2021 Preseason Pink Salmon Forecast

Sara Miller, Rich Brenner, Jim Murphy

October 16, 2020--draft

# Objective

To forecast the Southeast Alaska (SEAK) pink salmon harvest in 2021.

# Executive Summary

Forecasts were developed using an approach originally described in Wertheimer et al. (2006), and modified in Orsi et al. (2016) and Murphy et al. (2019). We use a similar approach to Murphy et al. (2019) but assume a log-normal error. The approach is based on a multiple regression model with juvenile pink salmon catch-per-unit-effort (CPUE) and temperature data from the Southeast Alaska Coastal Monitoring Survey (SECM; Murphy et al. 2020). The final model used for the forecast was:

where is ln(pink salmon harvest in SEAK), is the coefficient for CPUE using the 'pooled' vessel calibration coeffcient (see *Calibration Coefficient Discusssion*), is the coefficient for the environmental covariate water temperature, and represents the normal error term that is lognormal. CPUE data are log-transformed catches that are standardized to an effort of a 20 minute trawl set and are calibrated to the fishing power of the NOAA Ship *John N. Cobb.* Water temperature data is the average (May through July) temperature in the upper 20 m at eight stations in Icy Strait. This is similar to what has been identified as the Icy Strait Temperature Index ('ISTI'), but historically this index has included May through August temperature.

Leave-one-out cross validation (hindcast) and model performance metrics such as Mean and Median Absolute Percentage Error (MAPE, MEAPE), mean absolute scaled error (MASE) (Hyndman and Kohler 2006), and Akaike Information Criterion corrected for small sample sizes (AICc values; Burnham and Anderson 2004) were used to evaluate forecast accuracy of alternative models. Statistical analyses were performed with the R Project for Statistical computing version 3.6.3 (R Core Team 2020). Based on the AICc, the MASE metric, and significant coefficients in the models, the preferred model (i.e., the additive model with CPUE and temperature) predicted that the SEAK pink salmon harvest in 2021 will be in the strong range with a point estimate of 28.3 million fish (80% prediction interval: 19.1 to 41.9 million fish).

# Calibration Coefficent Discussion

## Background

Excerpted from Wertheimer et al. 2010:

"From 1997 to 2007, SECM used the NOAA ship *John N. Cobb* to accrue an 11 year time series of catches with a Nordic 264 rope trawl fished at the surface... (Orsi et al. 2000, 2008)... In 2007, in anticipation of the decommissioning of the *John N. Cobb*, the *Medeia* and the *John N. Cobb* fished synoptically for 28 pairs of trawl hauls to develop calibration factors in the event of differential catch rates between the two vessels (Wertheimer et al. 2008). In 2008, the *Medeia* fished synoptically with the chartered research vessel *Steller* to determine relative fishing efficiency so that *Steller* catches could then be compared and calibrated to the SECM data series from the *John N. Cobb* (Wertheimer et al. 2009). In 2009, the commercial trawler *Chellissa* was chartered to fish the SECM transects in the northern and southern regions of Southeast Alaska. The *Medeia* was again fished synoptically in the northern region transects to determine relative fishing efficiency (Table 1)."

Estimated fishing power coefficients for juvenile pink salmon catches by the different vessels used during the Southeast Alaska Coastal Monitoring survey (Wertheimer et al. 2008, 2009, and 2010).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Chellissa:Medeia | Medeia:Cobb | Chellissa:Cobb | mixed Chellissa:Cobb | Medeia:Steller |
| Pink | 1.27 | 1.13 | 1.44 | 1.51 | 1.18 |

*'Chellissa:Cobb'* was calculated from estimates for *'Chellisa:Medeia'* and *'Medeia:Cobb'*. *'Mixed Chellissa:Cobb'* was a mixture of species estimates for *'Chellisa:Medeia'* and a pooled species estimate for *'Medeia:Cobb'*.

For the 2021 SEAK pink salmon forecast, there was a discussion as to which vessel calibration coefficient to use going forward. Using the four potential vessel calibration coefficients (mixed1, mixed2, species, pooled; Table 2), the corresponding time-series of juvenile abundance was slightly different (i.e., CPUE; standardized pink salmon catch based on a 20 min trawl set by year; Table 3). Performance metrics were used to evaluate forecast accuracy of alternative vessel calibration coefficients (Table 4) using the same model. The model used for the comparison of the vessel calibration coeffcient model was:

where is ln(CPUE+1), juvenile pink salmon abundance index based on the different vessel calibration coefficients, and is the average temperature in Icy Strait in May, June, and July at eight stations in Icy Strait.

Calibration coefficients used to convert vessel-specific catches to *Cobb* units. Direct calibrations with the *Cobb* are estimated for the *Steller* and *Medeia*, therefore mixed coefficients are only applied to the *Chellissa* and *NW Explorer*. Species-specific or pooled coefficients could be used as the mixed coefficients for the *Steller* and *Medeia*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Vessel | mixed1 | mixed2 | species | pooled |
| Cobb | 1.00 | 1.00 | 1.00 | 1.00 |
| Chellissa | 0.66 | 0.66 | 0.70 | 0.74 |
| NW Explorer | 0.66 | 0.66 | 0.70 | 0.74 |
| Steller | 1.05 | 0.96 | 0.96 | 1.05 |
| Medeia | 0.84 | 0.88 | 0.88 | 0.84 |

