SEAK Pink Salmon 2022 Forecast–final

Sara Miller

November 3, 2021

Objective

To forecast the Southeast Alaska (SEAK) pink salmon commercial harvest in 2022.

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 used a similar approach to Murphy et al. (2019), but assumed a log-normal error. This 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; Piston et al. 2021). The final model used for the forecast was:

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

where y is log-transformed pink salmon harvest in SEAK, β_1 is the coefficient for CPUE using the pooled-species vessel calibration coefficient, β_2 is the coefficient for the environmental covariate water temperature, and ϵ represents the normally distributed error term. The CPUE data are the average log-transformed catches standardized to an effort of a 20 minute trawl set and calibrated to the fishing power of the NOAA Ship *John N. Cobb.* For each year, the standardized catch is either taken from June or July, whichever month had the highest average catches. Water temperature data is the average (May through July) temperature in the upper 20 m at the eight SECM stations in Icy Strait.

Leave-one-out cross validation (hindcast) and model performance metrics were used to evaluate the forecast accuracy of models. These metrics included Akaike Information Criterion corrected for small sample sizes (AICc values; Akaike 1973; Burnham and Anderson 2004), the mean absolute scaled error (MASE metric; Hyndman and Kohler 2006), the weighted mean absolute percentage error (wMAPE; based on the last 5 years), leave one out cross validation MAPE (MAPE_LOOCV), one step ahead forecasts (MAPE_one_step_ahead) for the last five years (years 2017 through 2021), and significant coefficients (i.e., covariates) in the model. Based on the all the performance metrics (AICc, MASE, wMAPE, MAPE_LOOCV, MAPE_one_step_ahead), the preferred model (i.e., the additive model with CPUE and temperature; model m2) predicted that the SEAK pink salmon harvest in 2022 will be in the weak range with a point estimate of 15.58 million fish (80% prediction interval: 10.30 to 23.57 million fish).

Analysis

Model data

The data used in the model are shown in table 1.

Table 1: Model data.

Year	Harvest	CPUE	Temperature
1998	42.50	2.48	9.28
1999	77.80	5.62	9.40
2000	20.30	1.60	8.56
2001	67.00	3.73	8.77
2002	45.30	2.87	9.03
2003	52.50	2.78	8.20
2004	45.30	3.08	9.31
2005	59.10	3.90	9.33
2006	11.60	2.04	10.21
2007	44.80	2.58	8.75
2008	15.90	1.17	8.94
2009	38.00	2.32	7.91
2010	24.00	2.33	9.36
2011	58.90	4.11	9.35
2012	21.30	1.51	8.65
2013	94.70	3.52	8.48
2014	37.20	2.14	8.83
2015	35.10	3.80	9.12
2016	18.40	2.45	9.61
2017	34.70	4.35	10.20
2018	8.10	0.35	8.56
2019	21.10	1.17	8.92
2020	8.07	1.14	9.91
2021	48.40	2.15	8.89
2022	NA	0.88	8.89

Hierarchical models

Two hierarchical models were investigated. The full model was:

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

where X_1 is the average CPUE for catches in either the June or July survey, whichever month had the highest average catches in a given year, and was based on the pooled-species vessel calibration coefficient, and X_2 is a temperature index (i.e, average May through July temperature in the upper 20 m at the eight SECM stations in Icy Strait; ISTI20_MJJ). The CPUE data were log-transformed in the model, but temperature data were not. The simplest model did not contain a temperature variable (model m1). Parameter estimates are shown in table 2.

Table 2: Parameter estimates for the 2 models.

model	term	estimate	std.error	statistic	p.value
m1	(Intercept)	2.3352743	0.211	11.051	0
m1	CPUE	0.4304903	0.073	5.892	0
m2	(Intercept)	7.2663302	0.983	7.388	0
m2	CPUE	0.4932528	0.052	9.549	0
m2	ISTI20_MJJ	-0.5621995	0.111	-5.069	0

Performance metrics

The model summary results using the performance metrics AICc, MASE, wMAPE, MAPE_LOOCV, and MAPE_one_step_ahead) are shown in table 3. For all of these metrics, the smallest value is the preferred model. Models with Δ_iAICc ≤ 2 have substantial support, those in which $4 \leq \Delta_i$AICc ≤ 7 have considerably less support, and models with Δ_iAICc > 10 have essentially no support (Burnham and Anderson 2004). The performance metric MAPE was calculated as:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} |\frac{A_t - F_t}{A_t}|$$

where A_t is the observed value and F_t is the predicted value. The performance metric wMAPE was calculated as:

$$\text{wMAPE} = \sum_{t=1}^{n} \frac{1}{w_t} \sum_{t=1}^{n} |\frac{A_t - F_t}{A_t}| w_t.$$

where w_t is the weight for each year. For the wMAPE metric, the last 5 years (juvenile years 2016-2020) were given a weight of 1 and all other years, a weight of 0.001. Therefore, compared to the performance metric MAPE_LOOCV, the performance of the model in the last 5 years was given more weight in the wMAPE metric.

