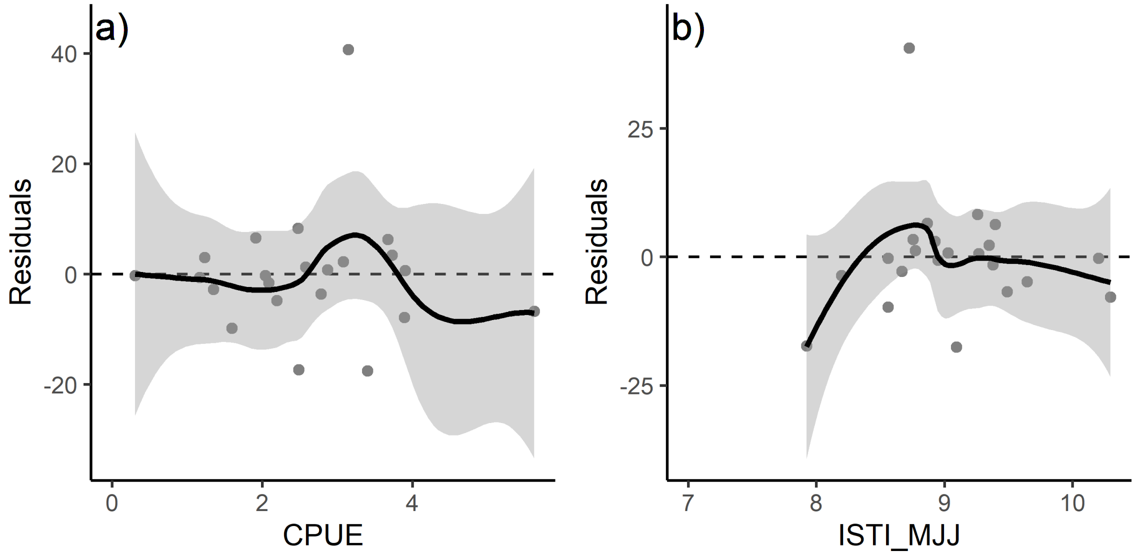

Figure 2: Residuals versus predicted plots for a)CPUE and b) temperature for model 2.

Influential Datapoints

The Bonferroni outlier test suggested that observaton 16 (juvenile year 2012) was an outlier. No terms were significant in the lack-of-fit curvature test, suggesting that the model was properly specified.

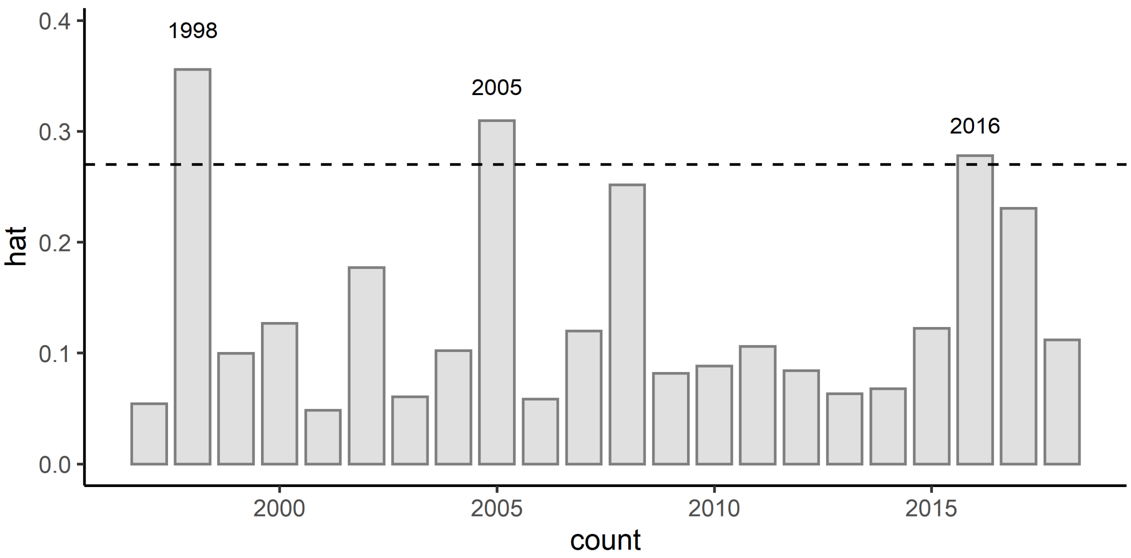

Figure 2: Diagnostics plots of influential observations including a) Cook's Distance (witha cut-off of 0.87), b) leverage values (with a cut-off value of 0.27).

