2020 Preseason Pink Salmon Forecast

Sara Miller and Rich Brenner

October 2019

# Objective

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

# Executive Summary

Forecasts were developed using an approach described in Wertheimer et al. (2011). A linear 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:

where is ln(harvest), is the coefficient for ln(CPUE+1), are the coefficients for environmental covariates (e.g., water temperatures, climate indices, fish size and condition), and represents the normally distributed error term. Backward/forward stepwise regression was used for environmental covariate selection ( value threshold 0.01) and models were ranked via small sample Akaike Information Criterion (AICc). 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).

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% confidence interval: 7.4–18.8 million fish).

# Analysis

Three hierarchical models were investigated. The full model was:

where was ln(CPUE+1) and was the natural log of 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 is not significant (Table 1). Therefore, only the first two models will be considered.

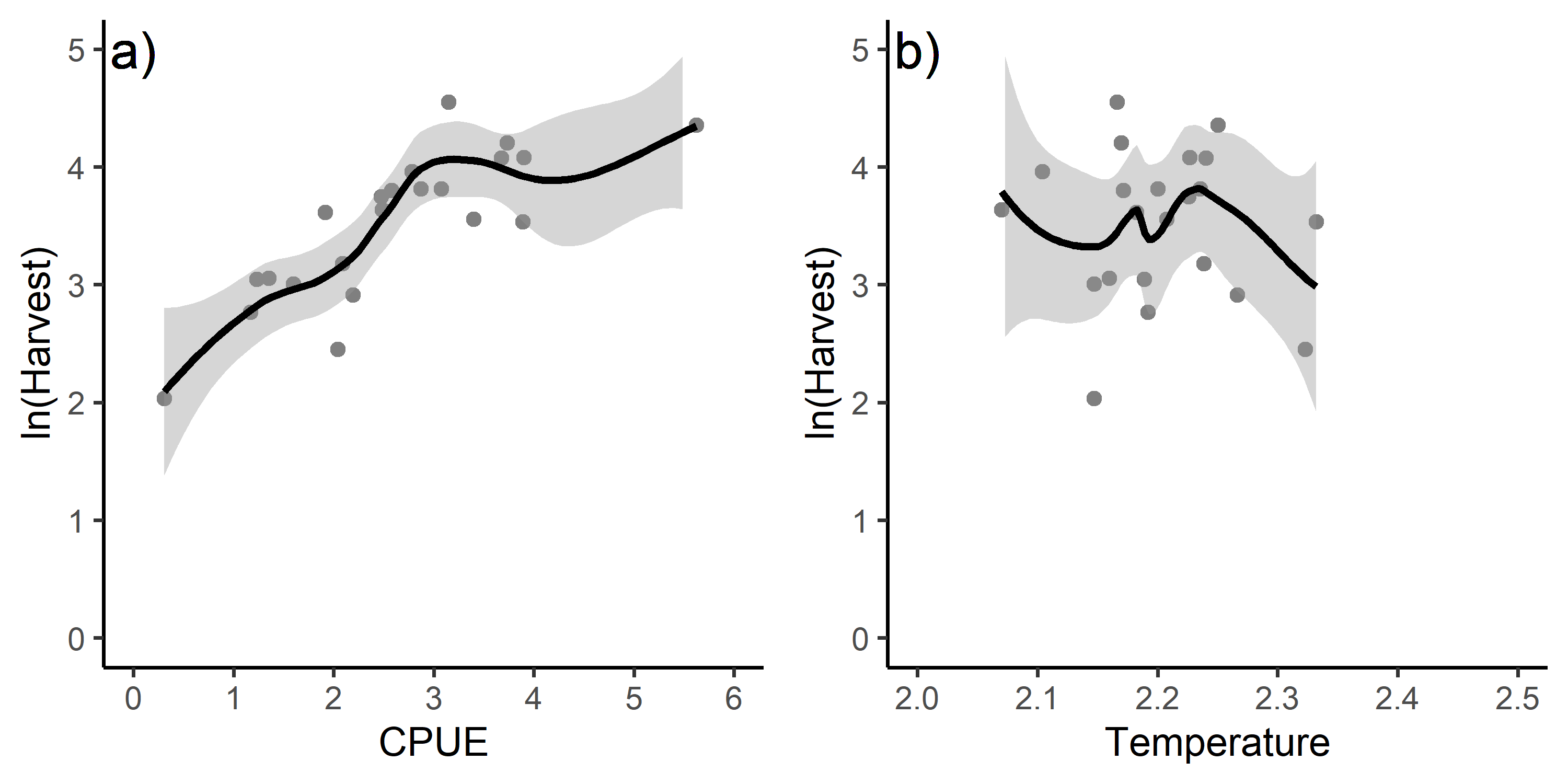


Figure 1: Relationship between a) cpue and ln(harvest) and b) temperature and harvest.

Parameter estimates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X1 | term | estimate | std.error | statistic | p.value |
| 1 | (Intercept) | 2.3738613 | 0.2038433 | 11.6455179 | 0.0000000 |
| 2 | CPUE | 0.4342078 | 0.0713387 | 6.0865706 | 0.0000060 |