The performance metric MAPE_LOOCV uses five steps.

1. The dataset is split into a training set. The training set uses all but one observation of the full dataset.
2. Run the regression model based on the training set.
3. Use the regression model based on the training set to predict F_t for the one observation left out of the model.
4. Repeat the process n times based on the number of observations in the dataset, leaving out a different observation from the training set each time.
5. Calculate MAPE, based on the average of all the training datasets (i.e., one MAPE is calculated for each training set and then these are averaged).

The performance metric MAPE_one_step_ahead involves three steps:

1. Estimate the regression parameters at time t from data up to time $t-1$.
2. Make a prediction of F_t at time t based on the predictor variables at time t and the estimate of the regression parameters at time t.
3. Calculate the MAPE based on the prediction of F_t at time t and the observed value of A_t at time t.
4. Repeat this for data up through year 2016 (e.g., data up through year 2016 is $t-1$ and the forecast is for year 2017; t), data up through year 2017 (e.g., data up through year 2017 is $t-1$ and the forecast is for year 2018; t), data up through year 2018 to forecast 2019, data up through year 2019 to forecast 2020, and data up through year 2020 to forecast 2021.
5. The MAPE_one_step_ahead will then be an average of the MAPE calculated from these five forecasts.

The AICc in table 3 is the AICc value and not the Δ_iAICc.

Table 3: Summary of model outputs and forecast error measures. These metrics included Akaike Information Criterion corrected for small sample sizes (AICc values), the mean absolute scaled error (MASE metric), the weighted mean absolute percentage error (wMAPE; based on the last 5 years), leave one out cross validation MAPE (MAPE_LOOCV), and one step ahead forecasts (MAPE_one_step_ahead).

model	AdjR2	AICc	MASE	wMAPE	MAPE_LOOCV	MAPE_one_step_ahead
m1	0.594	32.56	0.389	0.191	0.118	0.214
m2	0.809	16.28	0.255	0.117	0.079	0.133

Log transformation bias in a linear-model

To correct for log transformation bias in a linear-model, a bias correction (Miller 1984) was applied to the predicted 2022 SEAK harvest and its prediction interval (output from the car package (Fox and Weisberg 2019) in program R; R Core Team 2020; table 4) from each of the two models. The bias correction, applied to each value, is:

$$\hat{E}(Y_m) = \exp(\hat{\mu_{\mathrm{m}}} + \frac{\hat{\sigma_{\mathrm{m}}}^2}{2})$$

where $\hat{\mu}$ is the predicted value (or 80% upper or lower prediction interval value) from the individual model m.

Table 4: Summary of model forecasts including the 80 percent prediction intervals (corrected for log transformation bias in a linear-model).

model	terms	fit	fit_LPI	fit_UPI
m1	CPUE	16.513	9.041	30.161
m2	CPUE + ISTI20_MJJ	15.579	10.297	23.571

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.

Table 5: Detailed output for model m2. Juvenile years 1997, 1998, 2014, 2017, and 2020 (years 1998, 1999, 2015, 2018, 2021) show the largest standardized residual. Year refers to the forecast year. Fitted values are log-transformed.