The data for the variable ln(CPUE+1) (time-series of juvenile pink salmon abundance) using different vessel calibration coefficients. The variable ln(CPUE+1) is defined as the average ln(CPUE+1) for catches in either the June or July SECM survey, whichever month had the highest average catches in a given year.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| year | mixed1 | mixed2 | species | pooled |
| 1997 | 2.4777 | 2.4777 | 2.4777 | 2.4777 |
| 1998 | 5.6224 | 5.6224 | 5.6224 | 5.6224 |
| 1999 | 1.5977 | 1.5977 | 1.5977 | 1.5977 |
| 2000 | 3.7300 | 3.7300 | 3.7300 | 3.7300 |
| 2001 | 2.8688 | 2.8688 | 2.8688 | 2.8688 |
| 2002 | 2.7847 | 2.7847 | 2.7847 | 2.7847 |
| 2003 | 3.0778 | 3.0778 | 3.0778 | 3.0778 |
| 2004 | 3.8994 | 3.8994 | 3.8994 | 3.8994 |
| 2005 | 2.0403 | 2.0403 | 2.0403 | 2.0403 |
| 2006 | 2.5728 | 2.5728 | 2.5728 | 2.5728 |
| 2007 | 1.1676 | 1.1676 | 1.1676 | 1.1676 |
| 2008 | 2.5768 | 2.3432 | 2.3432 | 2.5768 |
| 2009 | 2.0942 | 2.0942 | 2.2054 | 2.3330 |
| 2010 | 3.6878 | 3.6878 | 3.8836 | 4.1083 |
| 2011 | 1.3059 | 1.3059 | 1.3753 | 1.4548 |
| 2012 | 3.1610 | 3.1610 | 3.3289 | 3.5215 |
| 2013 | 1.9234 | 1.9234 | 2.0256 | 2.1428 |
| 2014 | 3.4266 | 3.4266 | 3.6086 | 3.8174 |
| 2015 | 2.2016 | 2.2016 | 2.3185 | 2.4526 |
| 2016 | 3.9057 | 3.9057 | 4.1131 | 4.3511 |
| 2017 | 0.3104 | 0.3104 | 0.3269 | 0.3458 |
| 2018 | 1.1716 | 1.2338 | 1.2338 | 1.1716 |
| 2019 | 1.1420 | 1.2026 | 1.2026 | 1.1420 |
| 2020 | 2.1475 | 2.2615 | 2.2615 | 2.1475 |
| average | 2.5372 | 2.5374 | 2.5861 | 2.6419 |

## Conclusion

Based on *x,y,and z???*, the discussion was limited to the 'species' and pooled' vessel calibration coefficients. Although there is a bit more statistical support for the species-specific coefficient ('species'), there is not a very meaningful difference between the 'species' and 'pooled' coefficents. As the pooled-species ('pooled') coefficients are currently used for sockeye, coho and Chinook salmon catches, the 'pooled' vessel calibration coefficient will be used moving forward for calculating the time-series of juvenile pink salmon abundance.

Comparison of the performance metrics for a model based on calculating ln(CPUE+1) using diferent vessel calibration coefficients.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| AdjR2 | AICc | MAPE | MEAPE | MASE | vessel calibration coefficient index |
| 0.810 | 17 | 0.080 | 0.060 | 0.263 | mixed1 |
| 0.819 | 15 | 0.077 | 0.060 | 0.257 | mixed2 |
| 0.828 | 14 | 0.075 | 0.064 | 0.251 | species |
| 0.821 | 15 | 0.076 | 0.063 | 0.257 | pooled |

# Forecast Models ('Pooled' Vessel Calibration Coefficient)

## Analysis

Three hierarchical models were investigated. The full model was:

where is the average ln(CPUE+1) for catches in either the June or July survey, whichever month had the highest average catches in a given year, and based on the 'pooled' vessel calibration coefficent, and is the average temperature in Icy Strait in May, June, and July at eight stations in Icy Strait (Icy Strait and Upper Chatham transects; 'ISTI'), and is the interaction term between CPUE and the temperature index. 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. In the past, the 'ISTI' variable was the average temperature in the upper 20 m during May through August at eight stations in Icy Strait. For simplicity, although the definition of the variable has changed, the variable is still called 'ISTI.'

The regression coefficients CPUE and temperature ('ISTI') were significant in the first two models (m1, m2). The interaction term was not significant in model m3 (Table 5). Therefore, only the first two models will be considered further.

Parameter estimates for the five potential models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | term | estimate | std.error | statistic | p.value |
| m1 | (Intercept) | 2.275 | 0.207 | 10.967 | 0.000 |
| m1 | CPUE | 0.441 | 0.071 | 6.232 | 0.000 |
| m2 | (Intercept) | 6.893 | 0.970 | 7.109 | 0.000 |
| m2 | CPUE | 0.503 | 0.051 | 9.860 | 0.000 |
| m2 | ISTI | -0.526 | 0.109 | -4.817 | 0.000 |
| m3 | (Intercept) | 3.917 | 2.464 | 1.590 | 0.128 |
| m3 | CPUE | 1.653 | 0.880 | 1.879 | 0.076 |
| m3 | ISTI | -0.201 | 0.270 | -0.746 | 0.465 |
| m3 | CPUE:ISTI | -0.125 | 0.095 | -1.310 | 0.206 |

The model summary results using the metrics AICc, MAPE, MEAPE, and MASE (Hyndman and Kohler 2006) are shown in Table 6. For all of these metrics, the smallest value is the preferred model. The difference () 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 AICc 2 have substantial support, those in which 4 AICc 7 have considerably less support, and models with AICc > 10 have essentially no support (Burnham and Anderson 2004). These two metrics (AICc, MASE) suggest that model m2 is the preferred models. Model m2 (based on CPUE and average temperature in May through July) was used to forecast the 2021 pink salmon harvest (Figure 1).