| 3 | (Intercept) | 11.1295695 | 2.4428039 | 4.5560635 | 0.0002159 |
| 4 | CPUE | 0.5054830 | 0.0598672 | 8.4433996 | 0.0000001 |
| 5 | ISTI\_MJJ\_log | -4.0609645 | 1.1305155 | -3.5921351 | 0.0019429 |
| 6 | (Intercept) | -1.2818793 | 6.8493217 | -0.1871542 | 0.8536331 |
| 7 | CPUE | 5.1751159 | 2.4298380 | 2.1298193 | 0.0472408 |
| 8 | ISTI\_MJJ\_log | 1.5765629 | 3.1176953 | 0.5056822 | 0.6192162 |
| 9 | CPUE:ISTI\_MJJ\_log | -2.1125090 | 1.0989493 | -1.9222988 | 0.0705407 |

The model summary results using the metrics AICc, MAPE, MEAPE, and MASE(Hyndman and Kohler 2006) are shown in Tables 2 and 3. For all 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 2 have substantial support, those in which 4 7 have considerably less support, and models with > 10 have essentially no support (Burnham and Anderson 2004). These two metrics suggest that model two is the preferred model.

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 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 greater than two or three times , it may be a concern (where is the number of parameters (i.e., 3) and is the number of observations (i.e., 22); = 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,

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 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 and Weisberg (1994) suggest using the median of the F-distribution with and 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 (< 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.

