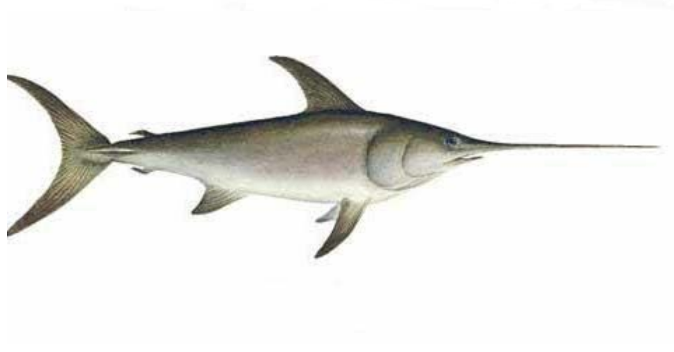
International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean

Using the EM Algorithm to Predict Catch-Per-Unit-Effort of Striped Marlin by the Japanese Distant Water Longline Fleet in Region 5[1]

Jon Brodziak
NOAA NMFS Pacific Islands Fisheries Science Center
Honolulu, Hawaii, USA



[1]Working document submitted to the ISC Billfish Working Group Workshop, 19-27 March 2007, Chinese Taipei. Document not to be cited without author's written permission.

Using the EM Algorithm to Predict Catch-Per-Unit Effort of Striped Marlin by the Japanese Distant Water Longline Fleet in Region 5

Jon Brodziak, Pacific Islands Fisheries Science Center, Honolulu, HI 96822-2326
Email: Jon.Brodziak@NOAA.GOV

Catch-per-unit effort (CPUE) data from the Japanese distant water longline fleet provide regional indices of relative abundance of striped marlin. There are consistent CPUE time series for regions 1 through 4 from the 1960s through the 2000s. For region 5, there is a gap in CPUE values from 1992-2005. The lack of CPUE in region 5 since 1992 limits the use of region 5 CPUE as a component of a stock-wide average CPUE for the longline fleet. In this paper, the expectation maximization (EM) algorithm is applied to predict missing values of striped marlin CPUE in region 5 using the covariance structure among regional CPUE series.

The Japanese CPUE data used in this analysis were provided by Yokawa (CPUE_STM0611.xls) to the striped marlin working group. The CPUE values were based on least squares means of standardized CPUE by area and region (Table 1). CPUE estimates for region 3 in 1992, 1994 and 1998-2000 were missing and were imputed using the average CPUE across regions 1, 2, and 4 in each year. The correlation matrix of the available data (Σ with elements ordered as regions 1, 2, ..., 5) showed that regional CPUE values were positively correlated.

$$\Sigma_{available} = \begin{bmatrix} 1 & 0.57 & 0.53 & 0.48 & 0.32 \\ 0.57 & 1 & 0.55 & 0.74 & 0.26 \\ 0.53 & 0.55 & 1 & 0.75 & 0.69 \\ 0.48 & 0.74 & 0.75 & 1 & 0.43 \\ 0.32 & 0.26 & 0.69 & 0.43 & 1 \end{bmatrix}$$

The expectation-maximization (EM) algorithm is a general iterative method for maximum likelihood estimation when data are incomplete with some missing observations (Dempster et al. 1977, Little and Rubin 1987). Two assumptions will be made to apply the EM algorithm to predict striped marlin CPUE in region 5. First, it is assumed that the regional CPUE data (X_{pxN}) are jointly distributed as a multivariate normal (MVN) random variable. with mean vector μ and covariance matrix Σ. Second, it is assumed that the missing observations are missing at random which implies that the missing values are not missing because of the value of their response (CPUE). The EM algorithm consists of paired prediction and estimation steps. Predict the contribution of the missing data to sufficient statistics for the complete data and then estimate the parameters of interest using the predicted sufficient statistics.

Sufficient statistics for the complete MVN CPUE data are the vector sum of all samples (T_1) and the outer product sum of all samples (T_2), where

(1) $$T_1 = \sum_{j=1}^{N} X_j \quad and \quad T_2 = \sum_{j=1}^{N} X_j X_j^T$$

In this application, the first variable (p=1) will be region 5 CPUE which has some missing values. We partition the mean vector μ into an incomplete (p=1) and complete data components (p=2,3,4,5) as $\mu = (\mu_1 \mid \mu_2 ... \mu_5)$. Similarly the covariance matrix Σ is partitioned into incomplete (Σ_{11}, Σ_{12}, Σ_{21}) and complete (Σ_{22}) data submatrices as

(2) $$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pn} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Given an initial guess for the missing values (in this case set equal to μ_1) the prediction step fills in missing values x_{1j} as