Summary of model outputs and forecast error measures.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | AdjR2 | AICc | MAPE | MEAPE | MASE |
| CPUE | 0.632 | 30 | 0.115 | 0.106 | 0.395 |
| CPUE+ISTI | 0.821 | 15 | 0.076 | 0.063 | 0.257 |

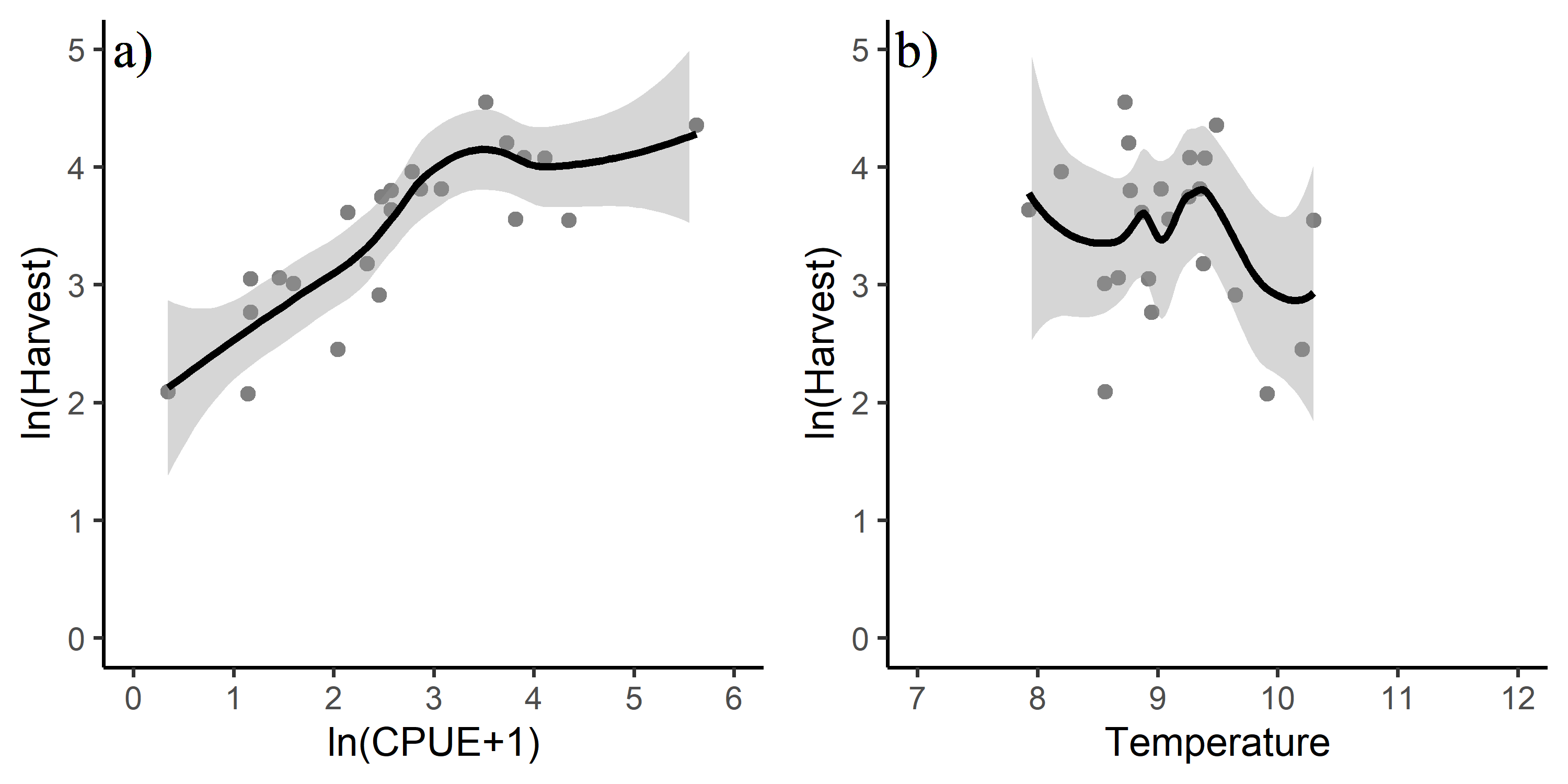


Figure 1: Relationship between a) ln(CPUE+1) and ln(harvest) and b) temperature in May through July (ISTI) and ln(harvest).

## Model Diagnostics

Model diagnostics for model m2 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.

### Cook's Distance

Cook's distance is a measure of influence, or the product of both leverage and outlier. Cook's distance,

where is the standardized Pearson residuals, are the hat values (measure of leverage), and is the number of predictor variables in the model, is a measure of overall influence of the data point on all fitted values (Fox and Weisburg 2019). A large value of Cook's distance indicates that the data point is an influential observation. Cook's distance values greater than , where is the number of observations (i.e., 23), was used as a benchmark for identifying the subset of influential observations (Ren et al. 2016). Therefore, a Cook's distance cut-off of 0.20 was used; observations with a Cook's distance greater than 0.20 were investigated further.

### Leverage

An observation that is distant from the average covariate pattern is considered to have high leverage. If an individual observation has a leverage value greater than 2 or 3 times (Ren et al. 2016), it may be a concern (where is the number of parameters in the model including the intercept (i.e., 3), and is the number of observations in the model (i.e., 23); = 3/23 = 0.13 for this study). Therefore, a leverage cut-off of 0.26 was used; observations with a leverage value greater than 0.26 were investigated further.