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

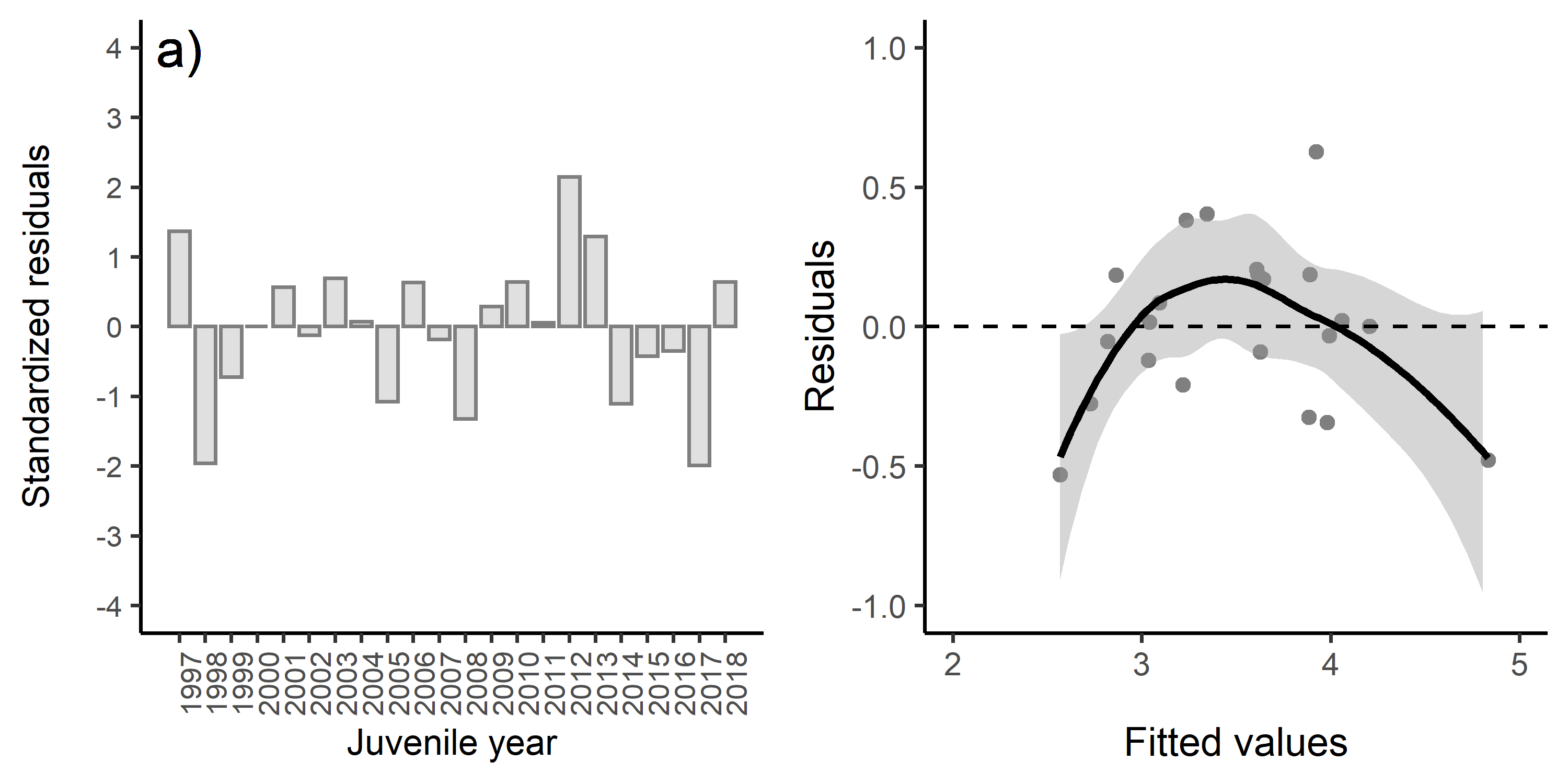


Figure 2: a) Standardized residuals versus juvenile year and b) residuals versus fitted values for model 2. Positive residuals indicate that the observed harvest was larger than predicted by the model.

