2020 Preseason Pink Salmon Forecast

Sara Miller and Rich Brenner

November 2019

Objective

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

Executive Summary

Forecasts were developed using an approach described in Murphy et al. in press. A multiple regression model was developed using monthly peak juvenile CPUE (standardized catch based on 20 minute trawl set) for the June and July surveys and associated environmental parameters. The model used was:

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

where y is ln(harvest), β_1 is the coefficient for ln(CPUE+1), β_2 is the coefficient for the natural log of the environmental covariate water temperature (either May through July average temperature 'ISTI_MJJ_log' or May through August average temperature 'ISTI_log' in the upper 20 m at eight stations in Icy Strait), β_3 is the interaction term, and ϵ represents the normal error term that is lognormal. Leave-one-out cross validation (hindcast) and model performance metrics such as Mean and Median Absolute Percentage Error (MAPE, MEAPE), and mean absolute scaled error (MASE) (Hyndman and Kohler 2006) were then used to evaluate forecast accuracy of alternative models. Statistical analyses were performed with the R Project for Statistical computing version 3.6.0 (R Core Team 2019).

Based on model 2, the SEAK pink salmon harvest in 2020 is predicted to be in the weak range with a point estimate of 11.8 million fish (80% prediction interval: 7.4 to 18.8 million fish).

Analysis

Five 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 ln(CPUE+1) and X_2 was either the natural log of the average temperature in Icy Strait in May, June, and July or the the natural log of the average temperature in Icy Strait in May through August, and β_3 is the interaction term between CPUE and one of the temperature indices.

The regression coefficients CPUE and temperature (ISTI_MJJ_log, ISTI_log) are significant in the first three models (m1, m2, m3). The interaction terms are not significant in either model 4 or 5 (m4 or m5; Table 1). Therefore, only the first three models will be considered further.

Table 1: Parameter estimates for the five potential models.

X1	model	term	estimate	std.error	statistic	p.value
1	m1	(Intercept)	2.3738613	0.2038433	11.6455179	0.0000000
2	m1	CPUE	0.4342078	0.0713387	6.0865706	0.0000060
3	m2	(Intercept)	11.1295695	2.4428039	4.5560635	0.0002159
4	m2	CPUE	0.5054830	0.0598672	8.4433996	0.0000001
5	m2	ISTI_MJJ_log	-4.0609645	1.1305155	-3.5921351	0.0019429
6	m3	(Intercept)	12.5308188	2.6606052	4.7097626	0.0001525
7	m3	CPUE	0.4962886	0.0573637	8.6516097	0.0000001
8	m3	ISTI_log	-4.6255051	1.2095296	-3.8242184	0.0011447
9	m4	(Intercept)	-1.2818793	6.8493217	-0.1871542	0.8536331
10	m4	CPUE	5.1751159	2.4298380	2.1298193	0.0472408
11	m4	ISTI_MJJ_log	1.5765629	3.1176953	0.5056822	0.6192162
12	m4	CPUE:ISTI_MJJ_log	-2.1125090	1.0989493	-1.9222988	0.0705407
13	m5	(Intercept)	-0.5695098	8.6905081	-0.0655324	0.9484725
14	m5	CPUE	5.3111554	3.0526234	1.7398659	0.0989525
15	m5	ISTI_log	1.2492744	3.9019002	0.3201708	0.7525256
16	m5	CPUE:ISTI_log	-2.1526780	1.3645734	-1.5775465	0.1320821

The model summary results using the metrics AICc, MAPE, MEAPE, and MASE (Hyndman and Kohler 2006) are shown in Table 2. For all these metrics, the smallest value is the preferred model. The difference (Δ_i) between a given model and the model with the lowest AICc value and the metric MASE were the primary statistics for choosing appropriate models in this analysis. Models with $\Delta_i \leq 2$ have substantial support, those in which $4 \leq \Delta_i \leq 7$ have considerably less support, and models with $\Delta_i > 10$ have essentially no support (Burnham and Anderson 2004). These two metrics (AICc, MASE) suggest that model two and three are the preferred models.

