Fishery stock assessment based on asymmetric logistic model

Osamu Komori　Department of Mathematical Analysis and Statistical Inference, Project Assistant Professor

1 Abstract

The stock assessment data such as RAM legacy data is essential for constructing an estimation model because the biomass in that data reflects the abundance of marine stock status properly. However, the available assessment data has limited sample size due to intensive data requirements and large amount of cost, and the ratio of collapsed stocks to non-collapsed stocks is highly imbalanced (very few collapse stocks in comparison with large number of non-collapse stocks). Moreover the stock status (collapse or non-collapse) is not deterministic, which is in fact estimated by stochastic biomass dynamics models such as age-structured model. To allow for the imbalancedness and uncertainty involved in the fishery data, we propose a new binary regression model with mixed effects for estimation of stock status by employing an asymmetric model. In the proposed model, we assume that the small part of observations of the collapsed stocks are distributed in the same way as those of the non-collapsed stocks, resulting in a mixture model of conditional probability of collapsed status given explanatory variables in fishery-related data. In the estimation equation, we observe that the weights for the non-collapse stocks are relatively reduced, which in turn puts more importance on the small numbers of observations of collapse stocks. As a result, the estimated collapse probabilities are much improved with a little degeneration of the estimated probabilities of non-collapsed stocks.

2 Materical and methods

We developed an aymmetric logistic regression with mixed effects to construct a prediction model of fishery status based on assessed stocks in RAM data and applied it to estimate the global fishery status based on unassessed stocks in FAO data. The asymmetric logistic regression means that we use an asymmetric logistic function as a link function in the generalized linear mixed model to allow for the imbalance in sample size and uncertainty of class labeling (collapse or non-collapse). The datset used for the prediction model consists of amout of catch, life history, major fishing areas as well as biomass information, which are commonly and widely investigated in preceding literature Thorson, Branch and Jensen (2012); Costello, Ovando, Hilborn, Gaines, Deschenes and Lester (2012). Some stocks in FAO data were identified to have high probability of being collapsed and those characteristics were clarified based on fish category and location information.

2.1 Asymmetric logistic regression

As seen in Figure 1, there are very few collapsed stocks in comparison with a large number of non-collpased stocks in RAM data. In this case, the typical statistical method such as logistic regression model does not work properly. That is, it causes that the prediction for non-collapse stocks is accurate; while the prediction for collapse stocks is not accurate. As a result, the over-all error rate could unresonably be estimated to be very low in a validation procedure.

Our aim is to propose a robust method to that unregular situation for the prediction of stock status. Let $y \in \{0, 1\}$ be a class label for non-collapse ($y = 0$) and collapse ($y = 1$), which is determined every year during the observation period, x and z be explanatory variables associated with the fixed and random effects, respectively. Then the conditional probability of y given (x, z) in a mixed-effect asymmetric logistic regression is formulated as

$$p(y|x, z, b, \delta) = \frac{(1-\delta)\exp\{y(x^\top\beta + z^\top b)\} + \delta}{1 + \delta + (1-\delta)\exp(x^\top\beta + z^\top b)}, \tag{1}$$

where β and b are fixed and random effects, respectively. Note that if $\delta = 0$, then it is reduced to a conditional probability in a usual logistic regressoin

$$p_{\mathrm{L}}(y|x, z, b) = \frac{\exp\{y(x^\top\beta + z^\top b)\}}{1 + \exp(x^\top\beta + z^\top b)}. \tag{2}$$

Then we have

$$p(y|x, z, b, \delta) = \begin{cases} w(x, z, \delta)p_{\mathrm{L}}(0|x, z, b) & \text{if } y = 0 \\ w(x, z, \delta)\{(1-\delta)p_{\mathrm{L}}(1|x, z, b) + \delta p_{\mathrm{L}}(0|x, z, b)\} & \text{if } y = 1 \end{cases}, \tag{3}$$

where

$$w(x, z, \delta) = \frac{1 + \exp(x^\top\beta + z^\top b)}{1 + \delta + (1-\delta)\exp(x^\top\beta + z^\top b)}. \tag{4}$$

We observe from (3) that $p(y|x, z, b, \delta)$ equals $p_{\mathrm{L}}(y|x, z, b)$ when $y = 0$ apart from the normalizing constant $w(x, z, \delta)$, while $p(y|x, z, b, \delta)$ is a contaminated model of $p_{\mathrm{L}}(y|x, z, b)$ with the mislabel probability δ Copas (1988); Takenouchi and Eguchi (2004); Hayashi (2012). We assume for the tuning parameter δ to satisfy $0 \le \delta \le 1$. The conditional probability (1) more correctly reflects the present situation in probabilistic manner. We confirm that the likelihood ration is given by

$$\frac{p(1|x, z, b, \delta)}{p(0|x, z, b, \delta)} = (1-\delta)\frac{p_{\mathrm{L}}(1|x, z, b)}{p_{\mathrm{L}}(0|x, z, b)} + \delta, \tag{5}$$

which implies that the linear discriminant function $x^\top\beta + z^\top b$ satisfies the Bayes risk consistency.