### 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; and
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/>.

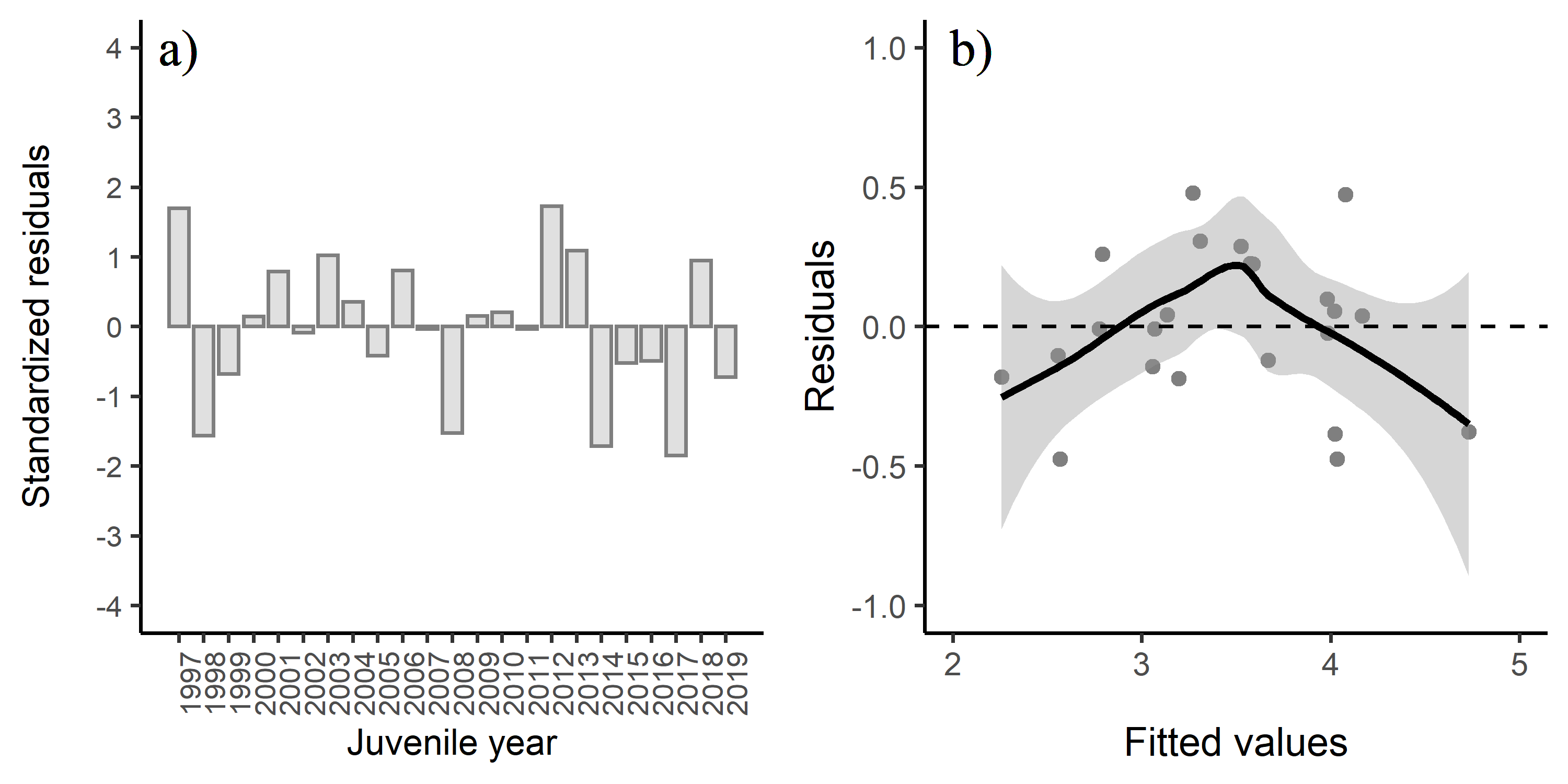


Figure 2: a) Standardized residuals versus juvenile year and b) residuals versus fitted values for model m2. Positive residuals indicate that the observed harvest was larger than predicted by the model.

### 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. In addition, 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/>.