Based on the AICc metrics, both models m2 and m3 have substantial support. Although the temperature index ISTI has been used in the past for forecasting pink salmon harvest, if temperature is actually altering how CPUE is related to abundance it makes sense to restrict the temperature data to the CPUE months in the forecast model (June and July). The month of May is included as there are important migratory dynamics prior to the time juveniles are actually sampled in Icy Strait. Therefore, the model m2 (based on average temperature in May through July) was used to forecast the 2020 pink salmon harvest.

Table 2: Summary of model outputs and forecast error measures

X1	model	AdjR2	AICc	MAPE	MEAPE	MASE
model.m1	CPUE	0.6318782	25.70547	0.1002820	0.0692999	0.3520559
model.m2	CPUE+ISTI_MJJ	0.7692275	17.32303	0.0810289	0.0549547	0.2670966
model.m3	CPUE+ISTI	0.7810405	16.16703	0.0810207	0.0610963	0.2668789

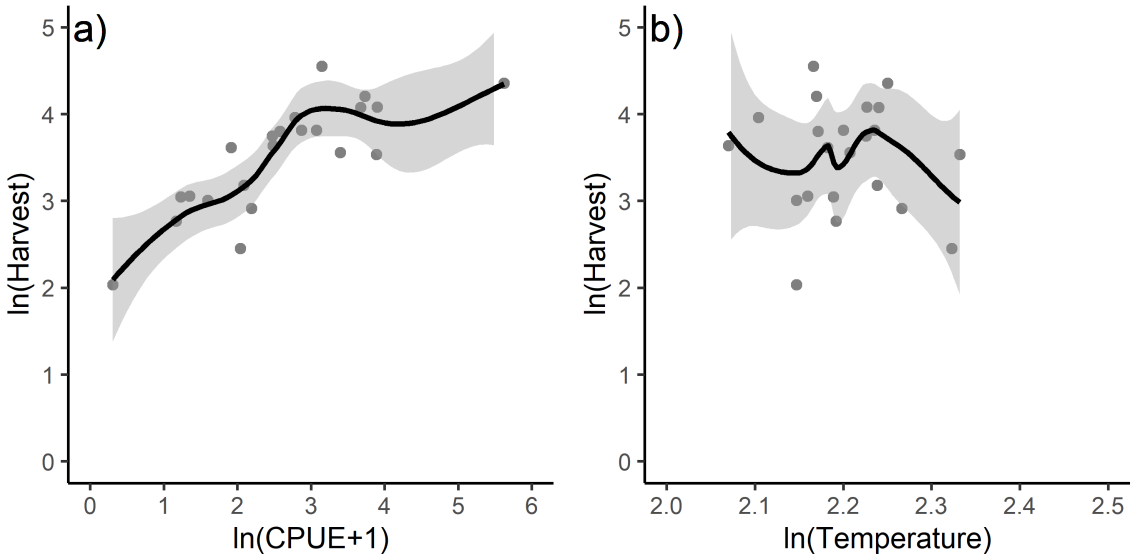

Figure 1: Relationship between a) ln(CPUE+1) and ln(harvest) and b) ln(temperature) in May through July (ISTI_MJJ_log) and ln(harvest).

Model Diagnostics

Model diagnostics for model 2 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 two or three times p/n, it may be a concern (where p is the number of parameters (i.e., 3) 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 0.27 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 i_{th} data point 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. 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.

Residuals vs. Fitted Plot

The characteristics of an unbiased residual vs. fitted plot and what they suggest about the appropriateness of the simple linear regression model include: 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. The above paragraph was taken almost directly from the source: https://newonlinecourses.science.psu.edu/stat462/node/117/