year	SEAKCatch	CPUE	ISTI20_MJJ	resid	hat_values	Cooks_distance	std_resid	fit_bias_corrected
1998	42.5	2.48	9.28	0.47	0.05	0.05	1.65	27.61
1999	77.8	5.62	9.40	-0.40	0.30	0.38	-1.63	121.32
2000	20.3	1.60	8.56	-0.23	0.09	0.02	-0.83	26.75
2001	67.0	3.73	8.77	0.03	0.10	0.00	0.10	67.96
2002	45.3	2.87	9.03	0.21	0.04	0.01	0.71	38.52
2003	52.5	2.78	8.20	-0.07	0.15	0.00	-0.25	58.62
2004	45.3	3.08	9.31	0.26	0.05	0.02	0.91	36.45
2005	59.1	3.90	9.33	0.14	0.09	0.01	0.49	53.85
2006	11.6	2.04	10.21	-0.08	0.26	0.01	-0.33	13.17
2007	44.8	2.58	8.75	0.18	0.06	0.01	0.64	38.96
2008	15.9	1.17	8.94	-0.05	0.10	0.00	-0.19	17.51
2009	38.0	2.32	7.91	-0.33	0.22	0.15	-1.25	54.92
2010	24.0	2.33	9.36	0.02	0.06	0.00	0.08	24.50
2011	58.9	4.11	9.35	0.04	0.11	0.00	0.15	59.05
2012	21.3	1.51	8.65	-0.09	0.09	0.00	-0.31	24.28
2013	94.7	3.52	8.48	0.31	0.13	0.07	1.14	72.26
2014	37.2	2.14	8.83	0.26	0.05	0.02	0.91	29.91
2015	35.1	3.80	9.12	-0.46	0.08	0.08	-1.62	57.78
2016	18.4	2.45	9.61	-0.16	0.09	0.01	-0.57	22.58
2017	34.7	4.35	10.20	-0.13	0.25	0.03	-0.52	41.33
2018	8.1	0.35	8.56	-0.53	0.20	0.35	-2.03	14.43
2019	21.1	1.17	8.92	0.22	0.10	0.02	0.80	17.62
2020	8.1	1.14	9.91	-0.17	0.25	0.05	-0.66	9.97
2021	48.4	2.15	8.89	0.55	0.05	0.07	1.92	29.13

Cook's distance

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

$$D_i = \frac{e^2_{PSi}}{k+1} * \frac{h_i}{1-h_i},$$

where e^2_{PSi} is the standardized Pearson residuals, h_i are the hat values (measure of leverage), and k 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's distance values greater than $4/(n-k-1)$, where n is the number of observations (i.e., 24), was used as a benchmark for identifying the subset of influential observations (Ren et al. 2016). Therefore, a Cook's distance cut-off of 0.19 was used; observations with a Cook's distance greater than 0.19 were investigated further (Figure 1a).

Leverage

An observation that is distant from the average covariate pattern is considered to have high leverage or hat-value. If an individual observation has a leverage value h_i greater than 2 or 3 times p/n (Ren et al. 2016), it may be a concern (where p is the number of parameters in the model including the intercept (i.e., 3), and n is the number of observations in the model (i.e., 24); $p/n = 3/24 = 0.13$ 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 (Figure 1b).

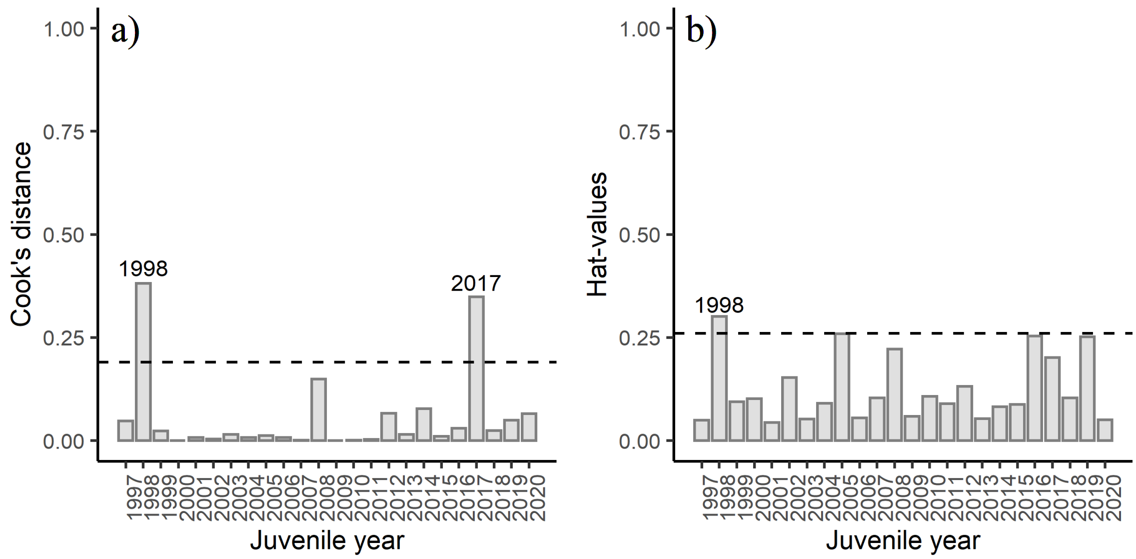

Figure 1: Diagnostics plots of influential observations including a) Cook's Distance (with a cut-off value of 0.19), and b) leverage values (with a cut-off value of 0.26) from model m2.