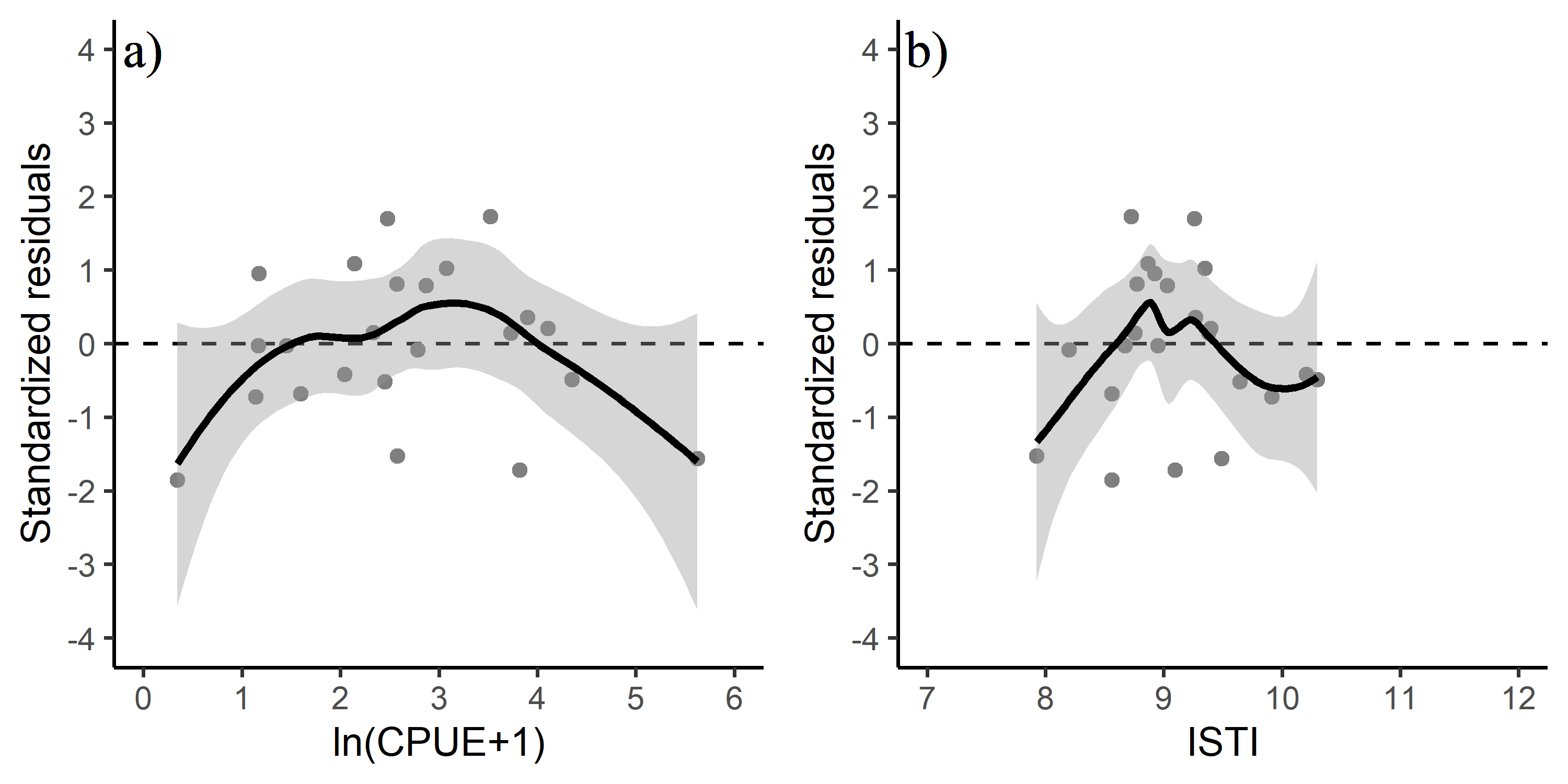


Figure 3: Standardized residuals versus predicted plots for a) CPUE and b) temperature.

### Influential Datapoints

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. The CPUE term was significant in the lack-of-fit curvature test (<0.05), suggesting some lack of fit for this term (Figure 3a). Diagnostics indicated that three of the data points were above the cut-off value for the Cook's distance (Figure 4a). Two observations had a high leverage value (Figure 4b), but none of the observations affected model fitting. Based on the Bonferroni outlier test, none of the data points had a studentized residual with a significant Bonferroni -value suggesting that none of the data points impacted the model fitting; although observations 16, 18 and 21 and were the most extreme (juvenile years 2012, 2014, and 2017 corresponding to years 2013, 2015, and 2018) based on standardized residuals (Figure 2a; Table 7). Based on the lightly curved fitted lines in the residual versus fitted plot (Figure 2b), the fitted plot shows some lack of fit of the model.