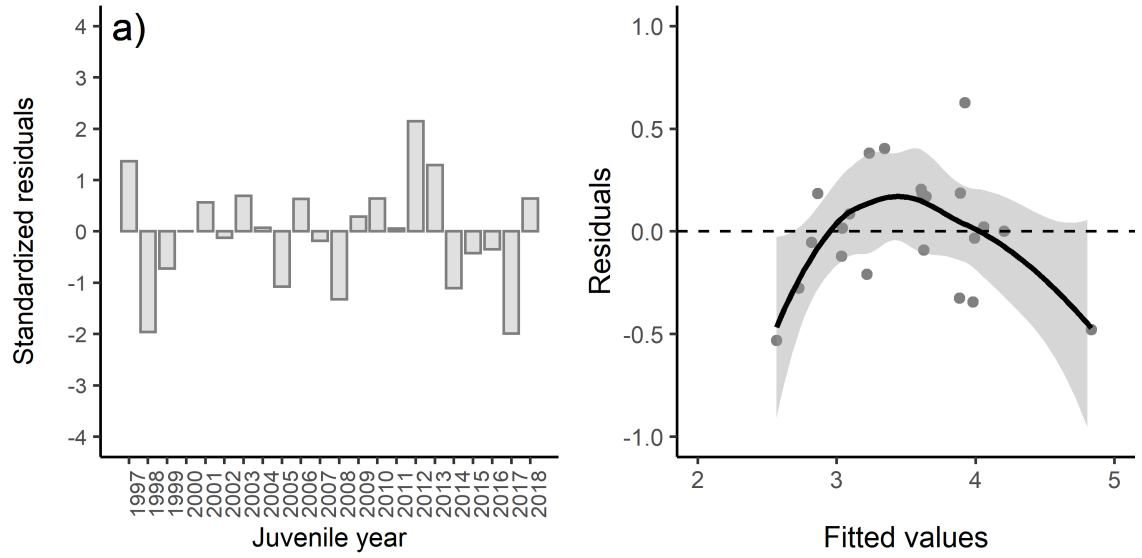

Figure 2: a) Standardized residuals versus juvenile year and b) residuals versus fitted values for model 2. Positive residuals indicate that the observed harvest was larger than predicted by the model.

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. The above paragraph was taken directly from the source: https://newonlinecourses.science.psu.edu/stat462/node/117/

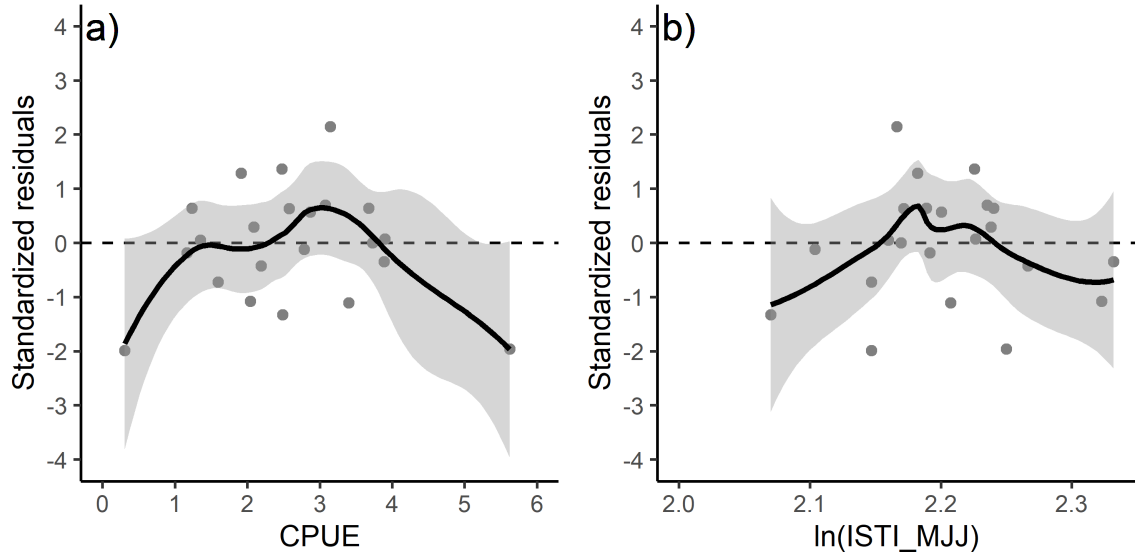

Figure 3: Standardized residuals versus predicted plots for a) CPUE and b) temperature for model 2.

Influential Datapoints

The Bonferroni outlier test for model 2 suggested that there were no outliers, although observation 16 was the most extreme (juvenile year 2012) based on standardized residuals (Table 3). The CPUE term was significant in the lack-of-fit curvature test ($P<0.05$), suggesting some lack of fit for this term.

