Journal of Forest Science
Vol. 24, No. 3, pp. 137~142, December 2008

Probability Estimation of Snow Damage on Sugi (*Cryptomeria japonica*) Forest Stands by Logistic Regression Model in Toyama Prefecture, Japan

Ken-ichi Kamo[1*], Hirokazu Yanagihara[2], Akio Kato[3], and Atsushi Yoshimoto[4]

[1]*Department of Liberal Arts and Sciences, Sapporo Medical University, S1 W16, Chuo-ku, Sapporo, Hokkaido, Japan*
[2]*Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan*
[3]*Toyama Agriculture, Forest Research Center, 3 Yoshimine, Nakaniikawa, Tateyama, Toyama, Japan*
[4]*Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan*

ABSTRACT : In this paper, we apply a logistic regression model to the data of snow damage on sugi (*Cryptomeria japonica*) occurred in Toyama prefecture (in Japan) in 2004 for estimating the risk probability. In order to specify the factors effecting snow damage, we apply a model selection procedure determining optimal subset of explanatory variables. In this process we consider the following 3 information criteria, 1) Akaike's information criterion, 2) Baysian information criterion, 3) Bias-corrected Akaike's information criterion. For the selected variables, we give a proper interpretation from the viewpoint of natural disaster.

Keywords : Information criterion, Logistic regression, Model selection, Snow damage

INTRODUCTION

Snow damage in forest stands causes by trunk breaking and bending, and uprooting by the weight of snow accumulating to leaf and branch. This would result in the economic loss of the management. So, if we can identify such variables for probability, we can manipulate the risk through some forestry activities, such as thinning, for better management scheme.

In order to identify the risk probability, we apply a logistic regression model, because it can deal with the binary responses expressed by some explanatory variables. In this case study we consider the model that response variable is the number of tree damaged by snow and explanatory variables are constructed by forest stand characteristics and geographic elements. The variable selection in regression analysis, which means the searching for the optimal subset of explanatory variables, is determined by information criterion.

In a general regression analysis, the expectation of response variables must be explained by the minimum essential number of variables from the statistical viewpoint, so we need to select the suitable subset of variables. We use 3 information criteria in searching the candidate model, and we decide the final model on the basis of the deviance between the candidates. In this case study, variable selection in regression model means to specify the risk factors.

This paper is organized as follows. In Section 2, we introduce the statistical method used in this study. In Section 3, we give the results on applying a logistic regression model to the data of snow damage occurring in Toyama prefecture, Japan. In Section 4, we discuss about the result obtained in Section 3.

* Corresponding author: (E-mail) kamo@sapmed.ac.jp

※ Advance in Forest Management and Inventory-selected papers from the international conference of IUFRO (Div. 4.01, 4.02, 4.04), Chuncheon, Korea, Oct. 13-17, 2008.

METHODS

Let the number of trees in sample area i (1, 2, ⋯ , m) be n_i, $\sum_i n_i = n$, and the risk probability for snow damage be p_i. Then the number of damaged tree y_i is distributed according to the binomial distribution, that is,

$$y_i \sim \mathrm{Bin}(n_i, p_i).$$

We set the observation vector as $\boldsymbol{y} = (y_1, \cdots, y_m)'$. The probability distribution function for damaged number y_i is known as

$$f(y_i) = {}_{ni}\mathrm{C}_{yi}\ p_i^{\ yi}\ (1-p_i)^{ni\text{-}yi}. \tag{1}$$

It is naturally expected that the probability p_i is affected by several factors, then these elements must be expressed by explanatory variables as $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ir})'$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)'$, where r is the number of variables. In a logistic regression model, logit $p_i = \log\{p_i / (1 - p_i)\}$ is expressed as the linear combination of explanatory variables, that is,

$$\text{logit } p_i = \sum_j \beta_j x_{ij} = \boldsymbol{x}_i'\ \beta,$$

which is equivalent to

$$p_i = \exp(\boldsymbol{x}_i'\ \beta) / (1 + \exp(\boldsymbol{x}_i'\ \beta)). \tag{2}$$

Here $\beta = (\beta_1, \cdots, \beta_r)'$ is unknown parameter vector.