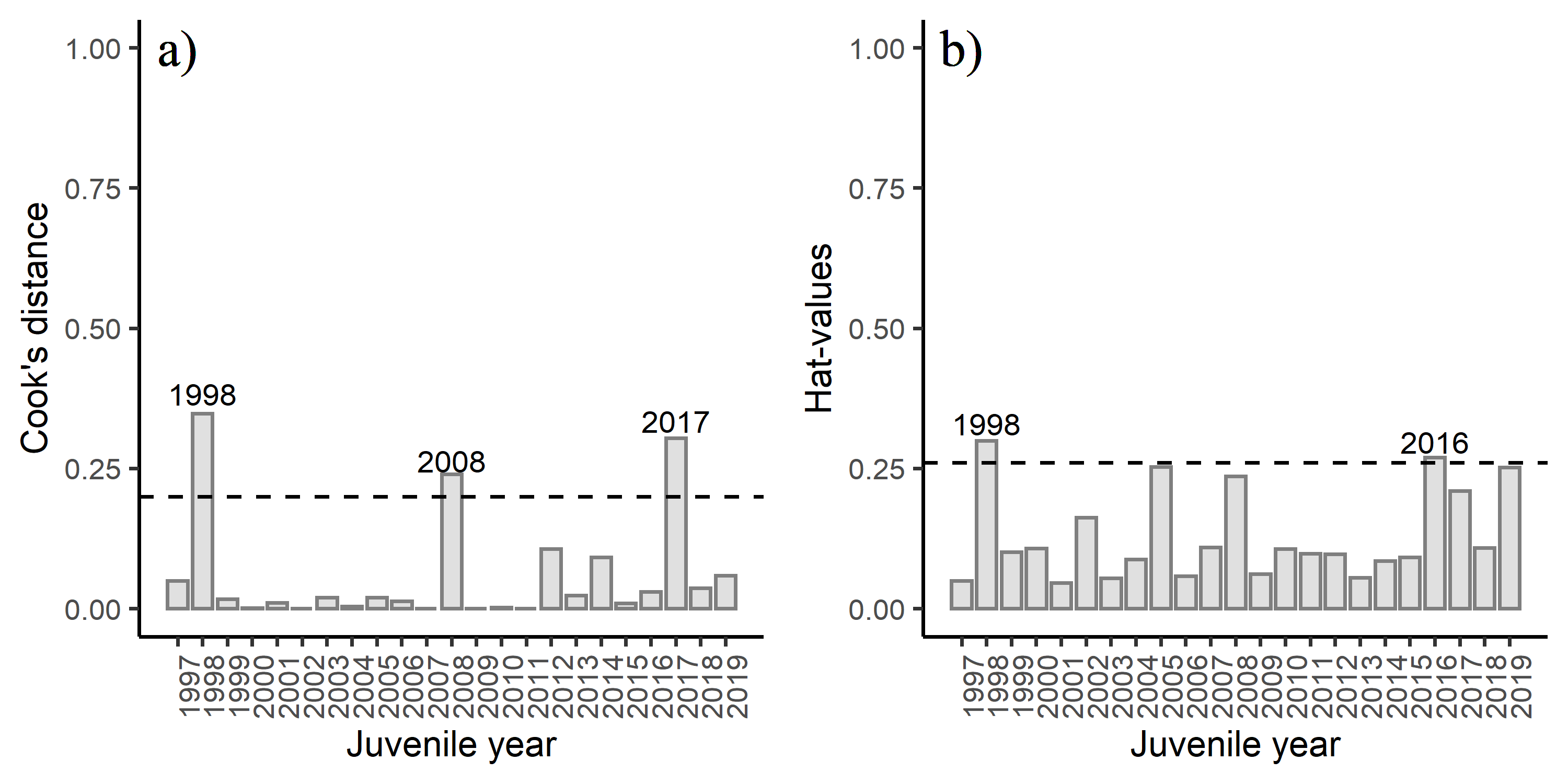


Figure 4: Diagnostics plots of influential observations including a) Cook's Distance (with a cut-off value of 0.20), and b) leverage values (with a cut-off value of 0.26) from model m2.

Detailed output for model m2. Juvenile year 2012, 2014, and 2017 (year 2013, 2015, and 2018) show the largest standardized residual. The variable SEAKCatch is in millions and the variable CPUE is ln(CPUE+1).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| year | juvenile\_year | SEAKCatch | CPUE | ISTI | resid | hat\_values | Cooks\_distance | std\_resid | fitted |
| 1998 | 1997 | 42.5 | 2.478 | 9.259 | 0.478 | 0.049 | 0.050 | 1.697 | 3.271 |
| 1999 | 1998 | 77.8 | 5.622 | 9.489 | -0.378 | 0.299 | 0.348 | -1.562 | 4.732 |
| 2000 | 1999 | 20.3 | 1.598 | 8.560 | -0.186 | 0.101 | 0.017 | -0.679 | 3.197 |
| 2001 | 2000 | 67.0 | 3.730 | 8.756 | 0.039 | 0.108 | 0.001 | 0.142 | 4.166 |
| 2002 | 2001 | 45.3 | 2.869 | 9.028 | 0.224 | 0.046 | 0.010 | 0.792 | 3.590 |
| 2003 | 2002 | 52.5 | 2.785 | 8.198 | -0.023 | 0.163 | 0.000 | -0.087 | 3.984 |
| 2004 | 2003 | 45.3 | 3.078 | 9.349 | 0.288 | 0.054 | 0.020 | 1.023 | 3.526 |
| 2005 | 2004 | 59.1 | 3.899 | 9.269 | 0.098 | 0.088 | 0.004 | 0.355 | 3.981 |
| 2006 | 2005 | 11.6 | 2.040 | 10.203 | -0.104 | 0.253 | 0.020 | -0.418 | 2.555 |
| 2007 | 2006 | 44.8 | 2.573 | 8.771 | 0.227 | 0.058 | 0.013 | 0.808 | 3.576 |
| 2008 | 2007 | 15.9 | 1.168 | 8.951 | -0.008 | 0.109 | 0.000 | -0.031 | 2.775 |
| 2009 | 2008 | 38.0 | 2.577 | 7.925 | -0.385 | 0.236 | 0.240 | -1.525 | 4.023 |
| 2010 | 2009 | 24.0 | 2.333 | 9.378 | 0.042 | 0.061 | 0.000 | 0.150 | 3.136 |
| 2011 | 2010 | 58.9 | 4.108 | 9.395 | 0.056 | 0.107 | 0.002 | 0.205 | 4.020 |
| 2012 | 2011 | 21.3 | 1.455 | 8.669 | -0.008 | 0.098 | 0.000 | -0.031 | 3.067 |
| 2013 | 2012 | 94.7 | 3.522 | 8.725 | 0.474 | 0.097 | 0.106 | 1.725 | 4.077 |
| 2014 | 2013 | 37.2 | 2.143 | 8.865 | 0.306 | 0.056 | 0.023 | 1.091 | 3.310 |
| 2015 | 2014 | 35.1 | 3.817 | 9.093 | -0.474 | 0.085 | 0.091 | -1.716 | 4.033 |
| 2016 | 2015 | 18.4 | 2.453 | 9.645 | -0.144 | 0.092 | 0.009 | -0.521 | 3.056 |
| 2017 | 2016 | 34.7 | 4.351 | 10.297 | -0.121 | 0.270 | 0.030 | -0.491 | 3.668 |
| 2018 | 2017 | 8.1 | 0.346 | 8.560 | -0.475 | 0.210 | 0.304 | -1.850 | 2.567 |
| 2019 | 2018 | 21.1 | 1.172 | 8.925 | 0.259 | 0.108 | 0.037 | 0.949 | 2.790 |
| 2020 | 2019 | 8.0 | 1.142 | 9.911 | -0.182 | 0.252 | 0.059 | -0.727 | 2.257 |

## Results

The best regression model based on the AICc value, the MASE metric, and significant coefficients in the model was model m2 (i.e. the model containing CPUE, and a May through July temperature variable). The adjusted value for model m2 was 0.82 (Table 6) indicating overall a good model fit.