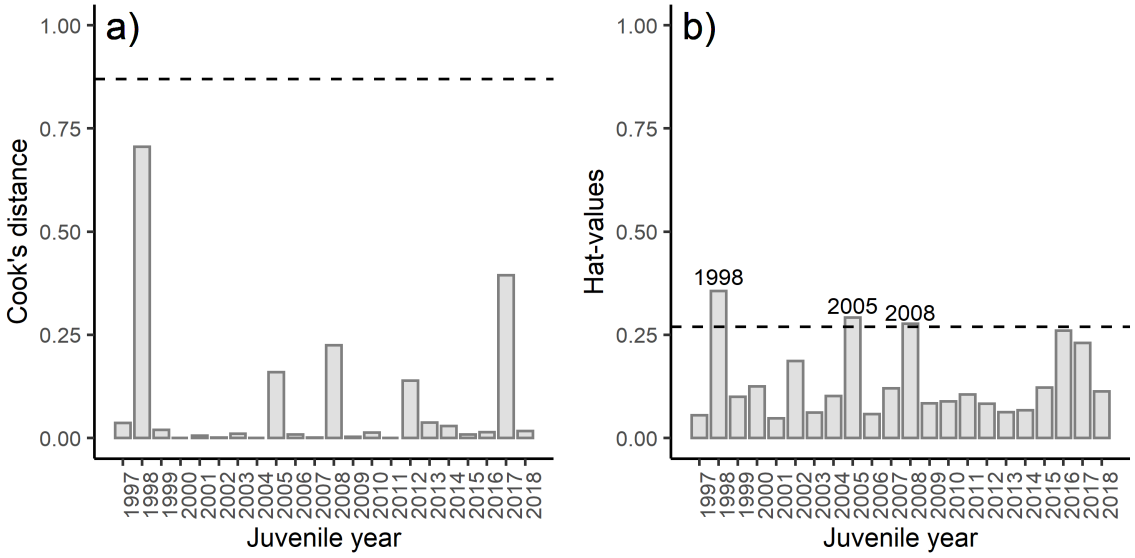

Figure 4: Diagnostics plots of influential observations including a) Cook's Distance (with a cut-off value of 0.87), and b) leverage values (with a cut-off value of 0.27) form model 2.

Results

The best regression model based on the MASE metric, significant coefficients in the models, and the argument for restricting temperature to the months when CPUE is sampled was model 2 (i.e. the model containing CPUE and May through July temperature). Diagnostics indicated some observations had high leverage values, but none of the observations affected model fitting and overall the model showed some 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, none of the data points had a studentized residual with a significant Bonferroni P-value suggesting that none of the 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, and the slightly curved fitted lines in the residual versus fitted plot, the fitted plot shows some lack of fit of the model. The adjusted R^2 value was 0.77 indicating overall 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 11.8 million fish (80% prediction interval: 7.4-18.8 million fish).