Let us consider the variable selection, which is the problem for searching the optimal subset of variables. In a regression analysis, this is determined by minimizing the risk function based on the Kullback-Leibler information (Kullback and Leibler, 1951) defined as

$$R = -2\ E_yE_u\ [L(\beta\text{-}hat; \boldsymbol{u})], \tag{3}$$

where E means the expectation under a candidate model and $\boldsymbol{u}$ is a future observation, which is independent of $\boldsymbol{y}$ and distributed according to the same distribution as $\boldsymbol{y}$. Here β-*hat* is the maximum likelihood estimator (MLE) of unknown parameter vector β under the candidate model, which is obtained by maximizing the log-likelihood function obtained from (1) as

$$L(\beta\ ; \boldsymbol{y}) = \sum_i \log f(y_i) = \sum_i \{\log_{ni} \mathrm{C}_{yi} + y_i \log p_i + (n_i - y_i) \log(1 - p_i)\}. \tag{4}$$

So MLE β-*hat* is obtained as

$$\begin{aligned}\beta\text{-}hat &= \mathrm{argmax}_\beta L(\beta\ ; \boldsymbol{y}) \\ &= \mathrm{argmax}_\beta \sum_i \{y_i\ \boldsymbol{x}_i'\beta - n_i \log(1 + \exp(x_i'\ \beta)\},\end{aligned}$$

here we use the relation (2). The expected probability in i-th sample is obtained as logit p_i-*hat* $= \boldsymbol{x}_i'\ \beta$-*hat*.

Generally, we cannot calculate (3) exactly, because it includes the unknown distribution. It is easy to obtain the simplest estimator of R by $-2L(\beta\text{-}hat\ ; \boldsymbol{y})$. However, when we estimate R by $-2L(\beta\text{-}hat\ ; \boldsymbol{y})$, the constant bias appears. Then Akaike (1973) propose the following AIC (Akaike’s information criterion), which is the effective estimator of (3) as

$$\mathrm{AIC} = -2L(\beta\text{-}hat; \boldsymbol{y}) + 2r. \tag{5}$$

This criterion is obtained by correcting the previous bias as the number of variables. Hence, AIC estimates the part of goodness for fitting as log-likelihood, and the part of penalty for the model complexity as the number of variables. On the other hand, Schwarz (1978) propose a criterion BIC (Baysian information criterion) based on Baysian framework as

$$\mathrm{BIC} = -2L(\beta\text{-}hat\ ; \boldsymbol{y}) + r \log n. \tag{6}$$

Recently, Yanagihara et al. (2003) conducted the new AIC-type information criterion derived by correcting the bias of AIC using the perturbation expansion of MLE. This criterion, which is named as CAIC (Bias-corrected AIC), is defined as

$$\begin{aligned} \mathrm{CAIC} &= -2L(\beta\text{-}hat;\ \boldsymbol{y}) + 2r + (a_1 + a_2 + a_3)/n \\ &= \mathrm{AIC} + (a_1 + a_2 + a_3)/n, \end{aligned} \tag{7}$$

where the definitions of a_1, a_2 and a_3 by matrix and vector notations are referred in Yanagihara et al. (2003). Although the bias correcting term of (7) is complicated, the order of the bias is improved to $O(n^{-2})$. In our study, we use these representative 3 information criteria in model selection.

When a different model is selected on the basis of different information criterion, the deviance (McCullagh and Nelder, 1989)

$$\begin{aligned} D &= -2\log\{(\text{Likelihood of candidate model}) / \\ &\quad (\text{Likelihood of full model})\} \\ &= 2\textstyle\sum_i [y_i \log \{y_i / (n_i\, p_i\text{-}hat)\} + (n_i - y_i) \\ &\quad \log\{(n_i - y_i) / (n_i - n_i\, p_i\text{-}hat)\}] \end{aligned} \tag{8}$$

is used to decide which model outperforms others. Here the full model is defined as the one fully fitting the observations. Since the resultant D follows the Chi-square distribution with m-k-1 degrees of freedom, we can use this statistics to judge which is the best, where k is the number of variables in candidate model.