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 (P<0.05), suggesting some lack of fit for this term (Figure 2a). Diagnostics indicated that two of the data points were above the cut-off value for the Cook's distance (Figure 1a). One observation had a high leverage value (Figure 1b). 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; although observations 1, 2, 18, 21, and 24 and were the most extreme (juvenile years 1997, 1998, 2014, 2017, and 2020 corresponding to years 1998, 1999, 2015, 2018, and 2021) based on standardized residuals (Figure 3a; Table 5). Based on the lightly curved fitted lines in the residual versus fitted plot (Figure 3b), the fitted plot shows some lack of fit of the model.

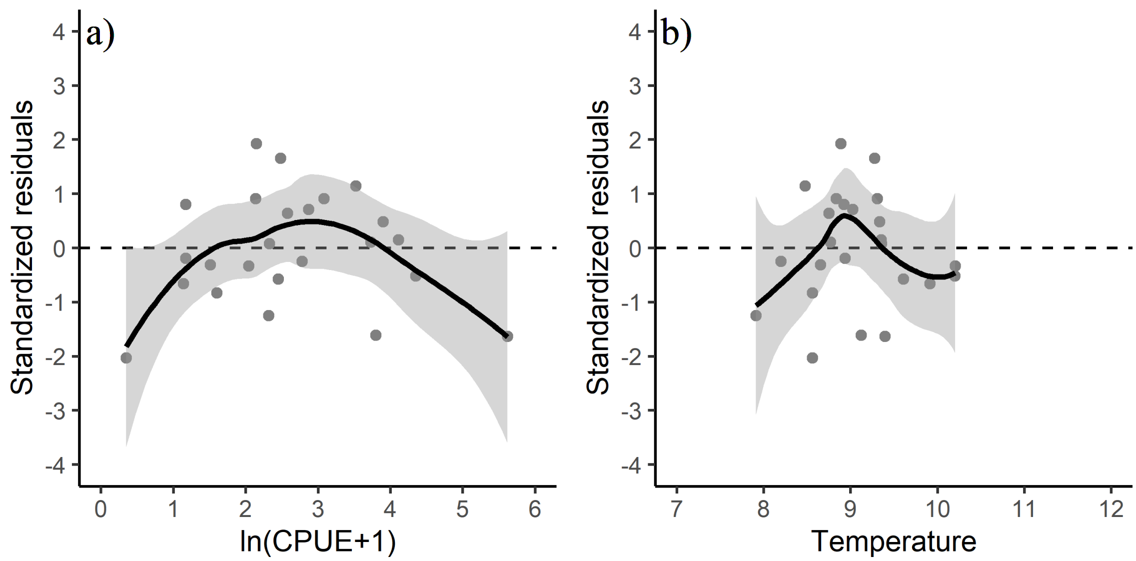

Figure 2: Standardized residuals versus predicted plots for a) CPUE and b) temperature.