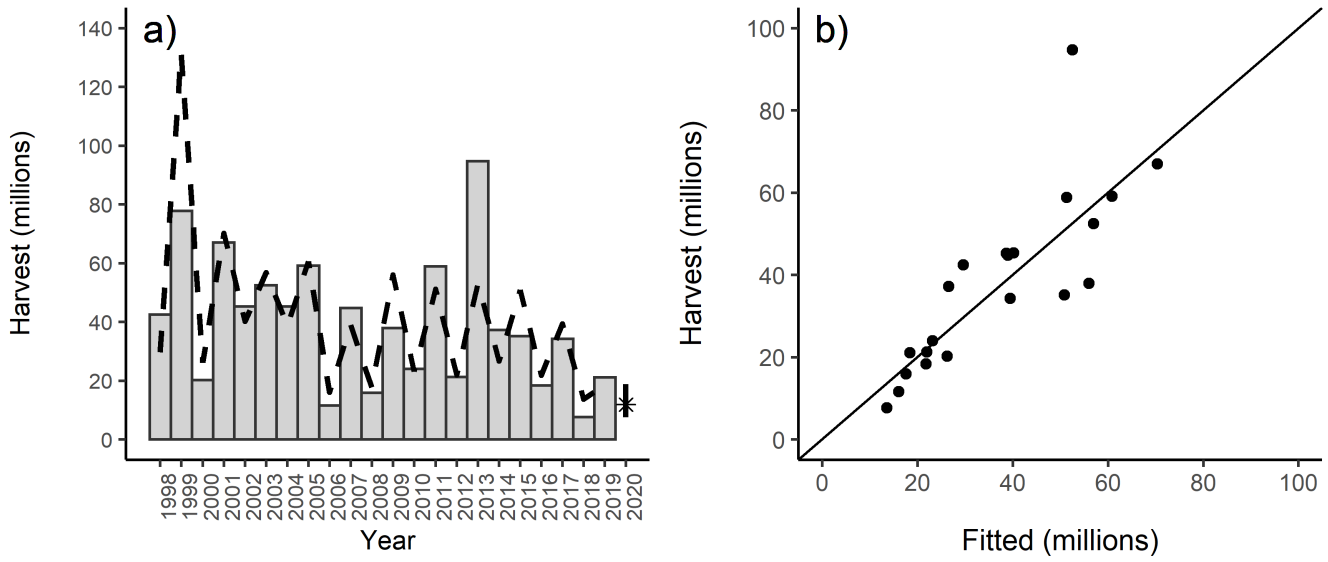

Figure 5: SEAK harvest (millions) a) by year and b) by the fitted values from model 2. The line in figure b is a one to one line. The predicted 2020 forecast is symbolized as a star with an 80% prediction interval (7.4-18.8 million fish) in figure a.

Table 3: Detailed output for model 2. Juvenile year 2012 (year 2013) shows the largest standardized residual.

X1	year	SEAKCatch_log	resid	hat_values	Cooks_distance	std_resid	fitted
1	1998	3.748327	0.4046601	0.0554505	0.0364524	1.3648456	3.343667
2	1999	4.354399	-0.4795375	0.3557974	0.7061429	-1.9584695	4.833936
3	2000	3.008155	-0.2097453	0.0997883	0.0194028	-0.7246454	3.217900
4	2001	4.204991	0.0005746	0.1250491	0.0000002	0.0020136	4.204416
5	2002	3.813748	0.1691461	0.0481844	0.0054502	0.5683177	3.644602
6	2003	3.960242	-0.0332751	0.1867656	0.0011199	-0.1209532	3.993517
7	2004	3.813528	0.2058933	0.0615330	0.0106083	0.6966878	3.607635
8	2005	4.079569	0.0205577	0.1021856	0.0001919	0.0711192	4.059012
9	2006	2.451867	-0.2767484	0.2917903	0.1595905	-1.0779775	2.728615
10	2007	3.802208	0.1876637	0.0577083	0.0081982	0.6337137	3.614545
11	2008	2.766319	-0.0526076	0.1205006	0.0015442	-0.1838808	2.818927
12	2009	3.636270	-0.3438873	0.2771450	0.2246606	-1.3258553	3.980157
13	2010	3.179303	0.0848265	0.0837078	0.0025695	0.2904824	3.094477
14	2011	4.075162	0.1858576	0.0891996	0.0133035	0.6383725	3.889304
15	2012	3.056357	0.0154383	0.1058205	0.0001130	0.0535172	3.040919
16	2013	4.550714	0.6267219	0.0831307	0.1391190	2.1454910	3.923992
17	2014	3.616309	0.3807868	0.0630647	0.0373096	1.2895334	3.235522
18	2015	3.558201	-0.3251765	0.0674788	0.0293885	-1.1038124	3.883378
19	2016	2.912351	-0.1216001	0.1226012	0.0084345	-0.4255405	3.033951
20	2017	3.535145	-0.0913536	0.2599768	0.0141901	-0.3481036	3.626499
21	2018	2.034841	-0.5318214	0.2305537	0.3944909	-1.9873860	2.566663
22	2019	3.047997	0.1836263	0.1125683	0.0172626	0.6389590	2.864371

References

Burnham, K. P., and Anderson, D. R. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods and Research 33: 261-304.

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.

Murphy, J. M., E.A. Fergusson, A. Piston, A. Gray, and E. Farley. In Press. Growth and harvest forecast models for Southeast Alaska pink salmon. North Pacific Anadromous Fish Commisson Technical Report No. 15:xx-xx.

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.