RESULTS

We apply logistic regression analysis to the data for snow damage of sugi forest in Toyama prefecture collected in 2004. The number of sample plots (20 m × 20 m) is 47. The total number of trees is 1761, and 599 trees are damaged by snow, then the estimated damage probability is 34%. On the other hand, we consider the following 16 elements which may affect snow damage. The 7 elements on forest stand characteristics are as follows;

S1. Forest stand age

S2. Average tree DBH: Diameter at breast height

S3. Average tree height

S4. Ratio of Height-DBH (H/D ratio): Defined as dividing height by DBH.

S5. Forest stand density.

S6. Stand volume

S7. Cultivar: Dummy variable expressing either boka- or kawaidani-sugi or not. These species are said to be weak to snow damage.

The 9 elements on geography are as follows;

G1. Altitude

G2. Slope gradient

G3. Contributing area: Index to show water flowing calculated by numerical examination. Attain 0 in ridge.

G4. Plan curvature: Curvature along with horizontal direction. The value becomes positive for concave surface, while it is negative for convex surface (Moore et al., 1993).

G5. Profile curvature: Same definition as plan curvature along vertical.

G6. Over ground openness: Degree of outlook. This attains large value near ridge.

G7. Under ground openness: Degree of interruption by the earth. This attains large value near gully.

G8. Topographic wetness index: Let *a* be contributing area and *b* be slope degree, then defined as log (*a* / tan *b*) (Beven, 1997). It is the index for wetness in the forest stand surface.

G9. Slope aspect: Combination of 4 dummy variables to identify 8 slope aspects (see Table 1). For

Table 1. Dummy variable expressing the slope aspect

	East	West	South	North
North	0	0	0	1
Northeast	1	0	0	1
East	1	0	0	0
Southeast	1	0	1	0
South	0	0	1	0
Southwest	0	1	1	0
West	0	1	0	0
Northeast	0	1	0	1

Eight slope aspects are expressed by the combination of four dummy variables. For example, “north” is (east, west, south, north)=(0, 0, 0, 1) and “northeast” is (1, 0, 0, 1).

example, “north” is identified by (east, west, south, north)=(0, 0, 0, 1), and “northwest” is (1, 0, 0, 1). By changing this combination, all aspects are covered.

Roughly saying, the variables G3-G8 express the convexity of the forest stand surface by several aspects.

Table 2 shows the results for selecting the optimal subset of explanatory variables by previous 3 information criteria (AIC, BIC and CAIC) in logistic regression model. Here the variable “slope aspect”, which is constructed by the combination of 4 dummy variables, is treated as one variable, that is, we do not consider the necessity of specific aspect only. Table 2 lists the best model in the fixed number of variables from 1 to 16.

For example, in the model with 4 variables, all the possible combination of 4 in 16 variables is $_{16}C_4$=1820, and in these the combination {cultivar, altitude, under ground openness, slope aspect} is selected by all information criteria. Over different number of variables, the model with 12 variables is selected by AIC and BIC, while one with 9 variables is selected by CAIC (Figure 1). The variables selected by CAIC are {age, DBH, cultivar, altitude, contributing area, over ground openness, under ground openness, wetness, slope aspect}. The model selected by AIC and BIC includes 3 more variables {density, stand volume, profile curvature} adding to the model by CAIC (Tables 2 and 3). Hence, the variable selected by CAIC is commonly selected by other criteria. These results enable us to predict the risk probability, for example, Figure 2 shows the fitting to estimated logistic curve, and Figure 3 shows the relation between the observed and predicted risk probability.