## 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/>

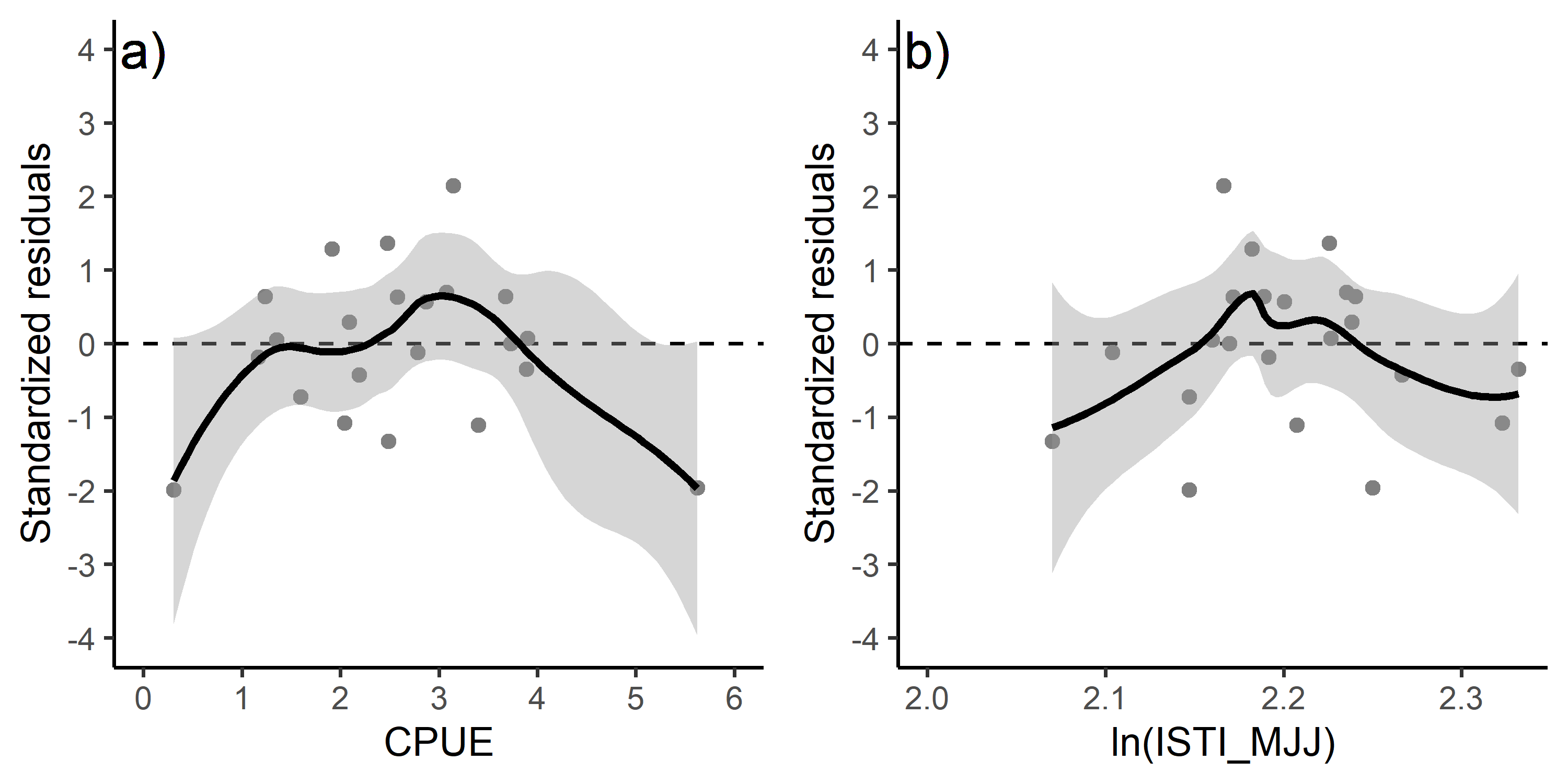


Figure 3: Standardized residuals versus predicted plots for a) CPUE and b) temperature for model 2.

## Influential Datapoints

The Bonferroni outlier test suggested that observaton 16 (juvenile year 2012) was an outlier. The cpue term was significant in the lack-of-fit curvature test (<0.05), suggesting some lack of fit for this term.