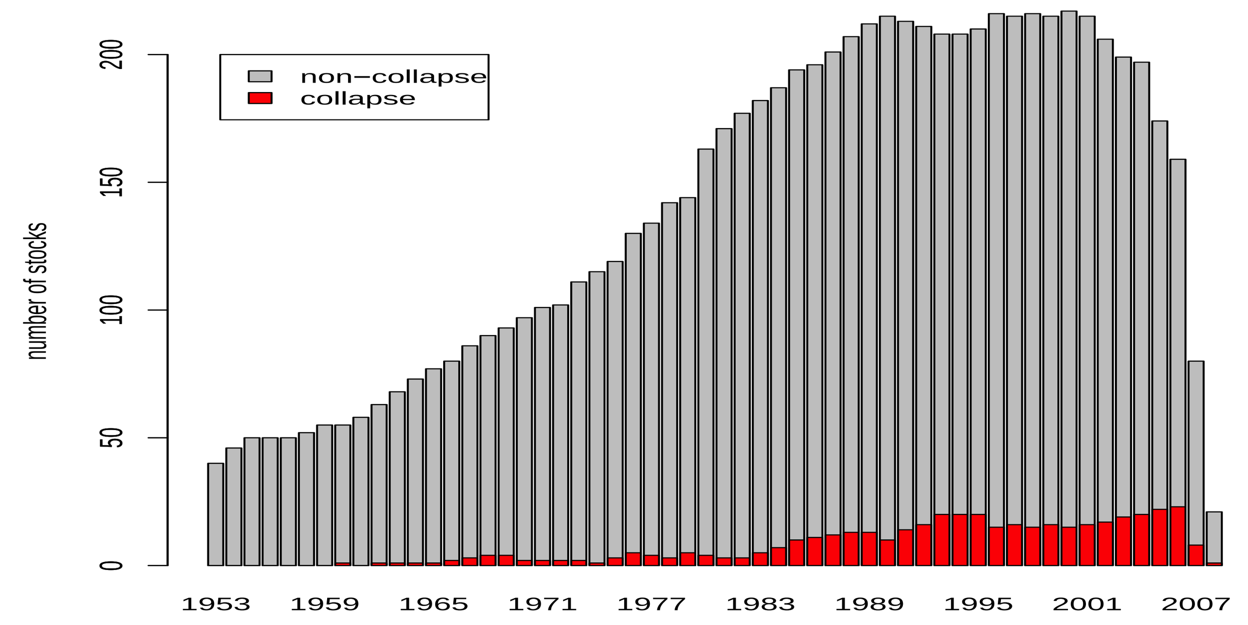

Figure 1: Barplots of numbers of non-collapsed stocks (gray) and collapsed stocks (red) during the observation years

Table 1: Mean 5-cross-validated AUC, TPR and TNR calculated for RAM legacy data set.

Year	**catch method**			**logistic model**			**asymmetric model**			
	AUC	TPR	TNR	AUC	TPR	TNR	AUC	TPR	TNR	δ
2000	**0.906**	0.013	**0.999**	0.891	0.093	0.988	0.892	**0.100**	0.984	0.004
2001	0.860	0	**1**	**0.879**	0.087	0.988	0.878	**0.093**	0.987	0.003
2002	0.877	0	**1**	**0.908**	**0.165**	0.982	0.905	**0.165**	0.981	0.003
2003	0.828	0	**1**	0.822	0.133	0.992	**0.858**	**0.378**	0.964	0.016
2004	0.839	0	**1**	0.825	0.10	0.987	**0.860**	**0.365**	0.969	0.016
2005	0.865	0.009	**0.997**	0.870	0.293	0.973	**0.884**	**0.475**	0.951	0.015
2006	**0.885**	0.134	**0.97**	0.879	0.267	0.966	0.884	**0.303**	0.956	0.008
2007	0.852	0	**0.99**	0.852	0	0.984	**0.853**	**0.01**	0.982	0.006

References

Copas, J. (1988). Binary Regression Models for Contaminated Data. *Journal of the Royal Statistical Society: Series B.* **50**, 225–265.

Costello, C., Ovando, D., Hilborn, R., Gaines, S. D., Deschenes, O. and Lester, S. E. (2012). Status and Solutions for the World 's Unassessed Fisheries. *Science* **338**, 517–520.

Hayashi, K. (2012). A boosting method with asymmetric mislabeling probabilities which depend on covariates. *Computational Statistics* **27**, 203–218.

Takenouchi, T. and Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate. *Neural Computation* **16**, 767–787.

Thorson, J. T., Branch, T. A. and Jensen, O. P. (2012). Using model-based inference to evaluate global fisheries status from landings, location, and life history data. *Can. J. Fish. Aquat. Sci* **69**, 645–655.