In this case study, the different model is selected as information criteria, that is, the model selected by AIC and BIC has 3 more variables than one by CAIC. This leads to the question as to whether these variables should be used in practice or not. Note that the difference of deviances between the model estimated by AIC and CAIC is 12.3 with the corresponding p value of 0.01. Thus, the additional 3 variables lead significant improve-

Table 2. Results of the selected models by each criterion

Variable	The number of variables^{*1}															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Age					○	○	○	○	○	○	○	○	○	○	○	○
DBH						*	○	○	○	○	○	○	○	○	○	○
Height																○
H/D ratio															○	○
Density											○	○	○	○	○	○
Cultivar			○	○	○	○	○	○	○	○	○	○	○	○	○	○
Stand volume											○	○	○	○	○	○
Altitude				○	○	○	○	○	○	○	○	○	○	○	○	○
Contributing area								○	○	○	○	○	○	○	○	○
Plan curvature														○	○	○
Profile curvature										○		○	○	○	○	○
Over openness									○	○	○	○	○	○	○	○
Under openness		○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Wetness						**	○	○	○	○	○	○	○	○	○	○
Slope aspect	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

*1) except for constant.
The notation “○” denotes the variable selected by all information criteria. The notation “*” denotes the variable selected by AIC and BIC, while “**” denotes one by CAIC only.

Table 3. Estimated parameter values of the selected model

Variable	AIC, BIC		CAIC	
	Coefficient	*p* value	Coefficient	*p* value
Constant	-2.048	0.55	-10.08	0.00
Age	0.041	0.00	0.048	0.00
DBH	-0.193	0.04	-0.046	0.01
Density	-0.002	0.00	-	-
Cultivar	-2.355	0.00	-1.991	0.00
Stand volume	0.003	0.00	-	-
Altitude	0.01	0.00	0.007	0.00
Contributing area	-0.003	0.00	-0.0002	0.03
Profile curvature	-23.627	0.05	-	-
Over ground openness	-0.08	0.05	-0.033	0.09
Under ground openness	0.11	0.00	0.115	0.00
Wetness	0.366	0.01	0.218	0.02
North	-1.971	0.00	0.011	0.96
East	1.327	0.00	1.539	0.00
South	-0.026	0.87	-0.075	0.61
West	0.113	0.63	0.259	0.23

The variables stand volume, density and profile curvature are not selected by CAIC.

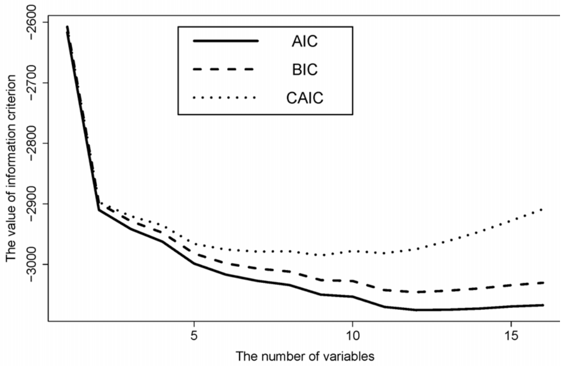

Fig. 1. Change in information criterion over different number of variables. The horizontal axis denotes the number of variables except for constant. The vertical axis denotes the value of information criterion of best model in the fixed number of variables.

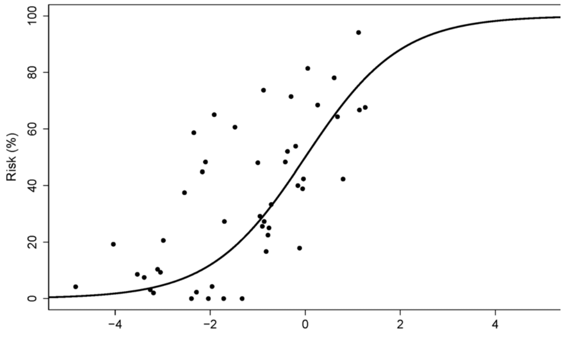

Fig. 2. Fitted logistic curve. The horizontal axis denotes the value of $x_i'\beta$, the vertical axis denotes the observed risk probability. The curve denotes the estimated logistic curve.