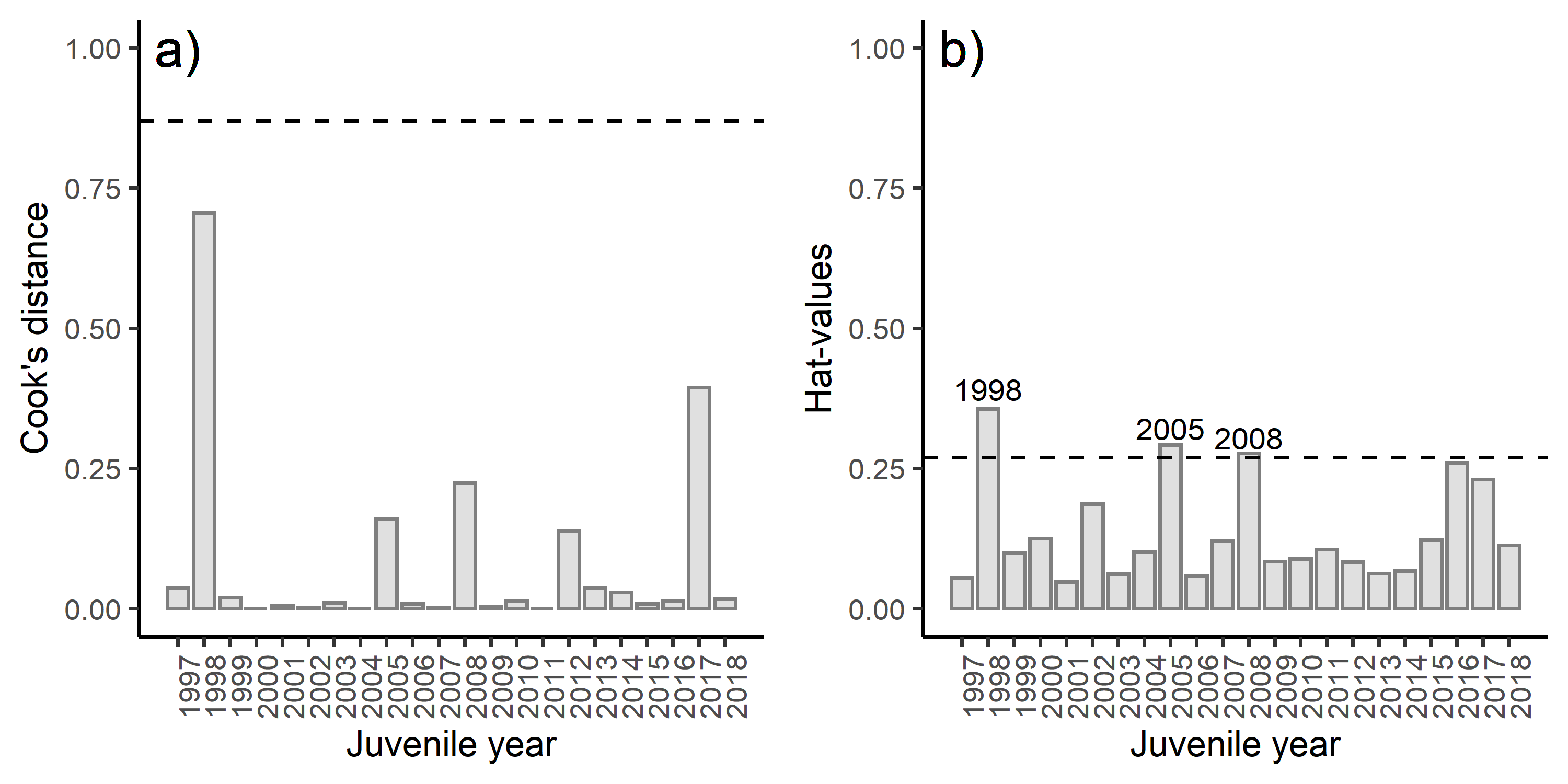


Figure 4: Diagnostics plots of influential observations including a) Cook’s Distance (with a cut-off value of 0.87), b) leverage values (with a cut-off value of 0.27).

# 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 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, one the data points had a studentized residual with a significant Bonferroni -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, 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 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% confidence interval: 7.4-18.8 million fish).