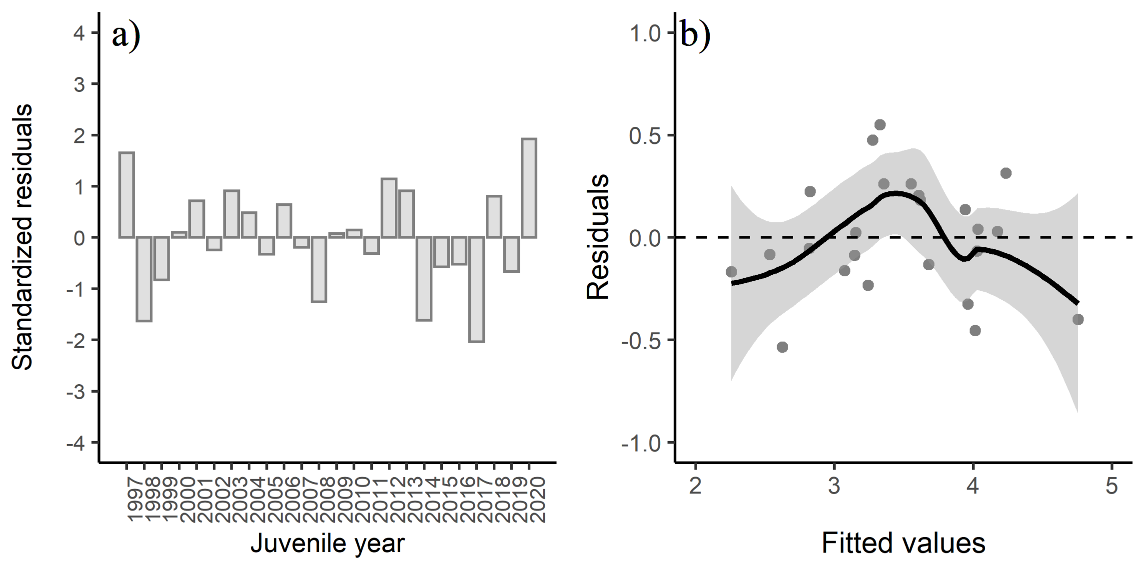

Figure 3: a) Standardized residuals versus juvenile year and b) residuals versus fitted values for model m2. Positive residuals indicate that the observed harvest was larger than predicted by the model.

Results

The best regression model based on the performance metrics looked at and significant coefficients in the model was model m2 (i.e., the model containing CPUE, and a May through July temperature variable). The adjusted R^2 value for model m2 was 0.81 (Table 3) indicating overall a good model fit. Based upon a model that includes juvenile pink salmon CPUE and May through July temperature (model m2), the 2022 SEAK pink salmon harvest is predicted to be in the weak range with a point estimate of 15.58 million fish (80% prediction interval: 10.30 to 23.57 million fish).

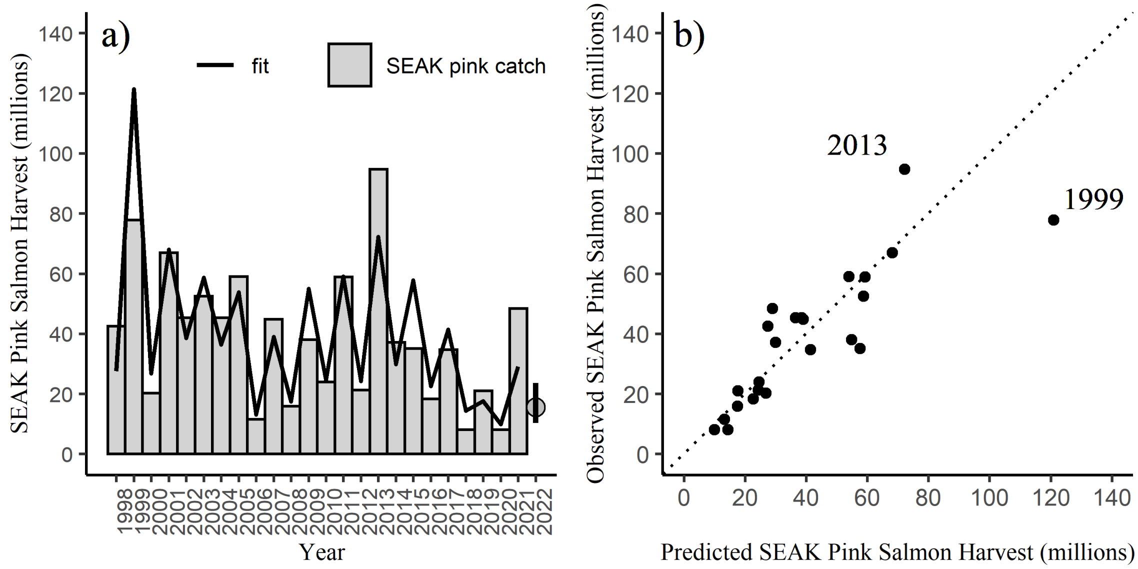

Figure 4: A) SEAK pink salmon harvest (millions) by year with the model fit (line). The predicted 2022 forecast is symbolized as a grey circle with an 80% prediction interval (10.30 to 23.57 million fish). B) SEAK pink salmon harvest (millions) against the fitted values from model m2 by year. The dotted line is a one to one line.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Pet rov & F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado.

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.

Miller, D. M. 1984. Reducing transformation bias in curve fitting. The American Statistician 38: 124-126.

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 Commission Technical Report No. 15: 75-81.

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)

Piston, A. W., J. Murphy, J. Moss, W. Strasburger, S. C. Heinl, E. Fergusson, S. Miller, A. Gray, and C. Waters. 2021. Operational Plan: Southeast coastal monitoring, 2021. ADF&G, Regional Operational Plan No. ROP.CF.1J.2021.02, Douglas.

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

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.

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