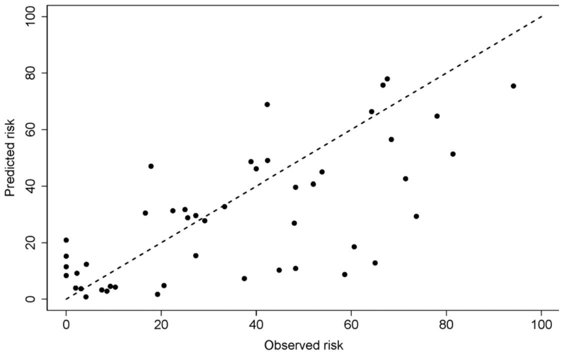

Fig. 3. Observed-predicted plot of the model by AIC and BIC. The horizontal and vertical axis denotes the observed and predicted risk probability, respectively. The broken line means the situation that predicted value is equal to observed one.

ment at the usual significance level 0.05, and it may be recommended to use the model estimated by AIC and BIC.

DISCUSSION

It is intuitively appropriate to expect that the risk varies continuously according to some continuous elements affecting snow damage. Motivating by these, we try to apply logistic regression model to snow damage in order to estimate the risk probability in continuous fashion. Another motivation for using this model is to automatically specify the elements of snow damage by variable selection procedure.

In the previous section, we obtained that the model selected by AIC and BIC being optimal, so let us note the properties of this model. We need to note the fact that it blew from southwest continuously when snow damage occurred, because the windy condition plays key rolu in the analysis of snow damage (Kato and Zushi, 2006). The sign of the estimated coefficient denotes whether the element increases or decreases the risk probability. The variables with positive coefficients are {age, stand volume, altitude, under ground openness, wetness, east and west}. It is naturally expected that the aged tree is fragile against the burden by snow, then this variable has an effect to increase the risk. For altitude, the amount of snow becomes greater near the ridge under the same condition for weather. In an area with high stand volume and wetness, the growth condition is better than other areas, so it makes the thin tree with high H/D ratio. For slope aspect, the opposite side is estimated to raises the risk, but we see that the east side has high risk by considering the significance of coefficients. In this case study, the east side receives a little effect for wind.

Next, the variables with negative coefficient are {DBH, density, cultivar, contributing area, profile curvature, over ground openness, north and south}. The tree with high DBH can put up with the pressure by snow. In a dense area, the total risk for snow is shared in many trees, so the risk per tree becomes slight. For cultivar, it is well known that boka- or kawaidani-sugi is weak to snow damage. The 3 variables contributing area, profile curvature and over ground openness denote the concave surface. In these areas, the effect for wind becomes little.

Totally considering, the elements raising the risk probability are as follows;

E1. Old stand,

E2. Tall figure,

E3. Geographic condition avoiding windy effect.

However, we need to note there exists the opposite case to E3, that is, wind raises the risk, for example, by pressure (Kato and Zushi, 2006).

Throughout our study, logistic regression model and model selection procedure can contribute to specify the elements against snow damage and to estimate the risk. Moreover the selected subset of variables is valid referring the results in the previous studies (Kato and Zushi, 2006). This means that our approach works well in the analysis for the risk of snow damage. For future problem, we need to make the detailed risk map by estimating continuous risk of snow damage. Owing to these results on risk probability, we hope to improve the management scheme against snow damage, for example, shinning or species selection.

LITERATURE CITED

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: 2nd International Symposium on Information Theory, Petrov B. N. and F. Csáki (eds.). Akadémiai Kiadó, Budapest, pp. 267-281.

Beven, K. 1997. Topmodel: A critique. Hydrogical Process 11: 1069-1085.

Kato, A. and K. Zushi. 2006. Relation between snow damage and geographic factor in Japanese cedar stands in Toyama Prefecture. FORMATH 6: 77-88 (in Japanese).

Kullback, S. and R. Leibler. 1951. On information and sufficiency. Ann. Math. Statist 22: 79-86.

McCullagh, P. and J. A. Nelder. 1989. Generalized linear models, 2nd edition. Chapman & Hall, London.

Moore, I. D., Gessler, P. E., Nielsen, G. A. and G. A. Peterson. 1993 Soil attribute predicting using terrain analysis. Soil Science Society of America J. 57: 443-452.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Statist. 6: 461-464.

Yanagihara, H., Sekiguchi, R. and Y. Fujikoshi. 2003. Bias correction of AIC in logistic regression models. J. Statist. Plann. Inference. 115: 349-360.