(3) $$\hat{x}_{1j} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1} \begin{bmatrix} x_{2j} - \mu_2 \\ x_{3j} - \mu_3 \\ x_{4j} - \mu_4 \\ x_{5j} - \mu_5 \end{bmatrix}, \quad \widehat{x_{1j}^2} = \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \left(\hat{x}_{1j}\right)^2, \quad \widehat{x_{1j}x_{pj}} = \left(\hat{x}_{1j}\right) x_{pj}$$

This fills in the values of the sufficient statistics T_1 and T_2. Next the estimation step produces new estimates of μ and Σ as

(4) $$\mu_{new} = \frac{1}{N}\widehat{T_1} \quad and \quad \Sigma_{new} = \frac{1}{N}\widehat{T_2} - \mu_{new}\mu_{new}^T$$

Equations (3) and (4) were iteratively applied to the regional CPUE data (Table 1) to predict missing CPUE values for region 5 (Figure 1). In this case, the EM algorithm converged in 29 iterations when the relative change in successive iterates $\left\| x_{1j}^{new} - x_{1j}^{old} \right\| \big/ \left\| x_{1j}^{old} \right\|$ was less than 0.0001.

References

Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). J. Royal Stat. Soc., Ser. B., 39:1-38.

Little, R., and D. Rubin. 1987. Statistical analysis with missing data. John Wiley and Sons, New York, 278 pp.

Table 1. Observed striped marlin CPUE for the Japanese distant water longline fleet for fishing regions 1 through 5 (R1-R5) along with predicted CPUE values for region 5 in 1980, 1985, 1990, and 1992-2005 based on the EM algorithm.

	Regional CPUE						
Year	R1	R2	R3	R4	Observed R5	Predicted R5	Missing Value?
1964	0.945	0.228	3.250	1.167	6.051		No
1965	0.749	0.220	2.827	1.071	4.061		No
1966	0.559	0.102	1.671	0.915	5.074		No
1967	0.528	0.121	2.164	0.564	5.335		No
1968	0.597	0.137	2.573	1.367	4.630		No
1969	0.572	0.115	1.377	0.995	4.082		No
1970	0.695	0.141	2.802	1.046	4.075		No
1971	0.727	0.070	2.281	0.651	5.425		No
1972	0.679	0.128	0.949	0.442	3.570		No
1973	0.660	0.179	1.262	0.705	2.042		No
1974	0.829	0.089	0.470	0.294	2.686		No
1975	0.553	0.063	1.384	0.459	3.378		No
1976	0.292	0.058	0.969	0.367	2.825		No
1977	0.170	0.022	1.075	0.345	1.094		No
1978	0.274	0.036	1.051	0.260	0.511		No
1979	0.443	0.100	1.213	0.609	2.931		No
1980	0.658	0.070	0.980	0.511		3.569	Yes
1981	0.392	0.053	0.810	0.297	2.574		No
1982	0.274	0.055	0.823	0.288	4.206		No
1983	0.234	0.071	0.667	0.357	3.425		No
1984	0.396	0.060	0.703	0.212	3.068		No
1985	0.539	0.130	1.059	0.752		3.043	Yes
1986	0.793	0.088	0.841	0.428	4.830		No
1987	0.411	0.068	1.541	0.616	6.165		No
1988	0.615	0.057	1.305	0.671	4.250		No
1989	0.537	0.063	1.039	0.486	3.895		No
1990	0.412	0.091	0.701	0.395		2.077	Yes
1991	0.595	0.069	1.066	0.452	0.312		No
1992	0.502	0.135	0.535	0.968		2.586	Yes
1993	0.696	0.075	1.436	0.514		4.171	Yes
1994	0.587	0.143	0.399	0.466		1.852	Yes
1995	0.817	0.051	1.136	0.430		4.320	Yes
1996	0.470	0.140	0.927	0.822		2.673	Yes
1997	0.541	0.089	0.337	0.336		1.992	Yes
1998	0.629	0.101	0.355	0.333		2.156	Yes
1999	0.475	0.070	0.335	0.459		2.162	Yes
2000	0.410	0.047	0.272	0.358		1.963	Yes
2001	0.298	0.068	0.615	0.466		1.961	Yes
2002	0.265	0.098	0.102	0.392		0.842	Yes
2003	0.411	0.091	0.303	0.387		1.595	Yes
2004	0.268	0.058	0.230	0.263		1.208	Yes
2005	0.185	0.030	0.240	0.270		1.256	Yes

Figure 1. Observed striped marlin CPUE for the Japanese distant water longline fleet in Regions 1 to 5 during 1964-2005 and predicted CPUE for Region 5 using the EM algorithm.