Results

The best model, based on MASE metric and significant coefficients was model 2 (i.e. the model containing cpue and temperature). Diagnostics indicated some observations had high leverage values, but none of the observations affected model fitting and overall the model did not show lack of fit. None of the data points were above the cut-off value for the Cook's distance. Based on the Bonferroni outlier test, one the data points had a studentized residual with a significant Bonferroni P-value suggesting one of data points impacted the model fitting. The conditional mean function in the residual plots should be constant across the plot in a "correct" model. Based on the results of the curvature test ($P > 0.05$), and the relatively flat (non-curved) fitted smooth lines in the Pearson residual plots, the plot does not show lack of fit of the model. The adjusted R^2 value was 0.72 indicating a good model fit.

Conclusion

The SEAK pink salmon harvest in 2020 is predicted to be in the weak range with a point estimate of 2 million fish (80% confidence interval: 0–12 million fish; Figure 3).

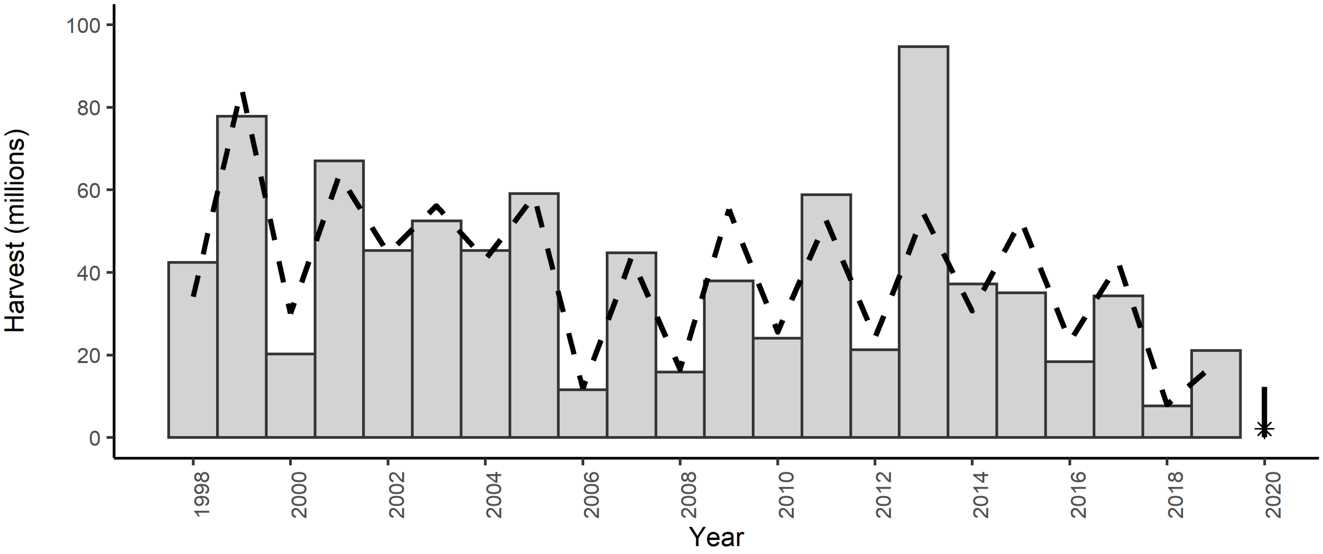

Figure 3: SEAK harvest (millions) by year with the fitted values from model 2 shown as a line in the figure. The predicted 2020 forecast is symbolized as a star with 80% bootstrap confidence intervals.

References

Cook, R. D. 1977. Detection of influential observations in linear regression. Technometrics 19: 15–18.

Cook, R. D. and S. Weisberg. 1994. An Introduction to Regression Graphics. New York: Wiley.

Dobson, A. J. 2002. An Introduction to Generalized Linear Models. Second Edition. New York: Chapman and Hall. 225 pp.

Fox, J. and S. Weisburg. 2019. An R Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage Publications, Inc.

Hyndman, R. J. and A. B. Koehler. 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22: 679-688.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Zhang, Z. 2016. Residuals and regression diagnostics: focusing on logistic regression. Annals of Translational Medicine 4: 195.