## Conclusion

The SEAK pink salmon harvest in 2021 is predicted to be in the strong range with a point estimate of 28.3 million fish (80% prediction interval: 19.1 to 41.9 million fish).

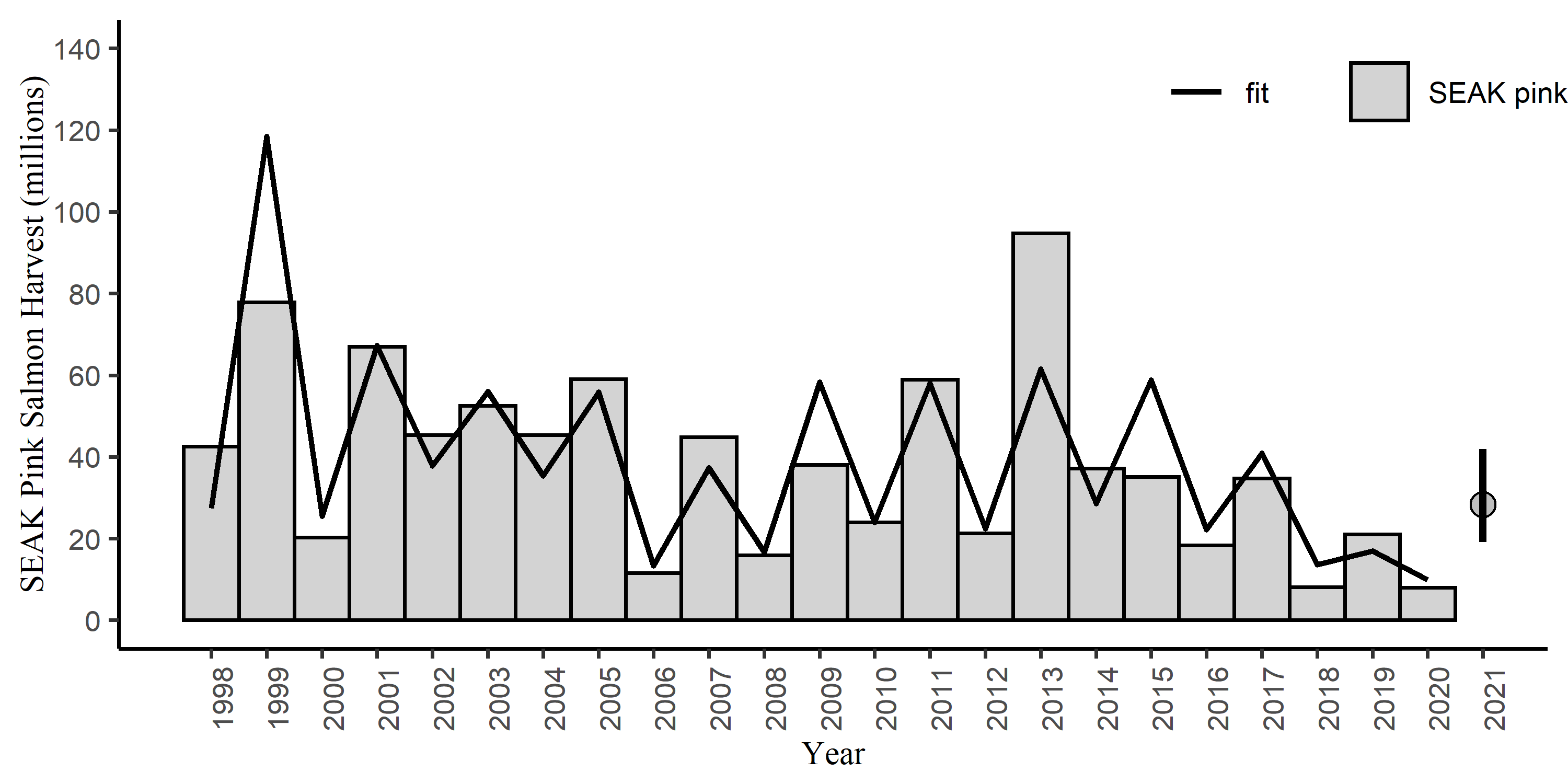


Figure 5: SEAK pink salmon harvest (millions) by year with the model fit (line). The predicted 2021 forecast is symbolized as a grey circle with an 80% prediction interval (19.1 to 41.9 million fish).