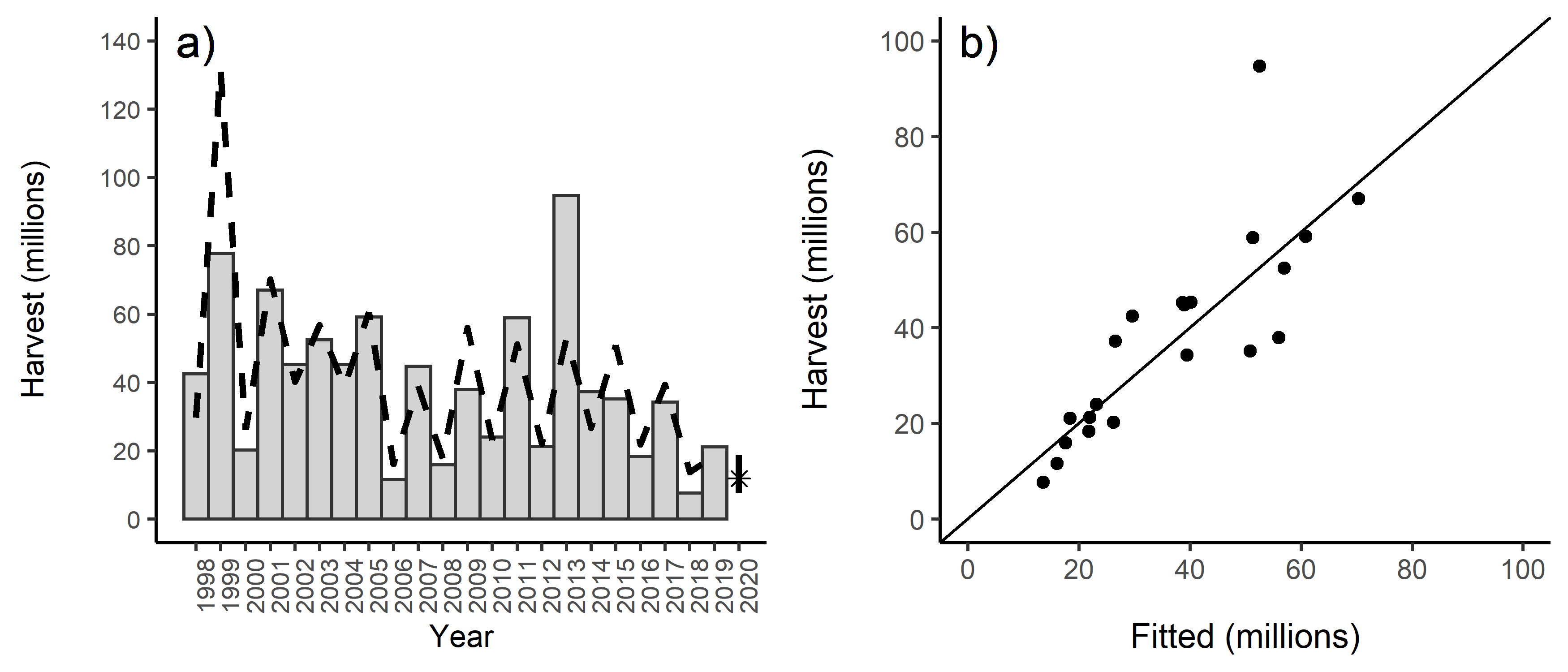


Figure 5: SEAK harvest (millions) a) by year and b) by the fitted values from model 2. The line in figure b is a one to one line. The predicted 2020 forecast is symbolized as a star with an 80% prediction interval (0-20 million fish).

# 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.

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/>.

Wertheimer, A. C., J. A. Orsi, E. A. Fergusson, and M. V. Sturdevant. 2011. Forecasting pink salmon harvest in Southeast Alaska from juvenile salmon abundance and associated environmental parameters: 2010 returns and 2011 forecast (NPAFC Doc. 1343) Auke Bay Lab., Alaska Fish. Sci. Cen., Nat. Mar. Fish. Serv., NOAA, 17109 Point Lena Loop Road, Juneau, AK 99801-8626, USA, 20 p.; <http://www.npafc.org/new/pub_documents.html>.

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