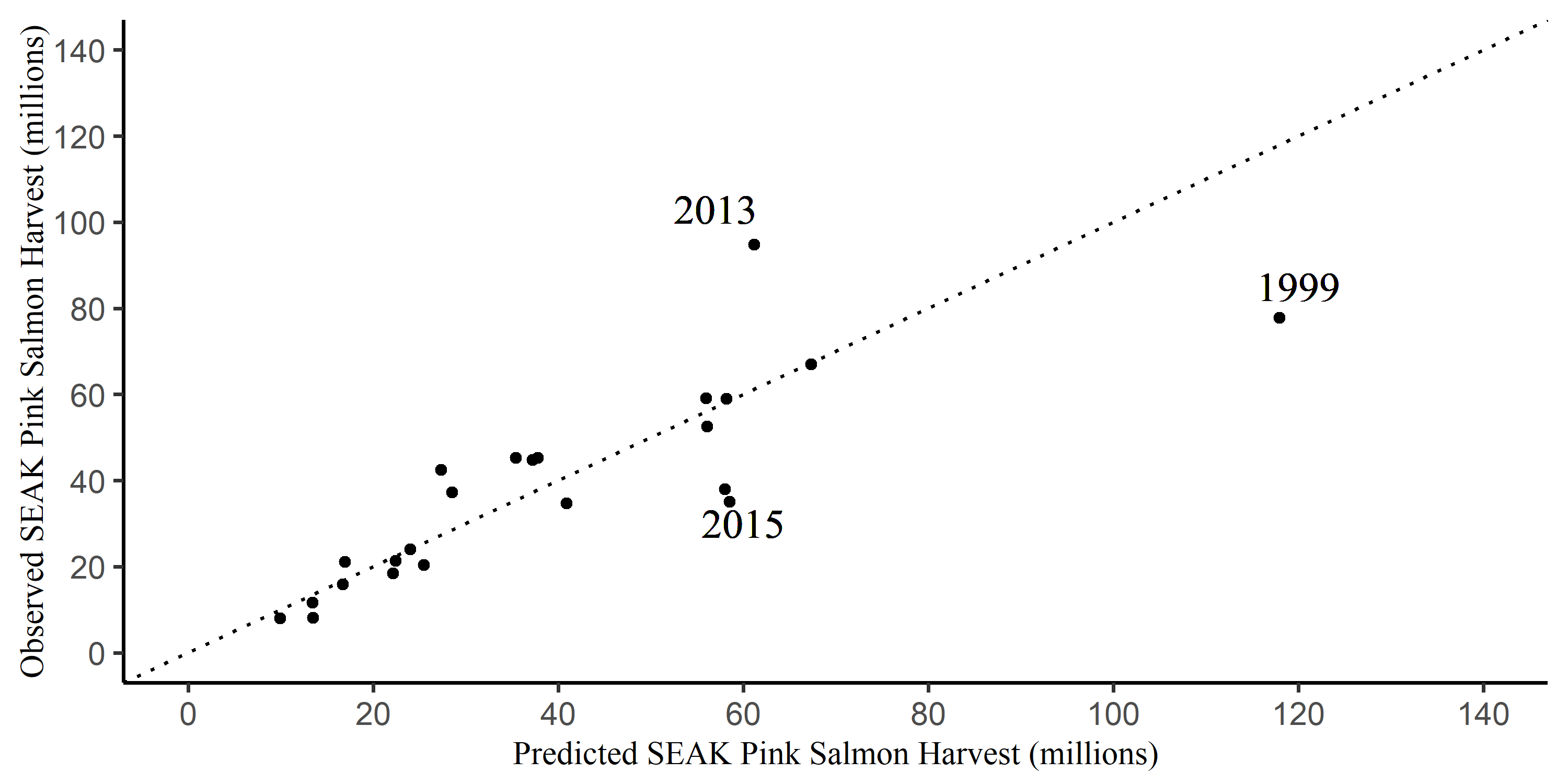


Figure 6: SEAK pink salmon harvest (millions) against the fitted values from model m2 by year. The dotted line is a one to one line.

# References

Burnham, K. P., and D. R. Anderson. 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.

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. 2019. Southeast Alaska pink salmon growth and harvest forecast models. North Pacific Anadromous Fish Commisson Technical Report No. 15: 75-81.

Orsi, J. A., M. V. Sturdevant, J. M. Murphy, D. G. Mortensen, and B. L. Wing. 2000. Seasonal habitat use and early marine ecology of juvenile Pacific salmon in southeastern Alaska. NPAFC Bull. 2:111-122.

Orsi, J. A., E. A. Fergusson, M. V. Sturdevant, B. L. Wing, A. C. Wertheimer, and W. R. Heard. 2008. Annual survey of juvenile salmon and ecologically related species and environmental factors in the marine waters of southeastern Alaska, May–August 2007. NPAFC Doc. 1110, 82 pp.

Orsi, J.A., E.A. Fergusson, A.C. Wertheimer, E.V. Farley, and P.R. Mundy. 2016. Forecasting pink salmon production in Southeast Alaska using ecosystem indicators in times of climate change. N. Pac. Anadr. Fish Comm. Bull. 6: 483–499. (Available at <https://npafc.org>) R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.r-project.org/index.html>

Ren, Y. Y., L. C. Zhou, L. Yang, P. Y. Liu, B. W. Zhao, and H. X. Liu. 2016. Predicting the aquatic toxicity mode of action using logistic regression and linear discriminant analysis. SAR and QSAR in Environmental Research 27(9). DOI: 10.1080/1062936X.2016.1229691

Wertheimer A. C., J. A. Orsi, M. V. Sturdevant, and E. A. Fergusson. 2006. Forecasting pink salmon harvest in Southeast Alaska from juvenile salmon abundance and associated environmental parameters. In Proceedings of the 22nd Northeast Pacific Pink and Chum Workshop. Edited by H. Geiger (Rapporteur). Pac. Salmon Comm. Vancouver, British Columbia. pp. 65–72.

Wertheimer, A. C., J. A. Orsi, E. A. Fergusson, and M. V. Sturdevant. 2008. Paired comparisons of juvenile salmon catches between two research vessels fishing Nordic 264 surface trawls in southeastern Alaska, July 2007. NPAFC Doc. 1112.,17 p.

Wertheimer, A. C., J. A. Orsi, E. A. Fergusson, and M. V. Sturdevant. 2009. Calibration of Juvenile Salmon Catches using Paired Comparisons between Two Research Vessels Fishing Nordic 264 Surface Trawls in Southeastern Alaska, July 2008. NPAFC Doc. 1180. 18 pp.

Wertheimer, A. C., J. A. Orsi, E. A. Fergusson, and M. V. Sturdevant. 2010. Calibration of Juvenile Salmon Catches using Paired Comparisons between Two Research Vessels Fishing Nordic 264 Surface Trawls in Southeast Alaska, July 2009. NPAFC Doc. 1177. 19 pp. (Available at <http://www.npafc.org>).

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